Bias in grote taalmodellen voor AI-toepassingen in de gezondheidszorg

Een framework voor detectie en mitigatie van bias in deeplearning-transformermodellen ter ondersteuning van de gezondheidszorg

Thierry Desot

Hogeschool Rotterdam Uitgeverij

1e druk, november 2024

Dit boek is een uitgave van Hogeschool Rotterdam Uitgeverij
Postbus 25035
3001 HA Rotterdam

Ontwerp:
Jargo Design

ISBN: 9789083481234
NUR: 984

Bias in grote taalmodellen voor AI-toepassingen in de gezondheidszorg

Een framework voor detectie en mitigatie van bias in deeplearning-transformermodellen ter ondersteuning van de gezondheidszorg

OPENBARE LES

dr. Thierry Desot

Lector toegepaste natuurlijke taalverwerving

Rotterdam, 28 november 2024

Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution

(ALBERT EINSTEIN)

Il sole non si muove –
The sun does not move

(LEONARDO DA VINCI)

Inhoudsopgave

Voorwoord

Mijn fascinatie (die ontstond ontstaan tijdens mijn middelbareschooltijd) voor de intrigerende morfologie en syntaxis van talen zoals Latijn en Grieks en een verdieping in Arabistiek tijdens mijn masteropleiding aan de Katholieke Universiteit Leuven (1992-1996) hebben mij ertoe gebracht computerlinguïstiek te studeren. Deze keuze was het gevolg van mijn interesse om machines taalkundig redeneervermogen bij te brengen. Het contrast tussen mijn interesse in Latijn en Grieks en in de meer wiskundige richtingen was opvallend te noemen. Maar er was in het curriculum van een richting als Latijn en Grieks geen plek voor computers en computerprogrammeren, wat op dat moment in opkomst was. Destijds kon ik me nauwelijks voorstellen dat mijn carrière zich zou ontwikkelen op het complexe, maar boeiende snijvlak van taal en computer in de disciplines van computerlinguïstiek en Natural Language Processing (natuurlijke taalverwerking). Natural Language Processing (NLP) is een tak van kunstmatige intelligentie (artificial intelligence of AI) die zich bezighoudt met het begrijpen en verwerken van menselijke taal door computers. Het doel van NLP is om machines te leren tekst en spraak te interpreteren, zodat ze taken kunnen uitvoeren zoals vertalen, samenvatten, vragen beantwoorden en sentiment analyseren. Dit gebeurt door middel van technieken die taalstructuren herkennen en betekenis uit woorden en zinnen halen, vergelijkbaar met hoe mensen taal begrijpen. De gedachte aan een carrière binnen deze vakgebieden leek toen nog verre van reëel, gezien de schijnbare onverenigbaarheid met mijn initiële achtergrond in Latijn en Grieks.

Een cruciaal aspect van de opleiding computerlinguïstiek is uiteraard het bekend zijn met computerprogrammeren. Doorgaans is deze competentie voorbehouden aan individuen met een achtergrond in computerwetenschappen, eerder dan voor pure taalkundigen. En toch zijn er overeenkomsten tussen computerprogrammeren en taalkunde. Een computerprogramma bestaat namelijk uit een gelaagde syntaxis en semantiek, waarbij een sterk analytisch vermogen op het gebied van taalkunde van essentieel belang is. Deze vaardigheid bleek dan ook bijzonder waardevol tijdens mijn studie en doctoraat in computerlinguïstiek en NLP.

Voor mij vertoonde het begrip van de morfologie en syntaxis van een complexe taal sterke overeenkomsten met het begrip van een computerprogramma. Beide domeinen vereisen abstract denken, analytisch vermogen en het vermogen om verbanden te zien over verschillende programmeerblokken heen.

Het eerste 'aha'-moment kwam voor mij tijdens mijn kennismaking met machine learning aan de Universiteit Antwerpen in 2008, waar ik als een magneet naartoe werd getrokken. Dit leidde tot mijn betrokkenheid bij de ontwikkeling van een toepassing voor een tekst-naar-spraak-applicatie, die automatische diakritisatie toepast – het invoegen van korte klinkertekens in de Arabische taal, waar de orthografie normaal geen korte klinkers bevat. In die tijd werden er al theoretische concepten van deep learning ontwikkeld, een tak van machine learning gemodelleerd naar de werking van de hersenen. In 2017 resulteerde dit in de opkomst van Large Language Models, namelijk de transformermodellen.

Een transformermodel is een type AI dat wordt gebruikt om patronen in data zoals tekst te herkennen en te verwerken. Het model plaatst woorden en zinnen in de context, zodat het de betekenis beter begrijpt, zelfs in lange teksten. Dit doet het met behulp van aandachtmechanismen die bepalen welke delen van de tekst belangrijk zijn; dit maakt het model snel en efficiënt in het begrijpen van taal. Het model wordt een transformermodel genoemd omdat het door middel van deze aandachtmechanismen de invoerdata 'transformeert' om complexe relaties in de context te begrijpen. Het transformermodel is eveneens de basis van de onderliggende technologie van het veelbesproken model ChatGPT. De toepassing van deze theoretische concepten in de praktijk was destijds echter beperkt door de algemene tekortkomingen in de rekenkracht van computers.

In 2017 kreeg ik aan de Université Grenoble Alpes de kans om een doctoraat te halen binnen het vakgebied van NLP, gericht op Spoken Language Understanding (SLU), een samensmelting van spraakherkenning en tekstbegrip. Het doel was een bijdrage te leveren aan de ontwikkeling van een spraakgestuurde module. Deze module moest senioren in staat stellen om commando's te geven aan apparaten in een slim huis, zoals bijvoorbeeld het commando om de deuren automatisch te sluiten of te vergrendelen of om de verwarming of een huishoudelijk toestel aan of uit te zetten. Met behulp van deze technologie kunnen senioren hun dagelijkse taken blijven uitvoeren en langer zelfstandig blijven wonen. De wetenschap dat ik mocht bijdragen aan technologie die de zorg ondersteunt en vooruithelpt, vormde een enorme drijfveer en motivatie tijdens dit doctoraat. Het schrijven van een puur theoretisch proefschrift, zonder praktische toepassing of maatschappelijke relevantie, zou voor mij uitermate moeilijk zijn geweest. Het doctoraat werd niet alleen een cruciale factor in mijn voortgezet onderzoek in toegepaste AI en NLP, maar ook in het uitvoeren van onderzoek dat professionals in de sociale en zorgsector direct ondersteunt en deze sectoren vooruithelpt. Het streven is dat deze professionals de leiding blijven houden, en niet vervangen worden, waarbij AI fungeert als ondersteuning en oplossingen biedt die zonder AI niet of nauwelijks denkbaar zouden zijn. In de latere fase van mijn doctoraat en tijdens mijn postdoc-periode aan de Universiteit Gent in 2021, maakte ik voor het eerst kennis met de opkomst van grote taalmodellen, die een ware revolutie

teweegbrachten binnen machine learning, met name binnen het domein van deep learning, en daarmee ook in de wereld van NLP.

De ideeën voor deze openbare les zijn mede tot stand gekomen door gesprekken en samenwerking met verschillende collega's. Ik wil in het bijzonder Hanneke Reuling, lid van het College van Bestuur van Hogeschool Rotterdam, Heleen Elferinck, directrice van Kenniscentrum Creating 010 en voormalig directeur Peter Troxler bedanken voor het mogelijk maken van deze openbare les en collega-lectoren Sunil, Maaike en Ben voor het meelezen en hun waardevolle feedback op eerdere versies. Daarnaast ben ik mijn ouders, mijn broer Miguel, andere familieleden en beste vrienden Peter, Jurgen en Petra dankbaar voor de feedback die ook zij hebben gegeven. Daarnaast wil ik mijn collega's Anne Marike, Elin, Sandra en Charlotte bedanken voor hun waardevolle advies ter voorbereiding van de finale tekstversie van deze openbare les. Ten slotte wil ik collega's Marianne, Tamara, Sjoerd en Lizzy bedanken voor hun uitstekende logistieke en praktische ondersteuning bij het vorm geven aan deze openbare les.

Inleiding

Tot nog toe blijft de interne werking van een deeplearningsysteem, zoals het transformermodel dat ook in ChatGPT wordt gebruikt, grotendeels een black box. Deep learning is een machinelearningbenadering die werkt zoals de hersenen, met lagen van 'neuronen' die samenwerken om patronen te herkennen en beslissingen te nemen. Deze lagen leren steeds complexere kenmerken kennen naarmate ze meer data verwerken, vergelijkbaar met hoe mensen leren door ervaring. Het is vaak nog steeds onduidelijk hoe een dergelijk systeem op basis van input tot een specifieke output komt. Dit gebrek aan transparantie wekt begrijpelijkerwijs wantrouwen bij gebruikers, die zich afvragen of de verkregen output wel betrouwbaar is, vooral in sectoren zoals de gezondheidszorg. Ondanks de potentiële voordelen brengen AI-ondersteunde gezondheidszorgsystemen risico's met zich mee, zoals onjuiste diagnoses en privacykwesties, wat leidt tot aarzeling bij gezondheidsorganisaties om AI in de zorg in te zetten. Deze voorzichtige benadering komt voort uit zorgen over aansprakelijkheid, gegevensbeveiliging, financiële investeringen en een neiging om vast te houden aan traditionele methodes. Het gebrek aan duidelijke regelgevende kaders voor het ethisch gebruik van AI in de gezondheidszorg versterkt deze terughoudendheid.

AI-systemen worden ingezet om afwijkingen (zoals tumoren of fracturen) op röntgenfoto's, MRI-scans of CT-scans te detecteren en te classificeren. Echter, wanneer een dergelijk systeem een diagnose stelt, blijft vaak onduidelijk op welke gronden het AI-systeem tot deze conclusie is gekomen. Een ander voorbeeld is de inzet van AI bij het voorspellen van ziekte-uitkomsten, zoals de kans op complicaties na een chirurgische ingreep of de kans op heropname in het ziekenhuis. Hoewel deze voorspellende modellen veelbelovend zijn, blijft het voor artsen te vaak onduidelijk welke factoren doorslaggevend zijn en waarom bepaalde voorspellingen worden gedaan. Deze onduidelijkheid kan leiden tot een gebrek aan vertrouwen in de aanbevelingen van de AI, terwijl vertrouwen natuurlijk cruciaal is in situaties waarin mensenlevens op het spel staan.

In april 2021 introduceerde de Europese Unie het eerste regelgevende kader en pionierswerk voor AI binnen de EU: de AI Act ofwel AI-verordening. Dit voorstel omvat een grondig onderzoek en de categorisatie van AI-systemen op diverse domeinen, afhankelijk van de risico's die ze met zich meebrengen voor gebruikers (European Parliament, 2023). Het voornaamste doel is ervoor te zorgen dat AI-systemen die in de EU worden ingezet, voldoen aan hoge normen op het gebied van veiligheid, transparantie, traceerbaarheid, non-discriminatie, milieubewustzijn en bijdragen aan het

bevorderen van eerlijkheid in AI, door ervoor te zorgen dat de systemen geen vooroordelen bevatten, rechtvaardige beslissingen nemen en geen groepen mensen ongelijk behandelen op basis van kenmerken zoals ras, geslacht of leeftijd.

Er wordt erkend dat verschillende risiconiveaus (beperkt, onaanvaardbaar en hoog risico) van AI-systemen verschillende graden van regulering vereisen. AI-systemen met een beperkt risico, waaronder systemen die zijn ontworpen voor het genereren of manipuleren van beeld-, audio- of videomateriaal, zoals deepfakes, moeten zich houden aan minimale transparantie-eisen. Deepfake-technologie heeft geleid tot een aanzienlijke controverse, met name op het gebied van deep fake porno, waarbij de gezichten van mensen – vaak zonder hun toestemming – worden geplaatst op de lichamen van pornoacteurs. Dat is wat er gebeurd is met de beelden van tientallen vrouwelijke BN'ers: er zijn nep-pornofilmpjes van ze gemaakt (Ravensbergen, 2024). Dit roept ernstige ethische en juridische vraagstukken op over privacy, toestemming en misbruik van beeldmateriaal. Veel landen worstelen met het reguleren van deze technologie omdat het moeilijk is om makers van illegale deepfakes te traceren en te vervolgen. Het psychologische effect op de slachtoffers kan diepgaand en langdurig zijn, wat de noodzaak onderstreept voor duidelijke wetgeving en strikte handhaving om misbruik tegen te gaan (Gamage et al., 2021).

AI-systemen die onaanvaardbare risico's met zich meebrengen of worden beschouwd als een bedreiging voor individuen, moeten worden verboden op grond van de Europese verordening. Het gaat hier dan bijvoorbeeld om spraakgestuurd speelgoed dat gevaarlijk gedrag bij kinderen kan aanmoedigen.

AI-toepassingen met een hoog risico, die impact hebben op veiligheid of fundamentele rechten, moeten een grondige beoordeling ondergaan vóór hun introductie op de markt en worden gedurende hun levenscyclus voortdurend geëvalueerd. Voorbeelden hiervan zijn toepassingen in de luchtvaart, auto's of medische apparaten. Ook medische systemen, die onder andere gebruikmaken van NLP, vallen binnen dit risicogebied voor AI; een voorbeeld daarvan is automatische ondersteuning van klinische besluitvorming bij het geven van medische adviezen. NLP is een deelgebied van AI dat gericht is op het begrijpen en verwerken van menselijke taal door computers. NLP heeft als doel machines in staat te stellen tekst en spraak te interpreteren, zodat ze taken kunnen uitvoeren, zoals vertalingen maken, samenvattingen genereren, vragen beantwoorden en sentimenten analyseren.

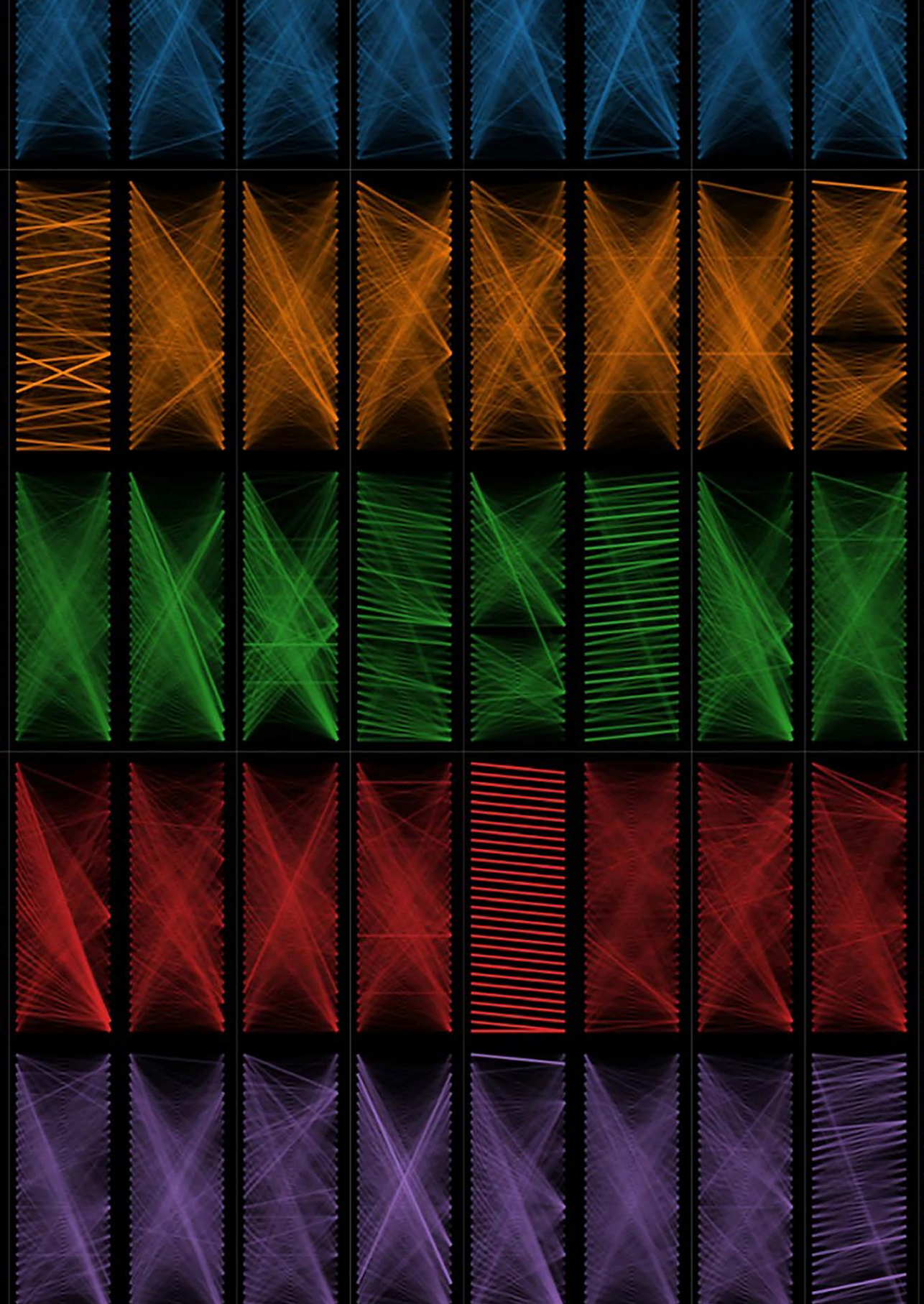
Achter ChatGPT en diverse andere NLP-toepassingen schuilt de architectuur van grote taalmodellen (Large Language Models of LLM's), die vooraf zijn 'gevoed' met uitgebreide en ongecontroleerde tekstdatasets, vaak afkomstig van het internet. Het grote risico op vooringenomenheid (bias) in deze grote taalmodellen ontstaat juist door die omvangrijke en ongefilterde gegevens waarmee ze zijn gecreëerd, nog eens versterkt door de

gebruikte algoritmes. Bias houdt in dat de modellen oneerlijke voorkeuren of uitsluitingen in zich dragen, bijvoorbeeld als een AI-systeem in klinische besluitvorming vaker foute diagnoses stelt voor minderheidsgroepen omdat de trainingsdata waarmee het gevoed is, niet representatief is. Dit kan leiden tot ongelijke behandeling, wat in strijd is met eerlijke en inclusieve AI. Deze vooringenomenheid, met name in de context van automatische klinische besluitvorming, kan leiden tot een onrechtvaardige behandeling van minderheidsgroeperingen in de samenleving, wat een ernstige schending vormt van het principe van eerlijke en inclusieve AI. Als dit systeem is getraind op historische gegevens waarin minderheidsgroepen minder vaak toegang kregen tot snelle zorg, kan het dezelfde bias reproduceren en patiënten uit deze groepen systematisch een lagere prioriteit geven voor dringende behandelingen, zelfs als hun medische situatie dat wel vereist.

In de volgende hoofdstukken wil ik laten zien hoe mijn lectoraat in toegepaste natuurlijke taalverwerking onderzoek mogelijk maakt waarin AI, en meer specifiek NLP met behulp van grote taalmodellen, kan bijdragen aan de ondersteuning en vooruitgang van de medische sector. Hierbij benadruk ik dat de inzet van AI niet bedreigend hoeft te zijn voor de professionals in de zorg, maar juist als ondersteuning en versterking van hun menselijke capaciteiten kan dienen, en niet als vervanging ervan. Dit wordt in volgende hoofdstukken geïllustreerd bij het gebruik van AI voor vroege herkenning van ernstige en complexe aandoeningen en ziektebeelden zoals sepsis en de ziekte van Alzheimer.

Daarnaast wil ik met een beschrijving van onderzoek naar en ontwikkeling van virtuele patiënten, aangedreven door AI, aantonen hoe de technologie mentale zorg en professionals in opleiding kan ondersteunen. Verder wil ik illustreren dat AI menselijke kennis en expertise kan aanvullen met diepe inzichten die zonder AI buiten bereik zouden blijven. Deze synergie tussen mens en machine kan aantonen dat AI, op voorwaarde dat deze goed wordt ingezet, een krachtig hulpmiddel is dat onze professionele en persoonlijke groei bevordert.

Deze technologie fungeert als een assistent voor professionals in de sociale of gezondheidssector, waardoor oplossingen mogelijk worden die zonder AI nauwelijks of niet haalbaar zouden zijn. Dit moet echter wel volgens principes van responsible AI gebeuren. Deze toepassingen dienen te voldoen aan normen op het gebied van veiligheid, transparantie, traceerbaarheid en non-discriminatie en ze dienen bij te dragen aan het bevorderen van eerlijkheid in AI. Hierbij zal de specifieke focus gericht zijn op het identificeren van vooringenomenheid (bias) in taalmodellen onder de motorkap van NLP-toepassingen in de gezondheidszorg en op het zoveel mogelijk verminderen van vooringenomenheid, waarvoor ik de krijtlijnen teken voor een kader voor toegepast onderzoek.



Een korte geschiedenis van het groot taalmodel

1

Om een beter begrip te krijgen van deep learning, en meer specifiek van transformer-modellen en grote taalmodellen (die een vorm van deep learning zijn), geef ik een beknopt historisch overzicht van AI, waarbij ik de rol van machine learning en vooral deep learning toelicht. Dit overzicht geeft inzicht in de evolutie van de technologieën en methodologieën die hebben geleid tot de huidige grote taalmodellen die onder de motorkap van ChatGPT zitten. Met de kennis van de historische context kunnen de principes en innovaties die aan de grondslag van deze moderne AI-technieken liggen, beter doorgrond en benut worden.

De plaats van deep learning in de geschiedenis van AI

Van golden age naar AI-winter

De fundamenten voor machine learning zijn in de vroege 20e eeuw gelegd. In de jaren '30 legde Alan Turing de basis voor AI met zijn concept van de 'universele machine' (later bekend als de Turingmachine), die elke berekening kon uitvoeren zoals die ook door een mens uitgevoerd kon worden. Hij toonde hiermee het vermogen van een machine om elke algoritmische procedure te simuleren, op voorwaarde dat er voldoende tijd en geheugen beschikbaar waren (Turing, 1936). In 1950 breidde Turing dit idee uit in zijn artikel 'Computing Machinery and Intelligence', waarin hij de Turingtest introduceerde, een criterium van intelligentie dat gebaseerd is op het vermogen van een machine om menselijk gedrag na te bootsen zodat een menselijke ondervrager het verschil niet kan zien tussen een machine en een mens (Turing, 1950). Searle stelde echter met zijn 'Chinese kamer' dat het slagen voor de Turingtest niet per definitie betekent dat de machine daadwerkelijk 'begrijpt' of 'bewustzijn heeft'. In zijn experiment bevindt een persoon die geen Chinees spreekt, zich in een gesloten kamer en gebruikt een set regels om Chinese karakters te manipuleren en vragen in het Chinees te beantwoorden, zonder de taal daadwerkelijk te begrijpen. Searle betoogde dat een computer die voor de Turingtest slaagt, net zoals de persoon in de kamer, simpelweg input manipuleert naar output volgens voorgeprogrammeerde regels, zonder enig echt begrip of bewustzijn (Searle,1980). Searle wil hiermee aantonen dat AI weliswaar schijnbaar intelligent gedrag kan vertonen, maar dat dit niet impliceert dat het daadwerkelijk bewustzijn heeft.

De formele geboorte van de term en de discipline van 'artificial intelligence' als academisch onderzoeksveld, vond plaats op het 'Dartmouth Summer Research Project on Artificial Intelligence' in 1956, een workshop van John McCarthy (wiskundige), samen met Marvin Minsky (cognitiewetenschapper), Nathaniel Rochester (computerwetenschapper) en Claude Shannon (bekend als de vader van de informatietechnologie). Het doel van de workshop was om in een groep van experts uit diverse wetenschappelijke disciplines te onderzoeken welke mogelijkheden machines hebben om na te denken en te leren. De organisatoren suggereerden namelijk dat elk aspect van leren of enige andere eigenschap van intelligentie zo nauwkeurig kan worden omschreven dat een machine kan gemaakt worden om het te simuleren. Dit zou inhouden dat intelligentie gerepliceerd kan worden door machines (McCarthy et al., 2006).

De term 'machine learning' werd geïntroduceerd door Arthur Samuel in 1959. Samuel werkte bij IBM en was een pionier op het gebied van AI en computer gaming. Hij gebruikte de term 'computer gaming' om een computerprogramma voor een damspel te beschrijven dat steeds meer bedreven raakte in dammen door ervaring op te doen en te leren uit verschillende partijtjes dammen. Het was een van de vroege succesvolle toepassingen van AI-technieken die lieten zien dat machines konden leren van data en hun prestaties door de tijd heen konden verbeteren zonder vooraf geprogrammeerd te worden om specifieke taken uit te voeren (Samuel, 1959).

Tijdens deze zogenaamde 'golden age' van AI in de jaren '60 en '70 speelden ook andere games zoals SHRDLU en Blocksworld een belangrijke rol in het demonstreren van de mogelijkheden van AI. SHRDLU (ontwikkeld door Terry Winograd aan het Massachusetts Institute of Technology – MIT) was een systeem voor natuurlijk taalbegrip, dat interageerde met een gesimuleerde blokkenwereld. Gebruikers konden in gewone taal instructies geven, waarop SHRDLU reageerde door blokken te verplaatsen en vragen te beantwoorden over de staat van deze blokkenwereld. Het systeem illustreerde hoe machines binnen een beperkte context menselijke taal konden begrijpen en erop konden reageren. Blocksworld bood een visueel en interactief platform waarop AI-systemen problemen konden oplossen door bijvoorbeeld blokken te stapelen of te sorteren. Deze games hielpen bij het verder ontwikkelen van technieken in het herkennen van patronen, ruimtelijk redeneren en het verwerken van natuurlijke taal, en speelden een sleutelrol in het vormgeven van het onderzoek naar AI (Winograd, 1971 en 1972).

De periode waarin SHRDLU en Blocksworld werden ontwikkeld, was ook een sleutelmoment voor de ontwikkeling van Natural Language Processing (NLP). Een vroeg hoogtepunt in NLP was het Georgetown-IBM-experiment in 1954, waarbij een computer meer dan zestig Russische zinnen succesvol automatisch naar het Engels vertaalde. Dit experiment illustreerde dat machines geprogrammeerd konden worden om eenvoudige vertaaltaken uit te voeren, wat verder onderzoek naar NLP stimuleerde.

Deze vroege projecten legden een stevige basis voor later onderzoek naar de syntaxis (de structuur van zinnen), semantiek (de betekenis van woorden en zinnen) en pragmatiek (het gebruik van taal in context) binnen AI-systemen, en benadrukten daarmee het belang van NLP als een essentieel onderdeel van AI (Hutchins, 1986).

In deze periode startten ook de eerste experimenten met perceptrons, nu beter bekend als neurale netwerken en deep learning, die geïnspireerd waren door de neurologische structuren in het menselijk brein en probeerden leerprocessen te simuleren. Een van de eerste modellen was het perceptron van Frank Rosenblatt in 1957, dat kon leren van inputgegevens door zijn gewichten — de waarden die bepalen hoe sterk een invoer bijdraagt aan het eindresultaat — automatisch aan te passen op basis van fouten in eerdere voorspellingen, waardoor het model geleidelijk beter werd in het herkennen van patronen (Rosenblatt, 1958).

De Dartmouth-conferentie legde dus de basis voor AI als een zelfstandige discipline en wakkerde het enthousiasme voor verder onderzoek naar AI aan. Hoewel de eerste verwachtingen niet meteen werden waargemaakt, legde deze bijeenkomst de basis voor een onderzoeksagenda die decennialang invloed zou uitoefenen. Tegelijkertijd effende zij de weg voor toekomstige vooruitgang in machine learning en de ontwikkeling van neurale netwerkarchitecturen.

De jaren '60 en '70 van de vorige eeuw vormden een belangrijke periode en een golden age in de ontwikkeling van AI, gekenmerkt door grote verwachtingen, maar uiteindelijk gevolgd door diepe teleurstellingen. Gedreven door het optimisme van de Dartmouth-conferentie, verwachtten onderzoekers en investeerders snelle en substantiële vooruitgang in het veld van AI. Dit enthousiasme leidde tot aanzienlijke investeringen in diverse projecten. Echter, tegen het midden van de jaren '60 en in de jaren '70 werd duidelijk dat veel van de ambitieuze doelstellingen, zoals het volledig begrijpen van natuurlijke taal door machines of het oplossen van problemen die het menselijk redeneren zou moeten evenaren, ver buiten het bereik bleven liggen van de toenmalige technologie.

Dit leidde tot een forse vermindering van de publieke en commerciële interesse en financiering midden jaren '70 tot begin jaren '80, een periode die bekend staat als de eerste 'AI-winter' (Wooldridge, 2021). De teleurstelling was grotendeels te wijten aan de beperkingen van de vroege AI-systemen, die niet in staat waren om te interageren met de complexiteit, onzekerheid en onvoorspelbaarheid van de *echte* wereld. De meeste vroege AI-programma's konden alleen functioneren onder gecontroleerde, specifieke, vooraf gedefinieerde omstandigheden en misten de mogelijkheid om te leren van nieuwe ervaringen uit de echte wereld of zich aan te passen aan onvoorziene omstandigheden. Veel van de vroege AI-programma's, zoals SHRDLU en Blocksworld, werkten slechts binnen strikt gedefinieerde en gecontroleerde omgevingen.

Bijvoorbeeld, in Blocksworld was de omgeving beperkt tot blokken die op elkaar gestapeld konden worden; de enige veranderingen die plaatsvonden, waren het resultaat van acties binnen dat systeem zelf. Dit kan vergeleken worden met de situatie waarin een persoon alleen woont en waarbij de locatie van objecten, zoals sleutels, niet verandert tenzij de persoon ze zelf verplaatst. Een huis waar meerdere mensen wonen en waar objecten onverwacht kunnen worden verplaatst door anderen, is een meer realistische weergave van hoe het werkt in de echte wereld (Crevier, 1993; Russel & Norvig, 2016; Wooldridge, 2021).

Knowledge-based AI en expertsystemen

Na de eerste AI-winter bloeide AI in de jaren '80 weer op. Deze opleving was voornamelijk te danken aan ontwikkelingen op het gebied van *knowledge-based AI* en *expert-systemen*. Het systeem van knowledge-based AI neemt beslissingen op basis van gestructureerde kennis en regels, zoals feiten, logica en ontologieën. Een ontologie is een gestructureerde manier om kennis in een bepaald domein te beschrijven. Het legt vast welke concepten, termen en relaties belangrijk zijn en hoe deze met elkaar in verband staan. Expertsystemen richten zich vooral op kennis over een bepaald domein in plaats van algemene kennis na te streven. Deze expertsystemen bootsen besluit-vorming van menselijke experts na met gebruik van uitgebreide domeinsets van regels en logica, die gedetailleerde kennis over een specifiek gebied vertegenwoordigen.

Parallel hieraan werd er vooruitgang geboekt op het gebied van *knowledge representation*, voor het structureren van domeinspecifieke informatie, als onderdeel van knowledge-based AI. Een van de bekendste voorbeelden is het CYC-project, in 1984 gestart door Douglas Lenat bij de Microelectronics and Computer Technology Corporation (MCC), een onderzoeksconsortium voor technologiebedrijven. Dit was een ambitieuze poging om een uitgebreide database van algemene kennis te creëren, die AI-systemen in staat zou stellen om te redeneren over alledaagse situaties, bijna zoals een mens dat doet. CYC probeerde een brede en rijke *knowledge graph* te bouwen, dat wil zeggen: een netwerk van verbonden feiten en concepten die de complexiteit van menselijke kennis nabootsen. Het uiteindelijke doel was om een AI-systeem te ontwikkelen, dat blijk zou geven van *gezond verstand* en op die manier de echte wereld beter zou begrijpen. Zo hanteert CYC brede categorieën, zoals tijd, ruimte, objecten en gebeurtenissen, waardoor zij een breed scala aan menselijke ervaringen en begrippen bevat. Door deze structuur kan CYC redeneren over alledaagse situaties, zoals begrijpen dat water nat is of dat mensen moeten eten om te overleven (Lenat, 1995).
Tegelijkertijd droeg de ontwikkeling van *logical AI* bij aan de creatie van systemen die konden redeneren met formele logica. Formele logica is een systeem van regels en symbolen dat gebruikt wordt om argumenten en redeneringen op een gestructureerde manier te analyseren. Deze aanpak bood een krachtig middel voor het op een voorspelbare en betrouwbare manier modelleren van complexe problemen en het genereren van nieuwe inzichten. Ten opzichte van knowledge-based AI richt logic-

based AI zich voor redenering en besluitvorming meer op het gebruik van logica (en niet zozeer op gestructureerde kennis en regels zoals knowledge-based AI). Logic-based AI maakt gebruik van formele logische systemen, zoals *propositielogica* en *eerste-orde logica*, om regels en relaties exact te definiëren en daarover te kunnen redeneren. Hoewel deze beide systemen voor logische redenering worden gebruikt, verschillen ze in wat ze kunnen uitdrukken. Propositielogica werkt met simpele uitspraken die waar of onwaar zijn, zoals "Het regent," en combineert deze met logische termen zoals "en," "of," en "niet." Hierbij wordt niet naar de inhoud van de uitspraken gekeken, maar alleen naar hun waarheidswaarde. Eerste-orde logica is complexer en werkt met uitspraken over objecten, hun eigenschappen en onderlinge relaties. Zij gebruikt variabelen, predikaten en kwantoren, zoals "voor alle" en "er bestaat", om gedetailleerdere redeneringen te maken, zoals "Alle mensen zijn sterfelijk" of "Er is een kat die zwart is". Hierdoor kan eerste-orde logica veel meer uitdrukken dan propositielogica (Sipser,1996).

Een iconisch voorbeeld van een medisch AI-expertsysteem is het in de jaren '70 aan Stanford University ontwikkelde MYCIN, om bacteriële infecties te diagnosticeren en hiervoor antibiotica voor te schrijven. Ondanks beperkingen in gebruikersinterface en verwerkingscapaciteit, toonde MYCIN aan dat machines complexe beslissingen konden nemen, gebaseerd op opgebouwde kennis en logische redenering (Buchanan & Shortliffe, 1984; Shortliffe, 2012). Een ander medisch expertsysteem is CADUCEUS, ontwikkeld in de vroege jaren '80 aan de universiteit van Pittsburgh, en ontworpen om menselijke expertise op een bepaald gebied te simuleren door middel van beslissingsondersteunende software die gebruik maakt van kennis en inferentieregels om problemen op te lossen. Het was de opvolger van het systeem INTERNIST-I, dat was ontworpen om de diagnostische redenering van een internist na te bootsen (Miller et al.,1986). CADUCEUS was bedoeld om nog complexere diagnostische problemen in de interne geneeskunde aan te pakken. Net als andere AI-expertsystemen stuitte CADUCEUS op enkele belangrijke beperkingen. Ten eerste was de complexiteit groot en het onderhoud van de kennisbasis zeer arbeidsintensief: het voortdurend updaten en onderhouden van de uitgebreide kennisbasis vereiste aanzienlijke inspanning. Ten tweede was de integratie met klinische workflows problematisch. Het systeem was moeilijk te integreren in de dagelijkse klinische praktijk, onder andere vanwege de technologische beperkingen van die tijd. Er was eveneens weerstand om dit systeem te accepteren onder medische professionals, vooral door scepticisme ten aanzien van de betrouwbaarheid van de diagnoses en de potentiële dehumanisering van de zorg (Miller, 1984).

De tweede AI-winter, in de late jaren '80 tot midden jaren '90 van de vorige eeuw, kenmerkte zich door een gevoelige afname in interesse en investeringen in AI. Deze afname werd voornamelijk opnieuw veroorzaakt door niet-ingeloste hoge verwachtingen van AI en beperkingen in technologie bij het focussen op symbolische AI-benaderingen, waarvoor het verwerken van *ongestructureerde data* een bottleneck bleek te zijn. Als gevolg hiervan schroefden zowel overheden als bedrijven hun investeringen terug, wat leidde tot het stopzetten van projecten en sluiten van onderzoekslaboratoria. Deze periode van tegenslag stimuleerde echter ook reflectie en innovatie en leidde uiteindelijk tot nieuwe AI-benaderingen, zoals machine learning en neurale netwerken aan het einde van de jaren '90.

Data gedreven benaderingen van AI

Na de tweede AI-winter, die begin jaren '90 eindigde, onderging het veld van AI een aanzienlijke gedaanteverwisselling, waarbij nieuwe methodologieën en toepassingen werden omarmd, die hielpen om de beperkingen die tot eerdere tegenslagen hadden geleid, te vermijden. Hier volgt een overzicht van markante ontwikkelingen in AI vanaf de periode na de tweede AI-winter tot de opkomst van deep learning.

Er kwam een verschuiving van traditionele AI-modellen die gebaseerd waren op symbolische redenering en op kennis gebaseerde systemen of op een top-down benadering, naar een meer bottom-up of datagedreven benadering, zoals *behavioral AI*. Rodney Brooks, een roboticus bij MIT, stelde dat intelligentie een eigenschap is die voortvloeit uit interacties met de omgeving, leidend tot gedragsregels; dit in tegenstelling tot de traditionele aanpak waarbij kennis en logica centraal stonden. Deze bottom-up benadering werd tegenover de opvattingen van John McCarthy gezet, die een meer top-down benadering van AI voorstond, waarbij systemen werden ontworpen met expliciete kennis en redeneervermogens.

Zo besefte Brooks dat insecten snel en efficiënt in hun omgeving kunnen navigeren en bewegen, terwijl ze weinig hersenen hebben. Dit inspireerde hem tot het ontwikkelen van robots die in staat waren om in echte, ongestructureerde omgevingen als in de echte wereld te functioneren, door te reageren op sensorische inputs in plaats van te vertrouwen op uitgebreide en gedetailleerde intern gerichte voorgeprogrammeerde modellen van de wereld. Dit betekende een aardverschuiving in hoe robots werden ontworpen en ingezet. Een bekend voorbeeld van Brooks' benadering is de Roomba stofzuigerrobot, ontwikkeld door iRobot, het bedrijf dat hij mede oprichtte. Roomba navigeert en reinigt zelfstandig ruimtes met behulp van een reeks eenvoudige gedragsregels en adaptieve reacties op zijn omgeving, via *reinforcement learning*. Reinforcement learning is een type van machine learning waarbij een 'agent' via trial-and-error zelf leert hoe het taken moet uitvoeren: een 'agent' ontvangt beloningen voor correcte acties en straffen voor niet-juiste acties (Brooks & Gomez, 2015; Wilmann & Sterling, 2005; Wooldridge, 2021).

In de jaren '90 van de vorige eeuw kwam er ook een hernieuwde interesse in *probabilistische* benaderingen voor het leren omgaan met de factor onzekerheid door AI-systemen in situaties als in de echte wereld, met name door het gebruik van Bayesiaanse netwerken. De probabilistische benaderingen bieden een wiskundig framework voor het incorporeren van onzekerheid en het leren van probabilistische relaties uit data (Pearl, 1988). Een Bayesiaans netwerk kan bijvoorbeeld helpen voorspellen of iemand regenlaarzen moet dragen. Dit systeem combineert factoren zoals het weer (regenachtig of droog) en de kans dat iemand bij regen regenlaarzen draagt. Als het regent, is de kans groot dat mensen regenlaarzen dragen, maar als het droog is, is die kans veel kleiner. Het netwerk gebruikt deze probabilistische relaties om met onzekerheid om te gaan en betere voorspellingen te doen.

Tussen 1990 en 2014 bereikte AI een niveau van volwassenheid, gekenmerkt door de integratie van verschillende van de hiervoor besproken technieken in robuustere systemen. De oprichting van DeepMind Technologies in 2010 en de latere ontwikkelingen van deep learning markeerden een belangrijk keerpunt. DeepMind Technologies is een Brits bedrijf gespecialiseerd in AI. Sinds de overname door Google in 2014 is DeepMind uitgegroeid tot een van de wereldleiders op het gebied van AI-onderzoek. Het bedrijf is vooral bekend vanwege de ontwikkeling van AlphaGo, een AI-programma dat in 2016 een wereldkampioen in het eeuwenoude Chinese bordspel Go versloeg, een belangrijke mijlpaal die de mogelijkheden van AI aantoonde. Deep learning maakt gebruik van grote neurale netwerken met vele lagen (diepe netwerken) om functies te leren afleiden uit grote hoeveelheden data, wat leidt tot ongekende prestaties in taken, variërend van beeldherkenning en natuurlijke taalverwerking tot zelfrijdende auto's (Silver et al., 2016).

De opkomst van deep learning ging gepaard met grote verbeteringen in hardware, zoals de toename in rekenkracht van processoren en de ontwikkeling van de grafische kaart (Graphics Processing Units of GPU's), die bijzonder goed geschikt zijn voor het uitvoeren van parallelle taken die essentieel zijn in AI-berekeningen. Het uitvoeren van parallelle taken in deep learning verwijst naar het gelijktijdig verwerken van meerdere berekeningen bij het trainen van neurale netwerken. In een neuraal netwerk worden veel berekeningen tegelijk uitgevoerd, zoals het vermenigvuldigen van grote hoeveelheden getallen en het bijwerken van duizenden gewichten in de neuronenlagen. Daarnaast leidden ontwikkelingen in algoritmes, zoals die in neurale netwerken of deeplearning-technieken, tot verbeterde prestaties in data-analyse en patroonherkenning.

De wederopstanding van AI in de gezondheidszorg kwam er met het doorbreken van machine learning en deep learning begin 2000, die via beeldherkenning voor significante verbeteringen in de diagnostische beeldvorming zorgden. Systemen zoals Google's DeepMind ontwikkelden in samenwerking met onderzoekers van Cancer Research UK Imperial Centre, Northwestern University en Royal Surrey County Hospital

algoritmes die in staat zijn om borstkanker te identificeren in mammogrammen, met een nauwkeurigheid die competitief is met die van radiologen. Dit markeerde een keerpunt in hoe AI werd ingezet voor diagnostische precisie (Massat, 2018).

In tegenstelling tot het eerder beschreven systeem dat zich focust op beeldherkenning, onderscheidt IBM Watson zich door toepassing van NLP. Gelanceerd in 2011 na een overwinning in 'Jeopardy!' (een Amerikaans quizprogramma, waarin deelnemers vragen moesten bedenken bij gegeven antwoorden), illustreert Watson een belangrijk tijdperk in de AI-evolutie. Watson nam het op tegen twee van de beste menselijke Jeopardy-spelers en won overtuigend. In dit spel moesten vragen in natuurlijke taal snel en nauwkeurig worden begrepen, waarna de juiste antwoorden in natuurlijke taal moesten worden geformuleerd (IBM, 2011).Watsons expertise in NLP en diverse vormen van machine learning, waaronder deep learning, vormt een directe verbetering ten opzichte van de vroege expertsystemen en zij wordt ook toegepast in de gezondheidszorg. In Watson for Oncology biedt deze expertise behandelopties door data-analyse en in Watson Imaging Clinical Review verbetert zij de precisie van medische diagnostiek op basis van machine- en deeplearningmodellen die zijn getraind op medische beelden (Park et al., 2023; Strickland, 2019). In volgende paragrafen wordt een overzicht gegeven van de verschillende vormen van deep learning en de werking ervan, die uiteindelijk leidde tot de opkomst van het transformermodel.

Deep learning

Het transformermodel is een vrij recent (2017) gecreëerde vorm van deep learning. Om een transformermodel doeltreffend te kunnen toepassen, is het van belang om de onderliggende deeplearningconcepten ervan te begrijpen. Ik verduidelijk dit in volgende paragrafen voor NLP.

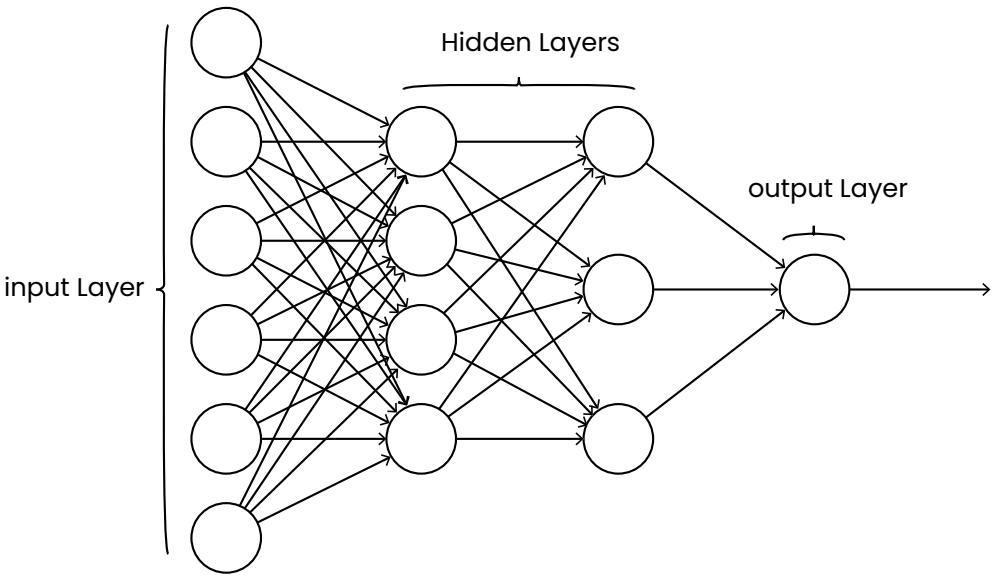
Deeplearningmodellen boden geleidelijk aan verbeterde oplossingen voor het modelleren van context in langere teksten, wat een van de grote uitdagingen is in het domein van NLP. Kortom, de automatische verwerking van uitgebreide tekst verbeterde gestaag, resulterend in het transformermodel als het hoogtepunt van deze evolutie.
Transformermodellen zijn een vorm van deep learning, dat op zijn beurt een onderdeel van machine learning is. Deep learning richt zich op diepe neurale netwerken, geïnspireerd door de manier waarop de hersenen informatie verwerken. De grondleggers van deep learning, onder wie Geoffrey Hinton, Yann LeCun en Yoshua Bengio, hebben bijgedragen aan de ontwikkeling van complexe algoritmes die leren van data zonder expliciete instructies te volgen (LeCun et al., 2015).

De impact van deep learning werd onder andere gedemonstreerd met AlexNet in 2012. Dit deeplearningmodel is ontwikkeld door Alex Krizhevsky, Ilya Sutskever en Geoffrey Hinton en domineerde de ImageNet-competitie, een challenge in beeldherkenning.

Het feit dat AlexNet de challenge won, toonde niet alleen de superioriteit van diepe neurale netwerken in beeldanalyse, maar het was ook een mijlpaal in de erkenning van hun potentieel over een breed spectrum van AI-toepassingen (Krizhevsky et al., 2017).

Architectuur van het deeplearningmodel

De architectuur en het paradigma van diepe neurale netwerken zijn gebaseerd op het functioneren van neuronen in de hersenen. Net zoals neuronen in de hersenen met elkaar communiceren, vormen kunstmatige lagen van neuronen in diepe neurale netwerken verbindingen met elkaar. Een deeplearningmodel verwerkt gegevens in die verschillende lagen (Figuur 1). Het begint met de *input layer*, waar de ruwe gegevens binnenkomen, zoals pixelwaarden van een afbeelding. Deze gegevens worden vervolgens doorgegeven aan één of meer *hidden layers*, die verantwoordelijk zijn voor het leren van complexe patronen en structuren in de data. Elke neuron in een hidden layer berekent een gewogen som van de inputs en past een activatiefunctie toe om te bepalen of de informatie moet worden doorgegeven aan de volgende laag. Uiteindelijk komt de informatie terecht in de *output layer*, die de uiteindelijke voorspelling of classificatie geeft, zoals het herkennen van een object in een afbeelding of het voorspellen van een waarde. Hoe meer *layers* het netwerk bevat, hoe dieper de informatie gaat; vandaar de naam *deep learning*.



Figuur 1. Architectuur van een neuraal netwerk (Goyal et al., 2018).

In de jaren '40 van de vorige eeuw beseften Warren McCulloch en Walter Pitts dat neuronen gemodelleerd konden worden als elektrische circuits, specifiek als eenvoudige logische circuits (McCulloch & Pitts, 1943). Ze gebruikten dit concept om een veelzijdig wiskundig model van neuronen te ontwikkelen. In de jaren '50 werd dit model verfijnd door Frank Rosenblatt, die het omvormde tot het neurale netwerkmodel (perceptron genaamd), dat gebaseerd is op de werking en interactie van neuronen in de hersenen (Rosenblatt, 1958). Het is het eerste neurale netwerkmodel dat daadwerkelijk werd geïmplementeerd en vandaag de dag nog steeds relevant is.
Diepe neurale netwerken zijn in staat om hiërarchische en niet-lineaire representaties te leren, wat hen geschikt maakt voor het uitvoeren van taken, zoals beeldherkenning, spraakherkenning en natuurlijke taalverwerking. Hiërarchische representaties stellen de netwerken in staat om informatie te verwerken op verschillende abstractieniveaus; bijvoorbeeld in beeldherkenning kan een laag zich focussen op het herkennen van randen en een andere laag op het herkennen van hele objecten. Niet-lineaire representaties maken het mogelijk om complexe relaties in de data te modelleren. Een eenvoudig voorbeeld van een niet-lineaire functie is de activatiefunctie, zoals de Rectified Linear Unit (ReLU) (Nair & Hinton, 2010), die helpt bij het nemen van beslissingen door alleen positieve waarden door te geven van de ene naar de andere neuronlaag en negatieve waarden op nul te zetten. Dit soort functies stelt neurale netwerken in staat om verder te gaan dan het gebruiken van simpele, lineaire beslisregels, en complexe patronen in data te herkennen en te gebruiken. Deze benadering heeft geleid tot opmerkelijke doorbraken in verschillende domeinen, met deep learning als krachtig instrument voor het automatisch leren van representaties uit grote hoeveelheden data.

De evolutie van deep learning door uitdagingen in contextmodellering

Binnen het domein van NLP werd de evolutie naar het transformermodel binnen deep learning vooral gedreven door de voortdurende uitdaging om context op een effectieve manier te modelleren, met name bij linguïstische taken. Nog voordat deeplearningmodellen aan hun opmars begonnen, startte deze reis met statistische traditionele modellen, zoals Markov-modellen (Markov, 2006) en Conditional Random Fields (CRF's) (Lafferty et al., 2001), die hun beperkingen hadden in het vastleggen van complexe contextuele relaties. Een Markov-model is een statistisch model dat de overgang van de ene toestand naar de andere beschrijft op basis van de huidige toestand. In de zin '*"Kunt u mij alstublieft vertellen welke kant ik op moet", vroeg Alice*', waarin Alice de spreker is, ligt de uitdaging voor een NLP-systeem in het begrijpen dat de woorden 'mij' en 'ik' naar 'Alice' verwijzen. Dit proces staat bekend als coreferentieresolutie. Een Markov-model bepaalt de waarschijnlijkheid van elk volgend woord op basis van *alleen* het voorgaande woord in de zin. Net als een Markov-model is een CRF eveneens een probabilistisch grafisch model dat de afhankelijkheden tussen variabelen kan modelleren, maar al meer dan een Markov-model rekening houdt met

de contextuele informatie. In dezelfde voorbeeldzin van Alice modelleert een CRF de onderlinge relaties tussen woorden beter, zodat het de meest waarschijnlijke reeks woorden kan voorspellen op basis van de gegeven context.

Convolutional Neural Networks

Contextmodellering ging sterk vooruit met de intrede van deeplearningmodellen. Deeplearningmodellen moeten wel met meer data moeten gevoed worden dan meer traditionele machinelearningmodellen. Het Convolutional Neural Network (CNN) is een deeplearningmodel dat oorspronkelijk ontworpen is door Yann LeCun in de jaren '80 van de vorige eeuw om visuele data te verwerken, zoals afbeeldingen. De architectuur van het netwerk is geïnspireerd op de manier waarop het menselijk visuele systeem werkt. CNN's maken gebruik van convolutionele lagen. Dit zijn lagen die patronen of kenmerken uit afbeeldingen halen, zoals randen, vormen en texturen. Deze lagen gebruiken zogenaamde 'filters' of 'kernels', die over de afbeelding schuiven (convolueren) om specifieke details te detecteren. In 1998 presenteren LeCun et al. LeNet, een CNN dat cijfers uit een handschrift automatisch detecteert (LeCun et al., 1998).

Na aanpassingen zijn CNN's nu ook in staat om tekstuele informatie te verwerken door gebruik te maken van de sequentiële aard van tekst. In plaats van 2D-convoluties, passen ze 1D-convoluties toe op tekstreeksen om lokale tekstuele patronen vast te leggen. De eerder gebruikte voorbeeldzin *'"Kunt u mij alstublieft vertellen welke kant ik op moet", vroeg Alice'* kan dus ook worden verwerkt door een tekstgeoriënteerd CNN. In dit geval zouden de convolutielagen leren om specifieke woordcombinaties of zinsstructuren te herkennen die belangrijk zijn voor het begrijpen van de betekenis van de zin. De hieruit voortgekomen representaties kunnen vervolgens worden gebruikt voor taken zoals sentimentanalyse, tekstclassificatie of andere toepassingen waarbij tekstuele informatie van belang is. Het idee is dus dat de CNN-architectuur, oorspronkelijk ontworpen voor visuele data, flexibel genoeg is om toegepast te worden op verschillende soorten gegevens, inclusief tekstuele informatie.

Recurrent Neural Networks

Terwijl CNN's goed zijn in het vastleggen van lokale patronen in gestructureerde data zoals afbeeldingen, schieten ze soms toch tekort bij het begrijpen van de sequentiële aard van taal. Hier komt de kracht van Recurrent Neural Networks (RNN's) naar voren. In tegenstelling tot CNN's kunnen RNN's effectief omgaan met de dynamische en sequentiële aard van tekstuele informatie, waarbij ze contextuele afhankelijkheden en langeafstandrelaties in tekst beter begrijpen, wat belangrijk is voor taken in NLP. Het is een type neurale netwerkarchitectuur dat is ontworpen om sequentiële informatie beter te verwerken en complexere context beter te begrijpen (Hopfield, 1982). In de eerder gebruikte voorbeeldzin *'"Kunt u mij alstublieft vertellen welke kant ik op moet", vroeg Alice',* zou een RNN de opeenvolgende woorden één voor één verwerken, waarbij

het rekening houdt met de context van elk woord in relatie tot het voorgaande. Hierdoor kan het model informatie over de chronologische volgorde vastleggen en de samenhang tussen woorden begrijpen binnen de gegeven zin. Echter, het RNN ondervindt problemen met de parallelle verwerking en het vasthouden van informatie over langere periodes. Met andere woorden, hoe langer de te verwerken zin, hoe meer geheugenproblemen het ondervond om informatie over langere sequenties te onthouden.

Long Short-Term Memory en Gated Recurrent Unit

Long Short-Term Memory (LSTM) en Gated Recurrent Unit (GRU) zijn varianten van het RNN. LSTM en GRU zijn ontworpen om tekortkomingen in het kortetermijngeheugen op te lossen en om beter om te kunnen gaan met het probleem van langeafstandsrelaties in sequentiële data, zoals lange zinnen.

De LSTM is een speciaal soort neuraal netwerk dat informatie voor langere tijd vast kan houden. Het is te vergelijken met een slim notitieboekje, dat beslist wat belangrijk genoeg is om te onthouden, wat bijgewerkt moet worden en wat vergeten kan worden. Dit wordt mogelijk gemaakt door drie speciale 'poorten': een *vergeetpoort* beslist welke informatie niet langer relevant is en dus uit het geheugen kan worden gewist, een *invoerpoort* voegt nieuwe, belangrijke informatie toe en een *uitvoerpoort* bepaalt welke gegevens uit de opgeslagen informatie op een bepaald moment gebruikt moeten worden.

Een van de grootste uitdagingen in het trainen van neurale netwerken is iets wat het *vanishing gradient problem* of *probleem van het verdwijnende gradiënt* (Lemaréchal, 2012) heet. De 'gradient' is een maat voor hoe snel de output van een model verandert als je de input of parameters aanpast. Bij het vanishing gradient-probleem worden de veranderingen zo klein dat het netwerk bijna niet meer leert. Dit doet zich voor wanneer het netwerk zo diep is dat het moeilijk wordt voor dat netwerk om te leren van fouten, omdat de signalen die helpen bij het leren, zwakker worden naarmate ze door meer lagen van het netwerk gaan. Dit is alsof instructies zo zacht worden uitgesproken op de bovenste verdieping van een gebouw dat tegen de tijd dat ze de onderste verdieping bereiken, ze bijna onhoorbaar zijn. Dit maakt het bijzonder lastig voor het netwerk om langetermijnafhankelijkheden, of de relaties tussen dingen die ver uit elkaar liggen in de tijd, te leren.

LSTM's zijn speciaal ontworpen om het vanishing gradient problem aan te pakken. Door hun bijzondere structuur kunnen ze langdurige relaties herkennen en behouden, waardoor ze geschikt zijn voor taken als taalmodellering en spraakherkenning, waar het begrijpen van langdurige context cruciaal is (Hochreiter & Schmidhuber, 1997).

Een GRU is een soort versimpelde versie van de LSTM. Ze zijn beide ontworpen om neurale netwerken te helpen leren van informatie over lange perioden. De GRU is op te vatten als een efficiënte assistent die precies weet wanneer oude notities belangrijk zijn om te bewaren en wanneer het tijd is om plaats te maken voor nieuwe. De GRU maakt dit mogelijk met twee 'poorten': een updatepoort en een resetpoort. De *update-poort* werkt als een filter die bepaalt hoeveel van de eerder opgedane ervaringen of kennis behouden moet blijven voor toekomstig gebruik. Het is alsof iemand een boekenplank opruimt en beslist welke boeken waardevol genoeg zijn om te bewaren voor later. De *resetpoort*, aan de andere kant, helpt beslissen hoeveel van de oude informatie gewist moet worden om ruimte te maken voor nieuwe lessen en inzichten. Dit is vergelijkbaar met de beslissing om sommige pagina's uit een vol notitieboek te scheuren omdat ze niet meer relevant zijn.

Een LSTM daarentegen werkt meer als een hoofdarchivaris. Bij elk boek dat binnenkomt, stelt hij drie vragen: "Moeten we dit boek überhaupt binnenlaten (invoerpoort)?", "Moeten we dit boek op de hoofdplank bewaren of naar het archief sturen (vergeetpoort)?" en "Welke boeken moeten we vooral naar voren halen (uitvoerpoort)?" De LSTM heeft als het ware meer controle over de informatiestroom, maar het maken van deze keuzes duurt ook langer. Bovendien kan de LSTM complexere relaties over tijd beter begrijpen, terwijl de GRU sneller beslist met minder vragen.

Door deze aanpak is de GRU zowel eenvoudiger te hanteren als effectief in het vastleggen van langdurige informatie. Een GRU is ideaal voor taken zoals taal begrijpen en voorspellingen doen op basis van voorgaande gebeurtenissen, omdat mooi het evenwicht bewaard wordt tussen het onthouden van nuttige informatie en het vrijmaken van ruimte voor nieuwe gegevens. Een GRU presteert vaak vergelijkbaar met een LSTM, maar is efficiënter in termen van rekenkracht (Cho et al., 2014).

Door een LSTM of GRU toe te voegen aan een RNN, kan dit model effectiever omgaan met langeafstandsrelaties in zinnen en context beter modelleren. Het systeem leert welke informatie moet worden vastgehouden, vergeten of bijgewerkt, waardoor het geschikter is voor taken waarbij het begrijpen van context over langere sequenties cruciaal is, zoals bij natuurlijke taalverwerking.

Het aandachtsmechanisme

Terwijl met een LSTM en GRU verbeteringen in de architectuur werden doorgevoerd, om het probleem van het verdwijnende gradiënt aan te pakken en langetermijnafhankelijkheden in sequentiële gegevens te verbeteren, heeft de evolutie van het RNN geleid tot het *aandachtsmechanisme* of *attention mechanism*. Dit mechanisme heeft als doel de prestaties van RNN's verder te optimaliseren door het mogelijk te maken dat ze zich meer concentreren op relevante delen van de invoersequenties (Bahdanau et al., 2014).

Voor een mens is het begrip 'aandacht' intuïtief en vanzelfsprekend; mensen kunnen zich automatisch richten op relevante informatie in hun omgeving. Interessant is dat vóór de opkomst van het aandachtsmechanisme deeplearningsystemen vaak bottom-up redeneerden, terwijl het aandachtsmechanisme een meer top-down-benadering mogelijk maakt, waarbij het model bij het doen van voorspellingen selectief aandacht kan schenken aan specifieke delen van de invoer. Met het integreren van een aandachtsmechanisme in deeplearningsystemen zijn deze dus in staat om zich te concentreren op relevante delen van de inputsequenties, wat vooral een rol speelt bij taken waar context-langeafstandafhankelijkheden een rol spelen, zoals lange zinnen of teksten bij natuurlijke taalverwerking. Bij context-langeafstand-afhankelijkheden gaat het erom dat een model woorden of informatie uit een eerdere zin of ver terug in de zin of tekst moet onthouden om de betekenis van het huidige woord goed te begrijpen. Het aandachtsmechanisme helpt hierbij door het model op belangrijke woorden te laten focussen, zelfs als er veel andere informatie tussen staat. Dit mechanisme biedt dus een manier om deeplearningsystemen op een meer mensachtige wijze informatie te laten verwerken.

Het toevoegen van het aandachtsmechanisme verbetert het vermogen van een RNN om context te modelleren in sequentiële data, zoals zinnen, aanzienlijk. In een standaard-RNN, zonder aandachtsmechanisme, verwerkt het netwerk elk woord in de zin sequentieel, waarbij de voorspelling van een woord afhankelijk is van het vorige woord. Dit kan echter problemen opleveren bij het vastleggen van langetermijn-afhankelijkheden en het begrijpen van de context van woorden die verder weg in de sequentie liggen. Het aandachtsmechanisme introduceert een flexibel mechanisme waarmee het model zich op specifieke delen van de inputsequentie kan concentreren. Bij elke stap berekent het aandachtsmechanisme een gewogen gemiddelde van alle eerdere verwerkingsstappen, waarbij het gewicht wordt bepaald door de relevantie van elk woord in relatie tot het huidige woord. Het model berekent met andere woorden een soort *belangrijkheidsscore* voor elk woord in relatie tot het woord dat het model op dat moment probeert te begrijpen of te vertalen. Woorden die meer relevant zijn krijgen een hogere score, en zo kan het model een gewogen gemiddelde nemen, waarbij het meer aandacht besteedt aan belangrijkere woorden.

Door dit aandachtsmechanisme kunnen RNN-modellen zich nu beter focussen op woorden die het meest relevant zijn voor het begrijpen van de context op een gegeven moment. Hierdoor kunnen ze de samenhang tussen woorden beter begrijpen, zelfs als deze woorden verder weg in de sequentie liggen. Bij het verwerken van de eerder gebruikte voorbeeldzin '"*Kunt u mij alstublieft vertellen welke kant ik op moet*", *vroeg Alice'*, zou het aandachtsmechanisme zich kunnen concentreren op 'vertellen' en 'kant', waardoor het model beter in staat is om de context en de onderlinge relaties tussen deze woorden vast te leggen.

De opkomst van het grote taalmodel en transformermodel

2

Het aandachtsmechanisme (zie hoofdstuk 1) verbeterde de wijze waarop Recurrent Neural Networks (RNN's) de relatie tussen woorden begrijpen, door selectief te focussen op slechts enkele woorden, of anders gezegd: op de meest informatieve delen van de input. Dit opende de deur naar nog meer vernieuwing binnen deep learning, namelijk met het *transformermodel*, waar het concept van 'attention', met name *self-attention*, centraal staat. Zowel attention- als self-attentionmechanismen worden in neurale netwerken gebruikt om de relevantie van verschillende delen van de inputsequenties te bepalen.

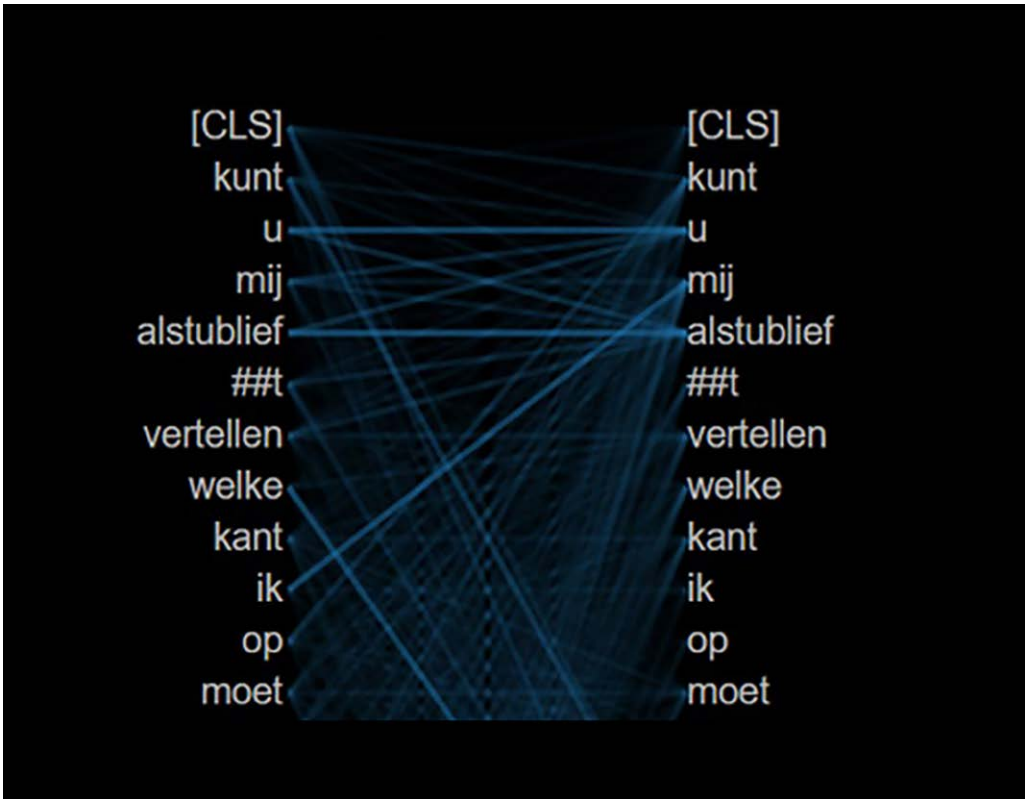
Attentionmechanisme en self-attentionmechanisme

Het attentionmechanisme stelt een computermodel in staat de belangrijkheid van woorden te identificeren bij taken zoals vertalen. Het model kijkt naar elk woord in de zin die vertaald moet worden en bepaalt vervolgens de relevantie van andere woorden in de originele zin voor een nauwkeurige vertaling. Op basis van deze relevantie combineert het model de woorden op een doordachte manier. Bij taken zoals vertalen maakt dit mechanisme het mogelijk dat het model zich focust op bepaalde delen van de zin, afhankelijk van het segment dat het vertaalt.

Self-attention is een sleutelbegrip in de architectuur van transformermodellen. Self-attention gaat echter nog een stap verder dan attention. Dit mechanisme bekijkt niet alleen sequentieel de relaties tussen bepaalde woorden in verschillende zinnen, of de relaties tussen enkele woorden binnen dezelfde zin, maar kijkt ook *parallel* naar *alle* woorden binnen een sequentie of meerdere sequenties. Deze aanpak verfijnt het inzicht in de verbanden tussen woorden, wat heeft geleid tot spectaculaire verbeteringen in de kwaliteit van NLP-taken (Rothman, 2022).

Door het self-attentionmechanisme is het mogelijk geworden om langetermijn-afhankelijkheden in sequenties en complexe relaties te modelleren door dynamisch en parallel de aandacht te richten op verschillende delen van de inputsequenties. Dit illustreer ik met nog eens dezelfde voorbeeldzin '*"Kunt u mij alstublieft vertellen welke kant ik op moet", vroeg Alice*'. Voorgangers van het transformermodel zouden er moeite mee hebben om in een coreferentieresolutie-taak de contextuele verbindingen tussen 'mij', 'ik' en 'Alice' te begrijpen. Het transformermodel kan dankzij self-attention

deze relaties modelleren en een dieper begrip ontwikkelen van de samenhang tussen deze woorden. De visualisatietechniek van Vig (2019) illustreert hoe een transformer-model coreferentieresolutie uitvoert. In Figuur 2 worden relaties tussen woorden binnen een zin aangeduid met blauwe lijnen. Hoe dikker de lijn, hoe sterker de attention of self-attention van het model. Een opvallend dikke lijn verbindt de woorden 'mij' en 'ik'; dit demonstreert het redeneerproces van het model voor de coreferentieresolutie-taak en het begrijpen dat de twee woorden aan elkaar gelinkt zijn en beide naar 'Alice' verwijzen.



Figuur 2. Self-attention in het transformermodel

De kracht van self-attention komt voort uit de manier waarop transformermodellen getraind worden. Transformermodellen maken gebruik van grootschalige unsupervised training op enorme hoeveelheden tekstdata, zoals beschreven in het baanbrekende wetenschappelijke artikel 'Attention is All You Need' van Vaswani et al. (2017).Tijdens dit trainingsproces leren de modellen patronen en relaties tussen woorden zonder expliciete labels. Vaswani ontwikkelde samen met zijn collega's bij Google Research het transformermodel. Bij unsupervised training worden transformermodellen, zonder

menselijke tussenkomst, blootgesteld aan grote hoeveelheden zinnen zonder specifieke annotaties die aangeven welke woorden belangrijk zijn of hoe ze aan elkaar gelinkt zijn. Door herhaaldelijk te worden geconfronteerd met verschillende contexten en zinsstructuren, leren de modellen *zelf* ontdekken welke woorden in welke contexten van belang zijn en hoe ze met elkaar in verband staan.

De rol van rekenkracht bij de creatie van transformermodellen

De indrukwekkende prestaties van transformermodellen zijn het gevolg van de grote rekenkracht van moderne hardware, zoals Graphics Processing Units (GPU's) en meer recentelijk de Tensor Processing Units (TPU's). Deze krachtige hardwarecomponenten kunnen de grootschalige parallelle berekeningen uitvoeren die noodzakelijk zijn voor het trainen van transformermodellen.

GPU's, oorspronkelijk ontworpen voor grafische verwerking, zijn bijzonder geschikt voor het uitvoeren van de vele matrixvermenigvuldigingen die aan de basis liggen van deeplearningalgoritmes. Hun vermogen om duizenden berekeningen tegelijkertijd uit te voeren, maakt het mogelijk om grote hoeveelheden data snel te verwerken en complexe modellen zoals transformermodellen efficiënt te trainen. Dit parallelle verwerkingsvermogen versnelt het trainingsproces aanzienlijk in vergelijking met traditionele Central Processing Units (CPU's), die minder goed zijn in parallelle verwerking en die doorgaans worden aangetroffen in standaardcomputers (Du et al., 2023).

In plaats van lokale hardware en servers te moeten aanschaffen, kunnen gebruikers via *cloudcomputingplatforms* toegang krijgen tot krachtige GPU's. Deze platforms bieden flexibele en schaalbare mogelijkheden om modellen sneller te trainen en te testen, zonder de kosten van fysieke hardware. Cloud-based servers zoals Google Colab[1], Amazon SageMaker[2] en Microsoft Azure[3] bieden elk unieke voordelen en nadelen in het gebruik van GPU's in deep learning. Google Colab is bijzonder aantrekkelijk vanwege de gratis toegang tot GPU's, wat het ideaal maakt voor beginners en kleinere projecten. Het platform is eenvoudig te gebruiken en integreert naadloos met Google Drive, maar de gratis versie heeft beperkingen in rekenkracht en sessieduur, waardoor het minder geschikt is voor grote of langdurige trainingssessies. Amazon SageMaker daarentegen biedt een robuust platform voor het ontwikkelen en implementeren van machinelearningmodellen op schaal, met krachtige GPU-instanties en een breed scala aan tools voor modelbeheer. Het nadeel van SageMaker zit in de kosten; aangezien het een pay-as-you-go-model gebruikt, kunnen de kosten snel

1 https://colab.research.google.com/

2 https://aws.amazon.com/sagemaker/

3 https://portal.azure.com/

oplopen, afhankelijk van het gebruik. Microsoft Azure biedt eveneens flexibele GPU-resources binnen een uitgebreide cloudinfrastructuur en is goed geïntegreerd met andere Microsoft-diensten, wat het aantrekkelijk maakt voor grote ondernemingen. Echter, net als bij SageMaker, kunnen de kosten van Azure hoog zijn, vooral voor langdurige of zeer intensieve taken. Hoewel Amazon SageMaker en Microsoft Azure krachtigere en meer schaalbare oplossingen bieden, zijn ze meestal duurder en complexer in gebruik, wat ze minder geschikt maakt voor het onderwijs, vooral op basis- of intermediate niveau. Google Colab biedt daarentegen een uitstekende balans tussen functionaliteit en toegankelijkheid voor educatieve doeleinden (Education Ecosystem, 2021).

Daarnaast hebben softwarebibliotheken, zoals de deeplearningframeworks TensorFlow[4] en PyTorch[5], de implementatie van transformermodellen vereenvoudigd en geoptimaliseerd, waardoor ze toegankelijker zijn voor onderzoekers en ontwikkelaars. Deze bibliotheken en frameworks benutten de rekenkracht van GPU's en TPU's volledig, wat noodzakelijk is voor het omgaan met de enorme datasets die nodig zijn voor het creëren van transformermodellen (Abadi et al., 2016; Paszke et al., 2019).

In essentie: zonder de vooruitgang in rekenkracht en de beschikbaarheid van gespecialiseerde hardware zou de doorbraak in NLP die door deeplearningmodellen en transformermodellen wordt aangedreven, veel moeilijker te realiseren zijn geweest. Het is de synergie tussen nieuwe algoritmes en krachtige hardware die de huidige prestaties en efficiëntie van NLP-toepassingen mogelijk maakt.

De toepassing van het transformermodel heeft een paradigmaverschuiving teweeggebracht in NLP, door een flexibele, parallelle benadering van contextmodellering mogelijk te maken en zo grote verbeteringen te bieden in het uitvoeren van diverse NLP-taken.

Algemene en domeinspecifieke grote taalmodellen

Grote en algemene taalmodellen, zoals Bidirectional Encoder Representations from Transformers (BERT) (Vaswani et al., 2017) vormden een doorbraak in NLP. Ze zijn beter in staat om complexere contextuele informatie te verwerken dan de eerdere neurale netwerkmethodes. NLP-toepassingen die worden aangedreven door grote taalmodellen, vertonen aanzienlijk verbeterde prestaties bij taken zoals automatische tekstclassificatie en vraagbeantwoording. De massa aan data waarmee een BERT-model wordt gecreëerd, stelt het model in staat ingewikkelde taalpatronen te identificeren en te begrijpen, waardoor het waardevol is voor allerlei concrete NLP-toepassingen (Devlin et al., 2019). Er bestaan algemene taalmodellen voor

4 https://www.tensorflow.org/

5 https://pytorch.org/

verschillende talen, zoals BERT voor het Engels, en RobBERT (Delobelle et al., 2020) en BERTje (De Vries et al., 2019) voor minder goed ondersteunde talen of *under-resourced languages* zoals het Nederlands.

Under-resourced languages zijn talen waarvoor weinig digitale data beschikbaar zijn, zoals tekstcorpora, annotaties, of taalkundige bronnen. Dit maakt het moeilijker om goed presterende taalmodellen te trainen, omdat er minder trainingsmateriaal is dan voor grote talen zoals het Engels. Voorbeelden hiervan zijn talen die door minder mensen worden gesproken of regionale dialecten, maar ook talen zoals het Nederlands kunnen als under-resourced worden beschouwd in vergelijking met het Engels.

Het Large Language Model Meta AI ofwel LLaMa (Touvron et al., 2023) behoort tot een nieuwe generatie taalmodellen, gebaseerd op een transformerarchitectuur. Het LLaMa-model heeft versies die meer parameters bevatten dan het BERT-model. BERT heeft, afhankelijk van de variant (zoals BERT-Base of BERT-Large), tussen de 110 miljoen en 340 miljoen parameters. Een parameter verwijst in deze context naar de variabelen van het model die worden aangepast (getraind) tijdens het leerproces, zodat het model patronen in data te leert herkennen. Deze parameters leren het model hoe het moet reageren op verschillende soorten input. Hoe meer parameters een model heeft, hoe complexer de taken zijn die het leert uitvoeren, maar ook hoe meer data en rekenkracht nodig zijn voor effectief trainen. LLaMA-modellen zijn beschikbaar in verschillende groottes, met minder of meer parameters, variërend van enkele tientallen miljoenen tot enkele tientallen miljarden parameters. Deze nieuwe generatie modellen heeft vaak minder extra data nodig om verder geoptimaliseerd te worden voor een specifieke NLP-taak dan BERT-modellen (Dunn et al., 2022).

Domeinspecifieke grote taalmodellen, zoals BioBERT voor de biomedische sector (Lee et al., 2020), SciBERT voor wetenschappelijke teksten (Beltagy et al., 2019) en Med-BERT (Rasmy et al., 2021), ClinicalBERT (Huang et al., 2019) en Med-PaLM (Singhal et al., 2023) voor de medische sector, zijn afgestemd op het domein van het elektronisch patiënten-dossier (EPD) en zijn meestal een kleinere versie dan het algemene grote taalmodel BERT-framework (He et al., 2023). Gemaakt op basis van een EPD-dataset, biedt een dergelijk model een genuanceerd en contextbewust begrip van gespecialiseerde terminologie, waarvoor meestal minder gegevens nodig zijn dan voor hun algemene tegenhangers. MedRoBERTa is een van de weinige domeinspecifieke medische taalmodellen in het Nederlands (Verkijk & Vossen, 2021), getraind op Nederlandse EPD's.

Multimodale grote taalmodellen

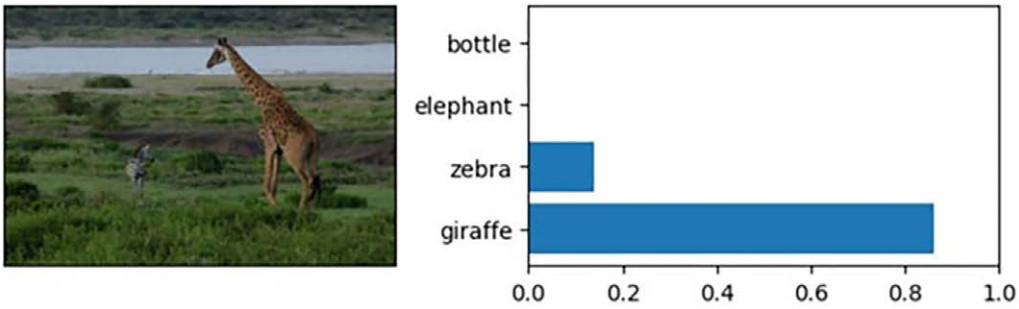
Multimodale grote taalmodellen vertegenwoordigen de nieuwste innovatie in de wereld van AI. Deze modellen zijn niet alleen in staat om complexe tekstuele informatie te begrijpen en te genereren, maar kunnen ook omgaan met andere vormen van data, zoals afbeeldingen of audiodata. Dit stelt deze modellen in staat om taken uit te voeren die een combinatie van verschillende soorten input vereisen, zoals het beantwoorden van vragen over de inhoud van een foto of het creëren van visuele content op basis van tekstuele beschrijvingen. Door de integratie van diverse gegevensbronnen bieden deze modellen een rijker en meer contextueel begrip, wat kan leiden tot meer accurate en relevante resultaten in een breed scala van toepassingen.

Een van deze multimodale modellen is Contrastive Language Image Pretraining (CLIP). CLIP is ontwikkeld met OpenAI[6], met als doel om de kloof tussen visuele beelden en tekstuele beschrijvingen te overbruggen. Door gebruik te maken van een *contrastieve leerbenadering*, leert CLIP directe verbanden te leggen tussen afbeeldingen en de bijbehorende tekstuele omschrijvingen.

Een contrastieve leerbenadering leert een model om onderscheid te maken tussen wat wel en niet bij elkaar hoort door paren van gegevens te vergelijken. In het geval van CLIP betekent dit dat het model leert dat een afbeelding en een bijbehorende tekst beschrijving bij elkaar horen (positief paar), terwijl andere combinaties (zoals een willekeurige tekst bij een afbeelding) niet passen (negatief paar). Het model wordt zo getraind om de overeenkomst tussen gerelateerde paren te maximaliseren en de overeenkomst tussen ongerelateerde paren te minimaliseren. Hierdoor kan CLIP beter begrijpen welke tekst het beste bij een afbeelding past, en andersom.

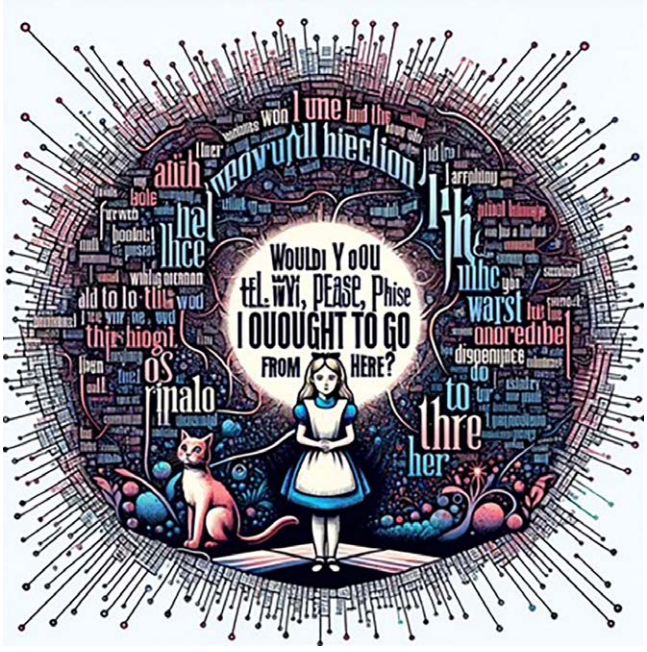
Dit stelt het model in staat om een breed scala aan taken uit te voeren, zoals het zoeken van afbeeldingen op basis van tekstuele zoekopdrachten, het genereren van een tekst die een afbeelding beschrijft en zelfs het begrijpen van complexe visuele concepten die in natuurlijke taal worden uitgedrukt. CLIP is getraind op een enorme dataset van 400 miljoen afbeeldingen en teksten van het internet, waardoor het een breed begrip heeft ontwikkeld van visuele en tekstuele data (Radford et al., 2021). Figuur 3 toont dat CLIP op een foto, waarop een giraf en een kleiner afgebeelde zebra staan, ‘giraf’ met een waarschijnlijkheid van 0,8 en ‘zebra’ met een waarschijnlijkheid van 0,2 correct identificeert.

6 https://openai.com/



Figuur 3. Identificatie van dieren door het multimodale CLIP-model

CLIP is niet een generatief model, omdat het voornamelijk is ontworpen om de relatie tussen afbeeldingen en tekstuele beschrijvingen te begrijpen en te interpreteren, niet om nieuwe inhoud te genereren. Het Dall-e-model van OpenAI is wel een generatief model, het genereert afbeeldingen op basis van tekst. Figuur 4 toont een door Dall-e gegenereerde afbeelding, gecreëerd in reactie op het verzoek om een visuele voorstelling te maken van een taalmodel dat de Engelse zin '"*Would you tell me, please, which way I ought to go from here?*"', uitgesproken door Alice, verwerkt. Merk op dat de woorden in de afbeelding niet correct zijn weergegeven.



Figuur 4. Door Dall-e gegenereerde afbeelding op basis van tekst

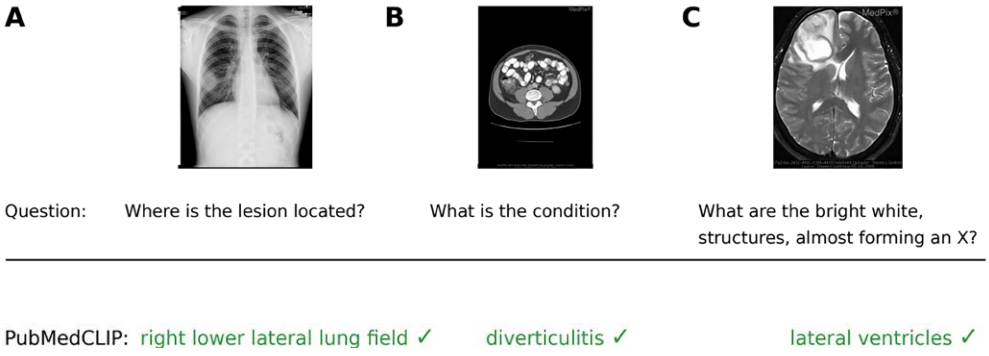
In tegenstelling tot generatieve modellen, die nieuwe gegevens kunnen produceren die lijken op hun trainingsdata, focust CLIP zich op het koppelen van bestaande afbeeldingen aan bijbehorende teksten. Het doel van CLIP is om een overeenkomst tussen afbeeldingen en tekst te vinden, terwijl generatieve modellen nieuw materiaal creëren, zoals afbeeldingen of tekst.

CLIP kent echter enkele beperkingen die verder onderzoek vereisen. Op het gebied van rekencapaciteit en efficiënt gebruik van data zijn er mogelijkheden voor verbetering, omdat CLIP momenteel een aanzienlijke hoeveelheid rekenkracht en grote hoeveelheden data nodig heeft.

Hoewel het model uitblinkt in zero-shot leertaken (wat wil zeggen dat het zonder voorafgaande training op een specifieke taak toch goede prestaties levert), presteert het minder goed bij taken die fine-tuning vereisen. Fine-tuning houdt in dat een model, dat al voorgetraind is, verder wordt aangepast met nieuwe data om specifieke taken beter uit te voeren, zoals het onderscheiden van verschillende automerken. Daarnaast heeft CLIP moeite met meer abstracte en systematische taken, zoals het tellen van objecten in afbeeldingen. Ook heeft CLIP moeite om goed te presteren op afbeeldingen die te veel afwijken van de beelden waarop het is getraind. Bovendien herkent het digitale tekst goed, maar heeft het moeite met handgeschreven tekst, zoals die in de MNIST-dataset; de MNIST-dataset is een veelgebruikte benchmark in de wereld van machine learning en *computer vision*, die bestaat uit een verzameling van 70.000 zwart-wit-afbeeldingen van handgeschreven cijfers (LeCun et al.,1998). Computer vision is een deelgebied van kunstmatige intelligentie dat machines in staat stelt om visuele informatie, zoals afbeeldingen of video's, te begrijpen en te interpreteren. Het doel is om computers te leren objecten, personen of patronen te herkennen, vergelijkbaar met hoe mensen dat doen. Toepassingen hiervan zijn gezichtsherkenning, objectdetectie en beeldclassificatie.

In de gezondheidszorg wordt inherent multimodaal gewerkt, er worden verschillende soorten data gecombineerd, zoals tekst en beeldmateriaal. Biomedische multimodale AI-systemen die deze uiteenlopende datatypes flexibel kunnen verwerken en integreren, openen de deur naar waardevolle toepassingen in de zorgverlening. Een voorbeeld hiervan is het CLIP-model dat artsen kan helpen bij het interpreteren van medische beelden. Daartoe kan CLIP verder gevoed of verfijnd worden met specifieke medische beeld-tekstparen. In een studie van Eslami et al. (2021) werd CLIP bijvoorbeeld verfijnd tot PubMedCLIP door het extra te voeden met 642 beelden en meer dan 7.000 vraag-antwoordparen om de nauwkeurigheid en relevantie in medische contexten te vergroten.

Figuur 5 (Eslami et al., 2021) toont hoe zo'n verfijnd CLIP-model correct vragen over medische beelden kan beantwoorden.



Figuur 5. Vragen over medische beelden beantwoord door CLIP verfijnd tot PubMedCLIP (Eslami et al., 2021)

Het MedPALM Multimodal (Tu et al., 2024) is een uitbreiding van MedPALM en een tweede voorbeeld van een multimodale AI, hier specifiek ontwikkeld voor de medische sector, die zowel tekstuele als visuele data kan verwerken en interpreteren. Dit model is getraind op een breed scala aan medische gegevens, zoals patiëntendossiers, röntgenfoto's en medische literatuur. Hierdoor is het model in staat om medisch personeel te assisteren bij het stellen van een diagnose, het geven van behandelingsopties en het begrijpen van complexe medische informatie. De kracht van MedPALM ligt in zijn vermogen om de nuances van medische terminologie te begrijpen en relevante verbanden te leggen tussen symptomen, diagnoses en behandelingen in een multimodale context.

Ten slotte

In dit hoofdstuk heb ik de werking en mogelijkheden van algemene en domeinspecifieke grote taalmodellen en multimodale modellen besproken. Er is bij deze technologieën al aanzienlijke vooruitgang geboekt in diverse domeinen. Een van de impactvolle toepassingen van deze modellen bevindt zich in de gezondheidszorg: MedPALM.

Om uiteindelijk een geschikt taalmodel te selecteren (hetzij een algemeen of een domeinspecifiek model, BERT- of LLaMa-model, mono- of multilingual), moet er vooral overleg gepleegd worden tussen de NLP-onderzoekers enerzijds en gezondheidszorg-professionals en stakeholders anderzijds. Afhankelijk van de specifieke medische behoefte zal in overleg uitgemaakt moeten worden welk taalmodel het best kan worden geselecteerd (He et al., 2023).

In het volgende hoofdstuk ga ik dieper in op concrete voorbeelden van hoe grote taalmodellen worden ingezet in de medische sector, waarbij ik begin met hun rol in de vroege detectie van ziektes.

Transformermodellen voor vroege ziektedetectie

3

Inleiding

De COVID-19-pandemie heeft het welzijn van mensen, de economie en de sociale stabiliteit zowel lokaal als wereldwijd ontwricht. De uitbraak en de snelle verspreiding ervan noopten tot de ontwikkeling van innovatieve strategieën voor effectieve, preventieve en therapeutische interventies (World Health Organization, 2020). Onderzoek bevestigt dat machine learning praktische en waardevolle oplossingen kan bieden in de ondersteuning van de volksgezondheid en het faciliteren van de bijbehorende besluitvormingsprocessen, variërend van voorspellende modellering en epidemiologisch onderzoek tot de ontwikkeling van medicijnen en vaccins (Martin-Moreno et al., 2022). Dit suggereert dat benaderingen op basis van machine learning ook een belangrijke rol kunnen spelen bij de voorbereiding en reactie op toekomstige pandemieën, en ook bij het overbruggen van de kennisachterstand voor de vroege detectie van complexe ziekten zoals Alzheimer, sepsis, alvleesklierkanker en multiple sclerose of het verbeteren van de selectie van gerichte bloedkweken, wat kan leiden tot vroegere en meer nauwkeurige diagnoses. De medische gemeenschap wordt zich steeds meer bewust van de sleutelrol die deze technologieën kunnen spelen bij het verbeteren van diagnostische processen en behandelingsstrategieën ter ondersteuning van de gezondheidszorg.

Machine learning kan bovendien heel waardevol zijn bij het detecteren van ziektes waarbij het vroeg stellen van een diagnose door een mens een uitdaging, moeilijk of zelfs onmogelijk is. Door gebruik te maken van machinelearningalgoritmes en grote hoeveelheden medische gegevens te analyseren, kunnen met machine learning subtiele patronen en indicatoren geïdentificeerd worden, die door mensen mogelijk over het hoofd zouden worden gezien of nooit zouden worden opgespoord. In situaties waarin vroege symptomen van een ziekte moeilijk waarneembaar zijn, kan voor de detectie daarvan de inzet van machine learning, deep learning en ook van grote taalmodellen een waardevol hulpmiddel zijn; daarmee kunnen potentiële gevallen van ziektes zoals kanker, neurologische aandoeningen of infectieziekten in een vroeg stadium worden geïdentificeerd. Het resultaat is een verbeterde kans op succesvolle behandeling en een positieve impact op de prognose van de patiënt. In de gezondheidszorg vertegenwoordigen deeplearningmodellen en transformermodellen de nieuwste technologische ontwikkelingen.

Ik heb ervoor gekozen dit te verduidelijken aan de hand van *sepsis*, een ziektebeeld waarbij vroege detectie noodzakelijk is (waar tijdens dit lectoraat al onderzoek voor wordt voorbereid) en waarbij ik me vooral wil concentreren op de inzet van grote taalmodellen. Vroege detectie van sepsis is uiterst belangrijk omdat dat behandelingen effectiever en meer gepersonaliseerd maakt, de progressie kan vertragen en levensbedreigende complicaties kan voorkomen.

Toch blijven vroege machinelearningtechnieken nog steeds een niet-onbelangrijke rol spelen, vooral bij de verwerking van *gestructureerde* data. In de gezondheidszorg omvatten vitale, gestructureerde data belangrijke metingen van bijvoorbeeld hartslag, bloeddruk, lichaamstemperatuur en ademhalingsfrequentie (Badawy et al., 2023).

Vroege machinelearningmethodes

Omdat de hier beschreven onderzoeksprojecten gebruikmaken van transformermodellen, een vorm van deep learning, wordt hier niet dieper ingegaan op traditionele machinelearningtechnieken, die voorafgingen aan deep learning. Er wordt hier alleen een overzicht gegeven van enkele veelgebruikte vroege machinelearningmethodes gevoed met gestructureerde data, en enkele voorbeelden van de toepassing ervan in de gezondheidszorg.

Logististic regression of logistische regressie is een statistische methode die de waarschijnlijkheid van een bepaalde uitkomst voorspelt op basis van één of meer voorspellende variabelen. Hoewel het vaak wordt gebruikt voor binaire uitkomsten (zoals 'ja/nee' of 'ziek/niet ziek'), kan het ook worden uitgebreid naar multinomiale uitkomsten, of uitkomsten met meer dan twee categorieën, om de waarschijnlijkheid van meer dan twee mogelijke categorieën te voorspellen (Cox, 1958). *Waarschijnlijkheid* geeft hier aan hoe groot de kans is dat een bepaalde gebeurtenis of uitkomst plaatsvindt, uitgedrukt als een getal tussen 0 en 1. Een waarschijnlijkheid van 0 betekent dat de gebeurtenis zeker niet gebeurt, terwijl een waarschijnlijkheid van 1 betekent dat de gebeurtenis met absolute zekerheid plaatsvindt. In het geval van logistic regression wordt dit getal gebruikt om te voorspellen hoe waarschijnlijk het is dat een bepaalde uitkomst hoort bij de gegeven waarden van de voorspellende variabelen.

In tegenstelling tot logistische regressie, waarbij een wiskundige formule wordt gebruikt om relaties tussen variabelen te modelleren en voorspellingen te doen, doet het *k-nearest neighbors* (KNN) algoritme voorspellingen door te kijken naar de meest vergelijkbare punten (buren) in de bestaande gegevens. Bij KNN wordt de categorie van een nieuw datapunt bepaald door te kijken naar de categorieën van de dichtstbijzijnde buren: het nieuwe datapunt krijgt de categorie die het meest voorkomt bij deze buren (Cover & Hart,1967).

Support Vector Machines (SVM's) werken door een *hyperplane* te vinden die de gegevens in verschillende klassen scheidt. Het doel van een SVM is om een maximale marge te creëren tussen de gegevenspunten van verschillende categorieën of klassen, waardoor de generalisatie naar nieuwe gegevens wordt verbeterd (Cortes & Vapnik, 1995). Dit maakt een SVM vooral nuttig voor het nauwkeurig classificeren van datapunten die dicht bij de scheidingslijn liggen en sterk op elkaar lijken, omdat zij de optimale grens tussen de klassen bepaalt (Cortes & Vapnik, 1995).

Decision trees vormen een machine learning techniek waarbij herhaaldelijk data worden opgesplitst op basis van bepaalde features (typische kenmerken), om een boomachtige structuur te creëren, waarbij elke tak een beslissingsregel vertegenwoordigt en elk blad een uitkomst. Het doel is om de data zo te splitsen dat de onderliggende patronen en relaties tussen de kenmerken duidelijk worden, wat kan leiden tot beslissingen die makkelijker interpreteerbaar zijn (Breiman, 2017). Meerdere decision trees worden gecombineerd in een *Random Forest* machinelearningtoepassing. Elke 'boom in het bos' wordt getraind op een willekeurige subset van de trainingsdata en bij elke splitsing wordt een willekeurige subset van features overwogen. De uiteindelijke voorspelling wordt bepaald door het gemiddelde van de voorspellingen (voor regressie) of de meerderheid van de stemmen of *votes* (voor classificatie) van alle bomen (Breiman, 2001) binnen deze *ensemble-learning-benadering*. Net als Random Forests, is *XGBoost* een ensemble-techniek, maar in plaats van *parallelle* bomen zoals in Random Forests, bouwt XGBoost bomen *sequentieel*. Elke nieuwe boom wordt getraind of gevoed met data, om de fouten van de vorige bomen te corrigeren (Chen & Guestrin, 2016).

Het Naive Bayes algoritme is gebaseerd op Bayesiaanse statistiek en gaat uit van de aanname dat de kenmerken van een dataset onafhankelijk van elkaar zijn, gegeven de categorie. Naive Bayes berekent de waarschijnlijkheid dat een gegeven datapunt tot een bepaalde klasse behoort door de individuele waarschijnlijkheden van elk kenmerk binnen die klasse te combineren. Dit gebeurt door de kans van elk kenmerk gegeven de klasse te vermenigvuldigen met de priorkans van die klasse. Hierdoor kan het model snel en effectief bepalen tot welke klasse een datapunt waarschijnlijk behoort (Duda & Hart, 1973).

Genetische algoritmes in machine learning vormen een optimalisatietechniek die de principes van natuurlijke selectie en genetica toepast om modellen te verbeteren en hyperparameters te optimaliseren. Ze beginnen met een populatie van mogelijke oplossingen (modellen of hyperparameters) en verbeteren deze iteratief door selectie van de beste oplossingen, kruising (combinatie van oplossingen) en mutatie (willekeurige veranderingen). Door dit proces herhaaldelijk uit te voeren, evolueren de oplossingen naar een steeds betere set van parameters of modellen die betere resultaten opleveren (Goldberg, 1994; Holland, 1992).

Voorspelling van ziektes via gestructureerde data met traditionele machine learning

Verschillende van de hiervoor beschreven meer traditionele machinelearningmodellen worden gebruikt om ziektes te voorspellen door te leren van patiëntgegevens en de relaties tussen risicofactoren, symptomen en uitkomsten te identificeren. Ze analyseren patronen in deze gegevens en gebruiken deze om het ontwikkelen van een specifieke aandoening, zoals diabetes of hartfalen, te voorspellen. Voor het voorspellen van diabetes wordt er bijvoorbeeld gebruikt gemaakt van machinelearningmodellen zoals logistic regression, KNN, SVM en Random Forest op de Pima Indian Diabetes Database (PIDD), voor 768 Indische patiënten op basis van 9 vitale data-features, waarbij het logisticregressionmodel de beste resultaten toonde (Krishnamoorthi et al., 2022).

In een ander geval hebben onderzoekers drie machinelearningmethodes (logistic regression, decison trees en Random Forest) gebruikt om COVID-19-uitkomsten te beoordelen op basis van patiëntgegevens uit Mexico en Brazilië. De features die aan het model werden meegegeven, waren gebaseerd op geografische, sociale en economische omstandigheden, en op klinische risicofactoren, medische rapporten en demografische gegevens van COVID-19-patiënten, voor het nauwkeurig voorspellen van herstel- en sterftecijfers (Iwendi et al., 2024).

Moturi et al. (2020) beschrijven een onderzoek dat is uitgevoerd om artsen te ondersteunen bij het voorspellen van hart- en vaatziektes via KNN, Random Forest, decision trees en Naive Bayes-algoritmes op de Cleveland Heart Disease Dataset, waarbij KNN de hoogste nauwkeurigheid behaalde.

Voor de diagnose van borstkanker, een ziekte die één op de 28 vrouwen in India treft, ontwikkelden onderzoekers een nauwkeurige classificatietechniek met een dataset van 569 Indische patiënten en 32 vitale features. Door genetische algoritmes te combineren met SVM-classificatie, breidden ze deze aanpak ook uit naar hartziekten en longkanker (Soni, 2020).

Voor de detectie van infecties via bloedanalyses, zijn bloedkweken vereist om de aanwezigheid van ziekteverwekkers in de bloedbaan te identificeren. Hierbij wordt een bloedmonster van de patiënt genomen en in een medium geplaatst dat microbiële groei bevordert, gevolgd door incubatie en observatie in het laboratorium. Hoewel bloedkweken de *gouden standaard* zijn voor de diagnose van bloedbaaninfecties, worden ze vaak overmatig gebruikt en leveren ze nog te vaak onvoldoende resultaten op. Bovendien zijn de verwerkingstijden (24 tot 72 uur) hiervan nog te lang, wat ze minder geschikt maakt voor een snelle diagnose bij spoedgevallen (Boerman et al., 2022). Gezien de toenemende hoeveelheid data in klinische laboratoria, kunnen machinelearningalgoritmes hier een oplossing bieden om tot meer gerichte bloedkweken te komen. Zo beschrijven McFadden et al. (2023) een machinelearningmodel

voor de voorspelling van bloedkweekresultaten met gestructureerde data van routinematig geanalyseerde bloedmonsters, zoals volledige bloedtelling, differentiële telling van witte bloedcellen en celpopulatiegegevens met behulp van Sysmex XN-2000 analyzers. De Sysmex XN-2000 is een hematologie-analyzer die gebruikt wordt in laboratoria voor het analyseren van bloedmonsters. Het apparaat is ontworpen voor hoge capaciteit en kan meerdere monsters tegelijkertijd verwerken. Twee machinelearningmodellen, XGBoost en Random Forest, werden getraind op data van 10.965 monsters uit de periode 2018-2019 en toonden veelbelovende resultaten.

Sepsis

Traditionele machinelearningtechnieken kunnen een aanzienlijke bijdrage leveren in de ondersteuning van de gezondheidszorg. Echter, bij de *vroege voorspelling* van ziektes zoals sepsis (een levensbedreigende aandoening) blijken deze technieken vaak tekort te schieten

Zevende jaarlijkse bijeenkomst van de European Sepsis Alliance op 18 maart 2024 te Brussel

Tijdens de 7e jaarlijkse bijeenkomst van de European Sepsis Alliance op 18 maart 2024 te Brussel werd inzichtelijke en praktische informatie gedeeld via presentaties en paneldiscussies. Sepsis is een levensbedreigende aandoening die voortkomt uit de reactie van het lichaam op een infectie, wat kan leiden tot weefselschade, orgaanfalen en mogelijk de dood. Deze aandoening ontstaat wanneer de reactie van het immuunsysteem op een infectie leidt tot wijdverspreide ontsteking, wat de bloedstroom belemmert en organen van voedingsstoffen en zuurstof berooft. Symptomen omvatten koorts, snelle hartslag, snel ademhalen, verwarring en extreme ongemakken. Vroege opsporing en behandeling met antibiotica en ondersteunende zorg zijn essentieel om de vooruitzichten te verbeteren. Sepsis kan zich snel ontwikkelen en vormt daarmee een medisch noodgeval.

Zoals tijdens de bijeenkomst werd vermeld, leidt sepsis elke minuut tot een sterfgeval in Europa. Wereldwijd werden in 2017 48,9 miljoen mensen door sepsis getroffen. In 2018 was deze aandoening verantwoordelijk voor 15 procent van alle neonatale sterfgevallen wereldwijd. Geschat wordt dat 15 van de 1000 gehospitaliseerde patiënten sepsis ontwikkelen.

In Europese ziekenhuizen ontbreekt het over het algemeen nog te vaak aan voldoende informatie over de vroege diagnose van sepsis. Zelfs als er een systeem voor vroege detectie aanwezig is, is de reactie van sommige zorg-professionals op deze signalen niet altijd adequaat. In België heeft de minister van Volksgezondheid dit probleem absolute voorrang gegeven, mede ten gevolge van een documentaire die op de nationale televisie werd uitgezonden. In deze documentaire deelde een sepsispatiënte haar ervaring met het falen van het gezondheidssysteem om sepsis op tijd te herkennen en te behandelen. Een andere overlevende belichtte hoe zijn toestand verkeerd werd gediagnosti-ceerd als neurocognitieve stoornissen, wat leidde tot vertraging in onmiddellijke en cruciale behandeling in een intensive care unit. Dit benadrukt de dringende behoefte aan verbeterde opleiding, training en protocollen voor vroege detectie en beheer van sepsis om dergelijke levensbedreigende misdiagnoses te voorkomen.

In tegenstelling tot de vooruitgang in de VS, verloopt de ontwikkeling van een Europees register voor gestructureerde gegevens van sepsispatiënten langzamer en dit vraagt om meer aandacht en onderzoeksinspanningen in dit gebied. Het oprichten van zo'n register is een cruciale stap voor het inzetten van AI bij de vroege detectie van sepsis.

Dit benadrukt de noodzaak van internationale samenwerking en investeringen in onderzoek om de vroege diagnose en behandeling van sepsis te verbeteren.

De discussies en panels tijdens deze conferentie beklemtoonden de groeiende noodzaak van wereldwijde interdisciplinaire netwerken en verbeterde financiering van onderzoek, om de vroege diagnose van deze ziekte te bevorderen. Er is een oproep gedaan om transdisciplinair, ziekte-overstijgend onderzoek te doen, waarbij inzichten in de ene aandoening ondersteuning kunnen bieden aan onderzoek naar een andere aandoening. Momenteel is het onderzoek te veel geconcentreerd op de kritische fase van deze ziekte bij gehospitaliseerde patiënten, die slechts 20 procent van de totale patiëntenpopulatie vertegenwoordigen. Dit benadrukt de noodzaak om de onderzoeksfocus te verbreden, zodat deze ook de *vroegere stadia van de ziekte* en een breder scala aan getroffen individuen omvat.

(Bron: European Sepsis Alliance, 2024)

Vroege diagnose bij Sepsis is moeilijk. De term sepsis (afkomstig van het Griekse woord 'σηψη', 'sēpsē' voor ontbinding) werd pas in de 19e eeuw, met de identificatie van micro-organismen als de oorzaak van infecties, gebruikt om een klinische ziekte veroorzaakt door ernstige infectie, te beschrijven. Momenteel wordt sepsis gedefinieerd als een levensbedreigende kettingreactie in een ontregeld gastorganisme op een infectie, wat kan leiden tot weefselschade en orgaanfalen, met mogelijk overlijden tot gevolg. Ondanks belangrijke medische vooruitgang blijft de behandeling en vooral een vroege diagnose van sepsis een grote uitdaging voor zowel clinici als onderzoekers. De behandeling is nog steeds beperkt tot het toedienen van antibiotica, vochttherapie en ondersteunende orgaantherapie. Het gebrek aan nieuwe behandelingen maakt deze ziekte moeilijk te beheersen en resulteert nog steeds in een aanzienlijke wereldwijde sterfte, waarbij alle leeftijdsgroepen worden getroffen.

Een recente studie heeft onthuld dat er in 2017 ongeveer 48,9 miljoen gevallen van sepsis en 11 miljoen aan sepsis gerelateerde sterfgevallen wereldwijd waren, wat neerkomt op bijna 20 procent van alle sterfgevallen wereldwijd. Er zijn significante regionale verschillen in de percentages van sepsisincidentie en -sterfte, waarbij ongeveer 85 procent van zowel de sepsisgevallen als de gerelateerde sterfgevallen zich voordoet in landen met een laag of gemiddeld inkomen. De Wereldgezondheidsorganisatie heeft sepsis uitgeroepen tot een wereldwijde prioriteit (Papathanakos et al., 2023; Rudd et al., 2020).

Jaarlijks worden in Nederland ongeveer 35.000 patiënten getroffen door sepsis. Elk jaar belanden zo'n 10.000 patiënten met sepsis op de intensive care (IC), wat het tot de belangrijkste doodsoorzaak op de IC maakt (Sepsis en daarna, 2020).

Typisch verloopt sepsis in drie stadia:

- Het eerste stadium is het stadium van het *Systemic Inflammatory Response Syndrome* (SIRS), waarin sepsis doorgaans wordt gekenmerkt door een zeer hoge of lage lichaamstemperatuur, hoge hartslag, ademhalingsproblemen, veranderingen in de witte bloedcellen en een bekende of vermoedelijke *infectie*. SIRS wordt effectief als sepsis beschouwd wanneer er daadwerkelijk een infectie aanwezig is.
- Het tweede stadium is het stadium van ernstige sepsis, waarin acute *orgaanstoornissen* beginnen. Ernstige sepsis kan ook optreden bij een lage bloeddruk of verminderde bloedstroom, gekenmerkt door symptomen als verminderde urineproductie, veranderde mentale status en ademhalingsproblemen.
- Het derde stadium van de *septische shock*. Dit is het meest ernstige stadium, gekenmerkt door een aanhoudende lage bloeddruk ondanks vochttoediening en perfusie-afwijkingen, of abnormale veranderingen en stoornissen in de bloedtoevoer naar een bepaald orgaan of weefsel, of verhoogde lactaatniveaus. Het lactaatgehalte in het bloed is hoger dan 2 mmol/L[7] en er is een aanhoudende lage

7 Mmol/L (millimol per liter) is een maateenheid voor de concentratie van een stof in een oplossing, waarbij 1 millimol de hoeveelheid van die stof is in één liter oplossing.

bloeddruk die het gebruik van medicijnen nodig maakt om de gemiddelde bloeddruk boven de 65 mmHg[8] te houden. De kans op overlijden in dit stadium wordt geschat op 30 tot 50 procent (Marik, 2015).

Bewezen is dat vroege detectie in het eerste stadium en waarschuwingssystemen effectief kunnen leiden tot het vroeg starten van een behandeling van sepsis en vermindering van het aantal sterfgevallen (Jiang et al., 2023).

Sepsis-2- en sepsis-3-definities

In de loop van de tijd zijn de criteria en definities voor sepsis aangepast om vroege identificatie en zorg te verbeteren. De twee primaire definities die brede acceptatie hebben gevonden, zijn *sepsis-2 en sepsis-3*, die zich kenmerken door enkele scores die vaak voorkomen in machinelearningexperimenten voor de voorspelling van sepsis.

Sepsis-2-definitie: SIRS-score

Geïntroduceerd in 2001, draait de definitie van sepsis-2 om het concept van het *Systemic Inflammatory Response Syndrome* (SIRS), dat wordt getriggerd door een infectie. Volgens deze definitie wordt sepsis geïdentificeerd door aan ten minste twee SIRS-criteria te voldoen. Septische shock onder sepsis-2 is een ernstige vorm van sepsis waarbij de patiënt een aanhoudende lage bloeddruk heeft ondanks pogingen tot vochtresuscitatie. Verder wordt ernstige sepsis beschreven als sepsis die escaleert en uiteindelijk orgaandisfunctie omvat (Bone et al., 1992; Levy et al., 2003).

Sepsis-3-definitie: SOFA- en qSOFA-scores

De sepsis-3-definitie uit 2016 legt de nadruk op orgaanfalen en identificeert sepsis als levensbedreigende orgaandisfunctie veroorzaakt door een ontregelde reactie op infectie. Deze definitie is grotendeels gebaseerd op de SOFA-score (*Sequential Organ Failure Assessment*), die disfunctie meet in verschillende vitale systemen: ademhaling, hart- en vaatstelsel, lever, nieren, bloedstolling en het centraal zenuwstelsel. Een toename van twee of meer punten in de SOFA-score als gevolg van een infectie kan op sepsis duiden (Seymour et al., 2016).

Daarnaast is er de qSOFA-score (*quick Sequential Organ Failure Assessment*), een variatie op de SOFA-score, als snelle toepassing *buiten* de IC. In tegenstelling tot de uitgebreide evaluatie die voor de SOFA-score vereist is, beoordeelt de qSOFA-score op drie specifieke criteria: veranderde mentale status, een systolische bloeddruk van 100 mmHg of minder en een ademhalingsfrequentie van 22 of meer ademhalingen per minuut. Een qSOFA-score van 2 of hoger suggereert een significant risico op ernstige uitkomsten, wat aanleiding geeft tot verder onderzoek naar orgaandisfunctie en de mogelijke aanwezigheid van sepsis (Amland & Sutariya, 2018; Koch et al., 2020).

8 MmHg (millimeter kwikdruk) is een maateenheid voor druk, gebruikt om bijvoorbeeld bloeddruk te meten, en geeft de hoogte aan van een kolom kwik die door die druk wordt ondersteund.

MEWS-score

Een andere score is de *Modified Early Warning Score* (MEWS-score), die niet specifiek voor de sepsis-2- of sepsis-3-definities geldt, maar eerder als een algemene score die kan worden toegepast om verschillende aandoeningen, inclusief sepsis, in een vroeg stadium te detecteren. De MEWS-score is ontworpen om proactief patiënten te identificeren die risico lopen op klinische achteruitgang, in tegenstelling tot de eerder genoemde SIRS- en SOFA-score, die worden gebruikt in de context van sepsis en orgaanfalen. Waar de SIRS-score zich richt op ontstekingsmarkers en de SOFA-score de mate van orgaanfalen beoordeelt, evalueert de MEWS-score systolische bloeddruk, hartslag, ademhalingsfrequentie, temperatuur en mentale status. Een MEWS-score van 5 of hoger suggereert een verhoogd risico op achteruitgang en mortaliteit (Guirgis et al., 2017; Hester et al., 2021; Subbe et al., 2001 en 2006).

De SOFA-, qSOFA-, MEWS- en SIRS-score hebben elk hun eigen specifieke toepassingsgebied en er is geen enkele score die als de 'beste' kan worden beschouwd. De effectiviteit van deze scores hangt af van factoren zoals de samenstelling van de patiëntenpopulatie, het type zorginstelling en het moment in het zorgproces waarop ze worden berekend. Voortbouwend op dit inzicht, zijn de genoemde scores van grote waarde in de toepassing van machine learning voor het vroegtijdig detecteren van sepsis. Door deze scores te integreren in machinelearningmodellen, is het mogelijk om te leren van historische patiëntgegevens en zo het moment waarop sepsis zich zal manifesteren beter te voorspellen.

Machine learning die traditionele sepsisdetectiemethodes en gestructureerde data van vitale functies gebruikt

Machinelearningtechnieken tonen veelbelovende resultaten bij detectie en automatische voorspelling van sepsis, onder andere gebruik makend van de SIRS-, MEWS- en qSOFA-score (Desautels et al., 2016; Islam et al., 2023; Kijpaisalratana et al., 2022; Lyra et al., 2019; Zhang et al., 2021). Als voorbeeld geef ik hier de studie van Kijpaisalratana et al. (2022), waarin drie traditionele machinelearningmodellen, te weten Random Forest Classification, Gradient Boost, logistic regression met deep learning neurale netwerken worden vergeleken, waarbij het Random Forest Model de beste resultaten liet zien. Met dit model worden *gestructureerde* gegevens over vitale data gebruikt (zoals lichaamstemperatuur, hartslag, ademhalingsfrequentie en bewustzijnsniveau) om de SIRS,-MEWS- en qSOFA-score te voorspellen. Alle machinelearningmodellen vertoonden betere prestaties in de voorspelling van de sepsisdiagnose in vergelijking met traditionele screeningsinstrumenten voor de voorspelling van de SIRS-, qSOFA- en MEWS-score.

Echter, uitsluitend vertrouwen op deze scores is misschien niet voldoende voor een vroege voorspelling van sepsis. Ondanks de verdienste van deze scores, leveren ze voor sepsisdetectie te vaak valse positieven op, waaruit blijkt dat andere aandoeningen

ook aan de basis kunnen liggen van deze scores, dus niet alleen sepsis. Hoewel de scores kunnen aangeven dat een patiënt risico loopt of systemische afwijkingen vertoont, kunnen ze mogelijk niet de vroege, subtiele tekenen van sepsis oppikken. Bovendien zijn deze machinelearningmodellen afhankelijk van laboratoriumresultaten, en het neemt vaak veel kostbare tijd om die te verkrijgen (Qin, Madan et al., 2021).
Sepsis is een heterogene aandoening met gevarieerde klinische presentaties. Vroege tekenen komen mogelijk niet altijd overeen met de parameters beoordeeld door deze scoresystemen, die een beperking kunnen vormen voor de detectie van sepsis in een vroeg stadium (Papathanakos et al., 2023).

Hoewel de prestaties van deep learning in de studie van Kijpaisalratana et al. (2022) de prestaties van traditionele modellen niet kon overtreffen, maakt het vermogen om zelfstandig features uit grote en gevarieerde datasets af te leiden, van deep learning een veelbelovende benadering voor de vroege voorspelling van sepsis, wat mogelijk de detectiecapaciteiten kan verbeteren door patronen te identificeren die aan traditionele machinelearningmodellen ontsnappen.

Automatische vroege detectie van sepsis door een combinatie van gestructureerde en ongestructureerde data

Een van de sleutels tot de vroege diagnose van sepsis is het vaststellen van een infectie. Zoals gezegd begint sepsis meestal als een lokale ontstekingsreactie op een infectie. Het stellen van de diagnose van een infectie is echter een uitdaging en vaak is hier onenigheid over tussen clinici. Zoals reeds vermeld wordt er voor de diagnose van sepsis onder andere gebruik gemaakt van de SIRS-score, maar die kan niet alle patiënten met deze ziekte identificeren. Sepsis begint doorgaans met een infectie, waarvan de bron veelvoudig kan zijn: longontsteking, urineweginfectie, buikinfectie of huidinfectie. De initiële infectie activeert dan het immuunsysteem. Aangezien het stellen van de diagnose van een infectie die sepsis veroorzaakt, verre van eenvoudig is, kunnen *ongestructureerde data* naast vitale data in de vorm van gestructureerde data, een rol spelen en soelaas bieden bij de vroege detectie van sepsis.

Een specifieke voorbeeldcase van sepsis gaat over de *gestructureerde gegevens* van een 54-jarige patiënt met een voorgeschiedenis van 10 dagen epigastrische pijn (pijn in het bovenste gedeelte van de buik), misselijkheid, geleidelijke kortademigheid in de voorbije 48 uur, een lichaamstemperatuur van 38,5 °C, een ademhalingsfrequentie van 28 ademhalingen per minuut, een hartslag van 115 slagen per minuut, een zuurstofsaturatie van 88 procent bij inademing en een bloeddruk van 94/65 mmHg. Deze patiënt had een voorgeschiedenis van door alcohol veroorzaakte alvleesklier-ontsteking. Deze *ongestructureerde* gegevens kunnen met de bekende gestructureerde gegevens worden gecombineerd in een machinelearningmodel voor de vroege voorspelling van sepsis, om op die manier beter in kaart te brengen welke vormen van ontsteking in meer of mindere mate risico op sepsis met zich meebrengen (Duncan et al., 2021).

Er is tot nu toe maar weinig onderzoek voorhanden dat voor de vroege voorspelling van sepsis gebruik maakt van de ongestructureerde informatie in elektronische patiëntendossiers (EPD's). Er is één studie met betrekking tot sepsis die alleen gebruik maakt van ongestructureerde data, en alleen ernstige sepsis (geen vroege detectie) heeft voorspeld met behulp van tekstfeatures, door patiëntennota's vectorieel voor te stellen (Culliton et al., 2017). Dit betekent dat tekstuele informatie (zoals woorden of zinnen uit patiëntennota's) wordt omgezet in numerieke waarden, zodat machinelearning-algoritmes deze kunnen verwerken. Deze studie vergeleek het model dat gevoed is met ongestructureerde data met een model dat was gevoed met *gestructureerde* numerieke kenmerken en toonde aan dat het op tekst of ongestructureerde data gebaseerde model beter presteerde dan de gestructureerde data. Voor zover bekend zijn er geen gelijkaardige studies voor Nederlandse EPD's.

Eveneens is er voorlopig weinig onderzoek dat *ongestructureerde en gestructureerde data combineert* voor de vroege detectie van sepsis (Goh et al., 2021; Pal et al., 2022; Qin, Madan et al., 2021; Wang et al., 2022). Een van die weinige onderzoeken naar vroege detectie van sepsis die gestructureerde en ongestructureerde data in een machine-learningbenadering combineert, is de studie van Qin, Madan et al. (2021). In dit onderzoek wordt aangetoond hoe ongestructureerde tekst uit EPD's een aanvulling kan vormen op gestructureerde gegevens voor de voorspelling van sepsis, gebruik makend van machine learning met een domeinspecifiek BERT-transformermodel, namelijk ClinicalBERT, dat vooraf is gevoed met klinische aantekeningen (Huang et al., 2019).

Dit model dat gestructureerde en ongestructureerde data combineert, werd geëvalueerd, gebruik makend van de Amerikaanse MIMIC-III-dataset als testdataset, die bestaat uit patiëntendossiers van ongeveer 60.000 opnames op de IC (Johnson, Pollard & Mark, 2016; Johnson, Pollard, Shen et al., 2016). De dataset bevat *gestructureerde* data, zoals demografische gegevens, vitale functies en laboratoriumtests. Deze informatie wordt gecombineerd met *ongestructureerde* informatie (in de vorm van rapporten van artsen, verplegers en radiologen) over het verloop van de ziekte en wat aan de ziekte voorafging, wat belangrijke aanwijzingen kunnen zijn bij het voorspellen van sepsis.

De *ongestructureerde* data werden omgezet in BERT-embeddings en samengevoegd (*geconcateneerd*) met de numerieke kenmerken van de gestructureerde gegevens, en werden geclassificeerd met behulp van XGBoost classificatie, een ensemble learning model (Chen & Guestrin, 2016). In dit onderzoek van Qin, Madan et al. (2021) krijgen alle *correcte vroege* sepsisvoorspellingen die plaatsvinden tot twaalf uur voor en tot drie uur na het begin van sepsis, een positieve score. Eveneens blijkt dat dit model beter presteert dan een machinelearningbenadering die alleen gestructureerde gegevens gebruikt. Helaas is in dit onderzoek niet onderzocht welk type informatie in

de ongestructureerde EPD's precies heeft bijgedragen aan de prestaties van dit model. Het onderzoek van Goh et al. (2021) toont eveneens aan dat het toevoegen van ongestructureerde data aan gestructureerde data de machinelearningvoorspelling 12 tot 48 uur voor de uitbraak van sepsis sterk doet stijgen. Het onderzoek van Pal et al. (2022) beschrijft eveneens een benadering met behulp van transformers voor de vroege voorspelling van sepsis. Het onderzoek bevat een analyse van ongestructureerde gegevens, waarbij de meest opvallende woorden worden benadrukt met behulp van een woordwolk. Echter, hier ontbreekt ook weer een meer gedetailleerde analyse van welke ongestructureerde gegevens vooral invloed hebben op de voorspelling. Er blijft dus een onontgonnen onderzoeksgebied dat verder verkend moet worden, namelijk een meer gedetailleerde analyse van welke specifieke ongestructureerde gegevens vooral invloed hebben op het vroeg kunnen voorspellen van sepsis.

Naar multimodale modellen voor vroege voorspelling van sepsis

Hoewel het gebruik van machinelearningmodellen die gevoed worden door zowel gestructureerde als ongestructureerde data voor de vroege voorspelling van sepsis, veelbelovend is, moet ook het inschakelen van multimodale modellen verkend worden. De snelle ontwikkeling van diagnostische technologieën in de gezondheidszorg biedt namelijk de mogelijkheid om heterogene gegevens uit verschillende kanalen te verwerken en te combineren. Dit biedt een betere, meer persoonlijke en patiëntgerichte medische diagnose. Zo kunnen bijvoorbeeld voor een meer persoonlijke diagnose data uit medische beeldvorming (bijvoorbeeld radiologie, pathologie en camera-beelden) en andere data – gestructureerde data (zoals laboratoriumtestresultaten, bloedtestresultaten) en/of ongestructureerde data (zoals medische rapporten) – gecombineerd worden in een *multimodaal* deeplearningmodel.

Deze combinatie van data uit medische beeldvorming en andere data wordt geïntegreerd in een multimodaal machinelearning- of deeplearningmodel; dat is een uitdaging omdat zulke heterogene data, zoals beeldinformatie, tekst en numerieke waarden, sterk variëren in structuur. Dit leidt tot een geheel van 2D-beelddata (zoals microscopisch gedetailleerde beelden van een tumor) en 3D-beelddata (zoals scans die gemaakt zijn met computertomografie [CT] en magnetische resonantiebeeld-vorming [MRI], die ruimtelijke informatie over diezelfde tumor geven). Deze heterogene formaten (namelijk beeld, ongestructureerde vrije tekst en gestructureerde data) vereisen verschillende pre-processingmethodes (Huang et al., 2020; Zhang et al., 2021). De meeste van deze deeplearningmethodes combineren data uit medische beeld-vorming met gestructureerde data (El-Sappagh et al., 2020; Holste et al., 2021; Kawahara, 2018; Lu et al., 2021; Mobadersany et al., 2018; Yan et al., 2021; Yap et al., 2018; Yoo et al., 2019), en minder met ongestructureerde data (Jacenków et al., 2022; Li et al., 2020; Wang et al., 2018). Voor de combinatie met ongestructureerde data wordt vaak een transformermodel gebruikt (Jacenków et al., 2022; Wang et al., 2018). Slechts weinig studies combineren data uit medische beeldvorming, gestructureerde data en

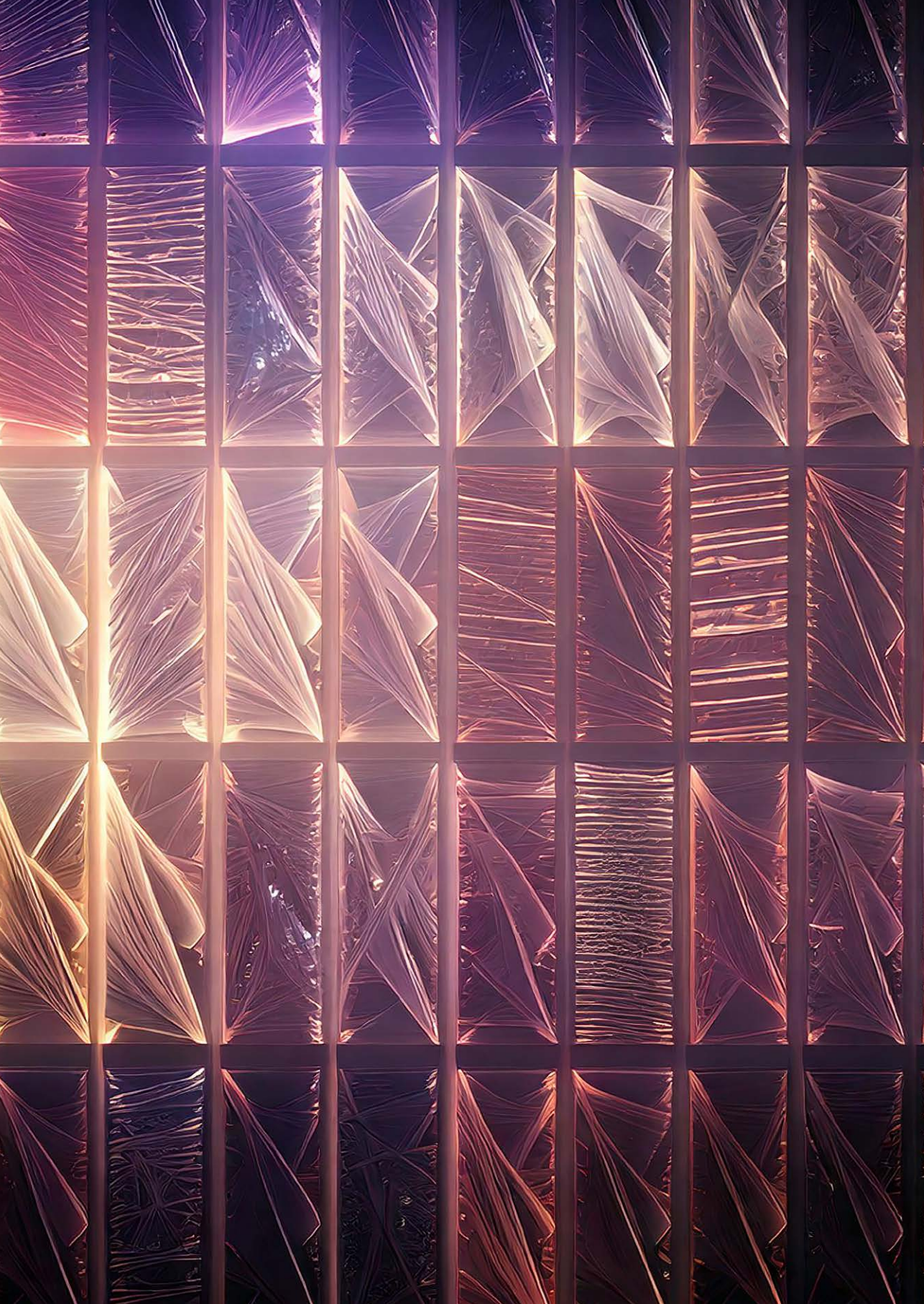
ongestructureerde data tegelijk binnen één machinelearningmodel. De studie van Zhou et al. (2023) beschrijft echter een multimodaal transformermodel voor de automatische diagnostisering van longontsteking, dat wel deze drie databronnen combineert.

Voor sepsis daarentegen is er voor zover bekend, nauwelijks of geen onderzoek dat medische beeldvorming integreert in een multimodaal machinelearningmodel (Duncan et al., 2021) dat gebruik maakt van transformermodellen, gecombineerd met gestructureerde data. Toch zou een dergelijk onderzoek kunnen bijdragen aan meer gerichte sepsisdetectie in een vroeg stadium. Hierbij kan bijvoorbeeld de borst-röntgenfoto een rol spelen, vooral gezien de frequent voorkomende respiratoire oorsprong van sepsis; de borst-röntgenfoto is een waardevolle diagnostische beeldvormingstechniek, die wordt toegepast voor de evaluatie van structuren binnen de thoracale holte, zoals longen, hart, bloedvaten en omliggende weefsels.

Net zoals bij machinelearningmodellen die gestructureerde en ongestructureerde data combineren, is er wel onderzoek gedaan naar multimodale machinelearningmodellen voor de gezondheidszorg, maar dit onderzoek bevat nauwelijks een diepe analyse over welke specifieke factoren sepsis kunnen verklaren. Ook is het gebruik van multimodale modellen voor vroege sepsisdetectie nog een onontgonnen terrein. Verdere verkenning en onderzoek op dit gebied kan aanzienlijk bijdragen aan de effectiviteit van vroege sepsisdiagnose en -behandeling.

Voor toekomstig sepsisonderzoek moet multimodaliteit verkend worden en moet er gestreefd worden naar een raamwerk dat gestructureerde en ongestructureerde data integreert. Innovatief onderzoek moet ernaar streven uit de ongestructureerde tekstgegevens van EPD's te identificeren welke delen van deze narratieve gegevens significant bijdragen aan de prestaties van het machinelearningmodel voor de vroege detectie van sepsis.

Voor verschillende eerdergenoemde studies (Culliton et al., 2017; Pal et al., 2023; Qin, Madan et al., 2021) werkten de onderzoekers met de Amerikaanse/Engelse MIMIC-datasets waarvan de al genoemde MIMIC-III-dataset deel van uitmaakt (Johnson, Pollard, Shen et al., 2016). Voor het toekomstig onderzoek van het lectoraat is het de bedoeling om vooral *Europese gegevens te gebruiken*, waarvoor samenwerking met nationale en internationale Europese medische instellingen noodzakelijk is. Deze samenwerking moet gericht zijn op het verzamelen en veilig beschikbaar stellen van gegevens over sepsispatiënten, ten behoeve van onderzoek dat AI toepast in de gezondheidszorg en het creëren van multimodale modellen mogelijk maakt.



4 Multimodale transformer modellen voor de diagnose van complexe ziektebeelden

Multimodale deeplearningmodellen kunnen de nauwkeurigheid van de voorspelling verbeteren door verschillende datatypes vanuit verschillende bronnen te integreren. Deze modellen kunnen een uitgebreider beeld van de toestand van de patiënt bieden door gelijktijdig meerdere gegevensbronnen te analyseren, wat leidt tot een betere contextuele interpretatie. Multimodale grote taalmodellen vormen de nieuwste innovatie in AI. Deze modellen kunnen niet alleen complexe tekst begrijpen en genereren, maar ook andere vormen van data, zoals beelden en audio, verwerken. Hierdoor kunnen ze taken uitvoeren die een combinatie van verschillende inputtypes vereisen, zoals vragen beantwoorden over de inhoud van een foto of visuele inhoud creëren op basis van tekstbeschrijvingen. Door diverse gegevensbronnen te integreren en te combineren, bieden deze modellen een rijkere context, wat leidt tot nauwkeurigere en relevantere resultaten voor een breed scala aan toepassingen. Dit maakt deze modellen niet alleen veelbelovend voor de vroege voorspelling van maar ook voor de *diagnose van complexe ziekten*. Ik licht dit toe aan de hand van onderzoek naar de ziekte van Alzheimer, waarvoor in het kader van dit lectoraat de krijtlijnen getekend worden.

Alzheimer

Het aantal mensen in de EU met de ziekte van Alzheimer wordt geschat op 7.850.000, terwijl dit aantal in de Europese landen die door Alzheimer Europe worden vertegenwoordigd, 9.780.000 bedraagt. Vrouwen worden nog steeds onevenredig getroffen door alzheimer: 6.650.000 vrouwen tegenover 3.130.000 mannen in Europa. Naar verwachting zal het aantal mensen met alzheimer in Europa tegen 2050 bijna verdubbeld zijn, tot 14.298.000 in de EU en 18.846.000 in de bredere Europese regio (Alzheimer Europe, 2019).

Momenteel zijn er in Nederland 280.000 mensen met alzheimer. Dit aantal zou de komende dertig jaar oplopen tot ruim een half miljoen, wat ook blijkt uit de nieuwste cijfers van Alzheimer Europe. In 2018 had 1,49 procent van de Nederlanders alzheimer, in 2050 zal dit 3,15 procent zijn. Wereldwijd zijn er ongeveer 20 miljoen mensen met alzheimer en verwacht wordt dat dat aantal in 2030 is verdubbeld (Nederlands Herseninstituut, 2018).

De meeste gevallen van alzheimer worden waargenomen bij mensen van 65 jaar of ouder. Terwijl het risico op het krijgen van alzheimer voor patiënten tussen de 65 en 74

jaar vijf procent is, neemt het risico met 50 procent toe boven de leeftijd van 85 jaar. Uit observatie is naar voren gekomen dat mensen met een hoge opleiding minder risico op alzheimer lopen omdat bij hen meer synaptische verbindingen in de hersenen worden gevormd. Dit creëert een synaptische reserve in de hersenen, waardoor patiënten het verlies van neuronen kunnen compenseren naarmate de ziekte vordert (Bhushan et al., 2018).

Helaas blijft wereldwijd bij bijna 75 procent van de mensen die lijden aan deze ziekte, alzheimer ongediagnosticeerd, onder andere vanwege het stigma dat eraan kleeft, waardoor de drempel tot hulp zoeken te hoog is (Schilling et al., 2022). Verder is er ook nog te veel onwetendheid over de vroege symptomen van deze ziekte.

De meest recente diagnostische richtlijnen classificeren het ziekteverloop in drie fasen op basis van de klinische symptomen van patiënten:

- eerste fase: een preklinische vorm van alzheimer (Jack Jr et al., 2011);
- tweede fase: een milde cognitieve stoornis (Albert et al., 2011);
- derde fase: dementie (McKhann et al., 2011).

Dementie (dus de derde fase) is een overkoepelende term voor een groep symptomen die gepaard gaan met een achteruitgang van het geheugen, denken en functioneren, die ernstig genoeg is om het dagelijks leven te beïnvloeden. Dementie wordt veroorzaakt door verschillende ziektes of aandoeningen die hersenschade veroorzaken, waaronder de ziekte van Alzheimer. Dementie als derde fase kan dan ook weer opgedeeld worden in drie fases, namelijk: milde, gematigde en ernstige dementie in de vorm van alzheimer (Arya et al., 2023).

Symptomen van geheugen- en/of andere denkproblemen kunnen zich al dan niet ontwikkelen tot alzheimer, die ernstig genoeg is om iemands vermogen om onafhankelijk te functioneren te belemmeren. Het tijdig voorspellen van het risico op progressie van een milde cognitieve stoornis naar alzheimer is dus uitermate belangrijk voor de klinische prognose, risico-inschatting en vroege interventie.

Alzheimer ontstaat door de ophoping van twee abnormale eiwitten in de hersenen: amyloïde plaques en tau tangles. *Amyloïde plaques* zijn kleverige opeenhopingen van eiwitfragmenten, die zich tussen de zenuwcellen nestelen en de onderlinge communicatie ervan verstoren. *Tau tangles* zijn verdraaide vezels van een ander eiwit dat zich in de hersencellen ophoopt, wat leidt tot een geleidelijk verlies van essentiële cellulaire functies. Deze eiwitophopingen beginnen jaren, soms decennia, voordat de eerste symptomen van geheugenverlies en cognitieve achteruitgang merkbaar worden. Het proces verloopt traag en onopgemerkt, waardoor de ziekte al ver gevorderd kan zijn tegen de tijd dat de diagnose wordt gesteld.

Het belang van vroege detectie van alzheimer mag, gezien het trage, sluipende begin van de ziekte, niet onderschat worden. Technieken die in staat zijn om de eerste tekenen van amyloïde accumulatie en andere vroege biomarkers op te sporen, zijn cruciaal. Vroege diagnose stelt patiënten in staat om eerder te beginnen met behandelingen die de progressie van de ziekte kunnen vertragen, wat resulteert in een aanzienlijke verbetering van de levenskwaliteit en verlenging van de periode van zelfstandig functioneren. Daarom is de ontwikkeling van innovatieve diagnostische methodes, zoals beeldvormende technieken en de analyse van lichaamsvloeistoffen, van groot belang in de strijd tegen alzheimer (Neve et al., 2000).

Bij het inschakelen van machine learning voor de automatische diagnose van alzheimer moet een *kwalitatieve verandering* in het ziekteproces worden voorspeld, dat houdt in: de classificatie en detectie van een milde cognitieve stoornis of van een van de fases van Alzheimer. Daarnaast moet ook een *kwantitatieve beoordeling* worden uitgevoerd, waarbij de ernst van de ziekte wordt vastgesteld aan de hand van een klinische score. Dit kan bijvoorbeeld gebeuren met behulp van de *Mini-Mental State Examination* (MMSE), een veelgebruikte test die de cognitieve functie beoordeelt en helpt bij het bepalen van de graad van cognitieve achteruitgang (Zhang et al., 2012).

De MMSE is een korte test die de cognitieve functie beoordeelt, voornamelijk bij ouderen, om dementie (zoals alzheimer) vast te stellen. Deze test onderzoekt oriëntatievermogen (plaats, tijd), geheugen (woorden onthouden en herhalen), aandacht, rekenvaardigheid, taalgebruik (voorwerpen benoemen, instructies volgen) en visueel-ruimtelijke vaardigheden (figuren tekenen). Met een maximale score van 30 punten, wordt 24 of meer als normaal beschouwd, terwijl lagere scores op mogelijke cognitieve problemen wijzen. De MMSE is een onderdeel van een breder onderzoek naar cognitieve functies en kan op zichzelf geen definitieve diagnose bieden. Het dient als screeningsinstrument om mogelijke cognitieve stoornissen te identificeren, maar moet worden gecombineerd met andere diagnostische tests en klinische evaluaties om een volledige en nauwkeurige diagnose te stellen (Cockrell & Folstein, 2002).

Computer vision en deep learning

Onderzoek heeft aangetoond dat computer vision gecombineerd met deep learning goede resultaten geeft bij het diagnosticeren van virale en niet-virale ziekten met behulp van medische beeldvorming (Verma et al., 2022). Computer vision is een AI-technologie die computers helpt beelden en video's te begrijpen, door bijvoorbeeld objecten, gezichten of teksten te herkennen. Positron-emissietomografie (PET) en Magnetic Resonance Imaging (MRI) zijn twee beeldvormingstechnieken die frequent gebruikt worden voor de diagnose van alzheimer. Bij PET worden radioactieve stoffen gebruikt om de activiteit van cellen in het lichaam te observeren voor het opsporen van kanker en

hersenstoornissen. Bij MRI worden sterke magneten en radiogolven gebruikt om gedetailleerde beelden van de binnenkant van het lichaam te maken, voor het analyseren van organen, spieren en gewrichten, zonder gebruik van straling.

Diffusion Tensor Imaging (DTI) is een vorm van MRI-technologie die wordt gebruikt om de richting en integriteit van witte stofbanen in de hersenen te evalueren. Deze witte stof bestaat voornamelijk uit axonen, die zijn omhuld door myeline, en is verantwoordelijk voor de communicatie tussen verschillende hersengebieden. Axonen zijn lange, dunne uitlopers van zenuwcellen die elektrische impulsen geleiden van het cellichaam naar andere zenuwcellen, spieren of klieren; daardoor hebben ze een sleutelpositie in het snel en efficiënt overdragen van informatie door het zenuwstelsel. DTI meet de diffusie van watermoleculen langs deze axonen. In gezond hersenweefsel bewegen watermoleculen voornamelijk langs de lengte van de axonen, maar in beschadigde of door ziekte getroffen gebieden kan deze beweging verstoord zijn. In tegenstelling tot traditionele MRI, die vooral de structuur van weefsels laat zien, biedt DTI inzicht in de richting en de snelheid van deze watermoleculenbeweging (Basser et al., 1994; Lu et al., 2018).

Bij alzheimer zijn er specifieke patronen van witte stof degeneratie die via DTI kunnen worden gedetecteerd, zelfs voordat klinische symptomen zich manifesteren. Veranderingen in de witte stof treden op in de vroege stadia van de ziekte, vooral in gebieden die betrokken zijn bij geheugen en cognitieve functies, zoals de temporale en pariëtale kwabben. Door de integriteit van de witte stof in de loop van de tijd te volgen, kunnen artsen en onderzoekers de progressie van alzheimer beoordelen en de effectiviteit van de behandeling evalueren. Integriteit van de witte stof verwijst naar de structurele kwaliteit en connectiviteit van de zenuwbanen in de hersenen.

DTI biedt dus een krachtig middel om de microstructurele veranderingen in de witte stof te onderzoeken, die geassocieerd zijn met alzheimer. Hoewel DTI op zichzelf geen definitieve diagnose biedt, draagt het wel bij aan een completer beeld van de hersengezondheid en helpt het bij het stellen van een diagnose en de behandeling van alzheimer en andere neurodegeneratieve aandoeningen. DTI wordt gebruikt in combinatie met andere diagnostische hulpmiddelen en klinische beoordelingen voor een nauwkeurige diagnose van alzheimer (Lu et al., 2018).

Deeplearningmodellen, zoals Convolutional Neural Networks (CNN's), hebben veelbelovende resultaten getoond ten opzichte van meer traditionele machinelearningbenaderingen zoals de Support Vector Machine (SVM) op het gebied van alzheimeronderzoek gevoed met PET- en MRI-scans. Toch vertoont een meer traditionele machinelearningbenadering, zoals een SVM, nog steeds scores die competitief zijn met die van een CNN.

In de studie van Subramoniam et al. (2022) wordt een model gepresenteerd waarin MRI-beelden worden gefragmenteerd en vervolgens als input worden gebruikt voor een CNN (ResNet-101) voor feature-extractie en -classificatie. Dit model classificeert en labelt beelden in vier klassen: niet-dement, zeer licht dement, licht dement en matig dement. Via drie CNN-lagen gevolgd door drie lagen van een standaard *fully connected deep neural network* (DNN) bereikt dit model een accuraatheid van 95,32 procent. Challis et al. (2015) gebruikten in hun studie een meer traditionele machinelearningbenadering, namelijk een SVM, die resultaten geeft die vergelijkbaar zijn met die van een CNN. Een nauwkeurigheid van 75 procent wordt bereikt voor detectie van een milde cognitieve stoornis ofwel Mild Cognitive Impairment (MCI), terwijl in het geval van conversie van een MCI naar alzheimer een nauwkeurigheid van 97 procent wordt bereikt.

Vroege diagnose via een audio-transformermodel

Naast visuele data, kunnen ook audiodata voor spraakherkenning gebruikt worden om alzheimer te detecteren. Alzheimer wordt namelijk gekenmerkt door een progressieve achteruitgang van cognitieve en functionele vermogens. Naarmate de ziekte vordert, treden symptomen zoals taalstoornissen, geheugenverlies, zelfverwaarlozing en gedragsproblemen op de voorgrond. Het vroeg herkennen van deze symptomen is noodzakelijk voor de diagnose en behandeling van Alzheimer. Vroege herkenning van alzheimer is belangrijk omdat het de mogelijkheid biedt om de progressie van de ziekte te vertragen voordat ernstige schade aan de hersenen optreedt.

Daarom is de analyse van spraak een belangrijke marker voor het opsporen van neurodegeneratieve aandoeningen, waaronder dus alzheimer. Zo wordt automatische spraakherkenning of *Automatic Speech Recognition* (ASR) steeds vaker toegepast voor de vroege voorspelling van alzheimer. Deze technologie biedt een snelle, kosten-effectieve, nauwkeurige en niet-invasieve methode voor de diagnose van alzheimer en voor klinische screenings (Qin, Liu et al., 2021).
Analyse van spraakdata voor alzheimer-detectie kan onderverdeeld worden in op *audio* gebaseerde en op *transcriptie* gebaseerde benaderingen. De op audio gebaseerde methodes maken gebruik van akoestische, articulatorische, fonetische en prosodische kenmerken uit het spraaksignaal (Haider et al., 2019; Ivanov et al., 2013; Yu et al., 2015).

Prosodische kenmerken zijn de aspecten van spraak die de *intonatie*, *ritme*, *toonhoogte* en *spreeksnelheid* van een spreker omvatten. Ze beklemtonen bepaalde woorden of zinnen en helpen bij het overbrengen van betekenis, zoals een vraag tegenover een bevestigende zin. Voorbeelden van prosodische kenmerken zijn pauzes in de spraak, toonhoogteveranderingen om een vraag te markeren, of een stijging en daling in de stem om nadruk te leggen op bepaalde delen van de zin. Een op audio gebaseerde analyse van de gesproken anamnese van een patiënt, zou gericht zijn op kenmerken zoals toonhoogte, spreektempo, pauzes en de duidelijkheid van de uitspraak. Als een

patiënt moeite heeft met het duidelijk uitspreken van bepaalde klanken of onregelmatige pauzes neemt, kan dit wijzen op cognitieve achteruitgang.

De op transcriptie gebaseerde methodes richten zich op kenmerken die zijn afgeleid van de tekstinhoud om taalstoornissen te karakteriseren (Ambrosini et al., 2019; Mirheidari et al., 2018). Bij een patiënt die een verhaal vertelt, zou in een op transcriptie gebaseerde analyse de gesproken tekst worden omgezet in geschreven woorden en dan worden gekeken naar aspecten zoals grammaticale fouten, woordkeuze, zinsstructuur en coherentie van het verhaal. Als de patiënt vaak verkeerde woorden gebruikt of moeite heeft om zinnen correct te vormen, kan dit ook een teken van alzheimer zijn.

Twee representatieve deeplearningmodellen die gebruik maken van ASR, zijn het *CTC-aandachtsmodel* (waarbij CTC staat voor: Connectionist Temporal Classification) en het *self-supervised transformer wav2vec model*. Het CTC-aandachtsmodel leert de juiste volgorde van gesproken woorden te voorspellen zonder dat de exacte timing van elk woord hoeft te worden gegeven. Het self-supervised transformer wav2vec model, daarentegen, maakt gebruik van een zelfsupervisie-aanpak, waarbij het leert van grote hoeveelheden ongelabelde spraakdata. Bij zelfsupervisie leert het model representaties van data door zichzelf te trainen met behulp van de inherente structuur van de data, zonder de noodzaak van handmatig gelabelde voorbeelden. Wav2vec bestaat uit twee fasen: eerst wordt een *speech waveform* omgezet in een *latente representatie*. Vervolgens wordt deze representatie verfijnd met behulp van contrasterend leren, waarbij het model onderscheid maakt tussen correct en incorrect gesynchroniseerde spraaksegmenten. Een speech waveform is een visuele weergave van het spraaksignaal in de tijd, waarbij de horizontale as de tijd weergeeft en de verticale as de amplitude (geluidsintensiteit) van het signaal. Een latente representatie is een compactere en abstractere vorm van het oorspronkelijke signaal, die complexe patronen en kenmerken van het spraaksignaal opslaat zonder alle ruwe details te behouden.

Dankzij deze aanpak kan wav2vec effectief patronen en kenmerken herkennen die relevant zijn voor de vroege signalering van alzheimer, zoals subtiele veranderingen in spraak en prosodie. Wav2vec overtreft zijn voorganger, CTC-aandachtsmodel, bij het vroegtijdig signaleren van alzheimer (Baevski et al., 2020; Zhang, Lu et al., 2020). Het diagnosticeren van alzheimer via machine learning gaat echter verder dan het analyseren van alleen maar *gestructureerde* tekstdata, audiodata of MRI- en PET-scans. Voor een vroege diagnose van deze ziekte is het essentieel dat deze data worden gecombineerd met *ongestructureerde* patiëntinformatie, zoals leeftijd, genetica, opleidingsniveau, medische historie en aanvullende gezondheidsproblemen. Deze multidimensionale aanpak kan het diagnostisch proces verrijken, waardoor het mogelijk wordt om een nauwkeurigere en holistische beoordeling van het risico op

alzheimer te geven. Dit maakt vroege interventie en het ontwikkelen van gepersonaliseerde behandelplannen beter mogelijk.

Het gebruik van ongestructureerde data voor de vroege detectie van de ziekte van Alzheimer

Alzheimer wordt veroorzaakt door een overmatige opbouw van proteïnen in de hersenen, wat hersencellen beschadigt. Deze ophopingen ontstaan vaak al jaren voordat de eerste symptomen (zoals geheugenverlies) zichtbaar worden, waardoor de ziekte al ver gevorderd kan zijn op het moment van de diagnose. Dit maakt vroege detectie van de allereerste symptomen van deze ziekte uiterst belangrijk.

Ongestructureerde data uit EPD's kunnen een belangrijke rol spelen bij de vroege detectie van alzheimer. Deze data kunnen klinische notities, rapporten en laboratoriumuitslagen bevatten, die niet in een gestandaardiseerd formaat zijn opgeslagen. Het analyseren van deze ongestructureerde gegevens met behulp van data-analyse en machinelearningtechnieken kan helpen bij het identificeren van subtiele patronen en aanwijzingen die wijzen op een vroeg stadium van alzheimer. Ook akoestische data kunnen hier een bijdrage leveren, door bijvoorbeeld veranderingen in spraakpatronen te signaleren (Mao, Xu et al., 2023). Wang et al. (2021) hebben een deeplearningmodel gebruikt om indicaties van cognitieve achteruitgang te detecteren uit klinische notities van vier jaar voor de eerste diagnose van een MCI.

In de afgelopen jaren hebben op transformers gebaseerde taalmodellen, zoals BERT en Generative Pre-trained Transformer (GPT), de grenzen verlegd voor diverse NLP-taken zoals het beantwoorden van vragen, classificeren van documenten en genereren van tekst. Na het observeren van het succes van deze modellen in het algemene domein, hebben onderzoekers domeinspecifieke biomedische varianten van BERT uitgebracht, zoals eerder al vermeld. De studie van Mao, Xu et al. (2023) schetst een domein-specifiek BERT-transformermodel voor de detectie van alzheimer (AD-BERT), dat gevoed is met ongestructureerde klinische notities uit EPD's om het risico op ziekteprogressie van een MCI naar alzheimer te voorspellen.

Naar een multimodale aanpak voor vroege detectie van de ziekte van Alzheimer

In eerdere paragrafen heb ik transformermodellen beschreven die zich richten op tekstuele en auditieve data voor de detectie van alzheimer. De op tekst (transcripties) gebaseerde modellen analyseren de inhoud van gesproken taal om taalstoornissen te identificeren, terwijl de op audio gebaseerde modellen akoestische en prosodische kenmerken van spraak analyseren. Gezien het feit dat gezondheidszorg inherent multimodaal is, waarbij verschillende soorten data samenkomen om een completer beeld van de patiënt te krijgen, is het logisch om de op tekst en audio gebaseerde

benaderingen te combineren. Dit leidt naar de ontwikkeling van multimodale transformermodellen die spraak, tekst en visuele data combineren. Deze modellen kunnen niet alleen taal- en spraakpatronen analyseren, maar ook visuele aanwijzingen, zoals gezichtsuitdrukkingen en motorische vaardigheden, waardoor een uitgebreidere en nauwkeurigere diagnose van alzheimer mogelijk wordt.

Toekomstig onderzoek naar de vroege detectie van alzheimer zou zich moeten richten op de ontwikkeling van multimodale machinelearningmodellen. Deze modellen integreren diverse datastromen, idealiter visuele en auditieve data, alsook gestructureerde en ongestructureerde data, om de diagnostische nauwkeurigheid te vergroten. De inzet van MRI- en PET-beeldvorming, gecombineerd met auditieve analyse van spraakpatronen, biedt een rijke dataset voor de identificatie van vroege indicatoren van alzheimer.

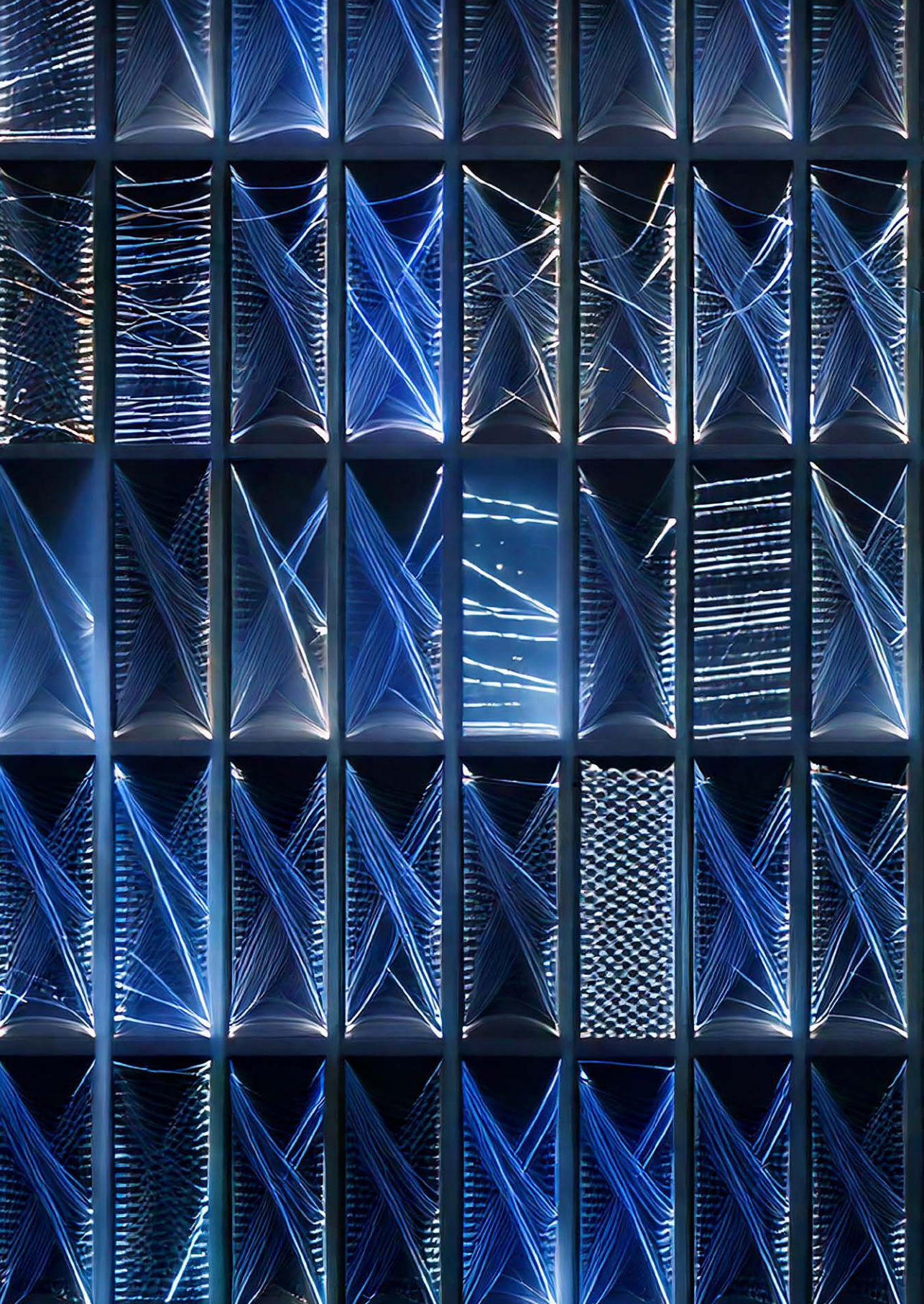
Met de multimodale benadering wordt gestreefd naar het onthullen van complexe patronen die indicatief zijn voor het beginstadium van alzheimer, mogelijk jaren voor de manifestatie van de eerste symptomen. De ontwikkeling van dergelijke modellen vereist echter een nauwe samenwerking tussen neurologen, data scientists en AI-specialisten, en kan een significante impact hebben op de vooruitgang in de vroege diagnose van alzheimer. Deze aanpak zou ook de deur kunnen openen naar een meer gepersonaliseerde behandeling, om de progressie van de ziekte te vertragen en de levenskwaliteit van patiënten te verbeteren. Een multimodale aanpak kan namelijk unieke kenmerken van elke patiënt in kaart brengen, waardoor behandelingen beter kunnen worden afgestemd op individuele ziektepatronen. Dit maakt gerichtere therapieën mogelijk, die beter aansluiten bij de specifieke behoeften van de patiënt.
De onderzoeksliteratuur vermeldt verschillende studies die in de richting gaan van een multimodale aanpak, maar die combineren meestal niet meer dan twee bronnen van informatie. Zo maken de studies van Zhang et al. (2023) en Gao et al. (2023) gebruik van een combinatie van PET- en MRI-scans in een multimodaal transformer model. De resultaten tonen voor deze gecombineerde benadering betere prestaties dan een op alleen MRI- of PET-scans getraind deeplearningmodel. De studie van lias en Askounis (2022) beschrijft een *multitask-benadering* die twee transformergebaseerde machinelearningmodellen combineert voor het gezamenlijk detecteren van alzheimer en het meten van de graad van progressie via het voorspellen van de mini-mental State Examination.

Een van de weinige studies over automatische vroege detectie van alzheimer die beeld-, gestructureerde en ongestructureerde data combineert, is het onderzoek van Chen et al. (2024). Dit onderzoek introduceert een transformermodel dat multimodale medische gegevens samenvoegt, waaronder MRI-scans, ongestructureerde teksten (zoals artsenrapporten en diagnoses) en gestructureerde data (zoals demografische gegevens en laboratoriumtestresultaten).

In meerdere studies over alzheimer (Challis et al., 2015; Chen et al., 2024; Gao et al., 2023; Zhang et al., 2023) werd met dezelfde Amerikaans/Engelse ADNI-dataset[9] (Mueller et al., 2005) gewerkt. In het onderzoek dat ik hierover wil opzetten met het lectoraat, is mijn streven om met Europese data te werken. Hiertoe moet er samengewerkt worden met nationale en internationale Europese medische instellingen om hun data over alzheimerpatiënten samen te brengen en in een veilige omgeving beschikbaar te stellen voor onderzoek dat AI toepast op de medische sector.

Naast de diagnose van alzheimer, kunnen machinelearningtechnieken een significante impact hebben op andere domeinen binnen de gezondheidszorg en het onderwijs. In het volgende hoofdstuk wordt verkend hoe machine learning wordt ingezet voor het verbeteren van de mentale gezondheidszorg en het verrijken van de leerervaring.

9 https://adni.loni.usc.edu/



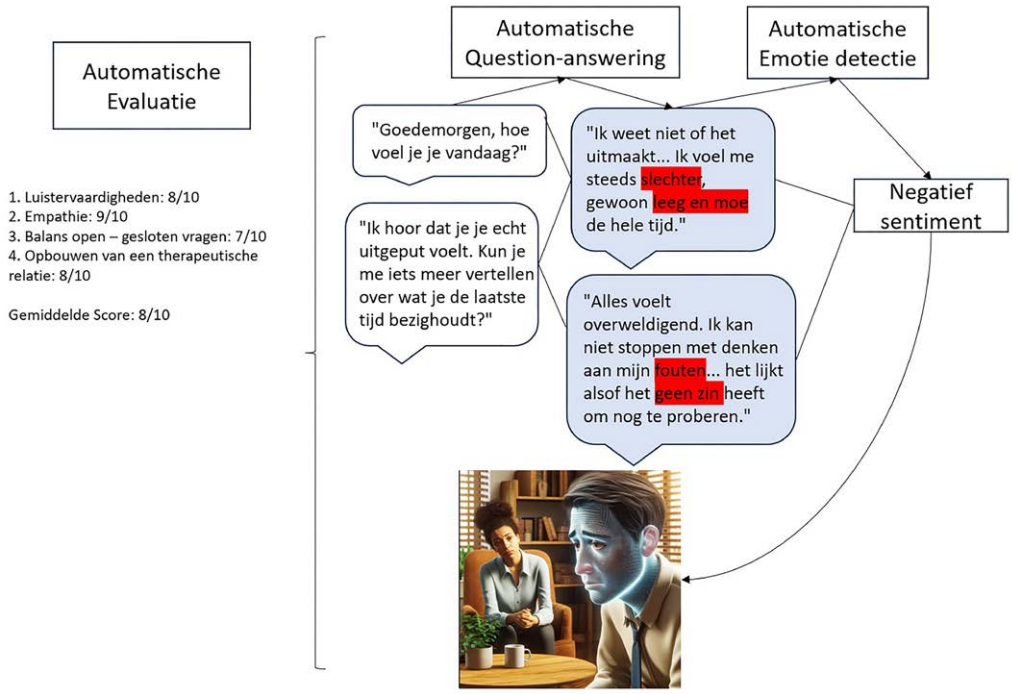
5 Emotiedetectie met taalmodellen voor mentale gezondheidszorg en onderwijs

Inleiding

Deeplearningtransformermodellen zijn niet alleen geschikt voor directe medische toepassingen, maar spelen ook een rol in het onderwijs en de mentale gezondheidszorg. De COVID-19-pandemie heeft de integratie van digitale en op machine learning gebaseerde oplossingen in deze domeinen versneld. Dit heeft niet alleen geleid tot verbeterde educatieve mogelijkheden, maar ook tot een verbeterde toegang tot en kwaliteit van geestelijke gezondheidszorg voor een breder publiek (Su et al., 2020). De toename van het aantal mensen met psychische problemen zoals angst, depressie en stress tijdens de pandemie, heeft geleid tot innovatieve benaderingen in de geestelijke gezondheidszorg, waaronder het gebruik van virtuele therapeuten en patiënten in psychologie- en psychotherapie-sessies.

Patiënten die anders geen toegang zouden hebben tot geestelijke gezondheidszorg, kunnen via teletherapie worden geholpen. Belemmeringen voor toegang tot deze zorg kunnen variëren van een gebrek aan beschikbare zorgverleners en financiële beperkingen tot mobiliteitsproblemen. Daarnaast kunnen stigma rond mentale gezondheid en taalbarrières mensen ontmoedigen om hulp te zoeken. Door de integratie van machine learning in virtuele patiënten kunnen deze continu leren en zich aanpassen aan de behoeften van individuele patiënten, wat leidt tot effectievere en efficiëntere behandelingen (Caponnetto & Casu, 2022; Stefan et al., 2021).

Virtuele patiënten kunnen psychologen in opleiding ondersteunen. Tijdens dit lectoraat wordt er onderzoek opgestart naar virtuele patiënten. Aangedreven door machine learning, kunnen zij een veilige en gecontroleerde omgeving bieden waarin studenten klinische vaardigheden kunnen oefenen en verfijnen.



Figuur 6 Systeem voor het houden van een dialoog met een virtuele patiënt, met emotiedetectie

Machine learning voor de interactie tussen een virtuele patiënt en psychotherapeut of psycholoog

Virtuele personages bieden over het algemeen interessante interactiemogelijkheden, wat ze des te meer geschikt maakt voor onderzoek naar gesimuleerde medische behandelingen en het effect ervan. Zo opent het ontwerpen van virtuele patiënten, die vaardigheden bezitten om te communiceren (zoals voorgesteld in figuur 6), mogelijkheden voor het oefenen en verbeteren van klinische interviewtechnieken, diagnostische beoordeling en therapietraining (Kenny et al., 2007). Op die manier kunnen studenten bepaalde scenario's herhaaldelijk doorlopen, waarbij ze fouten kunnen maken, van deze fouten kunnen leren, kunnen reflecteren, feedback krijgen, een meer gepersonaliseerd leertraject kunnen volgen en hun klinische vaardigheden kunnen aanscherpen zonder de veiligheid van patiënten in gevaar te brengen. Bovendien helpt het de stress te verminderen die komt kijken bij een eerste directe interactie met echte patiënten.

Voor een interactieve dialoog, als onderdeel van het oefenen met een virtuele patiënt, gebruikten NLP-systemen in een vroege fase een combinatie van regels en een kennisdatabase met vraag-antwoordparen. Lexicale bronnen zoals SentiWordNet (Baccianella et al., 2010), Pattern (Tulkens et al., 2016), Affect-WordNet (Strapparava & Valitutti, 2004) en SenticNet (Cambria et al., 2022) speelden hierin een rol. Deze databases en lexicale bronnen dienden dan als de bron voor het AI-systeem. Gebruikers konden tijdens simulaties vragen stellen aan een virtuele patiënt, waarna het systeem via algoritmes en zoekpatronen probeerde de vraag te begrijpen en te matchen met een geschikt antwoord uit de kennisbank. Het correcte antwoord werd dan getoond (Epstein et al., 2013).

Deze methodes schoten echter tekort in het omgaan met de complexiteit van interactie met patiënten. De overgang naar machine learning bood hier een antwoord, door de mogelijkheid voor onderzoekers om fijnere nuances in de communicatie met de patiënt vast te leggen via *feature engineering,* door specifieke features (kenmerken) van ziektebeelden (zoals concentratiestoornissen bij schizofrenie) aan een machine-learningsysteem mee te geven. Deze aanpak werd echter beperkt door de intensieve handmatige selectie van typische kenmerken en de noodzaak van uitgebreide datasets door de feature engineer. Een stroomversnelling kwam er met de opkomst van deep learning en grote taalmodellen binnen AI (Vaswani et al., 2017), zoals die gebruikt worden in ChatGPT en die zelf autonoom typische kenmerken van een bepaald ziektebeeld uit de communicatiestijl weten af te leiden, zonder tussenkomst van een feature engineer.

Emotiedetectie

Het vermogen om emoties nauwkeurig te detecteren en te begrijpen, speelt een fundamentele rol in zowel interactieve dialogen als in de ontwikkeling van virtuele patiënten voor de geestelijke gezondheidszorg. Emoties vormen een fundamenteel aspect van menselijke communicatie en welzijn, en de juiste interpretatie van emoties kan de effectiviteit van therapeutische interventies verbeteren. Door emotiedetectie via machine learning te integreren in virtuele patiënten, kunnen deze systemen empathischer reageren op en beter aansluiten bij de emotionele behoeften van gebruikers, waardoor ze waardevolle ondersteuning kunnen bieden in therapeutische contexten.

In de zorgsector kan emotiedetectie een belangrijke rol spelen, zowel in de patiëntenzorg als in klinisch onderzoek en beoordeling van de mentale gezondheid. Met inzicht in de emotionele staat van patiënten kunnen zorgverleners op de behoeften van die patiënten inspelen, wat leidt tot een grotere tevredenheid over de geleverde diensten. Dit aspect wordt ook geïntegreerd in trainingsprogramma's voor zorgpersoneel, waarbij communicatievaardigheden en empathie afgestemd worden op de

gedetecteerde emoties. In de geestelijke gezondheidszorg helpt het onderzoek naar emotieregulatie bij het opsporen van mentale stoornissen, wat een rol speelt bij het begrijpen en behandelen van deze aandoeningen (Dheera & Ramakrishnudu, 2021).
Grote taalmodellen, gevoed door tekstdialogen, kunnen niet alleen *inhoudelijk* reageren op gesprekken met de therapeut, maar ook de *emotionele staat* van een virtuele patiënt weergeven. Ze herkennen emotionele nuances, zoals 'angstig' en 'boos' bij angststoornissen, en 'slecht humeur' en 'slapeloosheid' bij depressie, wat leidt tot meer empathische en doelgerichte communicatie tijdens interacties met virtuele patiënten (Tao et al., 2023). Deze taalmodellen kunnen natuurlijke taal begrijpen en genereren, waarin ook emoties worden uitgedrukt (Ng et al., 2023; Yang et al., 2023).

In de studie van Llanes-Jurado et al. (2024) wordt aandacht besteed aan de belangrijke rol van een interactieve dialoogmodule door middel van een groot taalmodel. Het GPT-3-taalmodel van OpenAI werd hiervoor gebruikt, meer specifiek de text-davinci-002-variant. Deze module verwerkt een verscheidenheid aan inputs, zoals het emotionele verhaal dat aspecten zoals levensgeschiedenis, context, houdingen, stemming en motivatie omvat, en de voorgaande gesprekken.

Voor emotiedetectie maken zowel vooraf getrainde grote taalmodellen zoals BERT en zijn biomedische varianten (zoals BioBERT en ClinicalBERT) als ChatGPT van OpenAI hun opwachting in domeinspecifieke toepassingen. Een duidelijk voordeel van *open-source* grote taalmodellen ten opzichte van *commerciële* modellen zoals ChatGPT is hun aanpasbaarheid en mogelijkheid tot specialisatie voor specifieke domeinen, met name binnen biomedische en klinische contexten (Alshouha et al., 2024). Dit maakt open source modellen bij uitstek geschikt voor emotiedetectie en het verkennen van het volledige potentieel van NLP in medisch onderzoek en patiëntenzorg. Desondanks hebben open source modellen ook hun beperkingen, zoals een mogelijk gebrek aan initiatief voor verdere verduidelijking van vragen (Ray, 2023).

Multimodale taalmodellen die tekst en beeld combineren

Ten opzichte van transformermodellen die alleen met tekstdata omgaan, kunnen recente multimodale grote taalmodellen de interacties met virtuele patiënten verbeteren door niet alleen *verbale* en geschreven input te analyseren, maar ook te reageren op *non-verbale* signalen zoals gezichtsuitdrukkingen en lichaamsbewegingen, dankzij beeldherkenningstechnologie. Dit is heel relevant bij detectie van non-verbale kenmerken, zoals een negatieve aandachtsbias en veranderd bewegingsgedrag bij depressieve patiënten. Een negatieve aandachtsbias is een cognitieve neiging waarbij iemand meer aandacht besteedt aan negatieve stimuli, zoals verdrietige gezichten of negatieve woorden, en deze informatie ook sneller opmerkt of langer onthoudt. Bij depressieve patiënten kan dit leiden tot een focus op negatieve gebeurtenissen of gedachten, wat de symptomen van depressie kan versterken en de klachten in stand kan houden (Tao et al., 2023).

In de context van mentale gezondheidszorg, die intrinsiek multimodaal is, toont dit voorbeeld het belang aan van het integreren van diverse signalen en modaliteiten voor een diepgaander begrip en effectievere interventies.

De bekwaamheid van een arts om meerdere zintuiglijke inputkanalen te integreren, is uiteraard onmisbaar. Tijdens een consult let een arts zowel op verbale als visuele informatie. Een multimodaal taalmodel kan de verwerking van tekst en spraak combineren met de verwerking van visuele data, zoals gezichtsuitdrukkingen. Het model reageert dan niet alleen op wat er gezegd wordt, maar ook op hoe het wordt gezegd, of welke emoties er bijvoorbeeld verder via lichaamstaal en in het gezicht worden uitgedrukt (Safranek et al., 2023). Onderzoek en experimenten met multimodale modellen kunnen uitwijzen of ze geschikt zijn voor toepassingen zoals het simuleren van gesprekken met virtuele patiënten waarin emoties verbaal en visueel worden uitgedrukt.

Onderzoek heeft verder aangetoond dat patiënten met schizofrenie moeilijkheden ondervinden in hun non-verbale communicatie, wat invloed heeft op zowel hun vermogen om sociale signalen te begrijpen als de manier waarop ze gebaren gebruiken. Deze tekortkomingen in non-verbale communicatie, waaronder de interpretatie van gebaren, gezichtsuitdrukkingen en lichaamstaal, zijn deels te wijten aan een gebrekkige non-verbale sociale perceptie en aan negatieve symptomen en motorische afwijkingen. De bevindingen benadrukken het belang van sociale cognitie en de interpretatie van sociaal relevante stimuli, die ernstig zijn verstoord in het geval van schizofrenie (Walther et al., 2015). Virtuele patiënten met schizofrenie zouden dus op een multimodale manier gemodelleerd moeten worden, zodat een diagnose niet alleen gebaseerd hoeft te worden op interactieve dialogen, maar ook rekening houdt met de daarbij horende gezichtsuitdrukkingen en lichaamstaal.

Spraak in combinatie met tekst en beeld in een multimodaal model voor diagnose van depressie

Het vaststellen van een klinische depressie is een flinke uitdaging voor zorgprofessionals en het is dus ook een uitdaging om depressieve virtuele patiënten te modelleren. Er bestaat geen biologische gouden standaard voor; je kunt een depressie niet zien in een bloedtest of een scan (Kapur et al., 2012; Thomas-MacLean et al., 2005). Meestal wordt er gekeken naar *wat* iemand vertelt tijdens gesprekken en wat er uit vragenlijsten naar voren komt. Tegenwoordig wordt er ook gekeken naar *hoe* iemand praat en zich gedraagt. Bijzonder is dat de manier van spreken, dus niet alleen wát er gezegd wordt, heel waardevol blijkt te zijn voor het automatisch herkennen of iemand depressief is (Williamson et al., 2016).

De studies van Stehwien en Vu (2016) en Su en Tseng (2018) suggereren dat de manier waarop iets wordt gezegd (prosodische informatie) aanwijzingen kan bieden over waar de meest relevante semantische informatie in gesproken taal te vinden is. Prosodie is de studie van de melodie en het ritme van gesproken taal. Prosodie verwijst naar de variaties in toonhoogte, duur en intensiteit van lettergrepen, evenals het gebruik van pauzes binnen zinnen, die helpen om betekenis over te brengen, bepaalde woorden of zinsdelen te benadrukken en de structuur en het ritme van gesproken taal te organiseren. Het bepaalt de emotionele lading van een uiting en geeft luisteraars belangrijke informatie over de intentie van de spreker en de focus van de boodschap. Prosodische informatie helpt luisteraars bijvoorbeeld om vragende zinnen van bevestigende zinnen te onderscheiden, emotionele toestanden van de spreker te interpreteren en de semantisch meest prominente delen van een boodschap te identificeren.

Hier kunnen onder andere filterbank-features (fbank) en MFCC-features (waarbij MFCC staat voor: mel-frequentie cepstrale coëfficiënt) een rol spelen. Dit zijn audiofeatures die een rol spelen bij spraakherkenning. Door audio en video samen te brengen, is er een grotere kans dat er gedetecteerd kan worden of iemand depressief is (Muzammel et al., 2021). MFCC- en fbank-features bevatten de unieke eigenschappen van iemands stem, die beschouwd kunnen worden als een vingerafdruk. Deze audiofeatures maken het mogelijk om die vingerafdruk te analyseren, om op die manier meer te kunnen onthullen over de emotionele toestand van een persoon, zoals mogelijke tekenen van depressie, enkel door naar de stem te luisteren.

Fbank audiofeatures zijn gebaseerd op het groeperen van frequenties in banden of 'filters', die lijken op hoe het menselijk oor geluidsfrequenties waarneemt. Deze filters verdelen het geluid in verschillende frequentiebanden en meten de energie in elk van deze banden. Het idee is dat niet elke frequentie even belangrijk is voor hoe mensen geluid waarnemen, dus door te focussen op specifieke banden krijgen mensen nuttige informatie over het geluid. Dit is te vergelijken met het opsplitsen van een muziekstuk in verschillende instrumentgroepen, waarbij elke instrumentgroep een bepaalde typische klankkleur heeft. Fbank-features kijken naar hoeveel energie er in elke instrumentgroep zit.

De MFCC-features gaan een stap verder dan de filterbank-features. Nadat het geluid is opgedeeld in verschillende frequentiebanden, worden de energieën in deze banden omgezet naar een logaritmische schaal, zodat de representatie beter overeenkomt met de menselijke gehoorperceptie. Daarna wordt er een discrete cosinustransformatie (DCT) op toegepast. Dit proces resulteert in een set van waarden (de MFCC's), die een zeer compacte representatie van het geluid geven, waarbij vooral de meest belangrijke kenmerken van het geluidssignaal worden benadrukt. Het voornaamste verschil tussen fbank-features en MFCC-features is dus dat MFCC-features een extra stap van

transformatie ondergaan (de DCT), die helpt bij het verder comprimeren en benadrukken van de belangrijkste aspecten van de audio voor analyse. Fbank-features geven de ruwe energieën in de filterbanden aan, terwijl MFCC-features deze informatie verfijnen tot een set van coëfficiënten die de unieke kenmerken van het geluid beter vastleggen (Davis & Mermelstein, 1980). Dit maakt MFCC-features vaak effectiever voor taken zoals spraakherkenning, waarin het belangrijk is om de essentiële kenmerken van het geluidssignaal te identificeren en te isoleren. Echter voor de meest recente deep-learning-spraakherkenningsmodellen worden de fbank-features vaker gebruikt, die de ruwe energieën weergeven, maar tegelijk rijkere informatie bevatten.

Vormen van fusie van diverse informatiekanalen in een multimodaal model

Om depressie of schizofrenie te herkennen, halen multimodale modellen informatie uit meerdere kanalen: gezichtsuitdrukkingen, lichaamstaal, wat mensen zeggen en hoe ze dit zeggen. Verbale, audio- en visuele features kunnen dan samengevoegd worden in een multimodaal deeplearningmodel. Er zijn meerdere manieren om al deze informatie samen te brengen.

Bij een *vroege samenvoeging* (early fusion) van meerdere informatiekanalen zoals verbale, audio- en visuele bronnen, worden alle gegevens uit de verschillende bronnen al snel na het verzamelen samengebracht in één grote berg informatie (Zhang, Lin et al., 2020) om verder met machinelearningtechnieken verwerkt te worden.

Bij een *latere samenvoeging* (late fusion) worden eerst aparte beslissingen genomen op basis van de verschillende soorten informatie, zoals beeld en geluid, waarna deze beslissingen samen worden gevoegd (Samareh et al., 2018).

Een *gemengde aanpak* (hybrid fusion) combineert deze twee: er wordt vroeg informatie samengevoegd, maar ook de uitkomsten van verschillende tests worden op het einde samengevoegd (Yang, Jiang et al., 2017).

Bij *modelniveau-samenvoeging* (model fusion) worden de verschillende soorten informatie geïntegreerd door deeplearningtechnieken toe te passen op alle informatiebronnen tegelijk. Dit gebeurt nadat de afzonderlijke gegevens zijn verwerkt en gecombineerd tot één coherente dataset, waarop het model wordt getraind om patronen en correlaties te herkennen (Rejaibi et al., 2022; Tzirakis et al., 2017; Yang, Sahli et al., 2017).

Deze inzichten in multimodaliteit, namelijk het begrijpen en interpreteren van meerdere communicatiekanalen en de manier waarop ze worden samengebracht, openen deuren voor de ontwikkeling van multimodale machinelearningtechnieken. Op die manier kunnen diverse datastromen, zoals spraak, gebaren en gezichtsuitdrukkingen,

gecombineerd worden voor simulatie van de interactie tussen virtuele patiënt en zorg-professional.

Naar responsible AI toegepast op gezondheidszorg

Ik heb in dit en vorige hoofdstukken de onderzoeksmogelijkheden belicht van het gebruik van deep learning, transformer- en multimodale transformermodellen, voor de vroege voorspelling van sepsis en alzheimer en voor de creatie van virtuele patiënten. Hoewel deze onderzoekstechnieken veelbelovende perspectieven bieden, blijft deep learning een zogenaamde *black box*. Dit betekent dat het weliswaar mogelijk is om van input naar output te gaan, maar dat het niet volledig duidelijk is hoe deze deeplearningmodellen tot hun beslissingen komen. Bovendien kunnen de ontwikkelde applicaties vaak geen handelingskaders bieden die een gewetensvolle en competente inzet van AI in de medische praktijk waarborgen, wat risico's met zich meebrengt.
Daarom is een belangrijke kwestie die hierbij speelt, het concept van *responsible AI*, oftewel verantwoordelijke kunstmatige intelligentie. Dit houdt in dat AI-systemen niet alleen efficiënt moeten zijn, maar ook eerlijk, transparant en begrijpelijk voor de gebruikers. Ze moeten beslissingen nemen op een manier die ethisch verantwoord is en geen ongerechtvaardigde vooroordelen of discriminatie bevordert. Naast de beslissingen van het AI-systeem gaat het er ook om hoe de AI-applicatie in de context is ingezet. Zijn de mensen die werken met AI-toepassingen voldoende getraind? Hoe wordt het AI-systeem gemonitord zodat het ook na leren en updates functioneert zoals de bedoeling is? Welke controle hebben gebruikers en/of patiënten over het systeem?

Naast het concept van responsible AI zijn uiteraard ook dataveiligheid en privacy cruciale aspecten. AI-systemen verwerken vaak grote hoeveelheden gevoelige en persoonlijke gegevens, wat een risico kan vormen als deze gegevens niet adequaat worden beschermd. Het is essentieel dat AI-toepassingen voldoen aan strikte veiligheidsnormen en aan de privacywetgeving, zoals de Algemene Verordening Gegevensbescherming (AVG) in Nederland en de General Data Protection Regulation (GDPR; European Commission, 2018) in Europa. Deze regels betreffen het waarborgen van de anonimiteit van gebruikers, het toepassen van sterke versleutelingstechnieken en het beperken van de toegang tot persoonlijke gegevens tot alleen diegenen die deze absoluut nodig hebben. Gebruikers moeten niet alleen geïnformeerd worden over hoe hun gegevens worden gebruikt, maar ook controle hebben over hun eigen data, om hun privacyrechten te beschermen.

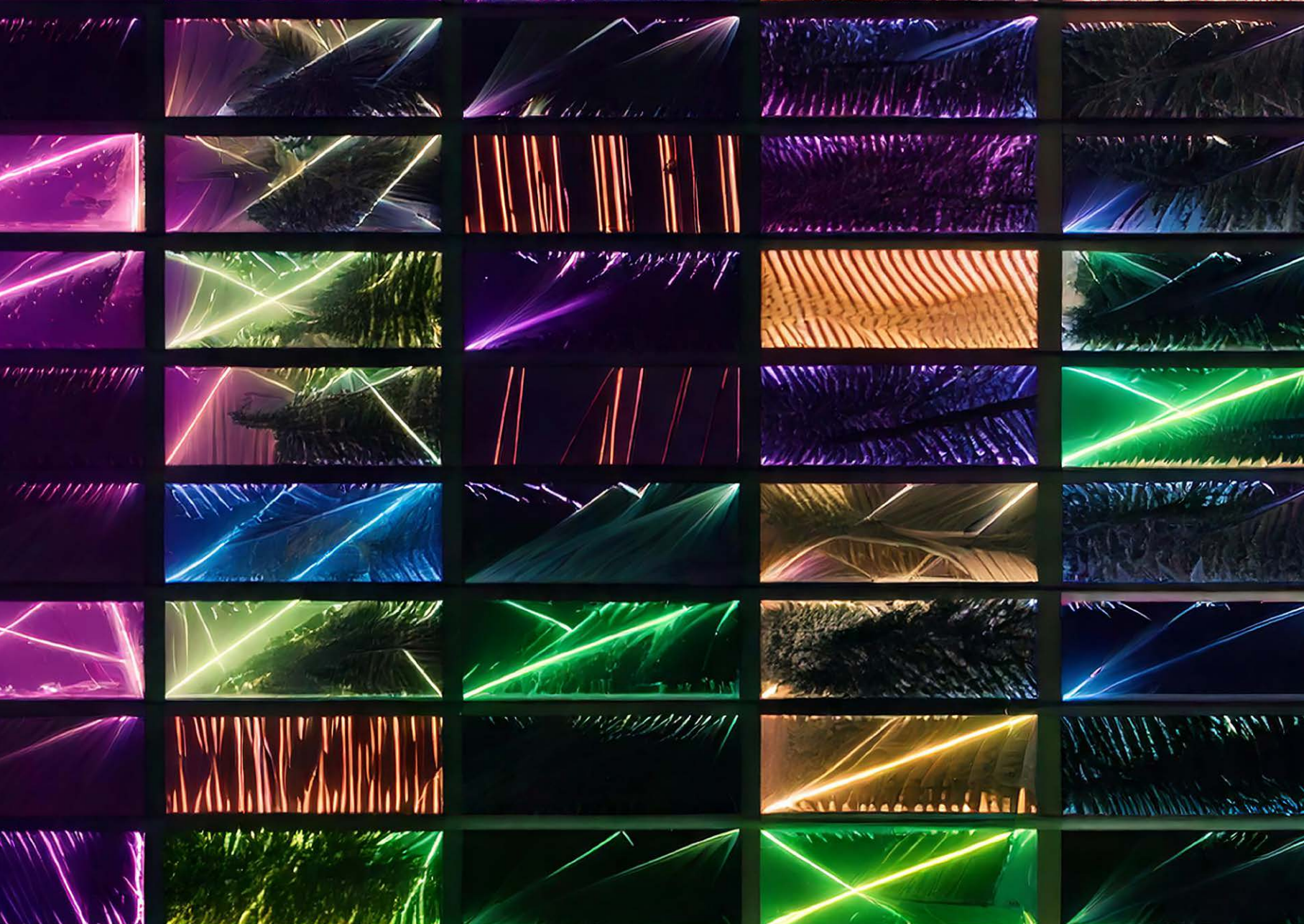
In het bijzonder voor de zorgsector is responsible AI noodzakelijk. De gezondheidszorg is een domein waar beslissingen letterlijk van levensbelang kunnen zijn en waar ethische overwegingen en patiëntveiligheid voorop staan. De inzet van AI voor het creëren van virtuele patiënten en voor de vroege voorspelling van alzheimer en sepsis, toont het potentieel om de kwaliteit van zorg te verbeteren en levens te redden. Echter, de black-box-aard van deep learning in AI brengt extra uitdagingen met zich mee binnen deze gevoelige context. Zonder volledige transparantie over hoe beslissingen

worden genomen, kunnen er vraagtekens gezet worden bij de betrouwbaarheid van diagnoses of behandeladviezen, wat kan leiden tot wantrouwen bij zowel zorgprofessionals als patiënten.

Het gebrek aan transparantie en het risico op vooringenomenheid (bias) in AI-modellen vormen serieuze uitdagingen bij het realiseren van responsible AI. Een van de grote uitdagingen binnen responsible AI is dan ook het *detecteren* en *mitigeren* (verminderen) van bias in machinelearningmodellen. Bias kan leiden tot oneerlijke of ongerechtvaardigde beslissingen, vooral in gevoelige sectoren zoals de gezondheidszorg.

Bias kan bijvoorbeeld leiden tot een AI-systeem dat vaker foute diagnoses stelt voor minderheidsgroepen omdat de trainingsdata voornamelijk afkomstig waren van een specifieke bevolkingsgroep. Dit kan resulteren in ongelijke toegang tot zorg of verkeerde behandelingen, wat leidt tot oneerlijke en ongerechtvaardigde beslissingen voor die patiënten. Daarom moet zeker aandacht besteed worden aan onderzoek naar *detectie en mitigatie (vermindering) van bias*. Door systematisch bias te identificeren en te verminderen, kunnen AI-systemen rechtvaardiger en meer betrouwbaar worden, wat de acceptatie en implementatie van deze technologieën in de gezondheidszorg kan bevorderen.

Onderzoek van dit lectoraat wil bijdragen aan de ethische inzet van AI en NLP in de gezondheidszorg, die de rechten van patiënten waarborgen en een inclusieve en eerlijke zorgomgeving bevorderen. Daarom zal ik in het volgende hoofdstuk een belangrijk aspect van responsible AI verkennen en toelichten, namelijk het detecteren en mitigeren van bias in grote taalmodellen voor de zorgsector. Hierbij leg ik eveneens de link met de vroege voorspelling van sepsis, diagnose van alzheimer en de ontwikkeling van virtuele patiënten.



6 Toegepast onderzoek naar biasdetectie en -mitigatie in grote taalmodellen in medische toepassingen

Elon Musk heeft raketten ontwikkeld en gelanceerd, met als eerste doel de mensheid in geval van wereldwijde rampen een vluchtroute naar Mars te bieden. Hij voegde later kunstmatige intelligentie toe aan de lijst van potentiële bedreigingen, waarschuwend voor scenario's waarin AI het menselijke intellect zou kunnen overtreffen en mogelijk vernietigen. Op een verjaardagsfeestje in 2013 had Musk een debat met Larry Page (voormalig CEO van Google), waarin Page betoogde dat superieure AI een natuurlijke evolutie is, terwijl Musk stelde dat het menselijk bewustzijn behouden moet blijven als een uniek licht in het universum. Bezorgd over mogelijke onveilige praktijken van techgiganten zoals Google, steunde Musk opensource-initiatieven voor AI (waaronder OpenAI), met als doel AI te democratiseren en te beschermen tegen monopolisering. In 2015 benaderde Musk de Amerikaanse president Obama om wetgeving voor AI-regulering en veiligheidsmechanismen te pleiten, maar Obama ondernam geen verdere actie (Isaacson, 2023). Echter, er verscheen een lichtpuntje aan de Europese horizon: in april 2021 introduceerde de Europese Commissie een regelgevend kader voor AI. Dit ambitieuze voorstel classificeerde AI-systemen op basis van de risico's die ze opleveren, met een focus op veiligheid, transparantie en eerlijkheid.

Onderzoekscontext voor biasdetectie en -mitigatie

De EU AI Act, het eerste regelgevende kader en pionierswerk voor AI, dat de Europese Commissie in april 2021 introduceerde, bevat categorieën voor AI-systemen die van toepassing zijn op verschillende domeinen, afhankelijk van de risico's die ze voor gebruikers met zich meebrengen (European Parliament, 2023). Het primaire doel is ervoor te zorgen dat in de EU gebruikte AI-systemen voldoen aan normen voor veiligheid, transparantie, traceerbaarheid, non-discriminatie en milieubewustzijn, en bijdragen aan het bevorderen van eerlijkheid (*fairness*) in AI. Het concept van *eerlijke* AI benadrukt de noodzaak om AI-systemen en -algoritmes zo te ontwerpen en te implementeren dat stereotypering, vooroordelen en discriminatie worden voorkomen. Het uiteindelijke doel is om individuen eerlijk en onpartijdig te behandelen, ongeacht persoonlijke kenmerken zoals geslacht en etniciteit.

Verschillende risiconiveaus van AI-systemen vereisen verschillende graden van regulering. AI-systemen met een *beperkt risico* moeten voldoen aan minimale transparantie-eisen, inclusief systemen die zijn ontworpen voor het genereren of manipuleren van beeld-, audio-, of videomateriaal, zoals deepfakes. AI-systemen die *onaanvaardbare risico's* met zich meebrengen of worden beschouwd als bedreigingen voor individuen, moeten worden verboden, zoals spraakgestuurd speelgoed dat gevaarlijk gedrag bij kinderen aanmoedigt.

AI-toepassingen met een *hoog risico*, die invloed hebben op veiligheid of fundamentele rechten (zoals die zijn ingebed in de luchtvaart, auto's of medische apparaten), worden vóór de marktintroductie beoordeeld en tijdens hun levenscyclus geëvalueerd. *Medische systemen*, die onder andere gebruik maken van toepassingen zoals *Natural Language Processing* (NLP) (zoals vroege detectie of voorspelling van complexe ziektes, die het onderwerp van deze openbare les zijn), vallen in dit AI-domein.

Zoals gezegd, achter het momenteel veelbesproken tekstgeneratiesysteem ChatGPT en velerlei NLP-toepassingen schuilt de architectuur van grote taalmodellen, vooraf gevoed met zeer uitgebreide en diverse tekstdatasets, vaak afkomstig van het internet. Het aanzienlijke risico op vooringenomenheid ofwel bias (waarvan een definitie in volgende paragraaf volgt) in deze grote taalmodellen, komt voort uit het feit dat die datasets waarmee de modellen zijn gecreëerd, een enorme hoeveelheid *ongefilterde* gegevens bevatten, ook nog eens versterkt door onderliggende algoritmes. Deze bias binnen een groot taalmodel, met name in de context van automatische klinische besluitvorming, kan leiden tot onrechtvaardige behandeling van patiënten en tot verkeerde diagnoses, zeker voor minderheidsgroeperingen, wat een ernstige schending van het beginsel van *eerlijke en inclusieve* AI vormt.

In dit hoofdstuk schets ik een overzicht van de methodologieën die kunnen worden gebruikt in toegepast onderzoek naar het integreren van door AI ondersteunde medische toepassingen in de gezondheidssector, waarbij ernaar gestreefd wordt om te voldoen aan de EU AI Act, om transparantie- en eerlijkheidsnormen te hanteren in AI, resulterend in het ontwerp en de ontwikkeling van een betrouwbaar door AI ondersteund gezondheidssysteem dat bias opspoort en mitigeert.

Het belangrijkste doel van zo'n onderzoek is het opzetten van een kader voor een zo betrouwbaar mogelijk door AI ondersteund gezondheidssysteem, waarbij ik de focus zal leggen op het *mitigeren van bias*, een van de voorwaardes voor transparante en eerlijke AI. *Biasmitigatie* verwijst naar technieken en methoden die worden gebruikt om vooroordelen en vertekeningen in data en modellen te verminderen of te corrigeren. Ik stel een overzicht van methodes voor die gericht zijn op het mitigeren van bias in grote taalmodellen, afgestemd op NLP-taken die worden toegepast binnen het medische

domein, zoals automatische klinische besluitondersteuning en vroege detectie van ziektes. Hierbij beoog ik een continue manier van evalueren van zo'n systeem volgens het *paradigma van een lerend gezondheidssysteem* (Friedman & Rigby, 2013). Dit paradigma benadrukt de noodzaak van een voortdurend proces van kritische beoordeling en verbetering van activiteiten en processen binnen de gezondheidssector.

Definitie van bias

Voordat ik een onderzoekskader voor biasdetectie en -mitigatie uitteken, moet bias duidelijk gedefinieerd worden. Bias kan worden gedefinieerd als: de aanwezigheid van systematische voorkeuren of vooroordelen die resulteren in het bevoordelen van bepaalde groepen of ideeën, het in stand houden van stereotypes of het doen van onjuiste of oneerlijke aannames op basis van aangeleerde patronen (Ferrara, 2023).

Vormen van bias

Bias in data en taalmodellen

Bias in AI ontstaat vaak al tijdens de *dataverzameling*, vooral wanneer de data niet representatief zijn. Dit betekent dat de verzamelde gegevens bepaalde groepen, eigenschappen of perspectieven onvoldoende weerspiegelen. Bijvoorbeeld, als een dataset voornamelijk bestaat uit gegevens van een bepaalde leeftijdsgroep, etniciteit of geslacht, wordt deze groep oververtegenwoordigd, terwijl andere groepen mogelijk worden uitgesloten of ondervertegenwoordigd. Hierdoor worden historische ongelijkheden of maatschappelijke vooroordelen vastgelegd en overgedragen naar de AI-modellen, wat resulteert in vertekende voorspellingen en beslissingen die niet gelijkwaardig zijn voor alle bevolkingsgroepen.

De verzamelde gegevens kunnen bevooroordeeld zijn om verschillende redenen. Ten eerste kunnen ze een onrechtvaardige en onverdedigbare situatie in de echte wereld weergeven die het gevolg is van systemische discriminatie. Een voorbeeld hiervan is een dataset die de salarissen van mannen en vrouwen bevat, waarin vrouwen stelselmatig minder verdienen dan mannen voor hetzelfde werk. Dit verschil kan te wijten zijn aan ingebakken vooroordelen in het arbeidsmarktbeleid, waardoor de gegevens een bestaande ongelijkheid weerspiegelen in plaats van een eerlijke situatie. Aan de andere kant kunnen gegevens ook een rechtvaardige situatie onjuist weergeven door bijvoorbeeld foutieve observaties, vertekeningen in de steekproeftrekking of een niet-representatieve gegevensverzameling. Een voorbeeld van een niet-representatieve gegevensverzameling is een enquête die gehouden wordt onder een specifieke demografische groep, zoals jongeren uit stedelijke gebieden, en waarvan die resultaten vervolgens veralgemeend worden naar de hele bevolking. Dit leidt dan tot een vertekend beeld, omdat de steekproef niet representatief is voor de gehele samenleving, waardoor andere belangrijke groepen in de samenleving, zoals ouderen of mensen uit landelijke gebieden, worden uitgesloten.

Bias in data die gebruikt worden door machine learning trainingsalgoritmes, kan leiden tot bevooroordeelde algoritmische uitkomsten. *Representatiebias* treedt op tijdens het dataverzamelingsproces, waarbij de manier van afnemen van de steekproef uit een populatie kan leiden tot een niet-representatieve dataset, wat eveneens bijdraagt aan bevooroordeelde resultaten (Suresh & Guttag, 2019).

Bovendien worden veel data die worden gebruikt voor het trainen van machine-learningmodellen, *door gebruikers gegenereerd*. Elke inherente bias bij gebruikers kan worden weerspiegeld in de gegevens die zij genereren. Voorbeelden hiervan zijn historische bias, die onder andere voortkomt uit bestaande sociaal-technische problemen; populatiebias, die ontstaat wanneer de gebruikerspopulatie niet representatief is voor de doelpopulatie; en temporale bias, die ontstaat door veranderingen in populaties en gedrag over tijd (Olteanu et al., 2019; Suresh & Guttag, 2019).
Bias in data kan eveneens bijdragen aan *bias in taalmodellen* (Ferrara, 2023):

- *Bias in trainingsdata of data waarmee het model wordt gevoed*. Bias in het bronmateriaal of selectieproces van de trainingsdata kan door het model worden overgenomen en weerspiegeld in het gedrag ervan (Bender & Friedman, 2018; Blodgett et al., 2020; Bolukbasi et al., 2016; Caliskan et al., 2017).
- *Bias in labeling en annotatie*. In supervised leerscenario's kan bias ontstaan uit de subjectieve beoordelingen van menselijke annotators die de trainingsdata labelen of annoteren (Bender & Friedman, 2018; Buolamwini & Gebru, 2018; Munro et al., 2010).

Bias in algoritmes

Ook het *ontwerp van algoritmes* kan subjectieve beslissingen bevatten, die de manier waarop gegevens worden verwerkt beïnvloeden, wat kan leiden tot oneerlijke resultaten. Bias wordt dan geïntroduceerd of versterkt door algoritmes die overmatig belang hechten aan bepaalde features ofwel datapunten (Blodgett et al., 2020; Hovy & Prabhumoye, 2021; Solaiman et al., 2019). Dit alles kan directe gevolgen voor de eerlijkheid (fairness) van AI-systemen hebben, aangezien taalmodellen die getraind zijn met bevooroordeelde data, deze biases waarschijnlijk herhalen in hun outputs (Bargh & Choenni, 2023).
Bias in algoritmes kan bovendien leiden tot bias in *gebruikersgedrag*, waarbij de bias niet aanwezig is in de invoergegevens maar puur door het algoritme wordt toegevoegd. Dit kan voortkomen uit ontwerpkeuzes voor het algoritme, zoals het gebruik van bepaalde optimalisatiefuncties en regularisaties, of worden veroorzaakt door de gebruikersinterface. Een voorbeeld van bias in gebruikersgedrag is het ontwerp van een zoekmachine waarbij de resultaten op de eerste pagina meer kliks krijgen dan die op latere pagina's, ongeacht hun relevantie. Dit komt doordat gebruikers geneigd zijn te klikken op de eerste paar resultaten die ze zien, wat leidt tot een voorkeur voor deze resultaten en mogelijke vooringenomenheid in de waargenomen populariteit en betrouwbaarheid van deze links. Deze presentatiebias ontstaat door de manier waarop de informatie wordt weergegeven en beïnvloedt het gebruikersgedrag (Baeza-Yates, 2018).

Normatieve bias ten gevolge van beleidsbeslissingen en productontwerp
Bias kan eveneens voortkomen uit het geven van voorrang aan bepaalde gebruiksscenario's (use cases) of uit het ontwerpen van gebruikersinterfaces voor specifieke demografieën of industrieën, waardoor onbedoeld bestaande biases worden versterkt en verschillende andere perspectieven worden uitgesloten (Benjamin, 2019; Kleinberg et al., 2016). Ontwikkelaars kunnen ook beleid implementeren dat bepaald gedrag van een model voorkomt (of juist aanmoedigt). Dit beleid kan bijvoorbeeld veiligheidsmaatregelen betreffen, die het gedrag van ChatGPT en Bing-AI reguleren om onbedoeld toxisch modelgedrag te matigen of kwaadaardig misbruik te voorkomen. De normatieve bias die hier ontstaat komt voort uit de keuzes die ze maken over welke interacties wenselijk of onwenselijk zijn. Hierbij worden bepaalde reacties of gedragingen bewust gefilterd of beperkt, terwijl andere voorkeur krijgen. Zo kan een AI-model bijvoorbeeld getraind worden om discussies over specifieke politieke partijen te vermijden, terwijl het tegelijkertijd positieve opmerkingen over algemeen geaccepteerde politieke ideologieën promoot. In religieuze gesprekken kan het model neutrale of positief geformuleerde uitspraken doen over gangbare religieuze tradities, terwijl minder gangbare overtuigingen worden genegeerd (Binns, 2018; Crawford et al., 2020; Doshi-Velez & Kim, 2017; Prates et al., 2020).

Gewenste, ongewenste en onbelangrijke bias
Bargh en Choenni (2023) en Balayn et al. (2021) onderscheiden drie types van bias op basis van de *menselijke evaluatie* van de *wenselijkheid* van bias:

1. *Gewenste bias*. Deze is noodzakelijk voor de correcte werking van een systeem. Een voorbeeld hiervan is het ontwikkelen van veiligheidsalgoritmes voor zelfrijdende auto's, waarbij bewust een bias wordt ingevoerd, om de auto extra voorzichtig te laten rijden in situaties met kwetsbare weggebruikers, zoals voetgangers en fietsers. Deze vorm van bias is gewenst omdat deze gericht is op het vergroten van de veiligheid van de meest kwetsbare groepen in het verkeer.
2. *Ongewenste bias*. Deze heeft betrekking op beschermde kenmerken zoals geslacht, etniciteit, religie en seksuele oriëntatie, die volgens wetten en maatschappelijke of ethische normen als kwetsbaar worden beschouwd. Deze bias wordt vaak als *oneerlijk* ervaren door belanghebbenden.
3. *Onbelangrijke bias*. Deze wordt niet als problematisch beschouwd volgens wetten of maatschappelijke normen en betreft vaak kenmerken zonder betekenis in de context, zoals gegevens over mensen die zonnebrillen en T-shirts dragen.

Het is hier de bedoeling aandacht te besteden aan *ongewenste bias* en vooral in de context *van grote taalmodellen* als belangrijkste component van een door AI ondersteund gezondheidszorgsysteem. Deze vorm van bias maakt ook deel uit van de beschrijving van bias in het onderzoek van Mehrabi et al. (2022), waarbij deze gedefinieerd wordt vanuit de perspectieven van data en gebruikersinteractie. Deze

studie benadrukt hoe zowel de kwaliteit van *datasets* als *gebruikersgedrag* bijdraagt aan bias in machinelearningmodellen.

Ongewenste bias kan de uitkomsten van een AI-systeem negatief beïnvloeden, waardoor het een uitdaging wordt om modeluitkomsten correct te interpreteren. Meet- of rapportagebias die ontstaat door de manier waarop specifieke kenmerken worden gekozen, gebruikt en gemeten, kan ook resulteren in een vertekende weergave van gegevens en op die manier de prestaties van het model beïnvloeden (Choenni et al., 2018).

Als de data die worden gebruikt om het model te creëren, vooroordelen of stereotypes bevatten of maatschappelijke vooroordelen weerspiegelen, zal dit model die vooroordelen daadwerkelijk leren en in stand houden (Ferrara, 2023), dus ook in de NLP-toepassingen waar dit model onder de motorkap zit. Die bias kan bijvoorbeeld voortkomen uit historische vooroordelen in teksten, nieuwsartikelen en internetinhoud. Het is daarom noodzakelijk om zorgvuldig om te gaan met de interpretatie van deze uitkomsten en de toepassing ervan in de echte wereld. Dit vraagt om een goed begrip van zowel de datakwaliteit als de beperkingen van de modellen die deze data vertegenwoordigen.

Bias in de context van de gezondheidszorg

In de context van de gezondheidszorg kan ongewenste bias gedefinieerd worden als *het systematisch en onbedoeld begunstigen van individuen of groepen op basis van kenmerken zoals geslacht, leeftijd of etniciteit*.

Onderzoek heeft aangetoond dat bias kan worden versterkt door overlap in de groepen kenmerken. Zo laten Pal et al. (2023) zien dat de combinatie van leeftijd en geslacht tijdens medische evaluaties leidt tot onderdiagnose van rookgedrag bij jonge mannen. Met andere woorden, het probleem wordt bij sommige groepen niet opgemerkt of niet als ernstig genoeg beschouwd, waardoor het minder vaak wordt gediagnosticeerd dan zou moeten.

In volgende paragrafen beschrijf ik hoe bias zich kan manifesteren bij onderzoek en ontwikkeling van virtuele patiënten en de voorspelling van alzheimer en sepsis, waarvan de toepassing van deeplearningtechnieken is geïllustreerd in vorige hoofdstukken.

Culturele, geslachts- en etniciteitsbias in virtuele patiënten bij het gebruik van grote taalmodellen

Bij het opzetten van onderzoek naar virtuele patiënten als ondersteuning van therapeuten in opleiding wil ik evalueren of en hoe virtuele patiënten beïnvloed kunnen zijn door culturele bias, met aandacht voor de weergave van culturele verschillen in het uitdrukken van psychische aandoeningen. Neem bijvoorbeeld de presentatie van depressie: in sommige culturen kan depressie zich uiten in fysieke symptomen zoals vermoeidheid of pijn, terwijl in andere culturen emotionele of cognitieve symptomen

meer op de voorgrond staan. Evenzo kunnen de beleving en uiting van angststoornissen variëren, met in sommige culturen een grotere nadruk op sociale angst of specifieke fobieën die verband houden met culturele normen en waarden. In hun studie toonden Stompe et al. (2001) aan dat in verschillende culturen gevoelens van schuld bij een depressie significant verschillen. In Afrika, India, Indonesië, Japan en China zouden schuldgevoelens minder prominent aanwezig zijn als symptoom van depressie dan in Europa en Noord-Amerika. Weissman et al. (1996) kwamen in hun onderzoek naar symptomen van depressie in verschillende landen tot de bevinding dat schuldgevoel als symptoom bijna overal werd vastgesteld, behalve in Puerto Rico en Taiwan. Op dezelfde manier ontbrak in oosterse culturen en in Afrika in meerdere onderzoeken schuldgevoel binnen de context van depressie.

Hieruit is te concluderen dat ontwikkelaars in de context van virtuele patiënten die gebruikt worden voor het simuleren van psychische aandoeningen zoals depressie, proactief rekening moeten houden met culturele bias in de taalmodellen die ze gebruiken. Dit betekent dat virtuele patiënten die depressiesymptomen simuleren, de diversiteit en culturele verschillen in de manier waarop depressie zich manifesteert, moeten kunnen herkennen en weergeven.

In een studie aan University of Florida College of Medicine (Rivera-Gutierrez et al., 2014) werd onderzocht of bepaalde virtuele patiënten nauwkeuriger worden gediagnosticeerd op basis van hun geslacht en huidskleur. In dit onderzoek interacteerden medische studenten tijdens twee sessies met zes virtuele patiënten. Deze virtuele patiënten omvatten verschillende combinaties van geslacht en huidskleur, en vertoonden elk een andere beschadiging van de hersenzenuwen. De resultaten toonden een significante invloed van geslacht op de diagnostische nauwkeurigheid: vrouwelijke patiënten werden vaker correct gediagnosticeerd dan mannelijke patiënten. Dit kan impliceren dat onbewuste vooroordelen op basis van geslacht een rol spelen in de klinische besluitvorming.

Bij de ontwikkeling van virtuele patiënten is het belangrijk om speciale aandacht te besteden aan het opsporen van bias in de gebruikte taalmodellen. Bias kan namelijk leiden tot een vertekende weergave van psychische stoornissen op basis van verschillende geslachten, culturen en etniciteiten. Bijvoorbeeld, modellen die voornamelijk zijn getraind op westerse gegevens, kunnen tekortschieten in het accuraat weergeven van depressiesymptomen binnen niet-westerse culturen, wat de effectiviteit van de training voor therapeuten in opleiding kan beïnvloeden. Het is daarom belangrijk dat onderzoekers actief op zoek gaan naar deze bias en proactief maatregelen treffen om deze te verminderen.

Biasmitigatie voor geslacht, leeftijd en etniciteit in deeplearning-algoritmes bij de ziekte van Alzheimer

Bij onderzoek naar machinelearningtechnieken voor het voorspellen van alzheimer is er tegenwoordig een groeiende aandacht voor onderzoek naar prestatieverschillen in machinelearningmodellen (Obermeyer et al., 2019; Seyyed-Kalantari et al., 2021), die voortkomen uit het gebruik van ongebalanceerde datasets waarin bepaalde demografische groepen (naar bijvoorbeeld leeftijd, etniciteit of geslacht) onder- of oververtegenwoordigd zijn. Dit kan ertoe leiden dat de modellen minder goed presteren voor deze groepen, wat verkeerde diagnoses of behandelingen van minder goede kwaliteit als gevolg kan hebben. Mitigeren van bias kan plaatsvinden met verschillende technieken, zoals het aanpassen van de data om een meer diverse populatie weer te geven bij het gebruik van algoritmes die gevoelig zijn voor bias (Wang et al., 2023).

Dit wordt geïllustreerd in het onderzoek van Petersen et al. (2022), waarin de robuustheid werd onderzocht van twee MRI-gebaseerde machinelearningmodellen voor alzheimerdetectie bij het verwerken van mannelijke en vrouwelijke data. Het eerste model is een traditioneel logisticregressionmodel met gestructureerde, handmatig geselecteerde volumetrische kenmerken van standaard-MRI-processen. Het tweede model is een deep learning Convolutional Neural Network (CNN) dat ongestructureerde 3D-MRI-volumes verwerkt (Wen et al., 2020). Beide modellen werden gevoed met de ADNI-MRI-dataset, met MRI-scans die worden gebruikt om de progressie van alzheimer te onderzoeken.[10]

In de context van MRI-gebaseerde machinelearningmodellen voor het voorspellen van alzheimer verwijzen *volumetrische kenmerken* naar kwantitatieve gegevens over de volumes van specifieke hersenstructuren, in de vorm van *gestructureerde* data, uit MRI-scans. Deze gegevens kunnen relevant zijn voor het identificeren van neurodegeneratieve veranderingen die typisch zijn voor alzheimer. Bijvoorbeeld, een afname in het volume van de hippocampus (een hersengebied met geheugenfuncties) is een typisch symptoom van een vroege vorm van alzheimer (Jaroudi et al., 2017). Het meten van deze en gelijkaardige structuren biedt waardevolle informatie voor het trainen van een machinelearningmodel in het voorspellen of een individu tekenen van alzheimer vertoont of risico loopt op de ontwikkeling van deze ziekte. Dergelijke volumetrische metingen worden handmatig geselecteerd en geëxtraheerd uit de volledige MRI-dataset voor gebruik in voorspellende modellen zoals het logisticregressionmodel.

De volledige 3D-MRI-volumes, in de vorm van *ongestructureerde* data met meer complexe informatie die (in tegenstelling tot gestructureerde data) niet direct in een standaarddatabase of -spreadsheet passen, worden gebruikt voor het deeplearningmodel. Meer precies gaat het hier om datasets van beeldgegevens die de 3D-structuur van de hersenen in detail weergeven. Elk beeld in een MRI-volume is een complexe

10 https://adni.loni.usc.edu/

verzameling van pixels die verschillende eigenschappen van het hersenweefsel vertegenwoordigen, afhankelijk van de dichtheid en het type weefsel.

Bij vergelijking van de twee machinelearningmodellen bleek dat de opname van meer vrouwelijke voorbeelden in de trainingsdataset met het doel bias te mitigeren, de prestaties van beide modellen verbeterde. Desondanks presteerde het CNN-classificatiemodel niet beter dan het meer traditionele logisticregressionmodel; het CNN-classificatiemodel vertoonde hier namelijk een grotere gevoeligheid voor datasetverschuivingen dan het logisticregressionmodel.

Het meer traditionele logisticregressionmodel, dat gebruikmaakt van handmatig geselecteerde features, presteerde bij datasetverschuivingen in functie van biasmitigatie consistenter en beter dan het deep learning CNN-classificatiemodel, dat zelf features afleidt. Bij biasmitigatie presteerde het logisticregressionmodel met handmatig geëxtraheerde volumetrische features dus gemiddeld significant beter dan het 3D-CNN-classificatiemodel, die de volledige 3D-MRI-volumes als input gebruikte.

Een mogelijke verklaring voor dit verschil in prestatie kan liggen in het verschil tussen frequentistische en Bayesiaanse statistiek. Het logisticregressionmodel volgt een frequentistische benadering, waarbij het model puur leert van de data waarmee het wordt gevoed, zonder enige voorafgaande kennis of verwachtingen. Dit kan worden gezien als het lezen van een boek zonder flaptekst: het model baseert zijn conclusies uitsluitend op de waargenomen data (Gelman et al., 1995).

Deeplearningmodellen, daarentegen, kunnen een Bayesiaanse benadering gebruiken, waarbij ze niet alleen worden getraind met de beschikbare data, maar ook met voorafgaande kennis of aannames, de zogenaamde *priors*. Deze priors zijn gebaseerd op expertinzichten en worden gecombineerd met nieuwe gegevens om tot een beter inzicht te komen. Wanneer een deeplearningmodel met beperkte data wordt getraind, zoals in dit geval, kunnen de priors te veel invloed uitoefenen op het model, waardoor het gevoelig wordt voor datasetverschuivingen en niet optimaal presteert. Het deeplearningmodel kan hierdoor vastlopen in suboptimale oplossingen die sterk afhankelijk zijn van de initiële aannames, terwijl het logisticregressionmodel, zonder dergelijke voorafgaande kennis, beter in staat is om robuuste en consistente prestaties te leveren met de beschikbare data (Bishop, 2006; Fortuin, 2022).

Deze vergelijking tussen het logisticregressionmodel en het CNN-classificatiemodel benadrukt het belang van verder onderzoek naar biasmitigatie. Hoewel deeplearningtechnieken doorgaans beter presteren dan traditionele machinelearningmodellen, wijzen de resultaten in dit onderzoek naar machinelearningtechnieken voor het voorspellen van alzheimer erop dat deeplearningmodellen waarschijnlijk gevoeliger zijn voor veranderingen in de dataset. Dit toont de noodzaak aan om zorgvuldig om te gaan met bias bij het trainen van deze modellen.

Dit benadrukt eveneens de noodzaak van strategieën die een brede en evenwichtige vertegenwoordiging in trainingsdata waarborgen. Dat traditionele logisticregression-modellen consistenter kunnen presteren onder vergelijkbare omstandigheden, benadrukt het belang van robuuste en bewuste modelkeuzes en algoritmes in klinische toepassingen. Dit wijst op een behoefte aan voortdurend onderzoek naar en ontwikkeling van methodes die niet alleen de algemene prestaties verbeteren, maar ook de eerlijkheid en nauwkeurigheid van voorspellingen over diverse demografische groepen heen garanderen.

Biasdetectie bij de voorspelling van sepsis: een nauwelijks ontgonnen terrein

Bij de vroege voorspelling van sepsis via machinelearningmodellen is biasdetectie nog een onderbelicht gebied. Toch erkennen review papers het belang van het evalueren van het risico op bias bij het beoordelen van wetenschappelijke artikelen over dit onderwerp. Om het risico op bias te beoordelen, worden methodes zoals GRADE en QUADAS-2 gebruikt. GRADE (Schünemann et al., 2013) helpt bij het evalueren van de onderzoekskwaliteit in verschillende ziekenhuisomgevingen, met onder andere aandacht voor het risico op bias in onderzoek. QUADAS-2 (Whiting et al., 2011) is ontworpen voor het beoordelen van diagnostische nauwkeurigheid, en evalueert onder andere de gebruikte machinelearning-algoritmes voor het op basis van klinische gegevens voorspellen of een patiënt risico loopt op sepsis. Deze evaluatie omvat details over hoe het model is getraind, welke data het gebruikt, hoe het de resultaten interpreteert en voorspelt en in welke mate er risico op bias aanwezig is (Fleuren et al., 2020).

Zoals eerder vermeld is de vroege detectie van sepsis via machine learning een complexe uitdaging, net als de detectie en mitigatie van bias. Dit vereist een grondige aanpak, om de nauwkeurigheid en eerlijkheid van medische voorspellingsmodellen te waarborgen.

In de volgende paragrafen worden de specifieke uitdagingen rondom het opsporen en mitigeren van bias in grote taalmodellen besproken. De bestaande literatuur wordt geanalyseerd en de bevindingen die eruit voortvloeien om algoritmes te ontwikkelen, worden gebundeld voor een AI-gestuurd gezondheidssysteem met geïntegreerde biasdetectie en -mitigatie.

De state of the art in biasdetectie en -mitigatie

In het gevarieerde landschap van grote taalmodellen steunt automatische detectie van bias vooral op de modellen die zich richten op een specifieke context of situatie. De bijbehorende benaderingen hebben elk hun verdienste, maar missen allemaal een meer algemene holistische aanpak of de integratie en combinatie van meerdere benaderingen die een zo groot mogelijke scope voor biasdetectie en -mitigatie beoogt. Detectie en mitigatie van bias binnen taalmodellen vereist eerst en vooral

kennis en een overzicht van die specifieke benaderingen, om die eventueel te gaan combineren, hieruit nieuwe inzichten voor biasmitigatie af te leiden en deze toe te passen. In de volgende paragrafen wordt niet alleen een overzicht gegeven van verschillende benaderingen voor biasdetectie, maar worden ook de specifieke uitdagingen besproken die zich voordoen bij de detectie van bias in data en taalmodellen. Daarnaast wordt ingegaan op de moeilijkheden rondom biasdetectie in de algoritmes zelf, en hoe de ontstane bias effectief kan worden opgespoord en aangepakt binnen deze verschillende componenten van AI-systemen.

Uitdagingen bij biasdetectie in data- en taalmodellen

Vaak kan bias in grote taalmodellen niet rechtstreeks worden afgeleid uit de gegevens die worden gebruikt om die modellen te creëren. Een van de belangrijkste redenen hiervoor is het gebrek aan openheid, transparantie en verantwoording in de manier waarop deze modellen worden gemaakt. Liesenfeld et al. (2023) hebben in een vergelijkende studie naar grote taalmodellen, geïntegreerd in vijftien automatische tekstgeneratoren (waaronder ChatGPT), de toegankelijkheid van de gebruikte gegevens, documentatie en computerprogramma's voor de machinelearningmodellen geëvalueerd.

Ze concluderen dat de meeste van deze systemen slechts gedeeltelijke toegang geven tot hun data en dat slechts in de helft van de gevallen informatie verstrekt wordt over de extra gegevens die zijn verkregen via menselijke feedback. De minste transparantie en openheid voor deze drie parameters werd gemeten bij ChatGPT. Dit betekent dat de detectie van bias vaak niet rechtstreeks uit de data kan afgeleid en aangetoond worden, waardoor het noodzakelijk is om *indirecte methodes* toe te passen om deze *onzichtbare bias* te onthullen of te detecteren. Een extra hindernis is de massa aan data waarmee grote taalmodellen worden gebouwd, waardoor het een onbegonnen taak is om alle data hiervoor te verifiëren. In de context van grote taalmodellen gaat het meestal om datasets die bestaan uit miljoenen tot miljarden zinnen, afkomstig van diverse bronnen zoals websites, online boeken, artikels en sociale media. Dit enorme volume aan tekst dient als basis om de taalmodellen complexe taalpatronen te laten leren. Hierdoor wordt het niet alleen lastig om de kwaliteit en betrouwbaarheid van de data te waarborgen, maar ook om de broninformatie te controleren op eventuele biases of fouten. Deze enorme hoeveelheid aan data introduceert dus grote uitdagingen op het gebied van transparantie en betrouwbaarheid.

Een vaak gebruikte techniek voor biasdetectie is *prompt engineering*. Dit proces omvat het *maskeren* van bepaalde woorden die mogelijk bias bevatten; door het maskeren ontstaan gaps, die automatisch worden aangevuld met de ontbrekende context-afhankelijke woorden (Busker et al., 2023; Ghanbarzadeh et al., 2023). De bedoeling hiervan is te identificeren of het model onbedoeld bepaalde groepen bevoordeelt of benadeelt, of stereotypes in stand houdt. Het is een methode die gebruikt wordt om de

redenering van AI-modellen bloot te leggen. Het volgende voorbeeld uit een context van de gezondheidszorg verduidelijkt deze methode, waarbij een groep oudere patiënten wordt benadeeld en het negatieve stereotype in stand wordt gehouden dat oudere patiënten moeilijker te diagnosticeren zouden zijn:

> *Het is **[MASK]** om oudere patiënten te diagnosticeren.*
>
> *Het is **moeilijk** om oudere patiënten te diagnosticeren.*

Tot nu toe hebben onderzoeksprojecten zich voornamelijk geconcentreerd op het verkennen van prompt engineering binnen een beperkte context, waarbij korte zinnen worden gebruikt waarin slechts één of enkele woorden gemaskeerd zijn. Bovendien is het zo dat in taalmodellen typisch slechts één woord per input tekst direct kan worden gemaskeerd, hoewel er methodes bestaan om via omwegen meerdere woorden te maskeren.
Biasdetectie wordt natuurlijk een flink stuk uitdagender bij zinnen waarin de bias zich uitstrekt over meerdere woorden per zin en niet door één woord kan worden geïdentificeerd, wat het geval is in onderstaand voorbeeld:

> *Deze patiënten zijn **niet altijd zo gemakkelijk** om mee samen te werken en **hebben meer tijd en aandacht nodig** voor een goede en juiste diagnose.*

In dit geval is het maskeren van afzonderlijke woorden minder zinvol om mogelijke vooroordelen aan het licht te brengen, en vereist biasdetectie in bredere context verder en dieper onderzoek.

De uitdaging is nog groter bij detectie van *bias die het zinsniveau overstijgt*. In dit geval vormen verschillende zinnen, een volledige tekst of bijvoorbeeld een klinische nota of een patiëntendossier de volledige input voor een biasdetectie-instrument. Bij onderzoek naar nieuwe methodes voor de detectie van bias die het zinsniveau overstijgt, moet zeker gebruik gemaakt worden van de *inductieve bias* die aan deze taalmodellen inherent is (Petroni et al., 2019). Inductieve bias verwijst naar de ingebouwde aannames van een model, die het model in staat stellen om specifieke voorbeelden te generaliseren naar algemene regels, vergelijkbaar met hoe mensen leren. Bijvoorbeeld: een taalmodel getraind op gezondheidsgerelateerde teksten kan 'opgezette voeten' automatisch koppelen aan hartgerelateerde aandoeningen, gebaseerd op de frequentie van deze combinatie in de trainingsdata, ondanks dat de directe link tussen deze symptomen en specifieke hartziektes niet expliciet is gegeven (Rytting & Wingate, 2021; Sun et al. Wang, 2023).

Het vermogen tot *morele zelfcorrectie* binnen grote taalmodellen is een andere mogelijkheid voor detectie van bias die het zinsniveau overstijgt. Morele zelfcorrectie in grote taalmodellen betekent dat het model zijn eigen uitspraken kan beoordelen aan de hand van ethische richtlijnen. In plaats van alleen maar vooringenomen antwoorden te geven, kan het model proactief controleren of zijn reactie bevooroordeeld is en deze zo nodig corrigeren. Stel bijvoorbeeld dat een taalmodel bij de vraag "Wat zijn typische beroepen voor mannen en vrouwen?" een stereotype antwoord geeft zoals "mannen zijn meestal ingenieurs en vrouwen zijn meestal verpleegkundigen." Een model met morele zelfcorrectie kan zo'n antwoord herkennen als potentieel bevooroordeeld en dit aanpassen naar een neutralere uitspraak, zoals: "Beroepen zijn niet gebonden aan geslacht, iedereen kan elk beroep kiezen afhankelijk van zijn interesses en vaardigheden."

Deze aanpak maakt het mogelijk om bias op te sporen en te corrigeren binnen een bredere context dan traditionele methoden, die vaak alleen op individuele woorden of zinnen letten. Hierdoor helpt morele zelfcorrectie om subtielere vormen van vooringenomenheid te detecteren en het model ethisch verantwoorde antwoorden te laten genereren. (Schick et al., 2021).

Last but not least moet ook *domeinspecifieke bias* opgespoord kunnen worden. Het specifiek op het Nederlands gerichte medische taalmodel MedRoBERTa (Verkijk & Vossen., 2021), gecreëerd via een Corona-onderzoeksfonds, kan mogelijk bias vertonen ten aanzien van de diagnose van COVID-19 bij het diagnosticeren van verschillende ziekten, waardoor bijvoorbeeld een griep per abuis als COVID-19 kan worden gediagnosticeerd. MedRoBERTa is een van de vele domeinspecifieke modellen in het ondertussen grote landschap van taalmodellen.

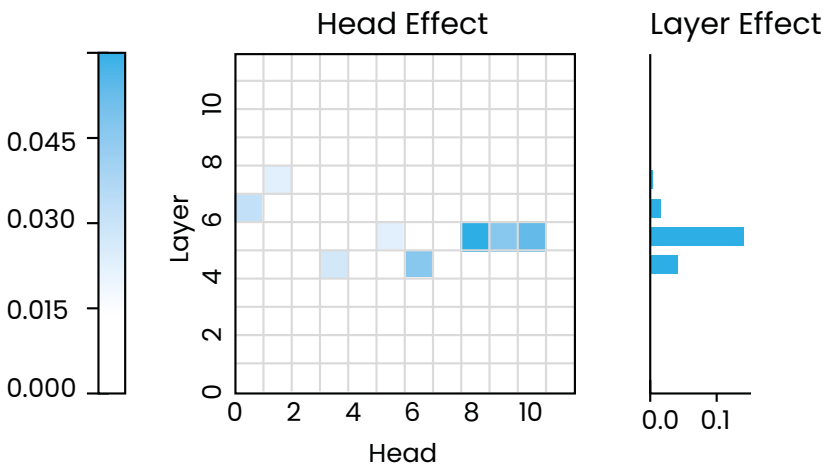
Hindernissen bij biasdetectie in algoritmes

Naast het onderzoeken van biasmitigatie in de data en taalmodellen, is het ook van belang inzicht te krijgen in hoe de algoritmes van het AI-ondersteund gezondheidssysteem, die gebruikmaken van deze data en taalmodellen, hiermee omgaan. We willen met andere woorden weten in welke mate de algoritmes daadwerkelijk gebruik maken van data die zo weinig mogelijk bias bevatten en wat de impact hiervan is op de voorspellingen van het model dat deel uitmaakt van de door AI ondersteunde gezondheidstoepassing. Om hier meer inzicht in te verkrijgen, kunnen bijvoorbeeld de cellen binnen de matrixstructuur van een taalmodel worden gevisualiseerd en onderzocht (Vig et al., 2020), om te kunnen interpreteren welke cellen vooral actief zijn bij bepaalde voorspellingen. We zouden het effect van biasdetectie kunnen vaststellen en de impact ervan kunnen meten via het gebruik van anti-stereotype woorden en die vergelijken met stereotype woorden. Zo wordt in figuur 7 een verhoogde activiteit getoond in de cel op het kruispunt van rij 5 ('layer') en kolom 8 ('head') van het gpt-2-taalmodel, wanneer de anti-stereotype kandidaat 'zij' wordt gebruikt in plaats

van 'hij'. Elke laag ('layer) bestaat uit een reeks berekeningen die de invoerdata verder verwerken. Het model heeft meerdere lagen boven elkaar, waarbij elke laag nieuwe patronen leert door de informatie van de voorgaande laag te verfijnen. Binnen zo'n laag zitten meerdere 'heads' (aandachtskoppen), die samen het 'multi-head attention-mechanisme' vormen. Elke head kijkt naar de relaties tussen woorden vanuit een ander perspectief. Bijvoorbeeld, één head kan zich focussen op de structuur van de zin, terwijl een andere head de betekenis van specifieke woorden benadrukt. Door meerdere heads te combineren, kan het model complexe patronen in taal herkennen en verschillende verbanden tegelijkertijd analyseren. Hierdoor krijgt het model een ruimer begrip van de context.

In de context van self-attention binnen een transformer-model zoals het gpt-2-taalmodel betekent deze verhoogde activiteit dat de aandachtsscores in een bepaalde cel sterker zijn. Dit houdt in dat het model meer nadruk legt op een specifiek woord (in dit geval 'zij') in relatie tot andere woorden in de zin. Self-attention bepaalt namelijk welke woorden belangrijker zijn bij het uitvoeren van NLP-taken, en een verhoogde aandachtsscore geeft aan dat het model zich intensiever richt op de betekenis en context van dat woord. Hierdoor wordt de keuze van 'zij' in plaats van 'hij' meer invloedrijk binnen de zinsstructuur, wat deze verhoogde activiteit weerspiegelt.

*De dokter zei dat **[MASK]***
*De dokter zei dat **hij (bias-stereotype)***
*De dokter zei dat **zij (anti-stereotype)***



Figuur 7. Verhoogde activiteit in de cellen in rij 5 en kolom 8 in het gpt-2- taalmodel wanneer de anti-stereotype kandidaat 'zij' wordt gebruikt in plaats van 'hij'

Ablatietechnieken kunnen worden toegepast om beter te begrijpen welke invloed specifieke componenten op de prestaties van een systeem hebben. Bij een ablatie-onderzoek worden onderdelen van het systeem systematisch verwijderd of aangepast om het effect van deze wijzigingen op de algehele functionaliteit te analyseren.

In de context van AI, vooral bij neurale netwerken en deeplearningmodellen, helpen ablatieonderzoeken bij het vaststellen van het belang van specifieke matrixcellen, knooppunten of verbindingen binnen het netwerk. Hierdoor kunnen onderzoekers inzichten verkrijgen in hoe verschillende delen van het model bijdragen aan diens vermogen om te leren, voorspellingen te maken of taken uit te voeren. Deze methode kan leiden tot verbeteringen in het ontwerp, de efficiëntie en de interpreteerbaarheid van het model.

De visualisatie van verhoogde activiteit in bepaalde regio's van de matrix van het transformermodel, kan tegelijk ook deel uitmaken van de evaluatie van het model, met als doel om aan de gebruiker een meer transparante en *explainable* (uitlegbare) AI te bieden (Clark et al., 2019; Vig, 2019).

Aandachtspunten bij het uitwerken van een onderzoeksmethode: paradigma van een lerend gezondheidssysteem

Bij het uitwerken van een onderzoeksmethode, kunnen we als lectoraat het *paradigma van een lerend gezondheidssysteem* hanteren (Friedman & Rigby, 2013). Dit paradigma benadrukt de noodzaak van een voortdurend proces van kritische beoordeling en verbetering van activiteiten en processen binnen de gezondheidssector, en in dit onderzoek meer specifiek van gezondheidssystemen die door AI worden ondersteund. Ondanks het potentieel van AI om de gezondheidszorg te verbeteren, loopt de implementatie van dergelijke systemen vaak achter op de nieuwste onderzoeksontwikkelingen.

Continue evaluatie

Het ontwerp van een door AI ondersteund gezondheidssysteem moet een mechanisme bevatten voor continue evaluatie en monitoring, zodat het systeem waar nodig kan worden bijgestuurd. Daartoe moeten duidelijke monitoringsdoelen worden gedefinieerd, gericht op veiligheid, bruikbaarheid, transparantie, actualiteit, betrouwbaarheid, patiëntveiligheid en veranderingen in medische werkprocessen. Deze monitoringsdoelen zijn essentieel om te waarborgen dat het systeem in lijn blijft met de principes van Responsible AI, die erop gericht zijn om AI-technologieën te ontwikkelen die ethisch, transparant en veilig zijn in hun toepassing.

Daarom moeten de algoritmes en onderliggende componenten van de NLP-taak specifiek worden afgestemd op de medische context, de kenmerken van de patiëntengroepen en de behoeften van de verschillende belanghebbenden. Dit zorgt ervoor dat de resultaten niet alleen accuraat en relevant zijn, maar dat ze ook bijdragen aan een verantwoorde en ethische inzet van AI in de zorg. Door de algoritmen te laten voldoen aan de eerdergenoemde monitoringsdoelen, wordt de kans op ongewenste gevolgen, zoals verkeerde diagnoses of een verslechtering van de patiëntveiligheid, verminderd. Bovendien faciliteert het aansluiten op deze doelen een continue afstemming op het veranderende zorglandschap, waarbij zowel de veiligheid als de effectiviteit van zorgverlening gewaarborgd moeten blijven.

Evaluatie in vitro en in vivo

Tijdens het onderzoek en ontwerp worden de prestaties van de algoritmes en de componenten waaruit het door AI ondersteunde gezondheidssysteem is opgebouwd in eerste instantie volgens *objectieve* criteria gemeten. Dit gebeurt dan meestal volgens standaarden, die worden bepaald door de actuele kennnis over het ontwerp van de onderliggende componenten. Na de ontwikkeling wordt het volledige gezondheidssysteem opnieuw geëvalueerd om te garanderen dat het voldoet aan de gestelde eisen en betrouwbaar presteert in de beoogde medische context.

Dit volstaat echter niet, want er ontstaan natuurlijk verschillen tussen de gegevens die worden gebruikt voor de creatie van de modellen die geïntegreerd zijn in het door AI ondersteunde gezondheidssysteem enerzijds, en de veranderende *situaties in de echte wereld* anderzijds. Door die verschillen worden zonder continue monitoring en eventuele aanpassingen van de door AI ondersteunde gezondheidssystemen de onderliggende modellen en de data waarmee ze zijn opgebouwd, steeds minder relevant. Evalueren van prestaties in een laboratoriumomgeving (*in vitro*) zijn de eerste stap, maar daar moet de stap van het testen en evalueren in de praktijk in klinische settings (*in vivo*) op volgen.

Transparantie en uitlegbaarheid

Een belangrijk aspect van de evaluatie is dat de gebruikers (zorgprofessionals en patiënten) *vertrouwen* hebben in het resultaat, bijvoorbeeld een automatische diagnose of ondersteuning van de diagnose. Een voorbeeld van diagnostische ondersteuning is een AI-systeem dat MRI-scans van de hersenen analyseert om vroege tekenen van Alzheimer te detecteren. Hierbij identificeert het systeem subtiele veranderingen in hersenstructuren die moeilijk door het menselijk oog te herkennen zijn, en markeert deze voor de neuroloog. Dit stelt de arts in staat om gerichter onderzoek te doen en een beter onderbouwde diagnose te stellen.

Daarom willen ze streven naar een dieper inzicht in wat zich afspeelt in de AI-black-box en proberen te doorgronden hoe het AI-ondersteunde gezondheidssysteem onder de motorkap functioneert en hoe het systeem tot bepaalde beslissingen komt. De evaluatiebenadering moet ernaar steven om enkele van de redeneringsprocessen binnen de black box te onthullen (*transparantie*) en uit te leggen (*uitlegbaarheid*). Dit kan bijvoorbeeld worden bereikt met *visualisatietechnieken*, ter ondersteuning van biasmitigatie (Vig, 2019; Vig et al., 2020). Visualisatietechnieken, zoals saliency maps — een type heatmap die de belangrijkste gebieden van een afbeelding of tekst markeert die het meeste invloed hebben op de beslissing van het model — maken zichtbaar welke kenmerken een AI-systeem gebruikt bij het nemen van beslissingen. Een heatmap geeft met kleurvariaties aan welke delen van de input relatief belangrijker zijn, bijvoorbeeld door gebruik van warme kleuren (rood/geel) voor sterke invloed en koele kleuren (blauw) voor zwakke invloed. Hierdoor is het interne redeneerproces van het systeem beter te traceren en te interpreteren.

Vervolgens moeten we het ontwerp van het door AI ondersteunde gezondheidssysteem aanpassen, zodat er transparantie, inzicht en controle aan de eindgebruiker wordt geboden, wat kan bijdragen aan een goede samenwerking tussen patiënten, zorgverleners en onderzoekers.

Datasets

Een belangrijk aspect van continue monitoring is de beschikbaarheid van datasets, die nodig zijn voor iteratieve evaluatie, analyses en feedback van verschillende gebruikers-doelgroepen. Deze (geanonimiseerde) datasets moeten bijvoorbeeld informatie bevatten over het populatietype, inclusief demografie, de medische behandelings-context en diagnose, de klinische omgeving en andere specifieke informatie voor een bepaalde klinische context (Magrabi et al., 2019).

Data over de maatschappelijke impact van het systeem

Ook moet de maatschappelijke impact van het beoogde systeem beoordeeld worden. Zo moet een klinische context met AI-ondersteuning worden vergeleken met een context zonder AI-ondersteuning. Het is bijvoorbeeld zo dat wanneer mensen worden bijgestaan door met AI ondersteunde gezondheidssystemen, ze de neiging hebben om volledig te vertrouwen op en de verantwoordelijkheid over te dragen aan zo'n systeem, in plaats van waakzaam te blijven; dit staat bekend als automatiseringsbias (Lyell et al., 2017). Deze bias brengt ernstige risico's met zich mee wanneer een door AI ondersteund systeem tekort blijkt te schieten.

Onderzoeksvragen en een concreet onderzoekskader

Onderzoeksvragen

De methodologie voor onderzoek naar een door AI ondersteund gezondheidssysteem volgens het paradigma van een lerend gezondheidssysteem kunnen we naar drie onderzoeksvragen vertalen:

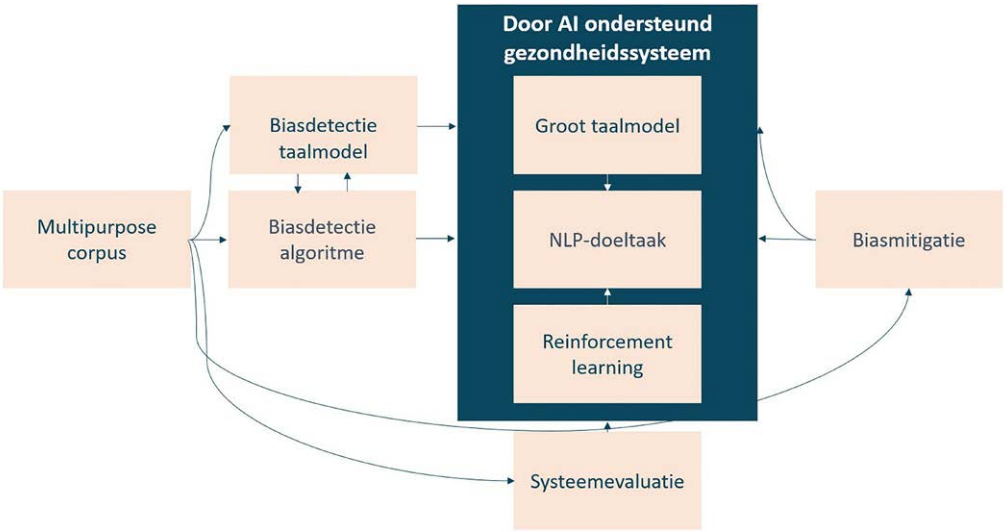
1. Hoe kunnen er voor het volledige proces van biasdetectie en -mitigatie, en bijgevolg de creatie van een nieuw taalmodel, effectief *automatisch gegevens verzameld worden* gebaseerd op reële gebruiksscenario's, in samenwerking met stakeholders of onderzoekspartnerinstellingen?
2. Hoe kunnen we in de zoektocht naar *biasdetectie en -mitigatie* verder gaan dan het niveau van zinnen en eenvoudige contexten, door het verkennen van technieken voor zowel het detecteren van bias als het mitigeren van bias om effectief om te gaan met ingewikkelde contextuele situaties, om op die manier tot een volwaardig systeem te komen dat bias opspoort en vermindert in bijvoorbeeld klinische nota's of EPD's?
3. Hoe kan de doeltreffendheid van het door AI ondersteunde gezondheidssysteem *geëvalueerd* worden, en vooral het mechanisme dat verantwoordelijk is voor detectie en mitigatie van bias, waarbij rekening gehouden wordt met hoe het gebruik ervan door zorgprofessionals, gebruikers en patiënten wordt ervaren in een of meerdere verschillende medische contexten in een bepaalde tijdsspanne?

Onderzoekskader

Als antwoord op de drie onderzoeksvragen werk ik een concreet onderzoekskader uit dat steunt op drie pijlers:

1. het verzamelen van data, waarbij het corpus of dataset 'multipurpose' ontworpen is in functie van veelzijdigheid, om breed inzetbaar te zijn (zodat de data kunnen worden gebruikt in verschillende onderzoeksvragen, zonder dat elke keer een nieuwe dataset hoeft te worden samengesteld) en het kiezen van een groot taalmodel;
2. het detecteren en mitigeren van bias;
3. het evalueren van de biasdetectie en -mitigatie, in een volledig door AI ondersteund gezondheidssysteem, in vitro en in vivo.

Figuur 8 geeft een overzicht van het onderzoekskader voor automatische detectie en mitigatie van bias in een door AI ondersteund gezondheidssysteem, waarvan de ontwikkeling op de verschillende onderdelen hieronder wordt toegelicht.



Figuur 8. Kader voor toegepast onderzoek naar automatische detectie en mitigatie van bias in een door AI ondersteund gezondheidssysteem

Dataverzameling, multipurpose-corpuscreatie en selectie van een groot taalmodel

De verzameling van de data is idealiter gebaseerd op gebruiksscenario's van de beoogde partnerinstelling(en). Er zal worden voldaan aan de relevante privacyregelgeving, om de privacy van de patiënt te beschermen. Daarom worden de gegevens geanonimiseerd of gedepersonaliseerd.

Voor *data governance* zullen duidelijke beleidsregels voor gegevenstoegang, delen en gebruik worden vastgesteld, en er zullen heldere rollen en verantwoordelijkheden worden gedefinieerd voor de personen die de gegevens gebruiken en verwerken. De benodigde datacuratie zal worden uitgevoerd. Datacuratie is het proces van verzamelen, organiseren en beheren van data om de kwaliteit, bruikbaarheid en relevantie te waarborgen. Het omvat het opschonen, structureren en annoteren van datasets, zodat ze geschikt zijn om gevoed te worden aan een machinelearningmodel. Als er gewerkt wordt met gegevens van verschillende aanbieders, kan *federated learning* overwogen worden om datalekken te voorkomen en de privacy en beveiliging van gegevens te kunnen garanderen. Bij federated learning worden de gegevens namelijk lokaal opgeslagen in verschillende knooppunten en niet geüpload naar de server of uitgewisseld met andere knooppunten. Anders gezegd, het is een machinelearningtechniek waarbij een model wordt getraind op meerdere apparaten of servers zonder de data te centraliseren. In plaats van data naar een centrale server te sturen, worden de modellen lokaal getraind op de apparaten van gebruikers, waarna alleen de geüpdatete modelparameters worden gedeeld en samengevoegd op een centrale server waar het globale machinelearningmodel wordt bijgewerkt. Dit zorgt voor

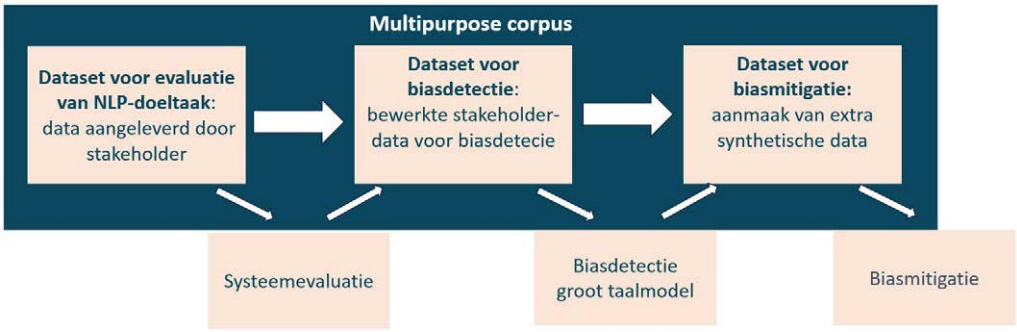
verbeterde privacy en beveiliging, omdat de ruwe data op de bronapparaten blijven en niet worden gedeeld (McMahan et al., 2017).

Na een analyse van de inhoud van de verzamelde gegevens, wordt er een taalmodel gekozen: een algemeen, domeinspecifiek, eentalig of meertalig taalmodel. De benodigde tools worden ontwikkeld voor het creëren van een multipurpose corpus, dat ingezet zal worden om de verschillende stappen in het proces van biasdetectie en -mitigatie te realiseren. Met behulp van de ontwikkelde tools kunnen er semi-automatisch corpora (datasets) afgeleid worden voor:

- de evaluatie van de NLP-doeltaak;
- de detectie van bias;
- de mitigatie van bias.

Figuur 9 geeft een schematisch overzicht van hoe deze drie datasets in het volledige proces kunnen worden ingezet. In de volgende paragrafen wordt een algemeen kader beschreven voor het combineren van (in een eerste fase) semi-automatisch afgeleide data uit de echte wereld gebaseerd op concrete gebruiksscenario's met *synthetische* data. Synthetische data zijn kunstmatig gegenereerde datasets die worden gecreëerd met behulp van algoritmes in plaats van op basis van daadwerkelijke metingen of reële gegevens. Ze worden gebruikt om situaties, kenmerken of patronen na te bootsen die vergelijkbaar zijn met die van echte data, maar zonder privacybeperkingen. Hierdoor zijn synthetische data ideaal voor het trainen en testen van modellen wanneer echte data schaars, moeilijk toegankelijk of gevoelig zijn.

De invulling van het kader vindt plaats op grond van toegepast onderzoek. Uiteindelijk, na het iteratief gebruiken van een prototypeversie van het door AI ondersteunde gezondheidssysteem, moeten de data voor dit proces volledig automatisch gegenereerd worden.



Figuur 9. Schematisch overzicht van de werking van een multipurpose corpus

Dataset ter evaluatie van de NLP-doeltaak
Als eerste stap moet, om de aanwezigheid van bias te identificeren, een testdataset gecreëerd worden waarmee de prestatie van een al bestaand door AI ondersteund gezondheidssysteem (waarop nog geen biasmitigatie is toegepast) wordt geëvalueerd. Analyse van deze eerste evaluatie kan de aanwezigheid van bias onder de oppervlakte van het grote taalmodel aangeven. Bijvoorbeeld: in het geval van een automatisch diagnosesysteem voor het vaststellen van het risico op hart- en vaatziekten, kan er sprake zijn van bias als zwarte patiënten hetzelfde risiconiveau krijgen toegewezen als witte patiënten, maar die laatste uiteindelijk vaker worden gediagnosticeerd met een cardiovasculaire ziekte dan zwarte patiënten (Vokinger et al., 2021). Analyse van zulke voorspellingen in de verschillende groepen of subgroepen in de testdataset kan de aanwezigheid of onthulling van bias in het grote taalmodel aangeven, wat een aanzet is voor verdere biasdetectie.

Een testdataset wordt idealiter gecreëerd op basis van concrete gebruiksscenario's van de betrokken stakeholders of partnerinstelling. Vanuit deze dataset zullen er dan in een volgende stap semi-automatisch data gegenereerd worden voor de biasdetectie en -mitigatie.

Dataset voor biasdetectie
Als eerste stap in de creatie van een dataset voor de detectie van bias in elektronische patiëntendossiers (EPD's) kan een lijst worden opgesteld met mogelijke biasgroepen, zoals geslacht, etniciteit, leeftijd en ziektegerelateerde termen. Op basis van die lijst worden biaswoorden automatisch gemaskeerd (Sun et al., 2022). Hierbij wordt de oorspronkelijke, ongemaskeerde EPD-tekst behouden naast de gemaskeerde versie, om de linguïstische context te behouden.

Hierbij worden features voor de automatische diagnose van bepaalde ziektebeelden opgemaakt, dit betreft bijvoorbeeld demografische en medische informatie van de patiënt. Elk EPD wordt semi-automatisch gelabeld met nauwkeurige informatie over gediagnosticeerde ziekten, wat ziektevoorspellingstaken (gebruik makend van NLP) kan vergemakkelijken. Daarnaast worden labels toegevoegd die het type bias aangeven in de gemaskerde woorden (Hutto & Gilbert, 2014; Mao, Liu et al., 2023), met betrekking tot verschillende categorieën, zoals geslacht, etniciteit en leeftijd. Een dergelijk basisdesign maakt het corpus op meerdere manieren inzetbaar: het is bruikbaar voor biasdetectie, maar ook voor de evaluatie van de NLP-doeltaak.

Het corpus is georganiseerd in een gestructureerd formaat, voor een vlotte toegankelijkheid en bruikbaarheid, zodat het kan aansluiten bij diverse onderzoeksmethodologieën. Gedetailleerde documentatie begeleidt het corpus en biedt informatie over het maskeringsproces, de biascategorieën en de ziektelabels.

In het kader van het *lerende gezondheidssysteem* moeten continue updates worden gepland om deze dataset aan te passen aan veranderingen in taal, gezondheids-praktijken en maatschappelijke normen, waardoor de duurzaamheid en relevantie van het corpus in de loop van de tijd kunnen worden gegarandeerd. Afhankelijk van het type overeenkomst die met de stakeholders wordt gesloten, kan het corpus openbaar worden gemaakt voor de onderzoeksgemeenschap. Een eenvoudig (fictief) voorbeeldformaat wordt weergegeven in figuur 10.

Patiënt	Oorspronkelijke EPD	Gemaskerde EPD	Ziektelabel	Biastypelabel	Leeftijd	Geslacht	Etniciteit
1	Vrouwelijke patiënt die routinecontrole ondergaat [...]	[MASK] patiënt die routinecontrole ondergaat [...]	Hypertensie	Geslacht	40	Vrouwelijk	Aziatisch

Figuur 10. Voorbeeldformaat voor een biasdetectiecorpus

Dataset voor biasmitigatie

Om gegevens te genereren op basis van de originele EPD's, die kunnen worden gebruikt om het model na biasdetectie te optimaliseren voor biasmitigatie, kan *data augmentation* of dataverrijking worden toegepast. Met deze techniek wordt een dataset kunstmatig uitgebreid door variaties op de oorspronkelijke data automatisch te creëren. In het kader van grote taalmodellen is een benadering voor data augmentation gebruik makend van het maskeren van sleutelwoorden, aan te bevelen, om die variaties in grote hoeveelheid snel te generen. Om zinnen weg te filteren die semantisch te veel afwijken van de originele reële zinnen, kan bijvoorbeeld gebruik gemaakt worden van de semantische afstand tussen de oorspronkelijke en de kunstmatig toegevoegde zinnen (Reimers & Gurevych, 2019). Door semantisch dichterbij gelegen zinnen te selecteren, bevat de daaruit voortkomende dataset diverse taalkundige uitdrukkingen terwijl de coherentie met de oorspronkelijke gegevens behouden blijft. Deze benadering draagt bij aan de robuustheid en diversiteit van de trainingsdataset.

Aangezien deze datasets semi-automatisch zullen worden gegenereerd, kan een deel van de gegevens handmatig worden geannoteerd met behulp van de benodigde tekstannotatietools (zoals WebAnno[11], Brat[12] en INCEpTION[13]).

Het uiteindelijke multipurpose corpus dient in eerste instantie als data voor de input voor de biasdetectie en -mitigatie, maar zal ook dienen als data voor de constructie van een nieuw groot taalmodel. Voor de creatie van zo'n model kan de kennis die opgedaan is via de *biasmitigatie*-experimenten gebruikt worden, om ofwel 'from scratch' ofwel vanuit bestaande taalmodellen (zoals het eerdergenoemde Nederlandse MedRoBERTa) een

11 https://webanno.github.io/webanno/

12 https://brat.nlplab.org/

13 https://inception-project.github.io/

taalmodel te creëren dat gericht is op de medische sector. Bij de creatie van een nieuw model wordt dan onmiddellijk biasmitigatie toegepast.

Detectie en mitigatie van bias in data, taalmodellen en algoritmes
Zoals al vermeld, zijn er tegenwoordig velerlei oplossingen voor biasdetectie en -mitigatie, maar die werken op een zeer gefragmenteerde manier. Vaak vormt zo'n onderzoek slechts een stukje van de puzzel. Onderzoeksprojecten die de hele puzzel oplossen en een volledige aanpak voor biasdetectie en -mitigatie uitwerken, zijn er veel minder. Het is de bedoeling in dit onderzoeksproject een meer holistische oplossing voor te stellen, en dit in een praktijkgerichte context.

Gebruik makend van het *biasdetectiecorpus*, dat gebaseerd is op de analyse van de gegevens en gebruiksscenario's van de partnerinstelling, wordt het bestaande door AI ondersteunde gezondheidssysteem getest, om op die manier een overzicht te verkrijgen van mogelijke biases in de gegevens. In eerste instantie zal dit in een semi-automatische opzet plaatsvinden. Na verschillende iteraties is het echter de bedoeling om het volledige proces te automatiseren.

Met behulp van het overzicht van de resulterende biaskandidaten, zullen de woorden die hiermee overeenstemmen in de testdataset, automatisch gemaskeerd worden. Vertrekkend van de al beschreven actuele technieken, zoals prompt engineering, wordt er verder onderzocht hoe een automatisch biasdetectiesysteem kan worden uitgewerkt. Het is in eerste instantie de bedoeling dat dit wordt toegepast op eenvoudige contexten, maar de uiteindelijke bedoeling is om meerdere technieken te combineren of nieuwe technieken af te leiden die biasdetectie toepassen op zowel eenvoudige als complexere contexten.

Nadat bias zoveel mogelijk uit de data en taalmodellen is verwijderd, willen we weten in welke mate de *algoritmes* gebruik maken van de data waaruit de bias zoveel mogelijk is verwijderd en wat de impact is op de voorspellingen van het model.
Nadat de cellen in de matrixstructuur van het model die het meest gevoelig zijn voor biasdetectie en -mitigatie, zijn geïdentificeerd, kan er gemeten worden in hoeverre minder relevante of niet-relevante cellen kunnen worden verwijderd om het biasmitigatieproces te verbeteren zonder dat de algemene prestaties slechter worden (Voita et al., 2019).

Om nog meer inzicht te krijgen in welke technieken voor biasdetectie moeten worden toegepast, in combinatie met de identificatie van de cellen die hiervoor het gevoeligst zijn, kan er in de analyse nog een stap verder worden gegaan en kan er worden onderzocht welk linguïstisch redeneerproces binnen het grote taalmodel het meest actief is. In onderzoek is aangetoond dat bepaalde cellen van het transformermodel specifieke linguïstische redeneerprocessen vertonen (Vig, 2019; Vig et al., 2020). Een

voorbeeld van een NLP-taak waarbij transformermodellen worden gebruikt, is de al beschreven coreferentieresolutie.

Coreferentieresolutie betreft de NLP-taak om te bepalen wanneer twee of meer uitdrukkingen in een tekst verwijzen naar dezelfde entiteit. Als het transformermodel hiervoor minder goed werkt, zou deze slechtere werking ten onrechte kunnen worden toegeschreven aan een slechtere werking van de cellen in de matrixstructuur van de transformer, die verantwoordelijk zijn voor het veroorzaken van bias. Dit wordt geïllustreerd met het volgende voorbeeld, gebruikmakend van de prompt-engineeringtechniek, zoals al uitgelegd in paragraaf 'De state of the art in biasdetectie en -mitigatie'.

*De dokter zei tegen de verpleegster dat **[MASK]** vandaag te laat is.*
*De dokter zei tegen de verpleegster dat **hij** vandaag te laat is.*

In deze zin wordt het voornaamwoord 'hij' gebruikt. Men zou hier kunnen aannemen dat dit voornaamwoord verwijst naar de arts, en dat hier sprake is van genderbias, op basis vanmaatschappelijke stereotypes. Hierbij wordt dan aangenomen dat het meer waarschijnlijk is dat artsen mannelijk zijn. De andere hypothese in dit geval is echter dat het model gewoon minder goed presteert op het gebied van coreferentieresolutie en hier niet in staat is om het ontbrekende persoonlijke voornaamwoord 'zij' te koppelen aan 'de verpleegster'. Dit kan in dit geval een indicatie zijn dat er meer zinnen met voorbeelden van coreferentie moeten worden toegevoegd aan de extra data voor optimalisatie in functie van biasmitigatie.

Evaluatie en continue monitoring

Sleutelbegrippen in de voorgestelde benadering voor het evalueren van het door AI ondersteunde gezondheidssysteem, zijn 'een lerend gezondheidssysteem' en 'continue monitoring', zoals beschreven in Friedman en Rigby (2013), Magrabi et al. (2019) en Hyppönen et al. (2013). Geïnspireerd door deze onderzoeken, kon ik het volgende schema (Tabel 1) opstellen als framework voor de evaluatie van een door AI ondersteund gezondheidssysteem, waarbij een *in vitro* en een *in vivo* onderzoekssituatie aan elkaar gelinkt blijven, wat in volgende paragrafen verduidelijkt wordt.

1. **Definieer de context voor evaluatie**
   - 1.1. De AI-applicatie
     - 1.1.1. Wat is de verwachte output van de componenten die er deel van uitmaken?
     - 1.1.2. Wat moet er precies gemeten worden?
   - 1.2. Leg de grenzen vast van wat acceptabele en niet-acceptabele resultaten zijn
   - 1.3. Definieer types van risico, afhankelijk van een specifieke medische context
   - 1.4. Identificeer de stakeholders, gebruikers, patiënten, medische professionals en hun verwachtingen

2. **Definieer de indicatoren voor monitoring**, zoals (patiënt)veiligheid, privacy, bruikbaarheid, transparantie, uitlegbaarheid, bias, actualiteit, betrouwbaarheid en wijzigingen in werkprocessen

3. **Definieer een mechanisme voor iteratieve evaluatie**
   - 3.1. Definieer de data en een mechanisme om die automatisch aan te leveren
   - 3.2. Definieer een mechanisme om automatisch gebruikersfeedback te analyseren
   - 3.3. Creëer automatisch een evaluatierapport en -documentatie

Tabel 1. Schema voor het evalueren van een door AI ondersteund gezondheidssysteem

Context voor evaluatie

Hoe *evaluatie* van een door AI ondersteund gezondheidssysteem eruitziet, hangt af van de precieze medische context. Dat geldt ook voor de vraag welke vormen van biasdetectie en -mitigatie die deel uitmaken van een dergelijk systeem, gebruikt worden. De output van het systeem zal bijvoorbeeld verschillen tussen vroege detectie van sepsis en vroege voorspelling van Alzheimer: de keuze voor risicotypes en de definiëring van een acceptabele foutmarge bij het automatisch vroeg voorspellen van deze ziektes zullen in deze twee contexten anders zijn.

Ook de aanpak van biasdetectie en -mitigatie kan variëren per medische toepassing. Bij sepsis kan de focus bijvoorbeeld liggen op het voorkomen van verschillen in voorspellingsnauwkeurigheid tussen leeftijdsgroepen, terwijl bij Alzheimer de nadruk kan liggen op het minimaliseren van vooroordelen op basis van geslacht of etniciteit. Afhankelijk van de ziekte en de betrokken patiëntengroepen is het mogelijk dat de methodes om bias te detecteren en te mitigeren anders moeten worden ingesteld om de nauwkeurigheid en eerlijkheid van het systeem te waarborgen.

De sociale context kan hier ook van invloed zijn: de context in een publiek toegankelijk ziekenhuis (met veel artsen en patiënten van veel verschillende nationaliteiten of etnische achtergronden) kan anders zijn dan die in een privaat ziekenhuis (waar dit helemaal niet het geval is).

Er moet verder ook aandacht besteed worden aan de prestatie van het complete biasmitigatiesysteem; een van de innovaties in dit onderzoek is namelijk het ontwerp van een *volledig* biasdetectie- en mitigatiesysteem, dat het niveau van ad-hoc-oplossingen voor biasdetectie moet overstijgen en zowel goed functioneert met eenvoudige korte zinnen als met meer complexe contexten als input. Eveneens gaat hier bijzonder veel aandacht naar de evaluatie van de automatische generatie van de corpora en de synthetische data die automatisch gegenereerd worden om bias op te sporen en sterk te verminderen.

Indicatoren voor monitoring

Met dit onderzoek wordt beoogd bias te detecteren en te mitigeren als een van de indicatoren voor *continue monitoring*. In het begin van dit hoofdstuk heb ik het begrip bias gedefinieerd, maar wat we heel precies als bias gaan beschouwen, zal ook afhangen van de *EPD-data* die we gaan gebruiken en van de specifieke types van bias waarop de participanten in een interdisciplinaire context, zoals ethische en juridische experts, medische specialisten en datawetenschappers, de nadruk willen leggen.
Ook de mate van *transparantie en uitlegbaarheid of explainability* van AI is een belangrijke indicator die zorgt voor het vertrouwen bij de gebruiker in het door AI ondersteunde gezondheidssysteem. De visualisatie van verhoogde activiteit in bepaalde regio's als deel van de matrix van het transformermodel (paragraaf 'De state of the art in biasdetectie en -mitigatie'), gelinkt aan specifieke biasmitigatieprocessen, kan deel uitmaken van de evaluatie van het model, om daarmee aan de gebruiker een meer transparante en explainable of uitlegbare AI te bieden.
Andere belangrijke indicatoren zijn de *betrouwbaarheid van het systeem, de veiligheid en de privacy*: op geen enkele manier mogen de veiligheid van de patiënt of de bescherming van persoonsgegevens in gevaar komen wanneer de gezondheidszorg door AI wordt ondersteund.

Eveneens moet ervoor gezorgd worden dat het *systeem up-to-date* blijft en er *iteratieve* processen lopen die voor periodieke updates zorgen, die gevoed worden met nieuwe kennis in een specifieke medische sector en de daarvoor noodzakelijke data. Telkens zullen hierbij automatisch evaluatierapporten gegenereerd worden.

Iteratieve evaluatie

In het kader van een *continue monitoring* willen we ons door AI ondersteund gezondheidssysteem in interactie met de gebruiker optimaliseren; daarvan kan een *reinforcementlearning*-module (Sutton & Barto, 2018) onderdeel uitmaken. Reinforcement learning is een AI-benadering, waarbij een agent leert door te interageren met een omgeving. Een agent is een term die binnen de context van reinforcement learning wordt gebruikt om een zelfstandige entiteit te beschrijven die beslissingen neemt en handelt binnen een bepaalde omgeving. Hij observeert de omgeving, kiest acties op basis van de huidige situatie en past zijn gedrag aan op basis van de feedback die hij ontvangt. Deze feedback krijgt hij in de vorm van een beloning of straf op basis van zijn acties; op die manier wordt de besluitvorming van het systeem geleidelijk aan geoptimaliseerd. In het geval van een dialoogsysteem leert het systeem van onze opmerkingen over de antwoorden van het systeem, zoals bijvoorbeeld bij ChatGPT het geval is. Via trial-and-error verfijnt de agent zijn strategie om de cumulatieve beloningen in de loop van de tijd te maximaliseren. Dit omvat het verkennen van verschillende acties, leren van gevolgen en dienovereenkomstig aanpassen van gedrag.

Het nadeel van reinforcement learning in de data waarmee het systeem is gevoed, is dat het mogelijk meer bias kan creëren. Daarom moeten als onderdeel van een continue monitoringsysteem, biasdetectie en -mitigatie op iteratieve wijze worden toegepast, dus ook op de nieuwe, toegevoegde data (Smith et al., 2024).

Bredere context van dit onderzoek

7

In eerdere hoofdstukken is de evolutie van deeplearningmethodes besproken, met speciale aandacht voor de opkomst van transformermodellen. Deze modellen zijn krachtig in het verwerken van zowel gestructureerde als ongestructureerde data, wat ze bijzonder nuttig maakt voor vroege ziekteherkenning, zoals bij sepsis en alzheimer. Transformers kunnen bijvoorbeeld subtiele sepsis-signalen herkennen die traditionele methodes vaak missen. En door data uit diverse bronnen te combineren, kunnen ze de diagnostische precisie voor alzheimer verbeteren. Verder heb ik het potentieel van een virtuele patiënt uitgelicht, die door multimodaliteit en emotiedetectie therapeuten kan helpen om hun interactieve vaardigheden te ontwikkelen.

Een belangrijk aspect van ons onderzoek is de detectie en mitigatie van bias in AI-systemen. AI-toepassingen in de gezondheidszorg moeten immers niet alleen efficiënt en effectief, maar ook eerlijk zijn. Bias kan ontstaan door ongelijkheden in trainingsdata of door de manier waarop algoritmes zijn ontworpen. Zonder adequate maatregelen kan bias leiden tot oneerlijke medische behandelingen en diagnoses. Daarom moet er gefocust worden op het ontwikkelen van methodes voor het identificeren en verminderen van bias. Deze methodes bieden technieken zoals het herzien van de trainingsdatasets, het aanpassen van algoritmes die gevoelig zijn voor onder- of overrepresentatie in data en het regelmatig evalueren van AI-modellen op bias. Deze inspanningen moeten bijdragen aan de implementatie van verantwoordelijke AI in de gezondheidszorg en ervoor zorgen dat deze technologieën zowel doeltreffend als rechtvaardig zijn.

Om onze benadering van een door responsible AI ondersteund gezondheidssysteem te kunnen realiseren, moet voldaan worden aan de vereisten van een optimale onderzoekscontext die de creatie van ‘een lerend gezondheidssysteem’ en ‘continue monitoring’ kan garanderen. Deze voorwaarden zijn belangrijk voor het succes van het systeem en zullen in de volgende paragrafen worden beschreven.

Het kenniscentrum: hart van onderzoek en verbinding

De geplande onderzoeksactiviteiten zullen zich voornamelijk afspelen vanuit het Kenniscentrum Creating 010 van Hogeschool Rotterdam. Dit kenniscentrum richt zich op praktijkgericht onderzoek naar de maatschappelijke transformaties door digitalisering en ontwikkelingen in informatie- en communicatietechnologie, met speciale aandacht voor gezondheid, digitalisering en AI. Vanuit deze visie pakt het kenniscentrum vier maatschappelijke opgaven aan: *een duurzame delta, een toekomstbestendige*

economie, een vitale gemeenschap en *een slimme en sociale stad*. Deze opgaven omvatten het versnellen van de transitie naar hernieuwbare energie, het ontwikkelen van circulaire economische modellen en het verduurzamen van logistieke ketens. Verder wordt gestreefd naar een vitale gemeenschap door verschillen in onderwijs, gezondheid, welvaart en welzijn te verkleinen en de zelfredzaamheid van burgers te vergroten. Tot slot richt het onderzoek van het kenniscentrum zich op het bevorderen van een slimme en sociale stad door verantwoorde digitalisering van productieprocessen, dienstverlening en infrastructuur, ondersteund door AI-toepassingen.

Het onderzoek van het kenniscentrum benadrukt de centrale rol van mensen binnen hun sociale context. Ontwerpers, ontwikkelaars en gebruikers van technologie hebben een cruciale invloed; hun keuzes kunnen de potentiële bedreigingen van technologieën verminderen, bijvoorbeeld door te kiezen voor veilige dataopslag en open-source-software. Zij beïnvloeden de toekomst door producten, diensten en stedelijke ruimtes te ontwerpen die rekening houden met de behoeften van gebruikers, de gezondheidszorg en de bredere samenleving.

AI en ethiek

Het kenniscentrum vormt ook een solide brug naar andere onderzoeksactoren. Aangezien het geplande onderzoek tijdens de zes jaar van dit lectoraat een kruisbestuiving kent tussen NLP en AI enerzijds en zorg en ethiek anderzijds, is er ook een sterke link met het onderzoeksprogramma AI & Ethiek van Hogeschool Rotterdam. Dit onderzoeksprogramma focust op het praktisch toepassen en toegankelijk maken van AI en data. Docenten, studenten en onderzoekers werken samen met praktijkpartners aan het ontwikkelen van zowel technische als ethische vaardigheden en kennis, wat gepaard gaat met technische implementaties.

Kenniscentrum en onderwijsveld: onlosmakelijk verbonden

Kenniscentrum Creating 010, dat nauw samenwerkt met de creatieve industrie en ICT, richt zich op sectoren zoals zorg, ondernemerschap, retail en stedelijke ontwikkeling. Er wordt samengewerkt met het Instituut voor Communicatie, Media en Informatietechnologie (CMI) en de WERKplaats Techniek. De samenwerkingen kennen een interdisciplinaire aanpak met maatschappelijke partners, opleidingsinstituten en kenniscentra binnen en buiten Hogeschool Rotterdam. Hierbij wordt zowel horizontaal als verticaal naar elkaar geluisterd en samengewerkt; dit betekent dat ideeën en feedback vrijelijk worden gedeeld en overwogen op alle niveaus.
In de opleiding Applied Data Science en AI van CMI kunnen de activiteiten voor onderzoeken zoals naar een kader voor biasdetectie en -mitigatie, geïntegreerd worden binnen deze interdisciplinaire, op de praktijk gerichte onderwijsaanpak. Daarbij wordt een goede balans tussen onderwijs en toegepaste onderzoeksactiviteiten nagestreefd. Deze activiteiten zijn nauw verbonden met het kenniscentrum en het

beroepenveld. Studenten leren dit kader toe te passen in de zorg en in samenwerkingen met andere instellingen en bedrijven, waardoor ze een breed en ethisch onderbouwd perspectief ontwikkelen, dat toepasbaar is in diverse professionele contexten. Hierbij kunnen samenwerkingen ontstaan met studenten, docenten, docent-onderzoekers en lectoren van andere onderwijsinstellingen en kenniscentra binnen Hogeschool Rotterdam, zoals met Kenniscentrum voor Zorginnovatie (KCZI), het Instituut voor Gezondheidszorg (IvG) en het Instituut voor Sociale Opleidingen (ISO).

Bedrijven, overheid, instellingen en andere partners: de brandstof voor toegepast onderzoek

Digitalisering stimuleert de vorming van netwerken tussen mensen, organisaties en bedrijven, waarbinnen innovaties kunnen bloeien. Creating 010 onderzoekt nieuwe modellen, die de waarde van deze netwerken voor burgers, instellingen en bedrijven maximaliseren.

De financiering van de geplande onderzoeksactiviteiten komt vooral uit fondsen. Deze fondsen zijn gericht op het vormen van consortia die interdisciplinaire samenwerking stimuleren. Een essentieel onderdeel van dit proces is de duidelijke vraagarticulatie vanuit de betrokken instellingen, bedrijven of overheidsorganen. Dit zorgt ervoor dat het onderzoek niet alleen theoretisch relevant is, maar ook praktisch toepasbaar en direct gericht op de specifieke behoeften en uitdagingen waarmee deze organisaties geconfronteerd worden.

Zo focust het onderzoeksproject 'TheraVatars: Emotionele Avatars ter ondersteuning van psychologen in opleiding' op de ontwikkeling van een virtuele patiënt, die psychologiestudenten een platform biedt voor het trainen van soft skills binnen een breed spectrum aan klinische scenario's. De conventionele trainingsmethodes, zoals rollenspellen en oefenen met acteurs, zijn vaak beperkt in scenario's, kunnen financieel belastend zijn en zijn niet altijd effectief in het volledig voorbereiden van studenten op de complexiteit van hun toekomstige beroepspraktijk.

De centrale onderzoeksvraag is of een virtuele patiënt een effectieve oplossing kan bieden. Het antwoord hierop vereist grondig onderzoek binnen een robuust consortium. Dit consortium zal bestaan uit Kenniscentrum Creating 010 (dat gespecialiseerd is in AI en Natural Language Processing (NLP) voor het ontwikkelen van interactieve dialogen), Erasmus School of Social and Behavioural Sciences (ESSB) van Erasmus Universiteit Rotterdam (die helpt bij het verfijnen en beoordelen van deze dialogen en het evalueren van een prototype) en het Rotterdamse mbk-bedrijf Kurtosis (dat verantwoordelijk is voor het bouwen van het prototype). Terwijl de laatste tekstpagina's voor deze openbare les werden geschreven, kwam het goede nieuws binnen dat de financiering voor dit onderzoeksproject is toegekend.

Datalab

In de medische sector is de aanwezigheid van een datalab met Graphics Processing Units (GPU's), zowel fysiek als cloud-based, essentieel bij diepgaand onderzoek naar complexe aandoeningen zoals voorspelling van sepsis en alzheimer via deep learning. GPU's zijn bijzonder effectief voor parallelle dataverwerking, wat nodig is voor het snel trainen van deeplearningmodellen (transformermodellen), gevoed met grote datasets. Deze technologie versnelt de verwerking van data voor het ontwikkelen van modellen die bijdragen aan nauwkeurigere diagnoses en gepersonaliseerde behandelplannen. Zo kan het model van federated learning ervoor zorgen dat de privacy van patiëntgegevens wordt beschermd terwijl het waardevolle inzichten deelt over instellingen heen. Dit verhoogt niet alleen de efficiëntie en effectiviteit van medisch onderzoek, maar verbetert ook de beveiliging van de patiëntgegevens en de privacy van de patiënten. Hogeschool Rotterdam ontwikkelt hiervoor de nodige infrastructuur in de vorm van datalabs (Datalab Healthcare, stadslab, datalab, VR-lab en Datalab Rotterdam).[14]

Toegankelijkheid van medische data

Een struikelblok bij het opzetten van de onderzoeksactiviteiten en het creëren van deeplearningmodellen is het verzamelen van de benodigde medische data. In vergelijking met de VS is de situatie in Europa nog meer uitdagend. De uitdagingen hebben onder andere te maken met de privacywetgeving, de versnippering van data, het gebrek aan gestandaardiseerde, opensourcedatasets en de beperkte toegang tot data. Specifiek voor sepsis en alzheimer houdt dat het volgende in:

- *Privacy en regelgeving.* In Europa is de privacywetgeving strikter dan in de VS, met name door de Algemene Verordening Gegevensbescherming (AVG) (Your Europe, z.d.), die strikte regels stelt aan het gebruik van persoonsgegevens. Dit heeft grote invloed op het verzamelen en gebruiken van patiëntengegevens voor onderzoek. Ziekenhuizen en onderzoekers moeten zorgen voor strenge compliance, wat vaak resulteert in vertragingen en hoge kosten.
- *Versnippering van data.* Europa kent een grote diversiteit aan gezondheidssystemen en talen, wat leidt tot een versnipperd landschap van medische data. In tegenstelling tot landen zoals de VS, waar grote, nationale datasets zoals de MIMIC-III-dataset[15] voor sepsis en de ADNI-dataset[16] voor alzheimer beschikbaar zijn, ontbreekt het in Europa aan vergelijkbare, grootschalige en toegankelijke datasets. Dit beperkt de mogelijkheden voor onderzoekers om met machine learning effectieve voorspellingsmodellen te ontwikkelen in een Europese context.

14 https://www.hogeschoolrotterdam.nl/onderzoek/projecten-en-publicaties/zorginnovatie/zorginnovatie-met-technologie/hr-datalab-healthcare/ en https://cmgt.hr.nl/stadslab-datalab-vrlab en https://docs.datalabrotterdam.nl/

15 https://physionet.org/

16 https://adni.loni.usc.edu/

- *Gebrek aan standaardisatie*. Er is een gebrek aan standaardisatie van de manier waarop medische gegevens worden verzameld en opgeslagen over verschillende Europese landen heen. Dit bemoeilijkt het combineren van datasets uit verschillende bronnen en vermindert de bruikbaarheid van de data voor machinelearningtoepassingen.
- *Gebrek aan toegang tot data*. Er lopen wel initiatieven om de toegang tot gezondheidsgegevens te verbeteren, zoals het European Health Data Space (EHDS)[17], dat beoogt gezondheidsdata binnen Europa beter toegankelijk en uitwisselbaar te maken, maar deze zijn nog volop in ontwikkeling. Het huidige gebrek aan toegankelijke, gestructureerde en geanonimiseerde datasets beperkt onderzoekers in hun mogelijkheden om geavanceerde analytische technieken toe te passen.

Een Nederlands initiatief dat hierin een oplossing voor beoogt te vinden, is CumuluZ[18], dat streeft naar het creëren naar een publieke data-infrastructuur voor de zorgsector. Dit project wordt geleid door diverse zorgkoepels en richt zich op het creëren van een toegankelijke, transparante en privacyvriendelijke omgeving voor het delen van gezondheidsdata. Het doel is om gezondheidsdata beschikbaar te stellen voor zowel directe zorgverlening als medisch wetenschappelijk onderzoek. CumuluZ streeft ernaar om de afhankelijkheid van zorg-IT-leveranciers te verminderen en wil ervoor zorgen dat patiënten meer zeggenschap krijgen over hun eigen gegevens.

Living labs: werken in realistische experimentele omgevingen

Het *living lab* is de ideale ontmoetingsplaats waar vertegenwoordigers uit kenniscentra, onderwijs, onderzoek, overheid, instellingen en bedrijven samenwerken binnen een interdisciplinaire omgeving. Daarnaast vormt een living lab een uitstekend platform voor het verder ontwikkelen van initiatieven zoals CumuluZ (zie vorige paragraaf), gericht op het toegankelijk maken van data, en voor het ontplooien van nieuwe initiatieven op dit gebied. Hopelijk wordt de impact van CumuluZ gedurende de looptijd van dit zesjarige lectoraat zichtbaar, zodat er meer kans is om met data uit Nederlandse ziekenhuizen te werken.

Een living lab is een onderzoeksomgeving waarin innovaties worden ontwikkeld en getest in realistische situaties. In een living lab werken gebruikers, producenten en andere belanghebbenden zoals onderzoekers en overheden samen aan het co-creëren, testen en evalueren van nieuwe technologieën, producten of diensten. Deze labs benadrukken een mensgerichte aanpak, waarbij eindgebruikers actief betrokken zijn bij het creëren van oplossingen die goed aansluiten bij hun daadwerkelijke behoeften en omstandigheden. Living labs worden gebruikt in diverse sectoren zoals stedelijke ontwikkeling, gezondheidszorg en onderwijs.

17 https://www.european-health-data-space.com/

18 https://www.cumuluz.org/

Voor onderzoek naar de vroege voorspelling van sepsis en alzheimer via AI en NLP wordt gestreefd naar aansluiting bij een living lab. In een dergelijke omgeving kunnen deze technologieën direct in de praktijk worden getest met echte patiënten, zorgverleners en andere relevante belanghebbenden. Dit stelt onderzoekers in staat om te observeren hoe de AI-modellen functioneren in realistische medische settings, wat cruciaal is voor het verfijnen van de technologieën om nauwkeuriger diagnoses te kunnen stellen.

Binnen een living lab kan het onderzoek naar AI en NLP profiteren van directe feedback vanuit de praktijk. Dit betekent dat modellen voor vroege voorspelling van sepsis en alzheimer niet alleen in laboratoriumomstandigheden worden ontwikkeld, maar dat ze worden getest en verfijnd in een omgeving die de dagelijkse praktijk van ziekenhuizen en zorginstellingen nabootst.

Het integreren van de onderzoeksprojecten in een living lab stelt onderzoekers in staat om de technische mogelijkheden en beperkingen van AI en NLP te onderzoeken, terwijl ze tegelijkertijd rekening houden met ethische overwegingen, gebruiksgemak en de acceptatie door zorgprofessionals en patiënten. Dit bevordert de ontwikkeling van betrouwbare, gebruiksvriendelijke en ethisch verantwoorde technologieën die effectief kunnen bijdragen aan het vroegtijdig herkennen van sepsis en alzheimer, wat uiteindelijk de patiëntenzorg kan verbeteren en levens kan redden.

Een concreet voorbeeld hiervan vormen de Medical Delta Living Labs en Fieldlabs[19], die een belangrijke rol spelen binnen de innovatieketen door het testen van technologische oplossingen van bedrijven en zorginstellingen in realistische omstandigheden, samen met zorgprofessionals en patiënten. Deze labs richten zich op praktijkvragen die de basis vormen voor de publiek-private projecten die zij uitvoeren, en dragen zo bij aan de realisatie van concrete innovaties in producten of processen. Hierdoor hebben deze labs niet alleen een significante maatschappelijke en economische impact op de regio maar ook daarbuiten, en vormen zij een verbindende schakel met de academische onderzoeken aan de kennisinstellingen van Medical Delta.

De werking van de Medical Delta Living Labs en Fieldlabs is gebaseerd op vijf kernpijlers: interdisciplinaire samenwerking, oplossingsgerichtheid, wetenschappelijke onderbouwing, inbedding in de zorgpraktijk en een focus op versnelling van innovaties. In co-creatie met relevante stakeholders binnen en buiten Medical Delta kan worden gewerkt aan het uitwerken van verschillende gebruiksscenario's, waaronder de vroege voorspelling van alzheimer en sepsis. Hierbij kan een (digitaal) ecosysteem (data-infrastructuur en leergemeenschap) worden ontwikkeld waarin deze gebruiksscenario's kunnen worden uitgevoerd met privacybeschermende technologie (federated learning).

19 https://www.medicaldelta.nl/

Conclusie

Tijdens de zes jaar van dit lectoraat hoop ik onderzoek te realiseren dat de mogelijkheden van AI in de medische sector verkent, met een bijzondere focus op de ontwikkeling en implementatie van transformermodellen voor vroege ziekteherkenning. Specifiek en ter illustratie hiervan focussen we op ziekten zoals sepsis en alzheimer. Daarbij gebruiken we deze modellen om subtiele signalen te detecteren die door traditionele methodes over het hoofd worden gezien, en waar AI menselijke kennishiaten verder aanvult. Verder zal dit onderzoek zich concentreren op de creatie en het gebruik van virtuele patiënten om zorgprofessionals te trainen in complexe interactieve en emotionele vaardigheden. Deze technologieën zullen niet alleen bijdragen aan betere diagnostische precisie, maar kunnen ook zorgverleners in hun opleiding ondersteunen. Daarnaast moet dit onderzoek leiden naar een focus op multimodale transformermodellen die in staat zijn om meerdere informatiestromen te combineren en te verwerken. Deze AI-modellen integreren en analyseren data uit diverse bronnen zoals tekst, audio en visuele inputs, waardoor er een holistischer en accurater beeld van de patiëntgesteldheid kan worden verkregen. Dit vermogen om complexe data samen te voegen is noodzakelijk voor de nauwkeurige herkenning en interpretatie van menselijke gedragingen en pathologische indicatoren in scenario's uit de praktijk.

Een belangrijk element van dit onderzoek is het aanpakken van bias in AI-toepassingen binnen de gezondheidszorg. Door bias in trainingsdata en algoritmes te identificeren en te corrigeren, streven we naar eerlijke en rechtvaardige AI-oplossingen. Dit vereist een continue evaluatie en aanpassing van AI-modellen, om bias te detecteren en te mitigeren.

De inspanningen hiervoor worden ondersteund door Kenniscentrum Creating 010 van Hogeschool Rotterdam, dat dient als het hart van dit onderzoek en van de verbinding: het kenniscentrum faciliteert de samenwerking tussen verschillende disciplines en belanghebbenden, zoals bedrijven, zorginstellingen en burgers, waardoor onderzoek en praktijk nauw met elkaar worden verweven. Hierdoor ontstaat een sterke verbinding tussen technologische ontwikkelingen en de maatschappelijke context waarin ze worden toegepast, wat bijdraagt aan oplossingen die daadwerkelijk aansluiten bij de behoeften van de samenleving. Dit centrum speelt een cruciale rol in de kruisbestuiving tussen technologische en ethische aspecten van AI, waarbij onderzoekers, docenten en studenten samenwerken met praktijkpartners. Deze samenwerkingen zijn noodzakelijk voor het ontwikkelen van praktisch toepasbare en ethisch verantwoorde technologieën, die niet alleen theoretisch relevant zijn, maar ook afgestemd op de daadwerkelijke behoeften in de gezondheidszorg.

Daarnaast zal het onderzoek voortdurend de interactie tussen onderwijs en praktische toepassing bevorderen, waarbij studenten direct betrokken zijn bij het onderzoek en de implementatie van AI in de gezondheidszorg. Dit zal niet alleen hun technische vaardig-

heden versterken, maar ook hun begrip van ethische en maatschappelijke implicaties van AI. Deze kruisbestuiving wordt vooral belichaamd in living labs.

Samengevat, de hier voorgestelde onderzoeksactiviteiten in de komende zes jaar zullen een dynamische mix zijn van technologische innovatie, ethische overwegingen en praktische implementatie, allemaal gericht op het verbeteren van de patiëntenzorg en het opleiden van de volgende generatie zorgprofessionals. Door deze holistische benadering hoop ik een significant positieve impact te hebben op de gezondheidszorg, gesteund door state-of-the-art AI-technologieën, diepgaande ethische reflectie en interdisciplinaire samenwerking.

Literatuurlijst

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX symposium on operating systems design and implementation* (pp. 265-283). USENIX Association.

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia, 7*(3), 270-279. http://dx.doi.org/10.1016/j.jalz.2011.03.008

Alshouha, B., Serrano-Guerrero, J., Chiclana, F., Romero, F. P., & Olivas, J. A. (2024). BioEmoDetector: A flexible platform for detecting emotions from health narratives. *SoftwareX, 26*, 101670. http://dx.doi.org/10.1016/j.softx.2024.101670

Alzheimer Europe. (2019). *Prevalence of dementia in Europe*. Geraadpleegd op 27 februari 2024, van: https://www.alzheimer-europe.org/dementia/prevalence-dementia-europe?language_content_entity=en

Ambrosini, E., Caielli, M., Milis, M., Loizou, C., Azzolino, D., Damanti, S., & Ferrante, S. (2019). Automatic speech analysis to early detect functional cognitive decline in elderly population. In *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 212-216). IEEE. http://dx.doi.org/10.1109/EMBC.2019.8856768

Amland, R. C., & Sutariya, B. B. (2018). Quick sequential [sepsis-related] organ failure assessment (qSOFA) and St. John Sepsis Surveillance Agent to detect patients at risk of sepsis: An observational cohort study. *American Journal of Medical Quality*, *33*(1), 50-57. http://dx.doi.org/10.1177/1062860617692034

Arya, A. D., Verma, S. S., Chakarabarti, P., Chakrabarti, T., Elngar, A. A., Kamali, A. M., & Nami, M. (2023). A systematic review on machine learning and deep learning techniques in the effective diagnosis of Alzheimer's disease. *Brain Informatics*, *10*(1), 17. http://dx.doi.org/10.1186/s40708-023-00195-7

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation* (pp. 2200-2204). European Language Resources Association (ELRA).

Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: A survey. *Journal of Electrical Systems and Information Technology*, *10*(1), 40. http://dx.doi.org/10.1186/s43067-023-00108-y

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54-61. http://dx.doi.org/10.1145/3209581

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Balayn, A., Lofi, C., & Houben, G. J. (2021). Managing bias and unfairness in data for decision support: A survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, *30*(5), 739-768. http://dx.doi.org/10.1007/s00778-021-00671-8

Bargh, M. S., & Choenni, S. (2022). Towards an Integrated Approach for Preserving Data Utility, Privacy and Fairness. In *International Conference on Multidisciplinary Research* (Vol. 2022, pp. 290-306). http://dx.doi.org/10.26803/MyRes.2022.24

Basser, P. J., Mattiello, J., & LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, *66*(1), 259-267. http://dx.doi.org/10.1016/S0006-3495(94)80775-1

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3615–3620). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D19-1371

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604. http://dx.doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). Association for Computing Machinery, New York. http://dx.doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons.

Bhushan, I., Kour, M., Kour, G., Gupta, S., Sharma, S., & Yadav, A. (2018). Alzheimer's disease: Causes & treatment: A review. *Annals of Biotechnology*, *1*(1), 1002. http://dx.doi.org/10.33582/2637-4927/1002

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In S. Friedler, & C. Wilson (Eds.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 149-159). PMLR.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.485

Boerman, A. W., Schinkel, M., Meijerink, L., van den Ende, E. S., Pladet, L. C., Scholtemeijer, M. G., & Nanayakkara, P. W. (2022). Using machine learning to predict blood culture outcomes in the emergency department: A single-centre, retrospective, observational study. *BMJ Open*, 12(1), e053332. http://dx.doi.org/10.1136/bmjopen-2021-053332

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A. & Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, *101*(6), 1644-1655 http://dx.doi.org/10.1378/chest.101.6.1644

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Breiman, L. (2017). *Classification and regression trees*. Routledge. http://dx.doi.org/10.1201/9781315139470

Brooks, R., & Gomez, K. (2015). Cover story: Rodney Brooks changes the face of robotics. *Create*, *1*(1), 30-36.

Buchanan, B. G., & Shortliffe, E. H. (1984). Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming project (the Addison-Wesley series in artificial intelligence). Addison-Wesley Longman.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Busker, T., Choenni, S., & Shoae Bargh, M. (2023). Stereotypes in ChatGPT: An empirical study. In *Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance: Digital Governance for Democratic, Equitable, and Inclusive Societies* (pp. 24-32). Association for Computing Machinery. http://dx.doi.org/10.1145/3614321.3614325

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186. http://dx.doi.org/10.1126/science.aal4230

Cambria, E., Liu, Q., Decherchi, S., Xing, F., & Kwok, K. (2022). SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3829-3839). European Language Resources Association.

Caponnetto, P., & Casu, M. (2022). Update on cyber health psychology: Virtual reality and mobile health tools in psychotherapy, clinical rehabilitation, and addiction treatment. *International Journal of Environmental Research and Public Health*, *19*(6), 3516. http://dx.doi.org/10.3390/ijerph19063516

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., & Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage*, *112*, 232-243. http://dx.doi.org/10.1016/j.neuroimage.2015.02.037

Chen, Q., Hu, X., Wang, Z., & Hong, Y. (2024). AliFuse: Aligning and fusing multi-modal medical data for computer-aided diagnosis. *arXiv preprint arXiv:2401.01074*.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery. http://dx.doi.org/10.1145/2939672.2939785

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1724-1734). Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/D14-1179

Choenni, S., Netten, N., Shoae-Bargh, M., & Choenni, R. (2018). On the usability of big (social) data. In *2018 IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications* (pp. 1167-1174). IEEE. http://dx.doi.org/10.1109/BDCloud.2018.00172

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.). *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 276–286). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W19-4828

Cockrell, J. R., & Folstein, M. F. (2002). Mini-mental state examination. In J. R. M. Copeland, M. T. Abou-Saleh, & D. G. Blazer (Eds.). *Principles and practice of geriatric psychiatry* (pp. 140-141). *Wiley*, 140-141.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273-297. http://dx.doi.org/10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-2. http://dx.doi.org/10.1109/TIT.1967.1053964

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *20*(2), 215-232. http://dx.doi.org/10.1111/j.2517-6161.1958.tb00292.x

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Nill Sánchez, A., Raji, D., Lisi Rankin, J., Richardson, R., Schultz, J., Myers West., & Whittaker, M. (2020). *AI Now 2019 Report*. AI Now Institute.

Crevier, D. (1993). *AI: The tumultuous history of the search for artificial intelligence*. Basic Books.

Culliton, P., Levinson, M., Ehresman, A., Wherry, J., Steingrub, J. S., & Gallant, S. I. (2017). Predicting severe sepsis using text from the electronic health record. *arXiv preprint arXiv:1711.11536*.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357-366. http://dx.doi.org/10.1109/TASSP.1980.1163420

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: A Dutch RoBERTa-based language model. In T. Cohn, Y. He, & Y. Liu (Eds.). *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3255–3265). Association for Computational Linguistics http://dx.doi.org/10.18653/v1/2020.findings-emnlp.292

Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., & Das, R. (2016). Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Medical Informatics*, *4*(3), e5909. http://dx.doi.org/10.2196/medinform.5909

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Du, J., Jiang, J., Zheng, J., Zhang, H., Huang, D., & Lu, Y. (2023). Improving computation and memory efficiency for real-world transformer inference on GPUs. *ACM Transactions on Architecture and Code Optimization*, *20*(4), 1-22. http://dx.doi.org/10.1145/3617689

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. Wiley.

Duncan, C. F., Youngstein, T., Kirrane, M. D., & Lonsdale, D. O. (2021). Diagnostic challenges in sepsis. *Current Infectious Disease Reports*, *23*, 1-14. http://dx.doi.org/10.1007/s11908-021-00765-y

Dunn, A., Dagdelen, J., Walker, N., Lee, S., Rosen, A. S., Ceder, G., & Jain, A. (2022). Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.

Education Ecosystem. (2021). *What is Google Colab?* Medium.com. Geraadpleegd op 10 mei 2024, van: https://ledutokens.medium.com/what-is-google-colab-281c5b59638f

El-Sappagh, S., Abuhmed, T., Islam, S. R., & Kwak, K. S. (2020). Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing*, *412*, 197-215. http://dx.doi.org/10.1016/j.neucom.2020.05.087

Epstein, J. H., Levin, M., & Jowell, M. S. (2013). Agent based simulation for training and assessing students in the field of anesthesiology. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* (pp. 332-336). IEEE. http://dx.doi.org/10.1109/CBMS.2013.6627811

Eslami, S., de Melo, G., & Meinel, C. (2021). Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*.

European Commission. (z.d.). Data protection: Rules for the protection of personal data inside and outside the EU. Geraadpleegd op 20 mei 2024, van: https://commission.europa.eu/law/law-topic/data-protection_en

European Parliament. (2023, 8 juni). EU AI Act: First regulation on artificial intelligence. Geraadpleegd op 10 oktober 2023, van: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

European Sepsis Alliance. (2024). *Is Europe ready to lead the global agenda on sepsis?* Geraadpleegd op 24 maart 2024, van: https://www.europeansepsisalliance.org/annualmeeting

Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday, 28(11).* http://dx.doi.org/10.5210/fm.v28i11.13346

Fleuren, L. M., Klausch, T. L., Zwager, C. L., Schoonmade, L. J., Guo, T., Roggeveen, L. F. & Elbers, P. W. (2020). Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, *46*, 383-400. http://dx.doi.org/10.1007/s00134-019-05872-y

Fortuin, V. (2022). Priors in Bayesian deep learning: A review. *International Statistical Review*, *90*(3), 563-591. http://dx.doi.org/10.1111/insr.12502

Friedman, C., & Rigby, M. (2013). Conceptualising and creating a global learning health system. *International Journal of Medical Informatics, 82*(4), e63-e71. http://dx.doi.org/10.1016/j.ijmedinf.2012.05.010

Gamage, D., Sasahara, K., & Chen, J. (2021). The emergence of deepfakes and its societal implications: A systematic review. *TTO*, 28-39.

Gao, X., Shi, F., Shen, D., & Liu, M. (2023). Multimodal transformer network for incomplete image generation and diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, *110*, 102303. http://dx.doi.org/10.1016/j.compmedimag.2023.102303

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC. http://dx.doi.org/10.1201/9780429258411

Ghanbarzadeh, S., Huang, Y., Palangi, H., Moreno, R. C., & Khanpour, H. (2023). Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.). *Findings of the Association for Computational Linguistics*: *ACL 2023,* (pp. 5448–5458). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2023.findings-acl.336

Goh, K. H., Wang, L., Yeow, A. Y. K., Poh, H., Li, K., Yeow, J. J. L., & Tan, G. Y. H. (2021). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, *12*(1), 711. http://dx.doi.org/10.1038/s41467-021-20910-4

Goldberg, D. E. (1994). Genetic and evolutionary algorithms come of age. *Communications of the ACM*, *37*(3), 113-120.

Goyal, P., Pandey, S., & Jain, K. (2018). *Deep learning for natural language processing*. Apress. http://dx.doi.org/10.1007/978-1-4842-3685-7

Guirgis, F. W., Jones, L., Esma, R., Weiss, A., McCurdy, K., Ferreira, J., & Gray-Eurom, K. (2017). Managing sepsis: Electronic recognition, rapid response teams, and standardized care save lives. *Journal of Critical Care*, *40*, 296-302. http://dx.doi.org/10.1016/j.jcrc.2017.04.005

Haider, F., de La Fuente, S., & Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 272-281. http://dx.doi.org/10.1109/JSTSP.2019.2955022

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., & Cambria, E. (2023). A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Hester, J., Youn, T. S., Trifilio, E., Robinson, C. P., Babi, M. A., Ameli, P., & Busl, K. M. (2021). The Modified Early Warning Score: A useful marker of neurological worsening but unreliable predictor of sepsis in the Neurocritically Ill: A retrospective cohort study. *Critical Care Explorations*, *3*(5), e0386. http://dx.doi.org/10.1097/CCE.0000000000000386

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735

Holland, J. (1992). *Adaptation in natural and artificial systems*: *An introductory analysis with applications to biology, control, and artificial intelligence*. The MIT Press. http://dx.doi.org/10.7551/mitpress/1090.001.0001

Holste, G., Partridge, S. C., Rahbar, H., Biswas, D., Lee, C. I., & Alessio, A. M. (2021). End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3294-3303). IEEE. http://dx.doi.org/10.1109/ICCVW54120.2021.00368

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554-2558. http://dx.doi.org/10.1073/pnas.79.8.2554

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432. http://dx.doi.org/10.1111/lnc3.12432

Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digital Medicine*, *3*(1), 136. http://dx.doi.org/10.1038/s41746-020-00341-z

Hutchins, W. J. (1986). *Machine translation*: *Past, present, future*. Ellis Horwood.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Web and Social Media* (pp. 216-225). AAAI Press. http://dx.doi.org/10.1609/icwsm.v8i1.14550

Hyppönen, H., Faxvaag, A., Gilstad, H., Hardardottir, G. A., Jerlvall, L., Kangas, M., & Vimarlund, V. (2013). Nordic eHealth indicators: Organisation of research, first results and plan for the future. In *MEDINFO 2013* (pp. 273-277). IOS Press.

IBM. (2011). *Watson, 'Jeopardy!' champion*. Geraadpleegd op 12 juni 2024, van: https://www.ibm.com/history/watson-jeopardy

Isaacson, W. (2023). *Elon Musk*. Fayard.

Islam, K. R., Prithula, J., Kumar, J., Tan, T. L., Reaz, M. B. I., Sumon, M. S. I., & Chowdhury, M. E. (2023). Machine learning-based early prediction of sepsis using electronic health records: A systematic review. *Journal of Clinical Medicine*, *12*(17), 5658. http://dx.doi.org/10.3390/jcm12175658

Ivanov, A. V., Jalalvand, S., Gretter, R., & Falavigna, D. (2013). Phonetic and anthropometric conditioning of MSA-KST cognitive impairment characterization system. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 228-233). IEEE. http://dx.doi.org/10.1109/ASRU.2013.6707734

Iwendi, C., Huescas, C. G. Y., Chakraborty, C., & Mohan, S. (2024). COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients. *Journal of Experimental & Theoretical Artificial Intelligence*, *36*(3), 315-335. http://dx.doi.org/10.1080/0952813X.2022.2058097

Jacenków, G., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Indication as prior knowledge for multimodal disease classification in chest radiographs with transformers. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (pp. 1019-1023). IEEE. http://dx.doi.org/10.1109/ISBI52829.2022.9761567

Jack Jr, C. R., Albert, M. S., Knopman, D. S., McKhann, G. M., Sperling, R. A., Carrillo, M. C., & Phelps, C. H. (2011). Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 257-262. http://dx.doi.org/10.1016/j.jalz.2011.03.004

Jaroudi, W., Garami, J., Garrido, S., Hornberger, M., Keri, S., & Moustafa, A. A. (2017). Factors underlying cognitive decline in old age and Alzheimer's disease: the role of the hippocampus. *Reviews in the Neurosciences*, *28*(7), 705-714. http://dx.doi.org/10.1515/revneuro-2016-0086

Jiang, Z., Bo, L., Wang, L., Xie, Y., Cao, J., Yao, Y., & Bian, J. (2023). Interpretable machine-learning model for real-time, clustered risk factor analysis of sepsis and septic death in critical care. *Computer Methods and Programs in Biomedicine*, *241*, 107772. http://dx.doi.org/10.1016/j.cmpb.2023.107772

Johnson, A., Pollard, T., & Mark, R. (2016, 4 september). *MIMIC-III Clinical Database*. PhysioNet. Geraadpleegd op 13 mei 2024, van: https://physionet.org/content/mimiciii/1.4

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 1-9. http://dx.doi.org/10.1038/sdata.2016.35

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, 17(12), 1174-1179. http://dx.doi.org/10.1038/mp.2012.105

Kawahara, J., Daneshvar, S., Argenziano, G., & Hamarneh, G. (2018). Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 538-546. http://dx.doi.org/10.1109/JBHI.2018.2824327

Kenny, P., Parsons, T. D., Gratch, J., Leuski, A., & Rizzo, A. A. (2007). Virtual patients for clinical therapist skills training. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (pp. 197-210). Springer. http://dx.doi.org/10.1007/978-3-540-74997-4_19

Kijpaisalratana, N., Sanglertsinlapachai, D., Techaratsami, S., Musikatavorn, K., & Saoraya, J. (2022). Machine learning algorithms for early sepsis detection in the emergency department: A retrospective study. *International Journal of Medical Informatics*, *160*, 104689. http://dx.doi.org/10.1016/j.ijmedinf.2022.104689

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Koch, C., Edinger, F., Fischer, T., Brenck, F., Hecker, A., Katzer, C., & Schneck, E. (2020). Comparison of qSOFA score, SOFA score, and SIRS criteria for the prediction of infection and mortality among surgical intermediate and intensive care patients. *World Journal of Emergency Surgery*, *15*, 1-10. http://dx.doi.org/10.1186/s13017-020-00343-y

Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). [Retracted] A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*, (1), 1684017. http://dx.doi.org/10.1155/2022/1684017

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90. http://dx.doi.org/10.1145/3065386

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 282-289). Morgan Kaufmann Publishers Inc.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. http://dx.doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE* (pp. 2278-2324). IEEE. http://dx.doi.org/10.1109/5.726791

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240. http://dx.doi.org/10.1093/bioinformatics/btz682

Lemaréchal, C. (2012). Cauchy and the gradient method. In M. Grötschel (Ed.), *Documenta Mathematica extra volume, Optimization Stories: 21st International symposium on Mathematical Programming, Berlin, August 19-24, 2012* (pp. 251-254) http://dx.doi.org/10.4171/dms/6/27

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*(11), 33-38. http://dx.doi.org/10.1145/219717.219745

Levy, M. M., Fink, M. P., Marshall, J. C., Abraham, E., Angus, D., Cook, D., & Ramsay, G. (2003). 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Critical Care Medicine*, *31*(4), 1250-1256.

Li, Y., Wang, H., & Luo, Y. (2020). A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 1999-2004). IEEE. http://dx.doi.org/10.1109/BIBM49941.2020.9313289

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In M. Lee, & C. Munteanu (Eds.). *Proceedings of the 5th International Conference on Conversational User Interfaces* (pp. 1-6). http://dx.doi.org/10.1145/3571884.3604316

Llanes-Jurado, J., Gómez-Zaragozá, L., Minissi, M. E., Alcañiz, M., & Marín-Morales, J. (2024). Developing conversational Virtual Humans for social emotion elicitation based on large language models. *Expert Systems with Applications*, *246*, 123261. http://dx.doi.org/10.1016/j.eswa.2024.123261

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., & Beg, M. F. (2018). Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Scientific Reports*, *8*(1), 5697. http://dx.doi.org/10.1038/s41598-018-22871-z

Lu, M. Y., Chen, T. Y., Williamson, D. F., Zhao, M., Shady, M., Lipkova, J., & Mahmood, F. (2021). AI-based pathology predicts origins for cancers of unknown primary. *Nature*, *594*(7861), 106-110. http://dx.doi.org/10.1038/s41586-021-03512-4

Lyell, D., Magrabi, F., Raban, M. Z., Pont, L. G., Baysari, M. T., Day, R. O., & Coiera, E. (2017). Automation bias in electronic prescribing. *BMC Medical Informatics and Decision Making*, *17*, 1-10. http://dx.doi.org/10.1186/s12911-017-0425-5

Lyra, S., Leonhardt, S., & Antink, C. H. (2019). Early prediction of sepsis using random forest classification for imbalanced clinical data. In *2019 Computing in Cardiology (CinC)* (pp. 1-4). IEEE. http://dx.doi.org/10.22489/CinC.2019.276

Magrabi, F., Ammenwerth, E., McNair, J. B., de Keizer, N. F., Hyppönen, H., Nykänen, P., & Georgiou, A. (2019). Artificial intelligence in clinical decision support: Challenges for evaluating AI and practical implications: A position paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. *Yearbook of Medical Informatics*, *28*(1), 128. http://dx.doi.org/10.1055/s-0039-1677903

Mao, C., Xu, J., Rasmussen, L., Li, Y., Adekkanattu, P., Pacheco, J., & Luo, Y. (2023). AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. *Journal of Biomedical Informatics*, *144*, 104442. http://dx.doi.org/10.1016/j.jbi.2023.104442

Mao, R., Liu, Q., He, K., Li, W., & Cambria, E. (2023). The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, *14*(3), 1743-1753. http://dx.doi.org/10.1109/TAFFC.2022.3204972

Marik, P. E. (2015). Sepsis. *Evidence-Based Critical Care*, 107-148. http://dx.doi.org/10.1007/978-3-319-11020-2

Markov, A. A. (2006). An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains. *Science in Context*, *19*(4), 591-600
.

Martin-Moreno, J. M., Alegre-Martinez, A., Martin-Gorgojo, V., Alfonso-Sanchez, J. L., Torres, F., & Pallares-Carratala, V. (2022). Predictive models for forecasting public health scenarios: Practical experiences applied during the first wave of the COVID-19 pandemic. *International Journal of Environmental Research and Public Health*, *19*(9), 5546. http://dx.doi.org/10.3390/ijerph19095546

Massat, M. B. (2018). A Promising future for AI in breast cancer screening. *Applied Radiology*, *47*(9), 22-5. http://dx.doi.org/10.37549/AR2521

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*, 115-133. http://dx.doi.org/10.1007/BF02478259

McFadden, B. R., Inglis, T. J., & Reynolds, M. (2023). Machine learning pipeline for blood culture outcome prediction using Sysmex XN-2000 blood sample results in Western Australia. *BMC Infectious Diseases*, *23*(1), 552. http://dx.doi.org/10.1186/s12879-023-08535-y

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263-269. http://dx.doi.org/10.1016/j.jalz.2011.03.005

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In A. Singh, & J. Zhu (Eds.). *Proceedings of the 20th International Conference on Artificial intelligence and statistics* (pp. 1273-1282). PMLR.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1-35. http://dx.doi.org/10.1145/3457607

Miller, R. A. (1984). INTERNIST-1/CADUCEUS: Problems facing expert consultant programs. *Methods of Information in Medicine*, *23*(01), 9-14. http://dx.doi.org/10.1055/s-0038-1635320

Miller, R. A., McNeil, M. A., Challinor, S. M., Masarie Jr, F. E., & Myers, J. D. (1986). The INTERNIST-1/quick medical REFERENCE project: Status report. *Western Journal of Medicine*, *145*(6), 816.

Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M., & Christensen, H. (2018). Detecting Signs of Dementia Using Word Vector Representations. In *Interspeech* (pp. 1893-1897). http://dx.doi.org/10.21437/Interspeech.2018-1764

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., & Cooper, L. A. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, *115*(13), E2970-E2979. http://dx.doi.org/10.1073/pnas.1717139115

Moturi, S., & Srikanth Vemuru, D. S. (2020). Classification model for prediction of heart disease using correlation coefficient technique. *International Journal of Advanced Trends in Computer Science and Engineering 9*(2), 2116-2123. http://dx.doi.org/10.30534/ijatcse/2020/185922020

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., & Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, *1*(1), 55-66.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., & Tily, H. (2010). Crowdsourcing and language studies: The new generation of linguistic data. In C. Callison-Burch, & M. Dredze (Eds.). *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 122-130). Association for Computational Linguistics.

Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine*, *211*, 106433. http://dx.doi.org/10.1016/j.cmpb.2021.106433

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In J. Ürnkranz, & T. Joachims (Eds.). *Proceedings of the 27th International Conference on Machine Learning* (pp. 807-814). Omnipress.

Nederlands Herseninstituut. (2018). *De ziekte van Alzheimer*. Geraadpleegd op 22 februari 2024, van: https://herseninstituut.nl/over-het-brein/de-ziekte-van-alzheimer

Neve, R. L., McPhie, D. L., & Chen, Y. (2000). Alzheimer's disease: A dysfunction of the amyloid precursor protein. *Brain Research*, *886*(1-2), 54-66.

Ng, H. W., Koh, A., Foong, A., & Ong, J. (2023). Real-time hybrid language model for virtual patient conversations. In *Artificial Intelligence in Education: 24th International Conference, AIED 2023* (pp. 780-785). Springer. http://dx.doi.org/10.1007/978-3-031-36272-9_71

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453. http://dx.doi.org/10.1126/science.aax2342

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13. http://dx.doi.org/10.3389/fdata.2019.00013

Pal, R., Patel, S., Bhatnagar, A., Garg, H., Singh, P., Soun, R. S., & Sethi, T. (2022). ShockModes: A multimodal model for prognosticating intensive care outcomes from physician notes and vitals [Preprint]. *medRxiv*, 2022-12.

Pal, R., Garg, H., Patel, S., & Sethi, T. (2023). Bias amplification in intersectional subpopulations for clinical phenotyping by large language models [Preprint]. *medRxiv*, 2023-03.

Papathanakos, G., Andrianopoulos, I., Xenikakis, M., Papathanasiou, A., Koulenti, D., Blot, S., & Koulouras, V. (2023). Clinical sepsis phenotypes in critically ill patients. *Microorganisms*, *11*(9), 2165. http://dx.doi.org/10.3390/microorganisms11092165

Park, T., Gu, P., Kim, C. H., Kim, K. T., Chung, K. J., Kim, T. B., & Oh, J. K. (2023). Artificial intelligence in urologic oncology: The actual clinical practice results of IBM Watson for Oncology in South Korea. *Prostate International*, *11*(4), 218-221. http://dx.doi.org/10.1016/j.prnil.2023.09.001

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Petersen, E., Feragen, A., da Costa Zemsch, M. L., Henriksen, A., Wiese Christensen, O. E., Ganz, M., & Alzheimer's Disease Neuroimaging Initiative. (2022). Feature robustness and sex differences in medical imaging: A case study in MRI-based Alzheimer's disease detection. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2022 - 25th International Conference, Proceedings* (pp. 88-98). Springer Science and Business Media Deutschland. http://dx.doi.org/10.1007/978-3-031-16431-6_9

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language,* (pp. 2463–2473). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D19-1250

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google translate. *Neural Computing and Applications*, *32*, 6363-6381. http://dx.doi.org/10.1007/s00521-019-04144-6

Qin, F., Madan, V., Ratan, U., Karnin, Z., Kapoor, V., Bhatia, P., & Kass-Hout, T. (2021). Improving early sepsis prediction with multi modal learning. *arXiv preprint arXiv:2107.11094*.

Qin, Y., Liu, W., Peng, Z., Ng, S. I., Li, J., Hu, H., & Lee, T. (2021). Exploiting pre-trained ASR models for Alzheimer's disease recognition through spontaneous speech. *arXiv preprint arXiv:2110.01493*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748-8763). PMLR.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, *4*(1), 86. http://dx.doi.org/10.1038/s41746-021-00455-y

Ravensbergen, R.. (2024, 19 maart). *Ophef om allerlei BN'ers in nep-pornofilmpjes*. Metronieuws. Geraadpleegd op 20 april 2024, van: https://www.metronieuws.nl/in-het-nieuws/binnenland/2024/03/ophef-deepfakes-porno-bners-deepfake

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121-154. http://dx.doi.org/10.1016/j.iotcps.2023.04.003

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D19-1410

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, *71*, 103107. http://dx.doi.org/10.1016/j.bspc.2021.103107

Rivera-Gutierrez, D. J., Kopper, R., Kleinsmith, A., Cendan, J., Finney, G., & Lok, B. (2014). Exploring gender biases with virtual patients for high stakes interpersonal skills training. In *Intelligent Virtual Agents*: *14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14* (pp. 385-396). Springer. http://dx.doi.org/10.1007/978-3-319-09767-1_50

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386. http://dx.doi.org/10.1037/h0042519

Rothman, D. (2022). *Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4*. Packt.

Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., & Naghavi, M. (2020). Global, regional, and national sepsis incidence and mortality, 1990-2017: Analysis for the Global Burden of Disease Study. *The Lancet*, *395*(10219), 200-211. http://dx.doi.org/10.1016/S0140-6736(19)32989-7

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson.

Rytting, C., & Wingate, D. (2021). Leveraging the inductive bias of large language models for abstract textual reasoning. *Advances in Neural Information Processing Systems*, *34*, 17111-17122.

Safranek, C. W., Sidamon-Eristoff, A. E., Gilson, A., & Chartash, D. (2023). The role of large language models in medical education: Applications and implications. *JMIR Medical Education*, *9*, e50945.

Samareh, A., Jin, Y., Wang, Z., Chang, X., & Huang, S. (2018). Predicting depression severity by multi-modal feature engineering and fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1). AAAI. http://dx.doi.org/10.1609/aaai.v32i1.12152

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. http://dx.doi.org/10.1147/rd.33.0210

Schick, T., Udupa, S., & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, *9*, 1408-1424. http://dx.doi.org/10.1162/tacl_a_00434

Schilling, L. P., Balthazar, M. L. F., Radanovic, M., Forlenza, O. V., Silagi, M. L., Smid, J., & Nitrini, R. (2022). Diagnosis of Alzheimer's disease: Recommendations of the Scientific Department of Cognitive Neurology and Aging of the Brazilian Academy of Neurology. *Dementia & Neuropsychologia*, *16*, 25-39. http://dx.doi.org/10.1590/1980-5764-dn-2022-s102pt

Schünemann, H. J. B. J., Brożek, J., Guyatt, G., & Oxman, A. (2013). Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. *Updated October 15th*, 2013.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417-424.

Sepsis en daarna. (2020). *Feiten over sepsis*. Geraadpleegd op 9 maart 2024, van: https://www.sepsis-en-daarna.nl/alles-over-sepsis/feiten-en-vragen/feiten-over-sepsisSeymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., & Angus, D. C. (2016). Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama*, *315*(8), 762-774. http://dx.doi.org/10.1001/jama.2016.0288

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, *27*(12), 2176-2182. http://dx.doi.org/10.1038/s41591-021-01595-0

Shortliffe, E. (Ed.). (2012). *Computer-based medical consultations: MYCIN* (Vol. 2). Elsevier.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489. http://dx.doi.org/10.1038/nature16961

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., & Natarajan, V. (2023). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Sipser, M. (1996). Introduction to the Theory of Computation. *ACM Sigact News*, *27*(1), 27-29. http://dx.doi.org/10.1145/230514.571645

Smith, B., Khojandi, A., & Vasudevan, R. (2024). Bias in reinforcement learning: A review in healthcare applications. *ACM Computing Surveys*, *56*(2), 1-17. http://dx.doi.org/10.1145/3609502

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Soni, V. D. (2020). Chronic disease detection model using machine learning techniques. *International Journal of Scientific & Technology Research*, *9*(9), 262-266.

Stefan, R., Mantl, G., Höfner, C., Stammer, J., Hochgerner, M., & Petersdorfer, K. (2021). Remote psychotherapy during the COVID-19 pandemic: Experiences with the transition and the therapeutic relationship: A longitudinal mixed-methods study. *Frontiers in Psychology*, *12*, 743430. http://dx.doi.org/10.3389/fpsyg.2021.743430

Stehwien, S., & Vu, N. T. (2016). Exploring the correlation of pitch accents and semantic slots for spoken language understanding. In *17th Annual Conference of the International Speech Communication Association (Interspeech 2016): Understanding Speech Processing in Humans and Machines* (pp. 730-734). International Speech Communication Association. http://dx.doi.org/10.21437/Interspeech.2016-511

Stompe, T., Ortwein-Swoboda, G., Chaudhry, H. R., Friedmann, A., Wenzel, T., & Schanda, H. (2001). Guilt and depression: A cross-cultural comparative study. *Psychopathology*, *34*(6), 289-298. http://dx.doi.org/10.1159/000049327

Strapparava, C., & Valitutti, A. (2004, May). WordNet Affect: An affective extension of WordNet. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.). *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (pp. 1083-1086). European Language Resources Association (ELRA).

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24-31. http://dx.doi.org/10.1109/MSPEC.2019.8678513

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, *10*(1), 116. http://dx.doi.org/10.1038/s41398-020-0780-3

Su, C. Y., & Tseng, C. Y. (2018). Perceivable information structure in discourse prosody-detecting prominent prosodic words in spoken discourse using F0 contour. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 424-428). IEEE. http://dx.doi.org/10.1109/ISCSLP.2018.8706643

Subbe, C. P., Kruger, M., Rutherford, P., & Gemmel, L. (2001). Validation of a modified Early Warning Score in medical admissions. *QJM*, 94(10), 521-526. http://dx.doi.org/10.1093/qjmed/94.10.521

Subbe, C. P., Slater, A., Menon, D., & Gemmell, L. (2006). Validation of physiological scoring systems in the accident and emergency department. *Emergency Medicine Journal*, *23*(11), 841. http://dx.doi.org/10.1136/emj.2006.035816

Subramoniam, M., Aparna, T. R., Anurenjan, P. R., & Sreeni, K. G. (2022). Deep learning-based prediction of Alzheimer's disease from magnetic resonance images [Withdrawn]. In M. Saraswat, H. Sharma, & K. Veer Arya (Eds.). *Intelligent vision in healthcare* (pp. 145-151). Springer. http://dx.doi.org/10.1007/978-981-16-7771-7_12

Sun, M., Oliwa, T., Peek, M. E., & Tung, E. L. (2022). Negative patient descriptors: Documenting racial bias in the electronic health record: Study examines racial bias in the patient descriptors used in the electronic health record. *Health Affairs*, *41*(2), 203-211. http://dx.doi.org/10.1377/hlthaff.2021.01423

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text classification via large language models. In H. Bouamor, J. Pino & K. Bali (Eds.). *Findings of the Association for Computational Linguistics*: *EMNLP 2023,* (pp. 8990–9005). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2023.findings-emnlp.603

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning*: *An introduction*. MIT press.

Tao, Y., Yang, M., Shen, H., Yang, Z., Weng, Z., & Hu, B. (2023). classifying anxiety and depression through LLMs virtual interactions: A case study with ChatGPT. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2259-2264). IEEE. http://dx.doi.org/10.1109/BIBM58861.2023.10385305

Thomas-MacLean, R., Stoppard, J., Miedema, B. B., & Tatemichi, S. (2005). Diagnosing depression: there is no blood test. *Canadian Family Physician*, *51*(8), 1102-1103.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P. C., & Natarajan, V. (2024). Towards generalist biomedical AI. *NEJM AI*, *1*(3). http://dx.doi.org/10.1056/AIoa2300138

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016). A dictionary-based approach to racism detection in Dutch social media. *arXiv preprint arXiv:1608.08738*.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungs-problem. In J. Brundan, S. Chatterjee, M. Chudnovsky, D. Isaksen, V. Marković, J. McKernan, J. Newton, H. Oh, M. del Pino, D. Schindler, S. Sheffield, M. Visan, D. T. Wise, & M. Yakimov (Eds.). *Proceedings of the London Mathematical Society, s2-42*(1), 230-265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433-460 http://dx.doi.org/10.1093/mind/LIX.236.433

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1301-1309. http://dx.doi.org/10.1109/JSTSP.2017.2764438

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, & I. Guyon (Eds.). *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. 6000-6010). Curran Associates Inc.

Verkijk, S., & Vossen, P. (2021). MedROBERTa.nl: A language model for Dutch electronic health records. *Computational Linguistics in the Netherlands Journal*, *11*, 141-159.

Verma, S. S., Prasad, A., & Kumar, A. (2022). CovXmlc: High performance COVID-19 detection on X-ray images using multi-model classification. *Biomedical Signal Processing and Control*, *71*, 103272. http://dx.doi.org/10.1016/j.bspc.2021.103272

Vig, J. (2019). A multiscale visualization of attention in the transformer model. In M. R. Costa-jussà, & E. Alfonseca (Eds.). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (pp. 37–42). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P19-3007

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, *33*, 12388-12401.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In A. Korhonen, D. Traum, & L. Màrquez (Eds.). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5797–5808). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P19-1580

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Communications Medicine*, *1*(1), 25. http://dx.doi.org/10.1038/s43856-021-00028-w

Walther, S., Stegmayer, K., Sulzbacher, J., Vanbellingen, T., Müri, R., Strik, W., & Bohlhalter, S. (2015). Nonverbal social communication and gesture control in schizophrenia. *Schizophrenia Bulletin*, *41*(2), 338-345. http://dx.doi.org/10.1093/schbul/sbu222

Wang, L., Laurentiev, J., Yang, J., Lo, Y. C., Amariglio, R. E., Blacker, D., & Zhou, L. (2021). Development and validation of a deep learning model for earlier detection of cognitive decline from clinical notes in electronic health records. *JAMA*, *4*(11), e2135174-e2135174. http://dx.doi.org/10.1001/jamanetworkopen.2021.35174

Wang, R., Chaudhari, P., & Davatzikos, C. (2023). Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proceedings of the National Academy of Sciences*, *120*(6), e2211613120. http://dx.doi.org/10.1073/pnas.2211613120

Wang, X., Peng, Y., Lu, L., Lu, Z., & Summers, R. M. (2018). TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9049-9058). IEEE. http://dx.doi.org/10.1109/CVPR.2018.00943

Wang, Y., Zhao, Y., Callcut, R., & Petzold, L. (2022). Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. *arXiv preprint arXiv:2203.14469*.

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H. G., & Yeh, E. K. (1996). Cross-national epidemiology of major depression and bipolar disorder. *JAMA*, *276*(4), 293-299. http://dx.doi.org/10.1001/jama.1996.03540040037030

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., & Alzheimer's Disease Neuroimaging Initiative. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, *63*, 101694. http://dx.doi.org/10.1016/j.media.2020.101694

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., & QUADAS-2 Group. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529-536. http://dx.doi.org/10.7326/0003-4819-155-8-201110180-00009

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., & Quatieri, T. F. (2016). Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* (pp. 11-18). Association for Computing Machinery. http://dx.doi.org/10.1145/2988257.2988263

Wilmann, D., & Sterling, L. (2005). Guiding agent-oriented requirements elicitation: HOMER. In *Fifth International Conference on Quality Software (QSIC'05)* (pp. 419-424). IEEE. http://dx.doi.org/10.1109/QSIC.2005.34

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language*. MIT

Winograd, T. (1972). Understanding Natural Language, *Cognitive Psychology*, *3*(1), 1-191. http://dx.doi.org/10.1016/0010-0285(72)90002-3

Wooldridge, M. (2021). *A brief history of Artificial Intelligence: What it is, where we are, and where we are going*. Flatiron Books.

World Health Organization. (2020, 13 oktober). *Impact of COVID-19 on people's livelihoods, their health and our food systems*. Geraadpleegd op 10 juni 2024, van: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems

Yan, R., Zhang, F., Rao, X., Lv, Z., Li, J., Zhang, L., & Liang, J. (2021). Richer fusion network for breast cancer classification based on multimodal data. *BMC Medical Informatics and Decision Making*, *21*(1), 1-15. http://dx.doi.org/10.1186/s12911-020-01340-6

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., & Sahli, H. (2017). Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (pp. 53-59). Association for Computing Machinery.

Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M. C., & Jiang, D. (2017). Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (pp. 45-51). Association for Computing Machinery. http://dx.doi.org/10.1145/3133944.3133950

Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., & Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, *2*(4), 255-263. http://dx.doi.org/10.1002/hcs2.61

Yap, J., Yolland, W., & Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. *Experimental Dermatology*, *27*(11), 1261-1267. http://dx.doi.org/10.1111/exd.13777

Yoo, Y., Tang, L. Y., Li, D. K., Metz, L., Kolind, S., Traboulsee, A. L., & Tam, R. C. (2019). Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *7*(3), 250-259. http://dx.doi.org/10.1080/21681163.2017.1356750

Your Europe. (2022). Algemene verordening gegevensbescherming (AVG). Geraadpleegd op 11 mei 2024, van: https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_nl.htm#:~:text=De%20AVG%20bevat%20gedetailleerde%20voorschriften,EU%20gevestigd%20zijn%2C%20of%20daarbuitenYu, B., Quatieri, T. F., Williamson, J. R., & Mundt, J. C. (2015). Cognitive impairment prediction in the elderly based on vocal biomarkers. In *Proceedings of Interspeech 2015* (pp. 3734-3738) http://dx.doi.org/10.21437/Interspeech.2015-741

Zhang, D., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLOS ONE*, *7*(3), e33182. http://dx.doi.org/10.1371/journal.pone.0033182

Zhang, D., Yin, C., Hunold, K. M., Jiang, X., Caterino, J. M., & Zhang, P. (2021). An interpretable deep-learning model for early prediction of sepsis in the emergency department. *Patterns*, *2*(2).

Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., & Kumar, S. (2020). Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. In (*ICASSP 2020*) *2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7829-7833). IEEE. http://dx.doi.org/10.1109/ICASSP40776.2020.9053896

Zhang, Y., Sun, K., Liu, Y., & Shen, D. (2023). Transformer-based multimodal fusion for early diagnosis of Alzheimer's disease using structural MRI and PET. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE. http://dx.doi.org/10.1109/ISBI53787.2023.10230577

Zhang, Z., Lin, W., Liu, M., & Mahmoud, M. (2020). Multimodal deep learning framework for mental disorder recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 344-350). IEEE. http://dx.doi.org/10.1109/FG47880.2020.00033

Zhou, H. Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., & Li, W. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, 7, 743-744. http://dx.doi.org/10.1038/s41551-023-01045-x

Over Thierry Desot

Thierry Desot is lector op het gebied van natuurlijke taalverwerking (Natural Language Processing - NLP) bij Kenniscentrum Creating 010. Hij zet zich in voor innovatief en toegepast onderzoek, waarbij hij NLP-technologieën implementeert in projecten die vallen onder verantwoorde AI, met een focus op de gezondheidszorg.

Hij onderzoekt vooral de invloed van vooringenomenheid (bias) in taalmodellen die in de medische sector worden gebruikt. Dergelijke bias, waaronder gender-, religieuze en politieke vooroordelen, kan leiden tot onnauwkeurige diagnoses en de kwaliteit van de zorg voor minderheidsgroepen beïnvloeden. Zijn onderzoek richt zich in de beginjaren van zijn lectoraat op het identificeren en aanpakken van deze bias in grote taalmodellen, waaronder ChatGPT, en andere modellen die gespecialiseerd zijn in medische terminologie. Daarnaast onderzoekt hij hoe AI menselijke besluitvorming kan ondersteunen in complexe situaties, vooral binnen de zorgsector. Dit omvat projecten gericht op de vroege opsporing van ziekten, een taak waarbij AI en NLP essentieel kunnen zijn vanwege de complexiteit van de taak en het feit dat het moeilijk is om vroege symptomen te herkennen zonder deze technologieën.

Thierry Desot behaalde zijn doctoraat aan de Universiteit Grenoble-Alpes (2017-2020), met als thema semantiek van gesproken taal (Spoken Language Understanding - SLU), als onderdeel van een project met sociale impact, gericht op de ontwikkeling van een stemgestuurde communicatiemodule voor een smart home voor senioren, gebruikmakend van deeplearningtechnieken. Vervolgens werkte hij als postdoctoraal onderzoeker aan de Universiteit Gent (2021-2023), waar hij zich richtte op informatie-extractie voor een nieuwsaanbevelingssysteem door het gebruik van grote taalmodellen.

Eerdere uitgaven

Hogeschool Rotterdam Uitgeverij

Dementie, geestelijke gezondheid en gedrag
Auteur Sjacko Sobczak
ISBN 9789083481227
Verschijningsdatum november 2024
Aantal pagina's 108

Samen werken aan een klimaatbestendig, waterrobuust en waterbewust Rotterdam
Auteur Ted Veldkamp
ISBN 9789083481203
Verschijningsdatum november 2024
Aantal pagina's 112

Lerend leiderschap in de lerende school
Auteur Annemarie Neeleman
ISBN 9789493012493
Verschijningsdatum mei 2024
Aantal pagina's 112

Voor ieder kind een stevig taalhuis
Auteur Martine van der Pluijm
ISBN 9789493012486
Verschijningsdatum februari 2024
Aantal pagina's 100

In stapjes radicaal digitaal circulair worden
Auteur Ton Kollenburg
ISBN 9789493012462
Verschijningsdatum november 2023
Aantal pagina's 112

Bouwen aan motiverende leeromgevingen
Auteur Petra Poelmans
ISBN 9789493012479
Verschijningsdatum november 2023
Aantal pagina's 92

Samen onderzoekend werken aan onderwijskwaliteit

Auteur Jeroen S. Rozendaal
ISBN 9789493012455
Verschijningsdatum november 2023
Aantal pagina's 124

Artificial Intelligence voor Duurzamere en Efficiëntere Logistiek

Auteur Dr. Raymond Hoogendoorn
ISBN 9789493012431
Verschijningsdatum mei 2023
Aantal pagina's 76

Versterken van het curriculaire denken en werken binnen Hogeschool Rotterdam

Auteur Dominique Sluijsmans
ISBN 9789493012417
Verschijningsdatum januari 2023
Aantal pagina's 94

In verband met taal

Auteur Jacqueline van Kruiningen
ISBN 9789493012400
Verschijningsdatum januari 2023
Aantal pagina's 136

Zelfregulerend leren gaat niet vanzelf

Auteur Patrick Sins
ISBN 9789493012424
Verschijningsdatum januari 2023
Aantal pagina's 124

Ontwerpen en produceren voor waardebehoud in een circulaire economie

Auteur Marcel den Hollander
ISBN 9789493012363
Verschijningsdatum januari 2023
Aantal pagina's 89

Thierry Desot

ISBN: 9789083481234

Bias in grote taalmodellen voor AI-toepassingen in de gezondheidszorg

Een framework voor detectie en mitigatie van bias in deeplearning-transformermodellen ter ondersteuning van de gezondheidszorg

In de gezondheidszorg biedt AI enorme kansen, maar ook aanzienlijke uitdagingen. Deeplearningsystemen, zoals de modellen achter ChatGPT, functioneren vaak als een 'black box', waarbij de weg van input naar output ondoorzichtig blijft. Dit gebrek aan transparantie wekt begrijpelijkerwijs wantrouwen bij gebruikers, vooral in sectoren zoals de gezondheidszorg, waar vertrouwen en nauwkeurigheid van levensbelang zijn. In deze openbare les onderzoekt Thierry Desot hoe AI en Natural Language Processing (NLP) op een verantwoorde manier kunnen worden ingezet om de medische sector te ondersteunen, met speciale aandacht voor de risico's van onjuiste diagnoses en de ethische implicaties van bias in taalmodellen.

AI-technologieën hebben de potentie om de zorg te verbeteren, bijvoorbeeld door het vroegtijdig herkennen van ernstige aandoeningen zoals sepsis en de ziekte van Alzheimer of door het ondersteunen van de geestelijke gezondheidszorg en van professionals tijdens hun opleiding. Toch blijft het vertrouwen in AI klein, deels vanwege de onduidelijkheid over hoe AI-systemen beslissingen nemen. In deze openbare les wordt behandeld hoe AI op verantwoorde wijze kan worden ingezet om menselijke capaciteiten te versterken, zonder deze te vervangen. De focus ligt op het identificeren en verminderen van bias in taalmodellen, wat cruciaal is om eerlijke en inclusieve zorg te waarborgen.

Thierry Desot, lector Natural Language Processing (NLP) bij Kenniscentrum Creating 010, combineert zijn expertise in deep learning met een sterke focus op sociale impact. Zijn huidige onderzoek richt zich op het ethisch inzetten van AI in de gezondheidszorg, met name door bias in taalmodellen te identificeren en aan te pakken.

praktijkgericht **onderzoek**

Hogeschool Rotterdam Uitgeverij

HOGESCHOOL ROTTERDAM