

Concept Scriptieverslag

Afstudeeropdracht

*Artificial intelligence voor het ontwikkelen van chemische producten*

Dennis de Braal | Student nr. 120152177 | HBO Informatica | Inzendcode 264S3 | *13-06-2022*

Inhoud

[1. Inleiding 2](#_Toc102039096)

[1.1 Aanleiding 2](#_Toc102039097)

[1.1.1 Dow 2](#_Toc102039098)

[1.1.2 Industrie 4.0 5](#_Toc102039099)

[1.1.3 Data Analytics & Geavanceerde Algoritmes 7](#_Toc102039100)

[1.1.4 Chemische Industrie en Data Analytics & Geavanceerde Algoritmes 7](#_Toc102039101)

[1.2 Probleemformulering 8](#_Toc102039102)

[1.2.1 Probleemstelling 8](#_Toc102039103)

[1.2.2 Doelstelling 10](#_Toc102039104)

[1.2.3 Onderzoeksvraag 10](#_Toc102039105)

[1.3 Aanpak 12](#_Toc102039106)

[2. Onderzoeksontwerp 12](#_Toc102039107)

[2.1 Onderzoeksmodel 16](#_Toc102039108)

[2.2 Technisch ontwerp 16](#_Toc102039109)

[2.2.1 Dataverzameling 17](#_Toc102039110)

[2.2.2 Data-analyse 20](#_Toc102039111)

[3. Theorie 23](#_Toc102039112)

[3.1 Materials Science 24](#_Toc102039113)

[3.2 Paradigma’s van de Wetenschap 27](#_Toc102039114)

[3.3 Machine Learning Workflow 29](#_Toc102039115)

[3.3.1 Data Verzameling 31](#_Toc102039116)

[3.3.2 Data Representatie 32](#_Toc102039117)

[3.3.3 Model Building 34](#_Toc102039118)

[3.3.4 Model Evaluation 38](#_Toc102039119)

[3.4 Machine Learning in Materials Science 39](#_Toc102039120)

[3.5 Benodigdheden inzet ML-Model/ -Techniek 45](#_Toc102039121)

[4. Literatuurlijst 48](#_Toc102039122)

[5. Bijlagen 52](#_Toc102039123)

# 1. Inleiding

## 1.1 Aanleiding

De aanleiding van dit onderzoek komt voort uit de noodzaak om deel uit te maken van de ‘Fourth Industrial Revolution’ en het implementeren van wat het Industrie 4.0 framework genoemd kan worden.

Dow heeft zichzelf als doel gesteld om het meest innovatieve chemische bedrijf ter wereld te zijn. Om dit te kunnen bereiken zal Dow Industrie 4.0 en de daarbij behorende technologieën moeten omarmen.

Als eerste volgt een beknopte beschrijving van Dow, het bedrijf waarbinnen dit onderzoek wordt uitgevoerd, en de doelgroep van dit onderzoek. In de daaropvolgende paragrafen wordt verder ingegaan op het concept Industrie 4.0, hoe het ontstaan is en wat het inhoudt, en de invloed hiervan op de chemische industrie.

Binnen het Industrie 4.0 framework zijn er veel verschillende ondersteunende technologieën. De focus van dit onderzoek ligt op het gebruik van Artificial Intelligence en Machine Learning voor het ontwikkelen van nieuwe producten. Dow heeft recentelijk een eerste versie van een door AI-/ML-technologie ondersteunde applicatie uitgebracht (PU Predictive Intelligence), wat heeft geleid tot interesse vanuit andere business groups en functies in de organisatie.

### Dow

Dit onderzoek is uitgevoerd binnen verschillende afdelingen van Dow. Dow is een chemisch bedrijf dat wereldwijd opereert en vestigingen heeft in zo’n 160 landen. Dow heeft zo’n 50.000 werknemers in dienst, wat het één van de grootste chemische bedrijven ter wereld maakt.

De uitvoering van het onderzoek is gedaan vanuit de Dow-vestiging in Terneuzen. Deze locatie is de grootste productiefaciliteit van Dow buiten de Verenigde Staten. Daarnaast is Terneuzen een zogenaamde hub-locatie, waar de administratieve werkzaamheden voor alle vestigingen in Europa, het Midden-Oosten en India worden uitgevoerd. Op deze locatie zijn ongeveer 3.500 werknemers werkzaam.

Dow is opgedeeld in verschillende business groups, gebaseerd op de product portfolio. Elk van deze business groups produceert en verkoopt een specifieke groep producten:

* **Industrial Intermediates & Infrastructure** (Industrial Solutions and Polyurethanes & Construction Chemicals)
* **Performance Materials & Coatings** (Coatings & Performance Monomers and Consumer Solutions)
* **Packaging & Specialty Plastics** (Olefins, Aromatics & Alternatives and Feedstocks & Energy)

De ambitie van Dow is “om de meest innovatieve, klantgerichte, inclusieve en duurzame ‘materials sciences company’ van de wereld te worden”. Op de website van Dow valt te lezen dat men dit wil bereiken door het “leveren van waardegroei en best-in-class prestaties” (Dow.com).

Om deze doelstelling te kunnen bereiken, is het noodzakelijk dat Dow meebeweegt met de (r)evolutie die gaande is op het gebied van digitalisering en automatisering (Industrie 4.0). In de volgende paragrafen zal dieper worden ingegaan op het Industrie 4.0-concept.

Daarnaast is het van groot belang om moderne technologieën, zoals AI en ML, te ontwikkelen en te implementeren in de organisatie. De industrie in het algemeen en de concurrentie van Dow in het bijzonder, hebben deze technologie omarmd. Om ervoor te zorgen dat Dow haar huidige positie in de industrie en de markt kan behouden en mogelijk te versterken, is het noodzakelijk de nodige aandacht aan deze onderwerpen te schenken.

Een veelgehoorde klacht, die ook vanuit klantonderzoek naar voren is gekomen, is dat Dow veel tijd nodig heeft om te kunnen reageren op specifieke klantverzoeken. Het gaat hierbij vooral om verzoeken voor producten waarvan de formulering niet (volledig) bekend is.

Om die formulering te creëren moet er extra onderzoek gedaan worden, wat gepaard gaat met laboratoriumproeven. Dit onderzoek is een kostbaar en vooral tijdrovend proces. Tijdens de duur van deze onderzoeken, kan het gebeuren dat een klant uiteindelijk overstapt naar een concurrent. Dit verlies van klanten kan grote gevolgen hebben voor de organisatie. Dow is dan ook naarstig op zoek naar manieren om dit proces te versnellen. Niet alleen op het gebied van chemische ontwikkelingen, maar ook met behulp van moderne IT-technologieën.

De doelgroep van dit onderzoek bestaat uit meerdere afdelingen binnen de organisatie.

De voornaamste doelgroep van dit onderzoek zijn de besluitvormers op het strategisch niveau binnen de verschillende business groups. Het gaat hierbij om de Research & Development-, en Marketing-afdeling van elke business group. De personen op deze posities (Leaders, Managers en Directors) zijn continu bezig met het zoeken naar manieren om ‘slimmer’ en effectiever te kunnen werken en met klanten om te gaan. Iets wat vandaag de dag steeds meer wordt bereikt door het inzetten van innovatieve technologieën.

Met behulp van de uitkomsten van dit onderzoek kunnen zij bepalen:

* Of AI/ML een geschikte oplossing is in hun specifieke omgeving en de specifieke toepassing.
* Welke voorwaarden er gelden en wat er nodig is om AI-/ML-technologie te kunnen inzetten.

Daarnaast is ook de IT-afdeling van Dow (Information Systems/ IS) en in het bijzonder het Digital Operations Center (DOC), een doelgroep van dit onderzoek.

Het DOC is een onderdeel van Information Systems dat zich richt op het digitaliseren van de verschillende productieorganisaties, door het inzetten van ‘advanced manufacturing’ en opkomende technologieën. Het DOC is als het ware de digitale Research & Development-afdeling van de verschillende business groups.

De uitkomsten van dit onderzoek biedt het DOC handvatten voor het implementeren van AI/ML-technologie voor het ontwikkelen van chemische producten.

Deze handvatten kan het DOC gebruiken om in samenwerking met de business groups, specifieke applicaties te ontwikkelen om zo het gebruik van deze technologie op grotere schaal toe te passen.

Daarnaast zijn de uitkomsten mogelijk te gebruiken in andere functies binnen de organisatie (Supply Chain, Logistiek, etc.) om inzicht te krijgen in de mogelijkheden van AI/ML en wat erbij komt kijken om deze technologie te kunnen toepassen. Echter met de kanttekening dat het onderzoek specifiek gericht is op de ontwikkeling van nieuwe producten. Mogelijk is er extra onderzoek nodig om de uitkomsten te kunnen generaliseren naar andere functies.

### 1.1.2 Industrie 4.0

De ‘Fourth Industrial Revolution’, oftewel Industrie 4.0 of 4IR, zal leiden tot een verandering in hoe wij leven en werken. Althans, dat stelt Schwab in zijn artikel over de aanstaande vierde industriële revolutie (Schwab, 2015).

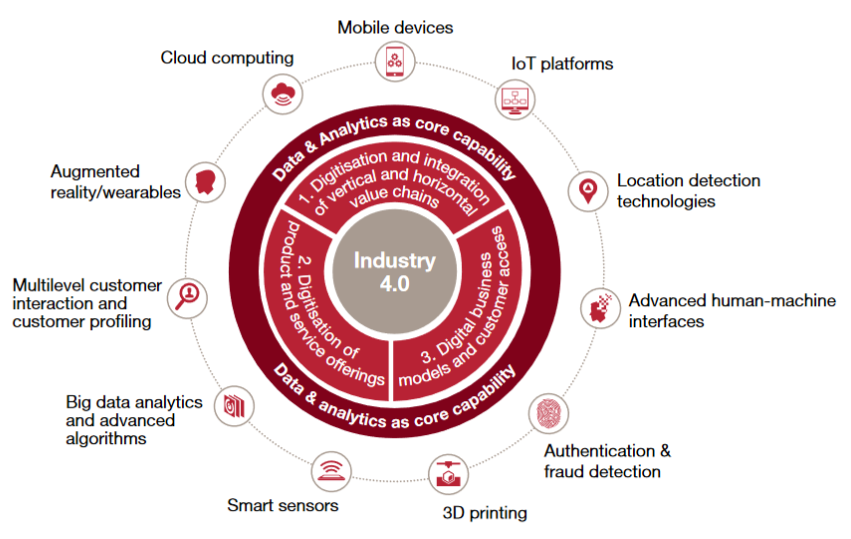
Ondanks dat er wel degelijk verschillen bestaan tussen de termen ‘Fourth Industrial Revolution’ en ‘Industrie 4.0’, worden deze door elkaar gebruikt in dit verslag. De essentie van beide concepten is het digitaliseren van de manufacturing- en automotive-industrie. Voor beide concepten gelden dezelfde technologieën, die als basis dienen voor de invulling van die digitalisering.

Na de First Industrial Revolution (water en stoom om de productieapparatuur aan te drijven), de Second Industrial Revolution (elektriciteit en massaproductie) en de Third Industrial Revolution (elektronica en informatietechnologie om productie te automatiseren), volgt nu dus de Fourth Industrial Revolution.

Volgens Schwab (2015) wordt er in deze vierde industriële revolutie voortgeborduurd op de derde revolutie met het verder digitaliseren van de verschillende industrieën. 4IR onderscheidt zich daarbij door het creëren van een fusie van verschillende technologieën, die de grens tussen de fysieke wereld en de digitale wereld doet vervagen.

Een ander belangrijk kenmerk van deze vierde revolutie is de snelheid waarmee deze zich ontwikkeld. In tegenstelling tot de voorgaande revoluties die zich op lineaire snelheid ontwikkelden, evolueert de vierde revolutie zich met exponentiële snelheid.

Geissbauer et all. (2016) vatten het concept, inclusief de ondersteunende digitale technologieën, samen in het volgende raamwerk:



Figuur 1: Industry 4.0 framework and contributing digital technologies. Overgenomen uit Industry 4.0: Building the Digital Enterprise, PwC, London, door Geissbauer, R., Vedso, J. & Schrauf, S., 2016, <https://www.pwc.com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016.pdf>

Het framework weergegeven in figuur 1 is een weergave van de technologieën, kerncapaciteiten en de digitaliserings-onderwerpen die tezamen Industry 4.0 vormen.

Het belang voor een organisatie om mee te kunnen gaan in deze industriële revolutie is groot. Zo stellen Bauernhansl et all. (2016) in het door Frauenhofer gepubliceerde rapport over de potentiële economische voordelen van Industrie 4.0, dat de implementatie van Industrie 4.0 kan leiden tot een verlaging van:

* Logistieke kosten (10-30%)
* Productiekosten (10-30%)
* Kwaliteitsmanagement gerelateerde kosten (10-20%)

Het voortbestaan van een organisatie kan in gevaar komen als men deze revolutie aan zich voorbij laat gaan.

### 1.1.3 Data Analytics & Geavanceerde Algoritmes

Uit een enquête uitgevoerd door PWC (Geissbauer et al., 2016) blijkt dat data analytics aan de basis staat voor Industrie 4.0. Dit gaat niet alleen over de besluitvormingsprocessen en de ondersteuning daarvan door (big) data, maar ook zeker over de ontwikkeling van nieuwe producten.

De voorbeelden die worden genoemd in de volgende paragraaf zijn gebaseerd op de mogelijkheid om geavanceerde analyses uit te voeren op de data, die in de loop der jaren is verzameld.

De volledige scope van het Industrie 4.0 framework is te breed om als onderwerp van het onderzoek te dienen. In samenspraak met de opdrachtgever is besloten om het onderzoek te richten op het onderwerp ‘Data Analytics en Geavanceerde Algoritmes’.

De keuze voor dit specifieke onderdeel van het framework, komt voort uit het initiatief dat Dow is gestart op dit gebied; AI en ML voor het produceren van PU-foams. Hierover meer in de volgende paragraaf. Dit onderzoek zal zich richten op een bredere toepassing van AI en ML voor het produceren van chemische producten.

### 1.1.4 Chemische Industrie en Data Analytics & Geavanceerde Algoritmes

De chemische industrie omarmt Industrie 4.0. Uit het onderzoek van PWC (Geissbauer et al., 2016) blijkt dat vooral deze vorm van industrie van plan is, om aanzienlijk te investeren in verdere digitalisering van de operationele processen en de implementatie van de aan Industrie 4.0 ondersteunende technologieën.

Dit komt ook tot uiting door verschillende initiatieven, die enkele van de grootste chemische bedrijven ter wereld, welke tevens concurrenten van Dow zijn, op dit moment aan het ontwikkelen zijn.

Zo heeft BASF een simulatie-tool ontwikkeld (Ultrasim) die gebruikmaakt van verschillende Industrie 4.0-technologieën om de ontwikkeling van nieuwe producten te kunnen versnellen.

Bayer maakt gebruikt van Machine Learning (data analytics en algoritmes) om de kwaliteit van batches al in een vroeg stadium in het productieproces te kunnen bepalen en op deze manier de efficiëntie van het proces te verhogen en afval te verminderen (QuaRTZ – Quality Release Time Zero).

Ook Dow maakt gebruik van Industrie 4.0-technologieën om het proces voor het creëren van nieuwe producten te verbeteren en te versnellen. Met behulp van Artificial Intelligence en Machine Learning kan de formulering van nieuwe PU-foams bepaald worden, wat moet leiden tot een kortere time-to-market en een reductie van de productiekosten.

## 1.2 Probleemformulering

### 1.2.1 Probleemstelling

Het ontwikkelen van nieuwe chemische producten is een tijdrovend en kostbaar proces. Door middel van trials en testen in het laboratorium wordt gezocht naar de juiste formulering om een product met de gewenste eigenschappen te ontwikkelen. Vanwege de lange duur van het proces, kan niet altijd tijdig aan de vraag van de klant worden voldaan. Het risico bestaat zelfs dat de klant haar business onderbrengt bij een concurrent.

Daarnaast is er de noodzaak om te blijven innoveren op het gebied van Research & Development en de digitalisering daarvan. Concurrenten van Dow zijn inmiddels al ver gevorderd in de digitale transformatie en Dow kan op dat gebied niet achterblijven.

Deze problematiek speelt binnen elke business group van Dow. De marketing-, IT- en R&D-functies van Dow en de business groups zijn actief bezig om een oplossing te vinden voor deze problemen.

De PU-business, onderdeel van de business groep Industrial Intermediates & Infrastructure, heeft een applicatie ontwikkeld die met behulp van AI/ML de formulering van het te ontwikkelen foam-product probeert te voorspellen.

Een klant kan in deze applicatie het type product aangeven met de bijbehorende eigenschappen die de klant wenst, en de applicatie gaat zoeken naar de chemische formulering om dat product te creëren.

Eerst zal de applicatie zoeken in een database met bekende formuleringen om te achterhalen of de formulering voor het gewenste product bekend is. Zo niet, dan zal een ML-algoritme op basis van door de klant aangegeven parameters (o.a. dichtheid van het product, omgevingstemperatuur, en vele andere), proberen een voorspelling te maken van de formulering om het product met die gewenste eigenschappen te kunnen creëren.

Het inzetten van deze toepassing van AI/ML leidt tot het sneller kunnen ontwikkelen van een formulering, met een aanzienlijke verlaging van het aantal experimenten en tests dat uitgevoerd dient te worden.

Eén van de grootste uitdagingen voor het succesvol voorspellen van de formulering, is het verzamelen en converteren van de data op basis waarvan het algoritme de voorspelling kan maken. Deze gegevens zijn veelal verspreid vastgelegd en zijn dan ook moeilijk te achterhalen. Tevens is er geen vaste structuur voor het vastleggen van die gegevens waardoor de conversie, waarbij de gegevens geschikt worden gemaakt om door een algoritme te worden gebruikt, bemoeilijkt wordt.

Daarnaast gaat het om vertrouwelijke gegevens, zogenaamde ‘Crown Jewel Data’. Deze gegevens vormen mede de basis van de concurrentiepositie van Dow en er dient dan ook voorzichtig mee om worden te gaan. Als deze gegevens in handen vallen van de concurrentie kan dit een impact hebben op Dow’s positie in de markt.

Dow heeft een strategie opgesteld (AI Strategy/ MLOps), waarin een route is uitgestippeld om te komen tot een infrastructuur, die het ontwikkelen van de AI/ML capaciteiten binnen Dow kan verbeteren en versnellen. Een ander onderdeel van die strategie gaat over het opzetten van een ‘data science workbench’ waarin geprobeerd wordt een platform te creëren, in een op het gebied van cyber security beveiligde omgeving, waarop data scientists kunnen werken met de crown jewel data.

De hiervoor beschreven toepassing is slecht één van de vele mogelijkheden die AI/ML kan bieden op het gebied van materials science. Daarnaast is de toepassing specifiek voor een bepaalde productsoort en business groep. De toepassing kan niet 1-op-1 worden ingezet binnen andere business groepen.

Het management van het Digital Operations Center en de verschillende business groups zijn op zoek naar verdere toepassingen van AI/ML op het gebied van het ontwikkelen van producten. Zij willen weten welke mogelijkheden AI/ML biedt en wat de benodigdheden zouden zijn om andere AI-/ML-toepassingen te kunnen inzetten in Dow.

Qua benodigdheden gaat het om de hoeveelheid data, die nodig is om een bepaalde AI-/ML-techniek te kunnen gebruiken en te gebruiken dataverzamelings- en dataconversie-methoden.

Daarnaast zal er ook aandacht moeten worden geschonken aan het veiligheidsaspect. Het product wordt gecreëerd op basis van een chemische reactie, waarbij er allerlei gevaren kunnen ontstaan als er in dat proces iets misgaat.

### 1.2.2 Doelstelling

Het doel van dit onderzoek is te achterhalen op welke wijze AI/ML ingezet kan worden, om het ontwikkelen van nieuwe producten in de chemische sector te bespoedigen. Het gaat hierbij om producten, die worden gecreëerd op basis van chemische formuleringen; dat wil zeggen een mengsel van verschillende chemische ingrediënten om zo een product met bepaalde eigenschappen te creëren. Het gaat vooral om producten voor verdere verwerking in andere producten. Hierbij te denken aan coatings voor o.a. verfproducten, plastics voor bijvoorbeeld de automotive-industrie, foams en nog vele andere producten en toepassingen.

Het gebruik van AI-/ML-technologieën voor het ontwikkelen en creëren van nieuwe producten zou moeten leiden tot de volgende voordelen:

* Het sneller kunnen ontwikkelen van nieuwe producten en sneller kunnen inspelen op de vraag van de klant.
* Het kunnen ontwikkelen van een grotere verscheidenheid aan producten.
* Mogelijkheid tot het simuleren van productieprocessen.

Dit onderzoek dient een overzicht op te leveren van AI-/ML-technieken en -methoden die binnen Dow zouden kunnen worden ingezet, en de daarbij behorende hardware-, software- en data-vereisten. Aan de hand van dit overzicht kan het management een besluit vormen over de manier waarop zij AI/ML willen toepassen binnen hun business groep.

### 1.2.3 Onderzoeksvraag

Om een oplossing te vinden voor de problematiek die geschetst is in de vorige paragraaf, is de volgende onderzoeksvraag opgesteld:

*Op welke wijze kan AI-/ML-technologie een bijdrage leveren aan het versnellen van de processen om nieuwe chemische producten te ontwikkelen, hierbij rekening houdend met de veiligheidsaspecten die van toepassing zijn bij deze processen?*

Deze onderzoeksvraag zal beantwoord worden aan de hand van de volgende deelvragen:

1. *Welke AI-/ML-technieken en -modellen worden er in de chemische sector gebruikt om het proces voor nieuwe chemische producten te ontwikkelen te ondersteunen?*
2. *Welke data-, hardware- en software-vereisten worden er gesteld om die technieken en modellen te kunnen gebruiken?*
3. *Welke veiligheidsrisico’s bestaan er bij het proces voor het creëren van nieuwe chemische producten? Welke relatie bestaat er tussen de geïdentificeerde risico’s en de inzet van AI/ML en hoe kunnen die risico’s worden gereduceerd?*
4. *Wat heeft Dow beschikbaar op het gebied van data, hardware en software, om de in de chemische sector gebruikte AI-/ML-technieken en -modellen, in de eigen organisatie te kunnen inzetten? Welke aanpassingen zijn er nodig in de bestaande IT-architectuur en -infrastructuur om AI/ML te kunnen inzetten binnen Dow, ter ondersteuning van de processen voor de ontwikkeling van nieuwe chemische producten?*

De relatie tussen de deelvragen kan als volgt worden uitgelegd.

Als bekend is voor welke toepassingen AI/ML ingezet kan worden in het ‘materials science’-werkveld, kan worden onderzocht welke specifieke AI-/ML-modellen en -algoritmes hiervoor gebruikt worden. Zodra dit inzichtelijk is gemaakt kan worden bepaald welke eisen er worden gesteld op het gebied van data, hardware en software. Op basis van deze eisen kan vervolgens worden onderzocht in welke mate de binnen Dow bestaande infrastructuren en architecturen m.b.t. data, hardware en software, toereikend zijn en of er eventueel aanpassingen gemaakt dienen te worden. Hierbij speelt het kunnen reduceren van eventuele risico’s op het gebied van de chemische processen een grote rol. Er dient immers gewaarborgd te worden dat de inzet van AI/ML niet leidt tot het vergroten van die risico’s.

## 1.3 Aanpak

In het vervolg van dit document, wordt ingegaan op de opzet van het onderzoek (hoofdstuk 2). Daarna volgt een beschrijving van de theoretische achtergrond van dit onderzoek, waarmee tevens een antwoord wordt gegeven op deelvragen 1 en 2.

# 2. Onderzoeksontwerp

Dit hoofdstuk bevat de beschrijving van de opzet van het onderzoek en de motivatie daarvoor. Er wordt eerst ingegaan op het type onderzoek dat uitgevoerd is en waarom. In de paragrafen daarna volgt het onderzoeksmodel en de details van de onderzoeksmethoden.

Binnen de basisprincipes van onderzoek, de zogenaamde methodologie, kan er onderscheid gemaakt worden tussen twee typen onderzoek:

* Fundamenteel onderzoek
* Praktijkgericht onderzoek

Het belangrijkste verschil tussen beide typen onderzoek, zit in het soort probleem dat met het onderzoek wordt geprobeerd op te lossen. Fundamenteel onderzoek gaat meestal om het ontwikkelen van kennis. Volgens Verhoeven (2018) is praktijkgericht onderzoek gericht op het vinden van oplossingen voor praktijkproblemen.

Uit de probleemstelling van dit onderzoek kan worden opgemaakt dat dit onderzoek een praktijkgericht onderzoek betreft.

Verhoeven (2018) maakt het onderscheid tussen kwantitatief en kwalitatief onderzoek. Kwantitatief onderzoek maakt gebruik van cijfermatige gegevens. Kenmerken over objecten, organisaties en personen worden met behulp van statistische technieken verwerkt. Saunders et al. (2019) geven aan dat bij dit type onderzoek centraal staat, dat het onderzoek is te generaliseren naar een grote populatie.

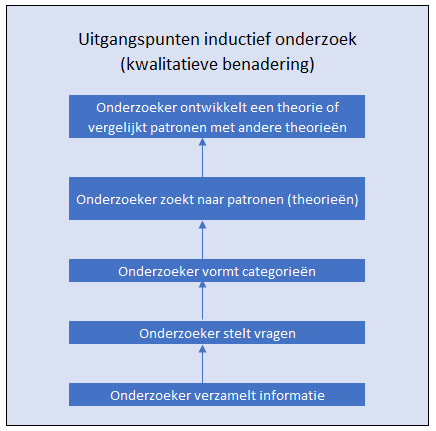
Verhoeven (2018) schrijft over kwalitatief onderzoek, dat dit wordt uitgevoerd met niet-cijfermatige gegevens. De gegevens worden verwerkt in alledaagse taal. Bij kwalitatief onderzoek gaat het vaak om inzicht krijgen in een probleem. De methoden voor het verzamelen van kwalitatieve data zijn flexibeler en gericht op het interpreteren van die data.

Het doel van dit onderzoek is het verkrijgen van inzicht in de manier waarop machine learning ingezet kan worden in het materials science werkveld. Hierbij speelt de context en de diepte van het onderzoek een grote rol. De kwalitatieve aanpak sluit hier het beste op aan.

De keuze voor een kwalitatieve aanpak komt ook voort uit de aard van het onderzoek. Het doel van het onderzoek is exploratief van aard. Er dient kennis gegenereerd te worden over de toepassing en inzet van machine learning in het materials science werkveld.

Vanwege deze inslag is dan ook gekozen voor een inductieve aanpak. Bij de inductieve methode wordt een theorie of model ontwikkelt op basis van het verzamelen van data. Dit in tegenstelling tot de inductieve methode, waarbij een theorie of model het uitgangspunt van het onderzoek is.

Figuur 2 is een weergave van de inductieve methode zoals beschreven door Saunders et. al (2019).



Figuur 2 - Uitgangspunten kwalitatief onderzoek. Overgenomen uit Methoden en technieken van onderzoek, Saunders, M. N. K., Lewis, P., Thornhill, A., Arnoldy, M., & Smitt, P. (2019). Methoden en technieken van onderzoek. Pearson

Naar aanleiding van de uitkomsten van de kritische literatuurstudie die is uitgevoerd in fase drie, is een model gecreëerd dat de basis heeft gevormd voor de verdere invulling en uitvoering van het onderzoek.

De uitvoering van dit praktijkgerichte, kwalitatieve onderzoek is gebaseerd op het onderzoeksproces zoals beschreven door Saunders et al. (2019) en is weergegeven in figuur 3:



Figuur 3 – Onderzoeksproces. Overgenomen uit Methoden en technieken van onderzoek, Saunders, M. N. K., Lewis, P., Thornhill, A., Arnoldy, M., & Smitt, P. (2019). Methoden en technieken van onderzoek. Pearson

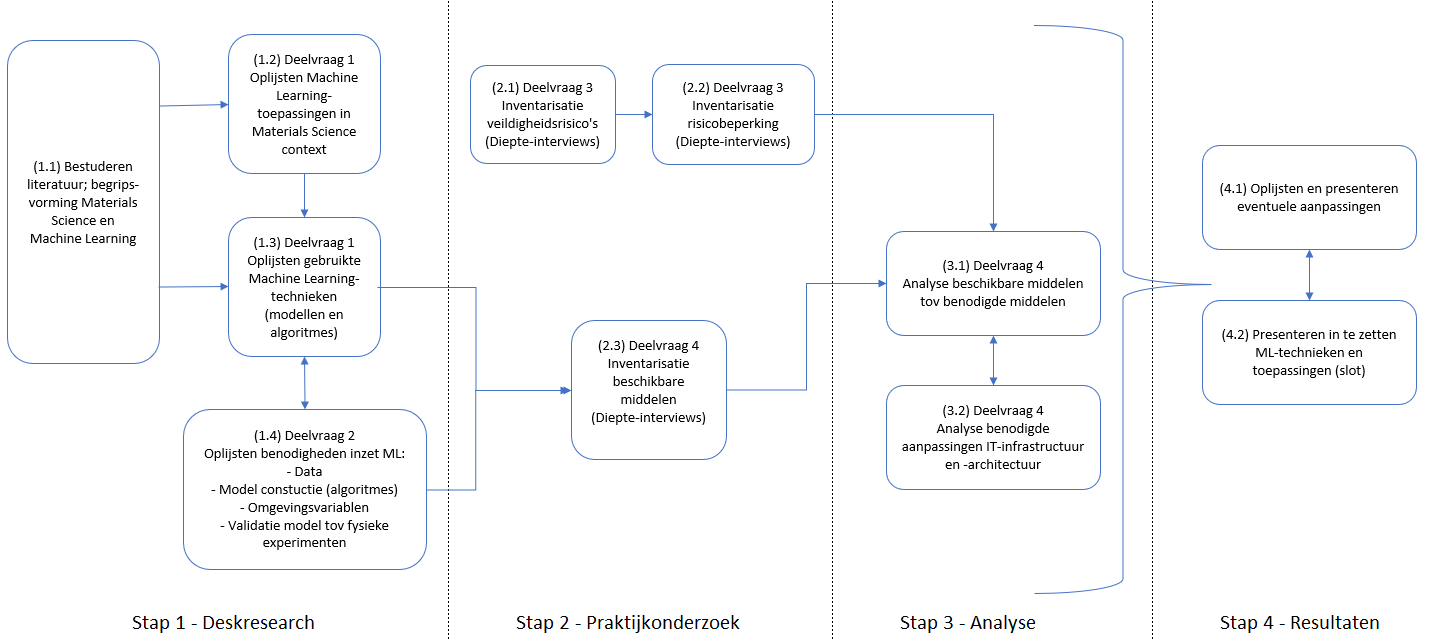
De eerste twee fasen ‘Oriëntatie op probleem en context’ en ‘Formuleren onderzoeksdoelstelling en vraagstelling’ zijn behandeld in het plan van aanpak. Dit hoofdstuk gaat in op de methodologische verantwoording en geeft een beschrijving van de aanpak van het onderzoek en de motivatie voor die aanpak.

De uitkomsten van fase drie, kritische literatuurstudie, zijn beschreven in hoofdstuk 3 en vormen het theoretisch kader van het onderzoek.

In paragraaf 2.1 wordt op basis van het onderzoeksmodel een globaal overzicht gegeven van de verschillende onderdelen van het onderzoek. Deze worden vervolgens uitgewerkt in paragraaf 2.2, waarbij verder wordt ingegaan op de details van de gebruikte onderzoeksmethoden.

## 2.1 Onderzoeksmodel

De uitvoering van dit onderzoek is schematisch weergegeven in figuur 4 – Onderzoeksopzet. Dit model is tevens opgenomen als bijlage bij dit document. Zie hiervoor bijlage 1.



Figuur 4 - Onderzoeksopzet

Het onderzoek is opgedeeld in 4 stappen, waarbij voor elke stap bepaalde activiteiten worden uitgevoerd. De nummering van de stappen geeft de volgorde van uitvoering aan.

In het volgende hoofdstuk wordt dieper ingegaan op de methoden voor dataverzameling en analyse, die voor dit onderzoek gebruikt zijn.

## 2.2 Technisch ontwerp

In deze paragraaf is beschreven welke methoden en technieken zullen worden gebruikt om de data te verzamelen en te analyseren. Er wordt eerst ingegaan op de methoden voor dataverzameling, en vervolgens komen de methoden aan bod om de verzamelde gegevens te analyseren.

### 2.2.1 Dataverzameling

De hoofdvraag is opgedeeld in vier deelvragen. Het antwoord op deelvragen 1 en 2 is gezocht door middel van een literatuuronderzoek. Voor beide deelvragen is een literatuuronderzoek uitgevoerd, waarbij de dataverzameling en analyse in samenhang met elkaar zijn uitgevoerd. Deelvragen 3 en 4 zullen worden beantwoord door het houden van interviews.

Doordat het literatuuronderzoek een groot deel van het totale onderzoek beslaat, is het belangrijk om dit literatuuronderzoek op gestructureerde wijze uit te voeren, en de uitvoering gedetailleerd vast te leggen.

#### 2.2.1.1 Literatuuronderzoek

Het doel van het literatuuronderzoek is het verschaffen van kennis en inzichten met betrekking tot het gebruik van machine learning in het materials science werkveld. De resultaten hebben geleid tot de structuur van het onderzoek, dat als basis dient voor het verdere vervolg van het onderzoek. Daarnaast heeft het literatuuronderzoek een antwoord gegeven op deelvragen 1 en 2.

De uitvoering van het literatuuronderzoek is gebaseerd op de methode Systematisch Literatuuronderzoek. Deze methode geeft een duidelijke structuur aan het onderzoek en verhoogt daarmee de betrouwbaarheid en herhaalbaarheid van het onderzoek. Systematisch literatuuronderzoek bestaat uit de volgende stappen:

* Probleemdefinitie;
* Zoekstrategie bepalen;
* Literatuur verzamelen;
* Waarderen en evalueren van de bronnen;
* Synthese.

Voor het doorlopen van de verschillende stappen van deze methode, is gebruikgemaakt van de checklist ‘Systematisch Literatuuronderzoek Protocol’ (Weber, 2011) [Weber].

Het gebruik van de checklist, waarin de aanpak en uitvoering van het literatuuronderzoek gedetailleerd zijn beschreven, komt de herhaalbaarheid, en daarmee de betrouwbaarheid, van het onderzoek ten goede.

Niet alle onderdelen van de checklist zijn uitgewerkt in het document zelf. Het onderzoeksgebied en beschrijving van het probleem zijn aan bod gekomen in hoofdstuk 1 en 2 van dit rapport.

De resultaten van de dataverzameling zijn vastgelegd in Excel wat verdere verwerking mogelijk maakt. Zie bijlage 3 – Werkblad Deelvraag 1, en bijlage 5 – Werkblad Deelvraag 2.

De voornaamste onderdelen van de checklist, met name op het gebied van betrouwbaarheid en herhaalbaarheid, zijn de hoofdstukken 6, 7, 10, 11 en 12.

* **Hoofdstuk 6:** de gehanteerde zoekstrategie met daarbij de geraadpleegde databases en gebruikte zoekwoorden.
* **Hoofdstuk 7:** de criteria op basis waarvan is besloten of een artikel in aanmerking is gekomen voor het onderzoek.
* **Hoofdstuk 10:** welke gegevens er uit de gevonden artikelen zijn onttrokken, en hoe deze verder gebruikt zijn.
* **Hoofdstuk 11:** gedetailleerd stappenplan van de uitgevoerde analyse. Dit hoofdstuk geldt tevens als een soort logboek voor de analyse fase.
* **Hoofstuk 12:** beschrijving van enkele belangrijke beslissingen die gedurende het literatuuronderzoek zijn genomen.

Het document Checklist Systematisch Literatuuronderzoek DV1 - DV2, biedt een volledig en gedetailleerd inzicht in de uitvoering van het literatuuronderzoek. Voor verdere details omtrent dit literatuuronderzoek, wordt dan ook verwezen naar bijlage 2.

#### 2.2.1.2 Interviews

Het praktijkgedeelte van dit onderzoek wordt uitgevoerd aan de hand van deelvragen 3 en 4. Deze deelvragen gaan in op de situatie bij Dow, waarbij geprobeerd wordt de theorie te verbinden met de praktijksituatie.

Het gebruik van AI/ML binnen de organisatie is een nieuw gebied. Er is al wat onderzoek uitgevoerd bij Dow, maar het is nog verre van ingeburgerd. Er zijn dan ook maar een beperkt aantal mensen beschikbaar, om een gedegen antwoord te kunnen geven op de onderzoeksvraag.

Daarnaast is de kennis van de onderzoeker op het gebied van AI/ML en materials science, te beperkt om een serie gesloten vragen op te stellen, om op basis daarvan een enquête of gestructureerde interviews te houden.

Er is dan ook gekozen voor het houden van diepte-interviews. De beperkte omvang van de populatie (hooguit 3 personen) wordt door Verhoeven (2018) ook als argument gegeven om te kiezen voor een diepte-interview.

Een ander belangrijk argument om te kiezen voor diepte-interviews, is dat de interviews een verkennend element bevatten (Saunders et al. 2019) [Sau]. Het doel is om de praktijksituatie te leren kennen, om zo een inschatting te kunnen maken van de mogelijke aanpassingen, die nodig zijn om AI/ML te kunnen inzetten. Door deze mate van verkenning, is een diepte-interview uitermate geschikt voor dit doel.

Saunders et al. (2019) beschrijven enkele uitdagingen met betrekking tot diepte-interviews, die een negatieve invloed kunnen hebben op de betrouwbaarheid en validiteit ervan [Sau]. De voornaamste uitdaging is het gebrek aan standaardisatie, en het probleem van bias.

Om het gebrek aan standaardisatie enigszins aan te pakken, zal er voor elk interview gebruik worden gemaakt van dezelfde topic-lijst. Deze lijst bevat de onderwerpen die zullen worden aangestipt tijdens het interview. De topic-lijst wordt opgesteld aan de hand van de te verzamelen gegevens (zie onderzoeksmodel in hoofdstuk 2.1 van dit document).

Het probleem van bias heeft te maken met interviewerbias en respondentenbias [Sau]. Interviewerbias wordt veroorzaakt door de manier waarop de interviewer de vragen stelt (verbaal en non-verbaal), wat een invloed kan hebben op de reacties van de geïnterviewde. Respondentenbias heeft verschillende oorzaken, en kan ertoe leiden dat een respondent een onvolledig beeld geeft van de situatie, of een sociaal wenselijk antwoord geeft. Een belangrijk middel om respondentenbias te voorkomen, is het zorgvuldig selecteren van de respondent [Sau].

Voorafgaand aan de diepte-interviews zal er een proef-interview worden gehouden met een Research Specialist van Dow. Dit proefinterview heeft als doel om te bevestigen dat het interview de gewenste resultaten zal opleveren. Daarnaast zal het proefinterview aantonen of er sprake is van interviewerbias.

Het selecteren van de respondenten zal gebeuren in samenspraak met een Research & Development Research Scientist van Dow. Deze persoon speelt ook een rol in het informeren en voorbereiden van de respondenten, om zo te proberen respondentenbias te voorkomen.

Om een antwoord te kunnen geven op de deelvragen drie en vier, is de input van experts in het veld nodig. Volgens Verhoeven is dit het belangrijkste argument, om te kiezen voor een selecte steekproef (Verhoeven, 2019). Voor de interviews zal gebruik worden gemaakt van een doelgerichte steekproef. Dit houdt in dat de respondenten dienen te voldoen aan bepaalde kenmerken (inclusiecriteria). In dit geval bestaan de inclusiecriteria uit de rol en werkzaamheden van de respondent binnen Dow. De respondenten dienen werkzaam te zijn in een chemische Research & Development afdeling, en betrokken zijn bij het onderzoeken en ontwikkelen van ML-modellen binnen hun werkgebied.  
Met behulp van de input van een Research & Development Research Scientist van Dow, is er een lijst opgesteld van personen die voor dit onderzoek zullen worden benaderd.

Het bepalen van de omvang van een selecte steekproef, is altijd een lastig punt. De grootte is afhankelijk van de onderzoeksvragen en het doel [Sau]. Voor dit onderzoek is echter de beschikbaarheid van experts de beperkende factor. Op dit moment zijn er vijf personen geïdentificeerd binnen Dow, die op basis van hun kennis, ervaring én rol binnen de organisatie, input kunnen leveren over het onderwerp. Het aantal respondenten voor de interviews is dan ook beperkt tot vijf.

Een nadeel bij een selecte steekproef, is dat deze een lage externe validiteit heeft. Dit houdt in dat de resultaten niet zomaar gegeneraliseerd kunnen worden naar de populatie. Dit nadeel wordt ten dele tenietgedaan, doordat alle personen waar Dow beschikking over heeft, zullen worden geïnterviewd. Echter zullen de resultaten beperkt blijven tot de organisatie zelf, en kunnen deze niet, of slechts ten dele generaliseerd worden.

### 2.2.2 Data-analyse

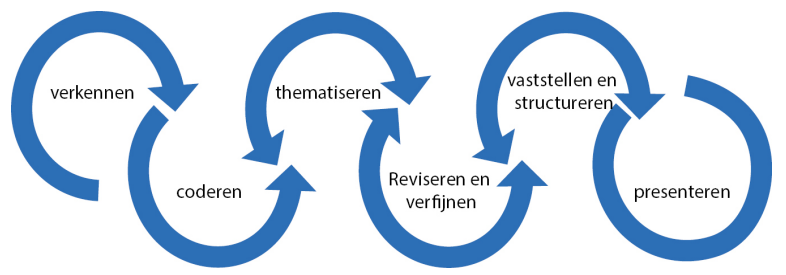
Deze paragraaf bevat een beschrijving van de data-analysemethoden die zijn gebruikt om de verzamelde data te ordenen en te analyseren.

#### 2.2.2.1 Analyse Literatuuronderzoek

Om de resultaten van de dataverzameling van het literatuuronderzoek op een ordelijke wijze te kunnen analyseren, is gebruikgemaakt van een variant van de methode ‘Thematische Analyse’.

Thematische analyse is een hulpmiddel om kwalitatieve data te identificeren, te analyseren en rapporteren. Met behulp van dit hulpmiddel kunnen gegevens worden geordend en geïnterpreteerd in het licht van de probleemstelling (Verhoeven, 2019) [Verhoev].

De methode bestaat uit een aantal stappen die zijn weergegeven in figuur 5.



Figuur 5 - Thematische analyse in zes stappen. Overgenomen uit Wat is onderzoek? Verhoeven, N. (2018). (6e editie). Boom Lemma.

De uitwerking van de analyse, met daarbij de gebruikte codering en structurering van de literatuur, is vastgelegd in bijlage 2 ‘Checklist Systematisch Literatuuronderzoek DV1 - DV2’.

Elke stap die is uitgevoerd in de analyse en verwerking van de resultaten, is gedetailleerd vastgelegd in hoofdstuk 11 van de checklist. Aan de hand van deze beschrijving, en de tussenproducten van de analyse, kan het onderzoek herhaald worden.

Daarnaast heeft het iteratieve karakter van de methode thematische analyse, een positieve invloed op de kwaliteit van het onderzoek. De kennis die wordt opgedaan tijdens de uitvoering van een iteratie, kan worden toegepast in een daaropvolgende iteratie.

Voor de resultaten omtrent de analyse van de literatuur, wordt verwezen naar hoofdstuk 3 van dit document.

#### 2.2.2.2 Analyse Interviews

De interviews zullen virtueel plaatsnemen, gebruikmakend van het softwareprogramma Microsoft Teams. Dit programma heeft de mogelijkheid om naast audio-0pnames, ook video op te nemen. Als de respondenten akkoord gaan, zullen zowel audio als video worden opgenomen.

Zoals aangegeven door Saunders et al. (2019) [Sau], is niet alleen de verbale reactie van belang, maar kan er ook veel informatie gehaald worden uit de non-verbale reacties. De video-opnamen kunnen hierbij een handig hulpmiddel zijn om de non-verbale reacties te identificeren en op de juiste manier te interpreteren.

Voor het verwerken en analyseren van de interviews, zullen de volgende stappen worden gevolgd:

* Transcriberen
* Open coderen
* Axiaal coderen
* Selectief coderen

***Transcriberen***

De interviews zullen woordelijk getranscribeerd worden. Dit houdt in dat alles wat gezegd wordt, zal worden vastgelegd. Aarzelingen, stopwoorden, etc. zullen worden genegeerd bij deze methode. Voor deze interviews is het minder van belang om vast te leggen *hoe* iets gezegd wordt. Van belang zijn de inhoudelijke tekst en informatie die door de respondenten wordt gegeven. De interpretatie van hoe de respondenten reageren, is van minder belang.

Voor het vastleggen van de transcripten, zal gebruik worden gemaakt van Word. Er zullen geen geautomatiseerde hulpmiddelen worden ingezet. De transcripten zullen als bijlage worden bijgevoegd.

De transcripten zullen ter controle naar de respondenten worden gestuurd. Het doel van deze controle, is verifiëren dat de respons op de juiste manier is geïnterpreteerd en vastgelegd. Dit verhoogt de betrouwbaarheid en validiteit van de interviews.

***Open coderen***

In deze stap zullen de tekstfragmenten worden voorzien van een codering. Het doel is voor ieder tekstfragment het hoofdthema aan te duiden middels een codering.

***Axiaal coderen***

Vervolgens zullen de toegekende coderingen worden vergeleken en samengevoegd waar logisch. Het uiteindelijke doel van deze stap, is het verkrijgen van enkele hoofdcategorieën. Deze hoofdcategorieën zijn gebaseerd op de deelvragen die middels de interviews, getracht worden te beantwoord.

Deze aanpak zorgt ervoor dat de analyse van de interviews, rechtstreeks gelinkt kan worden aan deelvragen.

***Selectief coderen***

In deze laatste stap van het coderingsproces, wordt de informatie van de hoofdcategorieën uit de vorige stap, bijeengebracht om zo relaties en verbindingen te leggen tussen de data. Hierna kunnen de resultaten worden vergeleken met elkaar en met de theorie, en kunnen conclusies worden geformuleerd.

De coderingen zullen worden toegekend met behulp van het softwareprogramma Atlas.ti.

# 3. Theorie

In dit hoofdstuk wordt de theoretische achtergrond van dit onderzoek gegeven. Er wordt eerst een introductie gegeven over het materials science werkveld, en hoe dit gebied is geëvolueerd tot een wetenschap die wordt gedreven door data.

Daarna volgt een beschrijving van hoe machine learning in dit werkveld wordt geïmplementeerd, waarbij ook ingegaan wordt op de werking van machine learning.

## 3.1 Materials Science

Materialen zijn overal om ons heen en hebben een grote invloed op hoe we als mens ons leven leiden, en de maatschappij zich als geheel ontwikkelt. Vanuit historisch perspectief is de vooruitgang van de mensheid sterk verbonden met haar vermogen om materialen te produceren en aan te passen. Deze relatie is zelfs zo sterk aanwezig, dat de eerste beschavingen worden genoemd naar de materiaalsoort die kenmerkend was voor die tijd; stenen tijdperk, bronzen tijdperk en ijzeren tijdperk.

In de hedendaagse tijd zijn de verschillende drankverpakkingen voor frisdranken, een mooi voorbeeld om de diversiteit aan materialen te illustreren. Zo zijn frisdranken te verkrijgen in blikjes (metaal), glazen flessen (keramiek) en plastic flessen (polymeren). Elk van deze verpakkingen heeft een set onderscheidende eigenschappen, wat de gebruikte materiaalsoort uitermate geschikt maakt voor het bewaren, en goedhouden van frisdrank.

Shackelford (2015) beschrijft materials science als een samenkomst van verschillende wetenschappen en technieken. Het is een multidisciplinair werkveld dat elementen bevat van o.a. metaalbewerking, keramische techniek, scheikunde en natuurkunde [Shack]. Askeland et al. (2010) voegen daar het doel van de materials science discipline aan toe. Namelijk het ontdekken van nieuwe materialen, en het verbeteren van bestaande materialen is Men tracht dit doel te bereiken door het onderzoeken en leren begrijpen van de structuur en samenstelling van materialen. [Ask]. Zowel Schackelford als Askeland et al. beschrijven Materials science aan de hand van een aantal componenten die zijn weergegeven in figuur 6.



Figuur 6 - Componenten van Materials Science. Overgenomen uit Jr., W. C. D., & Rethwisch, D. G. (2013). Materials Science and Engineering: An Introduction (9de editie). Wiley.

Deze opdeling naar de 4 componenten wordt gedeeld door en verder uitgeschreven in het werk van Callister en Rethwisch (2013) [Cal].

Processing oftewel verwerking slaat op de handelingen die nodig zijn om materialen te vervaardigen uit chemicaliën (middels een proces dat synthese wordt genoemd) en het fabriceren van materialen tot bruikbare objecten.

De structuur van een materiaal omvat de ordering van de interne componenten, en wordt onderverdeeld in 3 lagen: [Cal]

* **(Sub-)Atomische structuur**: de wijze waarop de atomen en moleculen zijn georganiseerd ten opzichte van elkaar. Sub-atomische structuur gaat over de interactie tussen een enkele atoom en de nuclei.
* **Microstructuur**: betreft de structuur van een materiaal dat kan worden waargenomen door middel van microscopische vergroting. De microstructuur bestaat uit een grote groep opeengestapelde atomen.
* **Macrostructuur**: deze laag gaat om de structurele elementen die met het blote oog kunnen worden waargenomen.

De structuur en de wijze van verwerking hebben een hele grote invloed op de eigenschappen van het materiaal. Door middel van het aanpassen van de structuur of het verwerkingsproces, kan er invloed uitgeoefend worden op de uiteindelijke eigenschappen van het materiaal.

De belangrijkste eigenschappen van een vast materiaal kunnen worden ingedeeld naar de volgende categorieën:

* Mechanische eigenschappen.
* Elektrische eigenschappen.
* Thermische eigenschappen.
* Magnetische eigenschappen.
* Optische eigenschappen.
* Chemische eigenschappen.

De prestaties van een materiaal (performance) kunnen worden gedefinieerd als de functie van de verschillende eigenschappen die het materiaal bezit. Eigenschappen en prestaties zijn sterk aan elkaar verbonden. Het zijn immers de eigenschappen van een materiaal, die de prestaties ervan bepalen.

Een andere belangrijke indeling die in materials science wordt gemaakt, is die van het soort materiaal. Voor vaste materialen wordt uitgegaan van vijf soorten. Deze indeling is gemaakt op basis van de chemische samenstelling en atomische structuur van de verschillende materialen binnen elke soort.

Askeland et al. (2010) geven hiervoor de meest volledige beschrijving:

* **Metalen**; materialen die bestaan uit één of meer metalen elementen (bijvoorbeeld ijzer, koper, aluminium, goud, etc.), en bevat vaak een toevoeging van een kleine hoeveelheid niet-metalen elementen. Deze soort materiaal is terug te vinden in objecten als zilverwaren, munten, sieraden, enz.
* **Keramieken**; een chemische verbinding (compound) tussen metaal- en niet-metaal-elementen. Enkele bekende voorbeelden van keramische materialen zijn onder andere glas, porselein en baksteen.
* **Polymeren**; de materialen in deze soort zijn overwegend organische materialen, die worden geproduceerd door een proces dat polymerisatie wordt genoemd. Polymerisatie voegt enkelvoudige moleculen (monomeren) samen tot een keten van zichzelf herhalende moleculen (polymeren). Polymeren zijn terug te vinden in onder andere rubber, allerlei soorten plastics en vele andere toepassingen, van kogelvrije vesten, Liquid Crystal Displays’s (LCD’s) tot kleren en koffiemokken.
* **Composieten**; een composiet bestaat uit een samenvoeging van twee of meer individuele materialen uit de hiervoor genoemde categorieën. Het idee achter het ontwikkelen van een composiet, is het mixen van de eigenschappen van de individuele materialen. De bekendste toepassingen van composiet zijn glasvezel en met koolstofvezel (carbon fiber) versterkte polymeren die worden gebruikt bij het bouwen van vliegtuigen.
* **Geavanceerde materialen**; deze soort bestaat uit onder andere halfgeleiders, biomaterialen en nano-materialen. Deze materialen worden gebruikt in high-tech toepassingen zoals computers, ruimtevaart en luchtvaart. In veel gevallen worden de geavanceerde materialen ingezet om de eigenschappen van bestaande materialen en producten sterk te verbeteren.

Callister en Rethwisch (2013) benoemen alleen halfgeleiders als geavanceerd materiaal en gaan verder niet in op de door Askeland genoemde additionele materialen. Ook Shackelford (2015) beperkt zich ook alleen tot het benoemen van halfgeleiders als geavanceerd materiaal.

## 3.2 Paradigma’s van de Wetenschap

Het materials science-werkveld is in de loop der jaren geëvolueerd tot een wetenschap die wordt gedreven door data. Lange tijd was deze wetenschap puur empirisch, gebaseerd op experimenteren en natuurlijke verschijnselen, en overeenkomend met het stenen en metalen tijdperk (brons, ijzer en staal). In de eeuwen daarna kwam het volgende paradigma op, gebaseerd op de wetten van de fysica en theoretische modellen en generalisatie. Dit tweede paradigma was tot de jaren ’50 van de twintigste eeuw de voornaamste vorm in de wetenschap.

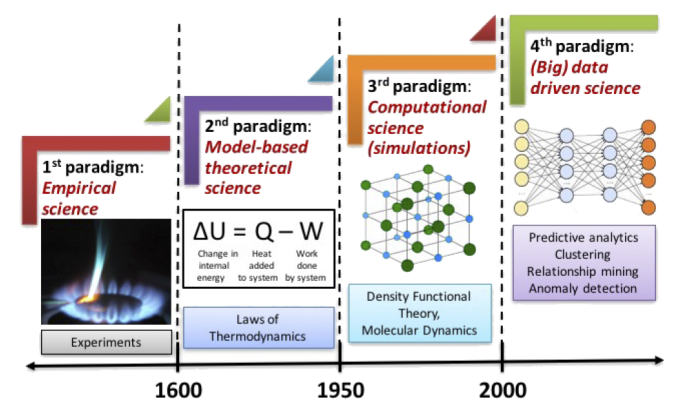
Door de opkomst van computers werd het derde paradigma, ‘computational science’, populair. Door de toegenomen rekenkracht werd het mogelijk om zeer complexe modellen te analyseren en op te lossen.

Schleder et al. (2019) geven in hun review aan, dat door het gebruik van computationele methoden een gigantische hoeveelheid data is gegenereerd. Ze stellen dat het vooral het gebruik van density functional theory (DFT) hierin een onderscheidende rol heeft gespeeld.

Agrawal en Choudhary (2016) benoemen ook molecular dynamics (MD) als één van de voornaamste methoden, die hebben geleid tot de enorme hoeveelheden data. MD wordt ook door Das et al. (2020) aangestipt in hun onderzoek naar de ontwikkeling van machine learning in het materials science werkveld.

Het toepassen van deze paradigma’s in de verschillende takken van de wetenschap, theorieën, experimenten en berekeningen, hebben geleid tot het genereren van een gigantische hoeveelheid data. Deze hoeveelheid data heeft de wetenschap verder doen evolueren naar het vierde paradigma; ‘(big) data driven science’.

Figuur 7 geeft een overzicht van de evolutie van de wetenschap en de vier genoemde paradigma’s.



Figuur 7 - De 4 Paradigma's van de wetenschap; Empirisch, Theoretisch, Computationeel en Data-gedreven. Overgenomen uit Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science door Agrawal, A., & Choudhary, A. (2016). APL Materials, 4(5), 053208. https://doi.org/10.1063/1.4946894

De hoeveelheid data die voortkomt uit het vormen van theorieën, uitvoeren van experimenten en simulaties, kan worden getypeerd als ‘big data’.

Big data kan worden beschreven aan de hand van de 5V’s van big data. De 5V’s worden door zowel Schleder et al. (2019), en Agrawal en Choudhary (2016) benoemd en beschreven in hun onderzoeken:

* Volume; de hoeveelheid data, de grootte van de dataset.
* Velocity; de snelheid waarmee de data gegenereerd wordt.
* Variety; de verschillende datatypen (zowel gestructureerd als ongestructureerd).
* Veracity; de accuraatheid en betrouwbaarheid van de data.
* Value; het vermogen om bruikbare informatie te kunnen afleiden uit de data.

Eén van de grootste uitdagingen met betrekking tot big data is het kunnen analyseren van die data. Volgens onderzoek van Kitchin (2014), heeft de opkomst van artificial intelligence en machine learning, een grote invloed op de mate waarop de data succesvol kan worden geanalyseerd, en er bruikbare informatie uit te kunnen onttrekken.

## 3.3 Machine Learning Workflow

Om te kunnen bepalen wat er nodig is om de in de vorige paragraaf gepresenteerde ML-technieken te kunnen inzetten (zie deelvraag 2 van dit onderzoek), is het van belang om inzicht te krijgen hoe deze technieken en modellen zijn geïmplementeerd.

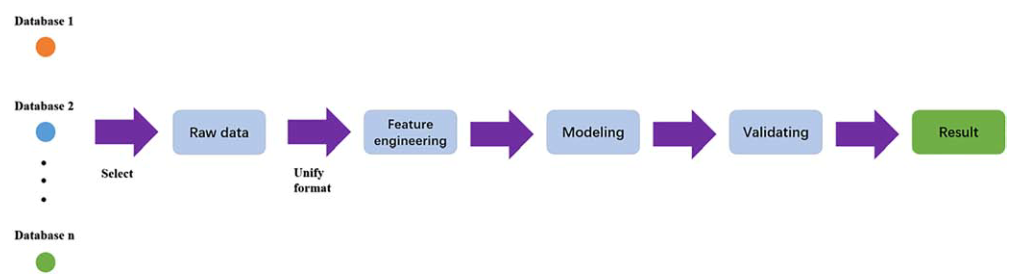
In de volgende paragrafen wordt ingegaan op het proces voor het ontwikkelen, trainen en valideren van een ML-model (de zogenaamde ML-workflow). Uit deze beschrijving kan in algemene termen worden opgemaakt, wat benodigd is om een ML-model in de praktijk te kunnen inzetten.

Op basis van deze informatie is bepaald welke gegevens er verzameld moesten worden, om een antwoord te kunnen geven op deelvraag 2. De resultaten van deze dataverzameling worden gepresenteerd in paragraaf 3.5.

Das et al. (2020) onderscheiden drie stappen in het proces voor het ontwikkelen van een ML-model:

* Dataverzameling en -representatie.
* Selecteren en valideren van het (ML-)model.
* Optimaliseren van het (ML-)model.

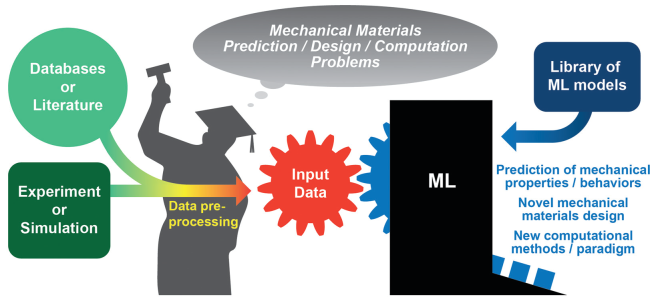
In hun onderzoek naar het gebruik van machine learning in materials science, geven Wei et al. (2019) een vergelijkbare indeling. Echter, nemen zij de stap optimaliseren van het model, niet op in hun overzicht. De stappen ‘modeling’ en ‘validating’ staan voor het selecteren en ontwikkelen van het model, en de evaluatie daarvan.



Figuur 8 - Machine Learning Workflow. Overgenomen uit Machine learning in materials science door Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., & Lei, M. (2019). InfoMat, 1(3), 338–358. https://doi.org/10.1002/inf2.12028

Deze indeling van de ML-workflow wordt gedeeld door Guo et al. (2021) in hun onderzoek naar het gebruik van ML voor het ontwerp van mechanische materialen. Hierbij hebben ze de workflow aangepast specifiek voor het toepassen in het material science domein, zie figuur 9. Zij stellen dat de voornaamste componenten van dit proces, bestaan uit:

* Een goed georganiseerde dataset met materiaal-data, achterhaald uit literatuur en bestaande databases, of verkregen door het uitvoeren van experimenten en simulaties.
* Een ML-model dat in staat is om de data-representatie van materialen en moleculen te kunnen ontleden.
* Een duidelijk gedefinieerd probleem dat door het ML-model dient te worden opgelost.



Figuur 9 - Schematische weergave van de ML-workflow in Materials Science. Overgenomen uit Artificial intelligence and machine learning in design of mechanical materials door Guo, K., Yang, Z., Yu, C. H., & Buehler, M. J. (2021). Materials Horizons, 8(4), 1153–1172. https://doi.org/10.1039/d0mh01451f

In de volgende paragrafen wordt verder ingegaan op de verschillende stappen in dit proces.

### 3.3.1 Data Verzameling

De eerste stap in het proces is het verzamelen van data. Vanuit een machine learning-oogpunt kan de data ingedeeld worden in twee categorieën:

* Training data; te gebruiken om het model te trainen/laten leren.
* Test data; te gebruiken om het getrainde model te testen met data van buiten de training set.

Wei et al. (2019) onderscheidden vier categorieën van data die van belang zijn op het gebied van materials science:

* Materiaaleigenschappen (fysieke eigenschappen of gegevens over de structuur van het materiaal).
* Data over de chemische reactie van grondstoffen.
* Grafische data of afbeeldingen (bijvoorbeeld een afbeelding van de oppervlakte van een materiaal).
* Data uit literatuur (tekstuele data).

Deze vier categorieën worden ook onderkend door Xi (2021) in het onderzoek naar het ontwerpen van nano-materialen met behulp van machine learning. Ook in andere onderzoeken, zoals in onder andere die van Sha et al. (2020), Guo et al. (2021) en Das et al. (2020), wordt gerefereerd aan deze vier categorieën.

Er lijkt consensus te bestaan over de verschillende bronnen waaruit deze data kan worden onttrokken. Zo worden in de onderzoeken van zowel Sha et al. (2020), Guo et al. (2021), Das et al. (2020) als die van Wei et al. (2019, de volgende bronnen onderkend:

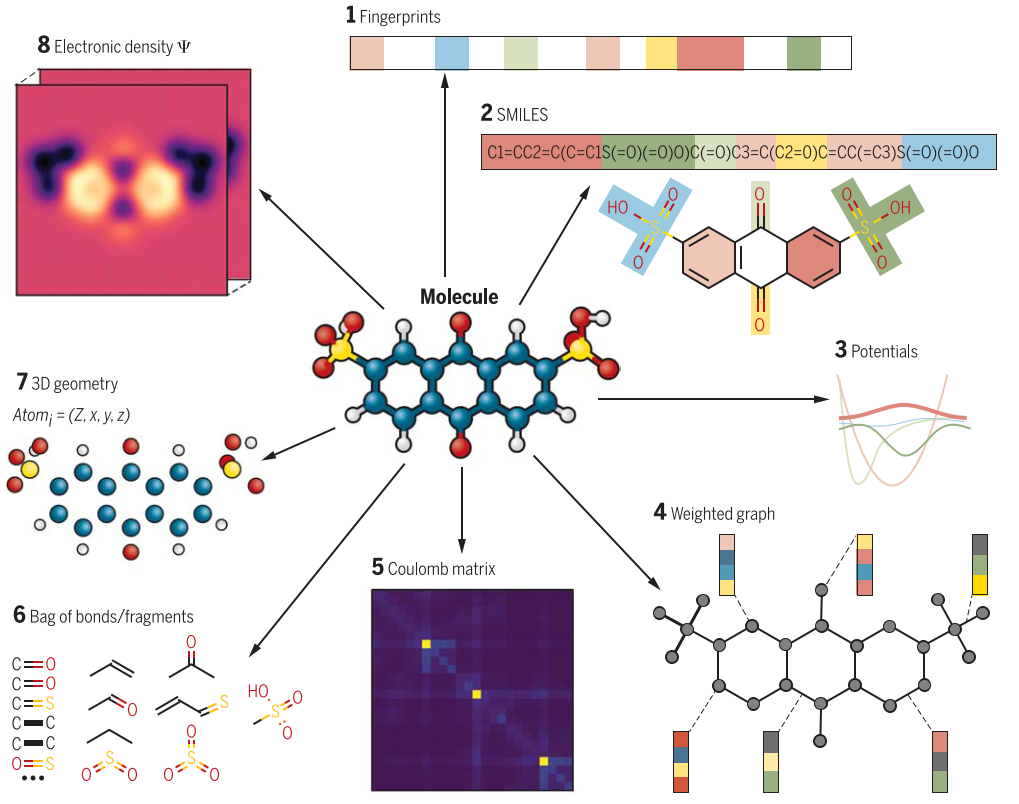
* Data voortgekomen uit experimenten, en simulaties (zoals DFT en MD).
* Online-databases met gegevens over materialen (zoals AFLOW, MATDAT, Material Project, en anderen).
* Literatuur; met behulp van text-mining technieken kunnen grote aantallen tekstuele bronnen worden doorzocht op bruikbare gegevens.

### 3.3.2 Data Representatie

Deze stap in het proces wordt ook wel ‘Feature Engineering’ genoemd. Feature Engineering heeft als doel om kenmerken (features) te onttrekken uit de ruwe data, om het ML-algoritme te kunnen laten werken, zo stellen Wei et al. (2019).

Schleder et al. (2019) gaan zelfs nog iets verder door te stellen dat de juiste representatie van een materiaal, cruciaal is voor de prestaties van het ML-model. Ook Butler et al. (2018) geven in hun onderzoek aan, dat de juiste vorm waarin de data aan het model wordt gepresenteerd, een grote invloed heeft op het trainen van het model; hoe beter de datarepresentatie geschikt is, hoe accurater de voorspelling van het model.

Er zijn verschillende manieren om een materiaal, of een molecuul weer te geven op een manier die geschikt is voor het trainen van een ML-model. Wei et al. (2019) geven het meest complete overzicht, dat is weergegeven in figuur 10.



Figuur 10 - Weergave van een molecuul. Overgenomen uit Machine learning in materials science door Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., & Lei, M. (2019). InfoMat, 1(3), 338–358. <https://doi.org/10.1002/inf2.12028>. Copyright 2018, The American Association for the Advancement of Science

Voor de traditionele ML-algoritmes is het vaak nodig om de kenmerken manueel te selecteren en te verwerken. Dit is echter een arbeidsintensief proces dat veel tijd kost, en waarvoor een grote mate van domeinkennis vereist is. Wei et al. (2019) stellen in hun onderzoek, dat het gebruik van deep learning-algoritmes voor het selecteren van de kenmerken, bezig is aan een opmars, en dat dit wel eens de trend kan worden op dit gebied. Dit is echter een recente ontwikkeling, die in andere onderzoeken niet direct wordt aangestipt.

Morgan en Jacobs (2020) geven in hun onderzoek enkele voorbeelden van deep learning-modellen, die veelbelovende prestaties laten zien vergelijkbaar met, of zelfs beter dan manuele methoden, zijn MEGNet, SchNet, en CGCNN.

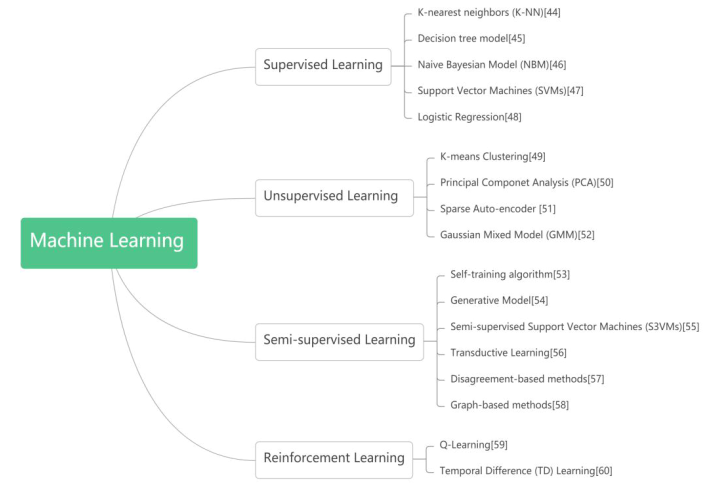
Zie de volgende paragraaf voor een korte introductie over enkele ML-algoritmes.

### 3.3.3 Model Building

De volgende stap in het proces, is het creëren van een ML-model. Belangrijk om op te merken, is het verschil tussen een algoritme en een model. Deze twee termen worden vaak door elkaar gebruikt, wat enigszins verwarrend kan zijn.

Een algoritme is de serie stappen die worden uitgevoerd op de dataset (training data), om een model te genereren. Het model is een vertegenwoordiging van die data, en kan worden gezien als een softwareprogramma, waarin data van buiten de training set kan worden ingevoerd. De output van dat model is een voorspelling over de, voor het model, nog niet eerder geziene data.

Er zijn verschillende categorieën ML-algoritmes om een model te bouwen. Een algemeen bekend onderscheid is dat tussen ‘supervised learning’, ‘unsupervised learning’, ‘semi-supervised learning’ en ‘reinforcement learning’. Xi (2021) heeft de veelgebruikte algoritmes samengevat in een overzicht, zie figuur 11.



Figuur 11 - Indeling en bijbehorende algoritmes per soort model. Overgenomen uit Functional Nanomaterials Design in the Workflow of Building Machine-Learning Models door Xi, Z. (2021). arXiv e-prints, arXiv-2108.

Naast deze vormen van ML, zijn er nog enkele typen leren te onderscheiden, waarvan in de vorige paragraaf is aangetoond dat deze ook gebruikt worden binnen het materials science domein. Zo beschrijft Settles (2009) in zijn survey de methode ‘Active Learning’. Zhu et al. (2021) geven in hun onderzoek een uitgebreide beschrijving van ‘Transfer Learning’, en Al-Sahaf et al. (2019) gaan dieper in op de methode ‘Evolutionary ML’:

* **Active Learning:** Om een alternatief te bieden voor het trainen op basis van (grote hoeveelheden) gelabelde data, bestaat er Active Learning. Deze vorm van ML, betrekt een zogenaamde ‘oracle’ (vaak in de vorm van een menselijke computergebruiker) om een label te voorzien voor niet-gelabelde trainingsdata. Denk bijvoorbeeld aan een captcha waarbij een afbeelding wordt getoond, en een gebruiker dient plaatjes aan te klikken die een bepaald voorwerp bevatten. De resultaten hiervan worden gebruikt om een ML-model te trainen (Settles, 2009).
* **Transfer Learning:** Methode die is gebaseerd op het principe dat een bestaand, getraind ML-model, ingezet kan worden in een nabij liggend domein. Binnen transfer learning wordt er onderscheid gemaakt tussen homogene en heterogene transfer. Bij homogene transfer delen het bron-domein en het doel-domein, dezelfde input features. Bij heterogene transfer dient er ook een aanpassing gemaakt te worden in de feature set (Zhu et al., 2021).
* **Evolutionary ML:** Gebaseerd op het principe van natuurlijke selectie. De algoritmes die voor deze soort ML zijn ontwikkeld, vinden hun oorsprong in de natuur. Zo is er bijvoorbeeld het Artificial Bee Colony algoritme, Particle Swarm Optimisation algoritme, en algoritmes die gebruikmaken van Genetic Operators. De evolutionaire algoritmes kunnen worden ingezet voor classificatie-, regressie- en clustering-problemen (Al-Sahaf et al., 2019).

Een ander onderscheid dat gemaakt kan worden tussen de verschillende algoritmes, is dat tussen regressie- en classificatie-algoritmes.

Regressie-algoritmes gaan op zoek naar de relatie tussen input X en output Y, op basis van gelabelde trainingsdata. Voorbeelden van dit soort algoritmes zijn Linear Regression, Logistic Regression en Support Vector Regression.

Guo et al. (2021) geven in hun onderzoek naar het gebruik van machine learning voor mechanische materialen, een inleiding naar regressie-algoritmes. Schleder et al. (2019) geven in hun artikel over computationele aanpakken voor materials science, een uitgebreide beschrijving van verschillende regressie-algoritmes, inclusief de werking van de algortimes op basis van haar functies.

In datzelfde artikel van Schleder et al. (2019) wordt ook ingegaan op classificatie algoritmes. Deze soort algoritmes trachten de dataset in te delen naar verschillende (al dan niet vooraf gedefinieerde) categorieën. Voorbeelden hiervan zijn onder andere Logistics Regression en Support Vector Machines.

Guo et al. (2021) benoemen ook nog enkele ML-algoritmes die in staat zijn de regressie- en classificatie-taak te combineren, zoals Decision Tree en Gradient Boosting. Echter wordt hier niet verder op ingegaan.

Bernhard Melig beschrijft in zijn werk ‘Machine learning with neural networks‘ (2021) een ander soort algoritme; namelijk dat van Artificial Neural Networks (ANN). Dit soort algoritme is gebaseerd op de werking van het menselijk brein. Een ANN heeft een input layer en een output layer, met daartussen één of meer hidden layers. Deze hidden layers bestaan uit activeringsfuncties (zoals een sigmoid function) die de input een bepaald gewicht geven en een output genereren voor de volgende laag in het netwerk. Als een netwerk bestaat uit meerdere hidden layers, spreekt men van een Deep Neural Network.

Mehlig geeft aan dat een ANN kan worden ingezet voor zowel classificatie-, als regressie-problemen, en op basis van zowel supervised en unsupervised learning getraind kan worden.

In hun onderzoek naar het gebruik van machine learning voor het modelleren van materialen, stellen Das et al. (2020) dat een ANN één van best werkende algoritmes voor Machine Learning is. Ze plaatsen daarbij wel de kanttekening, dat er veel data nodig is om het algoritme te trainen. Daarnaast geven zij aan dat een ANN ook rekenkundig intensief is, wat de nodige middelen vraagt. Er worden in het onderzoek echter ook enkele succesvolle initiatieven gegeven, van netwerken die ook goed presteren met kleinere datasets. Eén van de voorbeelden hiervan is Elemnet, dat beter presteert dan conventionele ML-modellen, op basis van een dataset van slechts 4000 verbindingen (uit honderdduizenden mogelijkheden).

Een groot voordeel van een ANN is volgens Das et al. (2020), dat deze in staat is om features te achterhalen uit de dataset. Dit neemt de noodzaak weg van de complexe taak om features te identificeren en ontwikkelen.

De beschikbare data bepaalt welke soort leren, geschikt is voor het creëren van het model. Als er de beschikking is over gelabelde data (d.w.z. van de input features is bekend wat de waarde van de output moet zijn), kan er gebruik worden gemaakt van ‘supervised learning’. Het model wordt getraind om de patronen te herkennen, die bepalen wat de output waarde is voor een gegeven input. Bij dit soort leren gaat het in het algemeen over regressie- of classificatie-problemen.

Indien de data niet gelabeld is, zal ‘unsupervised learning’ gebruikt moeten worden. Deze vorm van machine learning zoekt naar patronen en onderliggende relaties in de data. Deze patronen en relaties vormen het model en worden ingezet op nieuwe data, om daar voorspellingen voor te kunnen maken.

Ook voor Active Learning geldt dat de trainingsdata bestaat uit niet-gelabelde gegevens. Deze labels worden immers door een ‘oracle’ voorzien.

Voor transfer learning zijn er drie mogelijkheden met betrekking tot de trainingsdata en het al dan niet hebben van een label:

* **Transductive:** Alleen het bron-domein bevat gelabelde data.
* **Inductive:** Als ook het doel-domein gelabelde data bevat en kan inbrengen.
* **Unsupervised:** Er is geen gelabelde data beschikbaar voor zowel bron- als doel-domein.

Reinforcement learning wordt gezien als een derde ML-paradigma, naast supervised learning en unsupervised learning. Bij reinforcement learning wordt het model getraind door middel van het maximaleren van een numerieke beloning. Het model leert welke uitkomsten de hoogste beloning genereren. Er is dan ook geen sprake van al dan niet gelabelde data als input voor het trainen.

Evolutionary ML-models daarentegen, kunnen worden getraind met zowel gelabelde, als niet-gelabelde data.

Elk van de genoemde typen ML bevat een veelvoud aan algoritmes, die kunnen worden gebruikt om het model te creëren. Het is onmogelijk om binnen redelijke tijd, de details van de ML-algoritmes, op begrijpelijke wijze te beschrijven. Het betreft immers een veelvoud aan algoritmes, elk met hun eigen specifieke werking. Tevens is diepgaande kennis benodigd, om de details te begrijpen. Voor verdere informatie wordt verwezen naar het werk van Alpaydin (2016), Shalev-Schwartz (2014) en Kubat (2015).

### 3.3.4 Model Evaluation

De laatste stap in het proces is het valideren van het getrainde ML-model. Met behulp van een test- of validatie-dataset worden de prestaties en nauwkeurigheid van het model bepaald. Butler et al. (2018) stellen dat er drie bronnen van fouten bestaan, waar rekening mee moet worden gehouden:

* Model bias; het model bevat verkeerd aangeleerde aannames waardoor de juiste onderliggende relaties in de data niet worden gevonden.
* Model variance; het model levert verschillende outputs afhankelijk van welke training-dataset, of deel van de training-dataset wordt ingevoerd.
* Onherleidbare fouten.

Als een model niet goed presteert komt dit vaak door een hoge bias (underfitting) of een hoge variance (overfitting).

Zowel Butler et al. (2018) als Das et al. (2020) geven dezelfde uitleg over under- en overfitting.

Underfitting geeft vaak aan dat er onvoldoende kwalitatief goede data is gebruikt voor het trainen van het model, wat ervoor zorgt dat het model niet in staat is om de relatie in de data te kunnen vinden.  
Overfitting houdt in dat het model te complex is geworden, vaak voortgekomen uit een grote toename van het aantal parameters.

Das et al. (2020) geven verschillende meeteenheden om de prestatie en nauwkeurigheid van het model in uit te drukken:

* Loss function.
* Mean absolute error.
* Learning rate.
* Mean absolute relative error.
* Coefficient of determination R2.
* Receiver operating characteristic curve (ROC).

De meestgebruikte methoden om een model te evalueren zijn k-fold cross-validation en leave-one-out-cross-validation (LOOCV).

Deze conclusie wordt gedeeld door Butler et al. (2018) en Wei et al. (2019) in hun onderzoeken naar het gebruik van machine learning in materials science.

De betekenis en werking van deze meeteenheden en methoden is zeer technisch. Het is dan ook lastig, om de werking ervan te beschrijven, zonder diepgaande wiskundige en statistische kennis. Voor verdere informatie wordt verwezen naar het werk van Butler et al. (2018), Das et al. (2020) en Wei et al. (2019).

## 3.4 Machine Learning in Materials Science

In deze paragraaf wordt ingegaan op het gebruik van AI/ML in het materials science werkveld, en wordt een antwoord gegeven op deelvraag 1:

*Welke AI-/ML-technieken en -modellen worden er in de chemische sector gebruikt om het proces voor nieuwe chemische producten te ontwikkelen te ondersteunen?*

Het proces voor het ontwikkelen van nieuwe chemische producten, het onderwerp van dit onderzoek, bevat een breed scala aan handelingen en activiteiten. Alle vier de componenten waaruit materials science bestaat (zie beschrijving in paragraaf 3.1), spelen een rol in de ontwikkeling van nieuwe producten. Elk van deze componenten bestaat uit een groot aantal activiteiten en processen, die ieder een aandeel hebben in de ontwikkeling van een product.

Het gebruik van AI/ML in dit werkveld, is een populair onderwerp voor onderzoek. Een zoekopdracht met de woorden ‘Materials Science’ en ‘Machine Learning’ levert in Google Scholar, 19.700 resultaten op. En dat is alleen nog maar voor publicaties sinds 1 januari 2022. Wordt er iets verder terug in de tijd gekeken, zien we dat er sinds 1 januari 2017, ongeveer 369.000 artikelen zijn gepubliceerd. Dit gigantische aantal publicaties is een goede indicatie van de populariteit van dit onderwerp.

Een grote uitdaging bij het bestuderen van de literatuur over dit onderwerp, is dat veel van de publicaties zijn gericht op een zeer specifiek probleem, gebruikmakend van één of meer ML-technieken, om een model voor dat probleem te ontwikkelen.

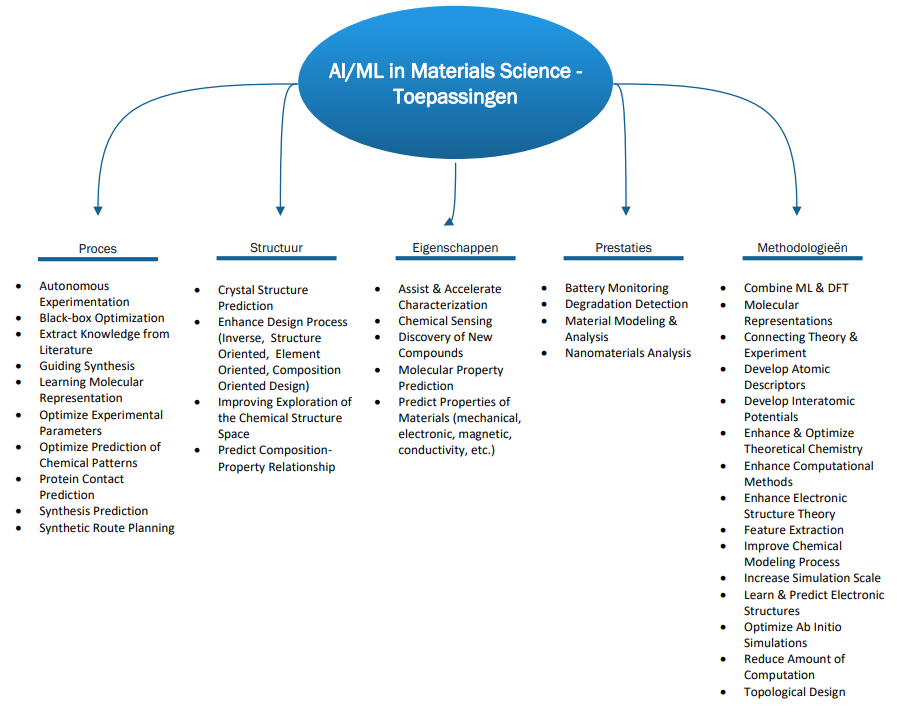
Bijvoorbeeld het ontwikkelen van een Deep Learning model, dat op basis van 2 moleculaire 'descriptors’, in staat is een DFT-berekening op te lossen, voor het bepalen welke energie er vrijkomt bij het samenvoegen van bepaalde moleculen (Ye et al. 2018). Of het onderzoek van Kim et a. (2019), waarbij een Graph Convolution Neural Network is ontwikkeld, voor het vinden van de best presterende katalysator voor de elektrochemische n2 (stikstof) reductie reactie.

Dit zijn slechts twee voorbeelden uit duizenden onderzoeken, die een zeer specifiek probleem proberen op te lossen met behulp van een specifieke ML-techniek.

Om een enigszins volledig beeld te krijgen van de ML-technieken, en de toepassingen daarvan, heeft het literatuuronderzoek zich gericht op de meer algemene publicaties. Met algemene publicaties wordt hier bedoeld; de artikelen die een overzicht geven van verschillende ML-toepassingen en -technieken, voor de verschillende (sub) processen van materials science. In deze artikelen (een soort ‘review’-artikelen) worden de verschillende toepassingen en technieken benoemd, maar wordt daar niet in detail over geschreven.

Op basis van deze artikelen is het mogelijk om een lijst samen te stellen van de verschillende ML-toepassingen in het materials science werkveld (zie figuur 12), en de gebruikte ML-technieken (zie figuur 13).

Alle artikelen die zijn geraadpleegd om de toepassingen te achterhalen zijn Engelstalig, daarom is ervoor gekozen om de toepassingen met de Engelse benaming op te nemen in de lijst. De benaming vertalen naar het Nederlands, levert in veel gevallen een onduidelijke beschrijving op. Een bijkomend voordeel voor het benoemen in het Engels, is dat de toepassing eenvoudiger kan worden terug herleid naar de oorspronkelijke tekst.



Figuur 12 - Overzicht toepassingen van AI/ML in het materials science domein.

De toepassingen zijn onderverdeeld in vijf categorieën; de vier componenten van materials science (zie paragraaf 3.1), aangevuld met de categorie ‘methodologieën’.

De categorie ‘methodologieën’ bestaat uit bestaande methodieken binnen de chemie, waarvoor AI/ML wordt ingezet om deze te versnellen en/of te verbeteren. Veel van deze methodieken komen voort uit het veld dat computationele scheikunde wordt genoemd.

Computationele scheikunde wordt door Jensen (2017) uitgelegd als een onderdeel van de chemie, waarbij gebruik wordt gemaakt van computers om theoretische scheikunde te ondersteunen en verbeteren. Jensen (2017) en Cramer en Cramer (2013) beschrijven theoretische scheikunde als het gebruikmaken van wiskundige methoden, gecombineerd met de wetten van de fysica, om chemische problemen en fenomenen te bestuderen.

Bekende voorbeelden van computationele scheikunde zijn onder andere Density Functional Theory (DFT) (Burke, 2012) en Molecular Dynamics (MD) (Attig, 2004).

Lezers die geïnteresseerd zijn in de details omtrent computationele scheikunde, worden verwezen naar het werk van Jensen (2017) en Cramer en Cramer (2013).

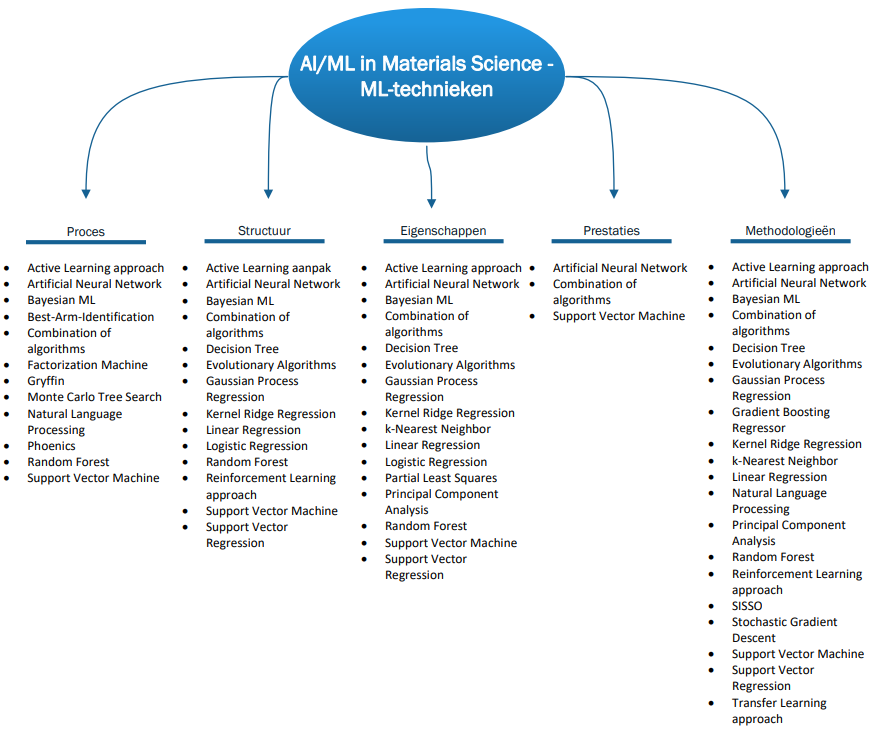
Door deze indeling kan op duidelijke wijze de relatie worden gelegd, naar de verschillende componenten waaruit materials science bestaat. Het toevoegen van de categorie ‘methodologieën’, is een logisch gevolg van de sterke opkomst van het gebruik van ML in de computationele scheikunde. De opkomst hiervan wordt onderstreept door het grote aantal toepassingen die tijdens dit onderzoek zijn geïdentificeerd voor die categorie.

Er dient te worden opgemerkt dat de toepassingen weergegeven in figuur 12, op een relatief hoog niveau zijn beschreven. Binnen elk van deze toepassingen zijn een veelvoud aan specifieke problemen te identificeren, waarvoor AI/ML kan worden ingezet.

De lijst met toepassingen is voortgekomen uit een analyse van 29 artikelen. In deze 29 artikelen wordt beschreven op welke wijze AI/ML een bijdrage levert aan het materials science werkveld, en voor welke toepassingen het gebruik voor AI/ML wordt onderzocht, of daadwerkelijk wordt ingezet. In de beschrijving van de toepassingen, wordt gerefereerd aan de specifieke onderzoeken voor die toepassing. Deze onderzoeken beschrijven de specifieke problemen, waarvan eerder in deze paragraaf twee voorbeelden zijn gegeven.

In totaal zijn ruim 700 referenties vastgelegd, van onderzoeken naar die specifieke problemen. Het is echter onmogelijk om van elk van die onderzoeken, de toepassing op dat niveau te identificeren en vast te leggen. Niet alleen is dit een omvangrijke taak qua aantallen, het vereist ook zeer diepgaande domeinkennis om de gedetailleerde toepassing te identificeren, en deze ook daadwerkelijk te begrijpen.

Naast de toepassingen die zijn opgenomen in figuur 12, zijn ook de gebruikte ML-technieken vastgelegd in een vergelijkbaar overzicht. Figuur 13 is een weergave van de lijst van ML-technieken, ingedeeld naar dezelfde vijf categorieën.



Figuur 13 - ML-technieken in het materials science werkveld.

Dit overzicht is voortgekomen uit de eerdergenoemde 29 artikelen, waarin wordt gerefereerd aan ruim 700 onderzoeken. De technieken die per categorie zijn genoemd in figuur 13, zijn gebruikt voor de toepassingen in diezelfde categorie gegeven in figuur 12.

De namen van de technieken zijn zoveel mogelijk overgenomen uit die artikelen. Er zijn echter wel enkele stappen genomen om bepaalde benamingen samen te voegen. Voor details over de hiervoor genomen stappen, wordt gerefereerd aan bijlage 2, hoofdstuk 11.

Bij dit overzicht dient te worden opgemerkt dat er een verschil in niveau van benaming en techniek zit. Zo is er enerzijds de naam van het ML-algoritme zelf (k-Nearest Neighbor, SISSO, Linear Regression, etc.), anderzijds een categorie ML-technieken (Active Learning, Reinforcement Learning, etc.) met elk een veelvoud aan algoritmes.

Dan is er ook nog de benaming ‘Artificial Neural Network’. Deze categorie bestaat uit een aantal verschillende neural networks zoals een Recurrent Neural Network, Convolutional Neural Network, Generative Adversarial Network, en anderen. De specifieke soort neural network is terug te vinden in bijlage 3, tabblad ‘Stap 1.6 ML-Technieken’.

Als laatste is er de benaming ‘Combinatie van algoritmes’. Deze categorie slaat op de onderzoeken waarvoor verschillende algoritmes zijn gebruikt om een specifiek model te ontwikkelen. De namen van de individuele algoritmes die gecombineerd zijn, zijn ook terug te vinden in bijlage 3, tabblad ‘Stap 1.6 ML-Technieken’.

Uit de lijst met ML-technieken en de toepassingen daarvan, kan worden geconcludeerd dat er waanzinnig veel potentie zit, in het gebruik van AI/ML binnen het materials science werkveld.

Het antwoord op deelvraag 1 is door middel van dit overzicht gegeven. Echter, een belangrijke conclusie die uit dit antwoord kan worden getrokken, is dat er niet één of slechts enkele technieken bestaan, waarop kan worden gefocust voor het gebruik ervan in het materials science werkveld.

De lijst met gebruikte ML-technieken is zeer divers. Er kan zelfs niet één specifieke aanpak (bijvoorbeeld supervised of unsupervised learning), worden aangemerkt als dé aanpak voor ML in het materials science werkveld.

Ook als er wordt gekeken naar het volgende niveau in classificatie in ML-technieken, dat van regressie, classificatie of een artificial neural network, kan er ook niet één specifieke categorie worden geïdentificeerd als zijnde de beste aanpak.

Het overzicht van ML-technieken gepresenteerd in figuur 13, geeft duidelijk weer dat ML-technieken gebruikt kunnen worden voor verschillende toepassingen. Dezelfde ML-technieken komen voor, in verschillende categorieën toepassingen.

Het omgekeerde is echter ook mogelijk. Een toepassing kan worden opgelost door verschillende ML-technieken. Dit wordt weergegeven in de tabel in bijlage 4 – ML-technieken per toepassing.

De tabel is dusdanig groot van omvang, dat deze niet kon worden opgenomen in het verslag en is dan ook als bijlage bijgevoegd.

In deze tabel is duidelijk zichtbaar dat een bepaalde toepassing van machine learning, kan worden ingevuld door verschillende technieken. Er zijn enkele toepassingen waarbij slechts één ML-techniek vermeldt staat. Er valt echter niet uit te sluiten, of er daadwerkelijk geen andere ML-techniek voor die toepassing gebruikt wordt, of dat deze toepassing slechts eenmalig is vermeld, in de onderzochte literatuur.

Hieruit valt op te maken dat er geen directe relatie bestaat, tussen de ML-techniek, en de toepassing waarvoor deze techniek wordt ingezet binnen materials science.

In de volgende paragraaf wordt ingegaan op wat er nodig is, om deze technieken daadwerkelijk te kunnen inzetten, waarmee een antwoord wordt gegeven op deelvraag 2.

## 3.5 Benodigdheden inzet ML-Model/ -Techniek

In deze paragraaf wordt ingegaan op het gedeelte van het onderzoek, waarin wordt gekeken naar wat er nodig is om de ML-technieken, die in de vorige paragraaf zijn gepresenteerd, daadwerkelijk te kunnen inzetten. Hiermee wordt een antwoord gegeven op deelvraag 2:

*Welke data-, hardware- en software-vereisten worden er gesteld om die technieken en modellen te kunnen gebruiken?*

Door de grote verscheidenheid aan ML-technieken, en daarbij behorende aanpakken, is het niet mogelijk om te focussen op één techniek of aanpak. Er is daarom gekozen om de onderzoeken, waaraan in de eerdergenoemde 29 artikelen wordt gerefereerd, te analyseren op de aanwezigheid van:

* De dataset waarmee het ML-model getraind is.
* Het ML-algoritme dat voor het trainen is gebruikt, of de stappen om het algoritme te reproduceren.
* De omgevingsvariabelen zoals softwareversie en gebruikte hardware, voor de ontwikkeling van het ML-model.
* Of de resultaten van het ML-model, zijn gevalideerd tegen resultaten uit fysieke experimenten.

Op basis van deze gegevens, kan worden bepaald of Dow gebruik kan maken van het ML-model, en op welke manier dit model geïmplementeerd kan worden.

De dataset dient beschikbaar te zijn om te kunnen bepalen of het model getraind is op basis van voldoende, en ook correcte data. Daarnaast is het belangrijk dat er inzicht is in het gebruikte algoritme. Als niet bekend is wat het algoritme precies doet, is het onmogelijk om te beoordelen hoe het model precies werkt.

De omgevingsvariabelen zijn van belang, om het algoritme op de juiste wijze te kunnen reproduceren. De gebruikte programmeertalen en bijbehorende libraries worden geregeld ge-update, wat van invloed kan zijn op de implementatie en werking ervan. Om het algoritme exact te kunnen reproduceren, dienen deze variabelen dan ook bekend te zijn.

Om een ML-model op industriële schaal te kunnen inzetten, is het noodzakelijk om te weten of het model is gevalideerd tegen de resultaten van experimenten, uitgevoerd in een laboratorium. Zoals beschreven in paragraaf 3.3.4, wordt de nauwkeurigheid van een ML-model gevalideerd tegen een bepaalde meeteenheid. Dit is echter een theoretische methode, waarbij wordt uitgegaan van historische data (data vergelijkbaar met de trainingsdata).

Om de daadwerkelijke, praktische nauwkeurigheid van het model te bepalen, dienen de voorspellingen die door het model worden gedaan, te worden gevalideerd tegen de resultaten van een fysiek experiment. Pas als een uitkomst of voorspelling van het model, is bevestigd door de resultaten die zijn verkregen uit het uitvoeren van een experiment, kan met zekerheid worden vastgesteld dat het model, een juiste uitkomst of voorspelling geeft.

Deze vier variabelen kunnen gerelateerd worden aan de ML-workflow, beschreven in hoofdstuk 3.3. De dataset is van toepassing op de stappen Data Verzameling en Data Representatie. Voor Model Building zijn de details van het algoritme en de omgevings-variabalen van belang. Het valideren tegen resultaten van experimenten, kan worden gezien als Model Evaluation.

Van de 29 artikelen die als basis hebben gediend om een antwoord te geven op deelvraag 1, zijn de referenties naar de specifieke onderzoeken vastgelegd. Dit heeft een overzicht opgeleverd van 709 referenties. Van deze 709 referenties zijn diegene die zijn gepubliceerd vóór 1 januari 2017, gemarkeerd als niet bruikbaar. Ook de dubbele waarden zijn eruit gefilterd, wat heeft geleid tot een uiteindelijke lijst van 310 unieke artikelen.

De lijst van 310 artikelen, is gesorteerd naar het aantal keren dat de desbetreffende referentie, in de 29 artikelen is genoemd. Deze sortering had als doel om te zorgen dat de artikelen, waar vaker aan gerefereerd is, onderdeel van de analyse zouden zijn. De sortering is nodig gebleken omdat niet alle 310 konden worden geanalyseerd.

De analyse is een zeer arbeidsintensieve en tijdrovende taak, waardoor het niet mogelijk is gebleken om alle 310 artikelen te analyseren. Er is besloten om ongeveer 2/3 van de artikelen te analyseren (200 artikelen). Op basis van dit aantal is het mogelijk om conclusies te verbinden aan de theorie.

De resultaten van de analyse zijn samengevat in tabel 1 – resultaten analyse DV2.

|  |  |
| --- | --- |
| **Variabelen** | **Aantal keer beschikbaar (uit een totaal van 200)** |
| Experimenten + Dataset + Algoritme + Omgevingsvariabelen | 1 |
| Experimenten | 5 |
| Dataset | 57 |
| Algoritme | 91 |
| Omgevingsvariabelen | 40 |
| Algoritme + Omgevingsvariabelen | 34 |
| Dataset + Algoritme | 41 |

*Tabel 1 – Resultaten Analyse DV2; aanwezigheid van de variabelen in de literatuur.*

Uit de resultaten blijkt, dat er slechts één geval is waarbij alle vier de variabelen aanwezig zijn. Daarnaast zijn er voor de geanalyseerde onderzoeken, maar vijf gevallen waarbij de resultaten van het model, zijn gevalideerd tegen de resultaten van een experiment.

De details van het algoritme worden wel vaker gedeeld. In bijna de helft van de gevallen is de code waaruit het algoritme bestaat beschikbaar gesteld, of is in detail beschreven hoe het algoritme is geïmplementeerd.

Ook de dataset wordt in een groot aantal gevallen gedeeld. Van de 57 gevallen waarbij dit het geval is, is de dataset volledig beschikbaar en te downloaden vanuit een openbare repository, of zijn de stappen om de dataset te genereren in detail beschreven.

Het vernoemen van de omgevingsvariabelen, heeft een vergelijkbaar resultaat als dat van de dataset. In 40 gevallen zijn deze omgevingsvariabelen expliciet vernoemd. Overigens is het zo dat als de code van het algoritme publiekelijk beschikbaar is, hier ook vaak de gebruikte softwareversies en libraries uit achterhaald kunnen worden.

Daarnaast zijn ook de resultaten van enkele combinaties van variabelen, in tabel 1 opgenomen.

De details van het algoritme en dat van de omgevingsvariabelen, zijn in 34 gevallen gegeven. Mogelijk kan op basis van die informatie, het ML-model gereproduceerd worden. Uiteraard dient dan wel onderzocht te worden welke dataset, er voor het trainen van dat model nodig is.

In 41 gevallen zijn zowel de dataset gedeeld, als de details van de implementatie van het algoritme. Dit is mogelijk al iets meer bruikbaar, omdat de werking van het algoritme gedurende de ontwikkeling, kan worden vergeleken met de resultaten van het desbetreffende onderzoek, om op die manier vast te stellen wat de invloed van de omgevingsvariabelen is.

De resultaten van deelvraag 1 en 2 leveren veel informatie op, die mede als input voor het praktijkonderzoek zullen worden gebruikt.

# 4. Empirie

Dit hoofdstuk bevat de gegevens en resultaten van het praktijkgedeelte van dit onderzoek. Eerst worden de onderzoeksgegevens beschreven, waarna de resultaten en analyse worden gepresenteerd.

## 4.1 Onderzoeksgegevens

Het praktijkgedeelte van dit onderzoek is uitgevoerd door middel van het houden van interviews, met experts die werkzaam zijn bij Dow. Met behulp van deze interviews is geprobeerd een antwoord te vinden op deelvragen 3 en 4:

Deelvraag 3:

*Welke veiligheidsrisico’s bestaan er bij het proces voor het creëren van nieuwe chemische producten? Welke relatie bestaat er tussen de geïdentificeerde risico’s en de inzet van AI/ML en hoe kunnen die risico’s worden gereduceerd?*

Deelvraag 4:

*Wat heeft Dow beschikbaar op het gebied van data, hardware en software, om de in de chemische sector gebruikte AI-/ML-technieken en -modellen, in de eigen organisatie te kunnen inzetten? Welke aanpassingen zijn er nodig in de bestaande IT-architectuur en -infrastructuur om AI/ML te kunnen inzetten binnen Dow, ter ondersteuning van de processen voor de ontwikkeling van nieuwe chemische producten?*

In de volgende paragraaf wordt beschreven welke soort interviews zijn afgenomen, en hoeveel. Er wordt uitgelegd hoe de respondenten zijn benaderd, en hoe de interviews zijn afgenomen. Tevens wordt een uitleg gegeven van de rollen die de respondenten uitvoeren, en de afdeling waarvoor zij werkzaam zijn. Als afsluiting van de paragraaf wordt beschreven hoe de vragenlijst is opgebouwd.

### 4.1.1 Werkwijze Interviews

Er zijn 5 interviews afgenomen met experts die verschillende rollen vervullen binnen Dow en R&D. Zie paragraaf 2.2.1.2 voor de criteria op basis waarvan de respondenten zijn geselecteerd. In paragraaf 4.1.1.1 wordt een beschrijving gegeven van de respondenten die hebben deelgenomen aan de interviews.

Er is gekozen om semi-gestructureerde interviews af te nemen. Deze vorm van interviewen biedt de mogelijkheid om te kunnen doorvragen, voor verdere verdieping of verduidelijking. Deze methode sluit ook goed aan op het verkennende karakter, dat de deelvragen kenmerkt.

De te beantwoorden deelvragen zijn zeer breed geformuleerd. Het is daarom niet mogelijk om de interviewvragen dusdanig af te bakenen, dat het houden van gestructureerde interviews of het uitzetten van een enquête, tot de mogelijkheden behoren.

Vanwege de brede formulering van de deelvragen, zijn ook de interviewvragen breed en algemeen geformuleerd. Er is geen specifieke context geschetst in de deelvragen. Een semi-gestructureerd interview biedt de ruimte om de context in te vullen tijdens het interview.

Om toevalligheden uit te sluiten, zijn voor alle interviews dezelfde volgorde en structuur aangehouden. Aan elke respondent zijn dezelfde vragen gesteld. Afhankelijk van de kennis, ervaring én rol van de respondent, kon er tijdens het interview specifiek worden doorgevraagd over bepaalde onderwerpen.

Om te waarborgen dat de respondenten vrijuit hun visie en mening over dit onderwerp kunnen geven, worden de namen van de respondenten niet vermeldt in dit document. Het vermelden van de namen heeft ook geen toegevoegde waarde voor de interpretatie en het presenteren van de resultaten.

De respondenten zijn via email uitgenodigd om deel te nemen aan het interview. In de uitnodiging is aangegeven, dat de respondent is geselecteerd op aanraden van een begeleider op de werkplek (een R&D research scientist). Ook is de vragenlijst meegestuurd in de uitnodiging, zodat de respondenten zich een beeld konden vormen van de inhoud van de interviews. Alle vijf de aangeschreven respondenten hebben de uitnodiging geaccepteerd, zonder verdere vragen te stellen

De interviews zijn virtueel afgenomen. De respondenten bevinden zich in een andere locatie dan de interviewer (de Verenigde Staten en België), waardoor persoonlijk interviewen niet mogelijk was. Om de interviews op afstand te kunnen afnemen, is gebruikgemaakt van MS Teams. Met deze software was het tevens mogelijk om de interviews op te nemen (zowel audio als video) en automatisch een transcript te genereren.

Bij aanvang van de interviews is een korte achtergrond gegeven over het onderzoek. Hierbij zijn de onderzoeksopzet, het doel van het onderzoek en de onderzoeksvragen aan bod gekomen.

De opnames en het vastleggen van de interviews in een transcript, is gestart vanaf het moment dat is begonnen met het doorlopen van de vragenlijst. De opnames zijn geëindigd na het beantwoorden van de laatste vraag van de lijst.

De duur van de interviews was vergelijkbaar. Alle vijf de interviews duurden tussen de 50 minuten en 1 uur. Dit betreft alleen het gedeelte van het interview, waarin de interviewvragen zijn besproken. De introductie en de uitleg over het interview en het onderzoek, zijn niet meegenomen in deze tijdsduur.

De transcripten die door MS Teams zijn gegenereerd, dienden opgeschoond te worden. Het doel was een woordelijk transcript te maken, waarbij stopwoorden zijn verwijderd. Daarnaast zijn de namen van de interviewer en respondenten, vervangen door de woorden ‘Interviewer’ en ‘Respondent’. Ook zijn de tijdstempels die door MS Teams automatisch zijn toegevoegd, verwijderd uit het transcript.

Na opschoning van het transcript is deze naar de desbetreffende respondent gestuurd, met de vraag of de inhoud van het interview juist is vastgelegd. Met uitzondering van enkele grammaticale aanpassingen, zijn alle respondenten akkoord gegaan met de vastlegging van de interviews.

Door deze inhoudelijke review, worden de betrouwbaarheid en validiteit van de interviewresultaten verhoogt.

De transcripten zijn als bijlage aan dit document bijgevoegd, zie bijlagen 7 t/m 11.

#### 4.1.1.1 Respondenten

Om diepgaande inzichten te verkrijgen en rijke data te verzamelen, was het noodzakelijk dat de respondenten over een grote mate van kennis en ervaring beschikken, met betrekking tot de R&D-processen. Dit gecombineerd met computationele methoden voor het ontwikkelen van modellen, ter ondersteuning van de R&D-processen.

Daarnaast was het van belang dat deze personen betrokken zijn bij de initiatieven die Dow onderneemt, om het gebruik van machine learning binnen R&D te onderzoeken en te implementeren.

Een factor die een grote invloed heeft gehad op het bepalen van de populatie, is het beperkt aantal mensen binnen Dow, die over deze kennis en ervaring beschikken. Er zijn uiteindelijk vijf personen bereid gevonden om mee te werken aan de interviews.

De respondenten zijn werkzaam binnen twee afdelingen van de R&D-organisatie:

* **Core R&D**

De afdeling die zich bezighoudt met onderzoek doen naar en het verbeteren van de mogelijkheden waarover Dow beschikt, om materialen te ontwikkelen, te verbeteren en te ontdekken. Core R&D werkt samen met de R&D-afdelingen van de verschillende business-groepen van Dow, om waarde te creëren voor de gehele organisatie.

* **Information Research**

Information Research (IR) is een onderdeel van de R&D-organisatie, dat zich bezighoudt met het voorzien van, en verder ontwikkelen van digitale mogelijkheden en initiatieven, om de R&D-organisatie te helpen haar strategische doelen te bewerkstelligen. IR zet haar middelen en kennis in, om R&D-mogelijkheden te ontwikkelen en te analyseren.

Voor dit onderzoek zijn twee personen van de IR-afdeling geïnterviewd en drie personen van Core R&D.

In tabel 2 is een korte beschrijving gegeven van de rollen die de respondenten binnen Dow uitvoeren, en de afdeling waarvoor de respondent werkzaam is.

|  |  |
| --- | --- |
| **Afdeling** | **Rol** |
| Information Research | Research Scientist IR. Houdt zich bezig met de ontwikkeling en implementatie van een AI/ML-strategie en -proces, voor de R&D-organisatie. |
| Solution Manager High Performance Computing (HPC) Cluster. Verantwoordelijk voor de ontwikkeling en inzet van het HPC-Cluster. |
| Core R&D | Research Scientist R&D (2 respondenten). Onderzoekt mogelijkheden voor ontwikkeling en verbetering van (nieuwe) materialen, met behulp van modellen en computationele methoden. Zowel op molecuul-niveau als op basis van formuleringen. |
| Sr. Research Scientist R&D. Zie boven, aangevuld met strategische activiteiten. |

*Tabel 2 – Afdeling en rol van de respondenten.*

#### 4.1.1.2 Vragenlijst

De interviews zijn afgenomen aan de hand van een vragenlijst die als bijlage is bijgevoegd, zie bijlage 6 – Vragenlijst Interviews. De vragenlijst is opgesteld in het Engels omdat alle respondenten Engelstalig zijn. De interviews zijn dan ook in het Engels afgenomen.

De vragen zijn opgedeeld naar de onderwerpen waaruit de deelvragen bestaan. Zie tabel 4 voor de onderwerpen per deelvraag. Het onderwerp is genoemd in de vragenlijst, zodat de link kan worden gelegd tussen het onderwerp en de interviewvraag.

|  |  |  |
| --- | --- | --- |
| **Deelvraag** | | **Onderwerpen** |
| *Deelvraag 3* | *Welke veiligheidsrisico’s bestaan er bij het proces voor het creëren van nieuwe chemische producten? Welke relatie bestaat er tussen de geïdentificeerde risico’s en de inzet van AI/ML en hoe kunnen die risico’s worden gereduceerd?* | Veiligheidsrisico’s |
| Relatie tussen veiligheidsrisico’s en inzet van machine learning |
| Risicobeperking |
| *Deelvraag 4* | *Wat heeft Dow beschikbaar op het gebied van data, hardware en software, om de in de chemische sector gebruikte AI-/ML-technieken en -modellen, in de eigen organisatie te kunnen inzetten? Welke aanpassingen zijn er nodig in de bestaande IT-architectuur en -infrastructuur om AI/ML te kunnen inzetten binnen Dow, ter ondersteuning van de processen voor de ontwikkeling van nieuwe chemische producten?* | Data |
| Hardware |
| Software |
| Externe modellen |
| Externe data |
| Aanpassingen IT-infrastructuur en -architectuur |

*Tabel 4 – Onderwerpen van deelvraag 3 en 4*

De vragen zijn algemeen en breed geformuleerd en bieden veel ruimte voor de respondenten om vanuit hun eigen kennis en ervaring, een gedetailleerd antwoord te kunnen geven. Deze manier van formuleren, sluit goed aan op het karakter van semi-gestructureerde interviews en biedt een goede basis voor het verzamelen van diepgaande inzichten. Tevens voorkomt de algemene formulering, dat de vragen suggestief zijn.

### 4.1.2 Analyse Interviews

De transcripten zijn geanalyseerd met behulp van coderingen. Om de coderingen toe te wijzen aan de tekstgedeeltes, zijn de stappen ‘open coderen’, ‘axiaal coderen’ en ‘selectief coderen’ doorlopen.

Voor de codering van de transcripten, is gebruikgemaakt van het softwareprogramma Atlas.ti.

De codes die tijdens de stap ‘axiaal coderen’ zijn toegekend, zijn gebaseerd op de deelvragen die worden beantwoord met de interviews. De coderingen volgen de onderwerpen per deelvraag, die zijn gepresenteerd in tabel 4.

Voor de stap ‘selectief coderen’ is gebruikgemaakt van de functionaliteit in Atlas.ti, om een rapport te genereren, waarbij de tekstgedeeltes van de verschillende transcripten, per code worden samengevoegd.

Met behulp van dit rapport zijn de resultaten van de verschillende interviews, per onderwerp met elkaar vergeleken en geanalyseerd. Dit rapport, inclusief de gebruikte codes, is bijgevoegd aan dit document, zie bijlage 11 - Codering transcripten.

## 4.2 Onderzoeksresultaten

In deze paragraaf worden de resultaten van het praktijkonderzoek gepresenteerd. Hierbij worden de deelvragen als leidraad gebruikt. Per deelvraag worden de relevante uitspraken weergegeven. In de volgende paragraaf worden deze uitspraken geanalyseerd.

### 4.2.1 Deelvraag 3

Tabel 5 geeft de drie onderwerpen aan waarin deelvraag 3 is opgedeeld. De resultaten worden gepresenteerd aan de hand van deze drie onderwerpen.

|  |  |
| --- | --- |
| **Deelvraag 3** | **Onderwerpen** |
| *Welke veiligheidsrisico’s bestaan er bij het proces voor het creëren van nieuwe chemische producten? Welke relatie bestaat er tussen de geïdentificeerde risico’s en de inzet van AI/ML en hoe kunnen die risico’s worden gereduceerd?* | Veiligheidsrisico’s |
| Relatie tussen veiligheidsrisico’s en inzet van machine learning |
| Risicobeperking |

*Tabel 5 – Onderwerpen van deelvraag 3*

* **Veiligheidsrisico’s**

Het gevaar dat bestaat bij het ontwikkelen van nieuwe materialen, is dat er onbedoelde reacties kunnen ontstaan bij het mixen van verschillende componenten.

Met betrekking tot formuleringen, gaf respondent A aan dat: *“Obviously, some of the ingredients that are used in formulations are hazardous often.”*

Bij de chemische reactie tussen verschillende componenten, spelen meerdere factoren een rol. Respondent B legt uit welke factoren een risico kunnen vormen, bij het samenvoegen van de componenten waaruit het materiaal wordt gemaakt: *“you have your ingredients, you have your side product, you have your expected conversion, you have the heat and temperature, like the temperature and pressure that you would want to run your reactions at. So, all of that are factors where you can have potentially safety risks. Because you can design for something that just puts you up for potential unplanned events.”*

* **Relatie tussen veiligheidsrisico’s en inzet van machine learning**

De mate van risico, wordt bepaald door het soort voorspellingen, dat het model maakt. Een model dat interpoleert, dat wil zeggen een voorspelling maken binnen het bereik van de data, zal minder snel tot een risico leiden dan een model dat extrapoleert.   
Bij extrapolatie worden voorspellingen gemaakt die buiten het bereik van de trainings-data liggen, wat erop neer komt dat de voorspelling een grote mate van onbekendheid en onzekerheid bevat.

Respondent C zegt hierover het volgende: *“Another risk that we have with machine learning is essentially on the extrapolations that are a bit on the edge of our training space domain. So, let's say we are in the process of exploring new materials. That is to say that we are trying to seek solutions in a domain space of your solutions where you have little experience.*

*So, your model is more likely to have been trained on a small zone of your solutions, and then you want to reach something that is outside that zone.”*

Respondent A gaat in op de vergelijking met een model dat interpoleert en geeft aan waarom dit minder risicovol is: *“One of the things, and it's only really a concern for extrapolation, right, because most of the algorithms and optimization types that we use so far, are really more for interpolation. Within like small changes to products within a sort of known product space, that's what we're most successful at so far.*

*And the safety risks within that are quite a bit lower, right, cause the all the products are similar to previous products. So, we wouldn't expect any really weird ingredient combinations that could be reactive in an unexpected way”*

Een ander aspect dat een rol speelt bij de inzet van machine learning voor materials science processen, zijn de factoren die in beschouwing worden genomen bij de ontwikkeling van een model.

Respondent B zegt hierover, dat op dit moment de modellen worden gedreven door de te voorspellen eigenschappen, en niet zozeer door de factoren die een rol spelen op het gebied van veiligheid: *“But the problem is that very often the optimization is driven, based on performance at this point in time and the safety considerations aren't something that we have sufficient data to really actually incorporate that and build that in as a constraint.*

*Like most generative models are purely property driven. And we have a hard enough time with that, let alone now incorporating constraints like, can we try to make something that's like, either biobased or like renewables based or like, biodegradable”.*

* **Risicobeperking**

Er zijn twee mogelijkheden benoemd om de risico’s te kunnen beperken.

De eerste mogelijkheid is om de uitkomsten van een model, te laten valideren door een SME in het veld. Respondent C geeft dit proces weer aan de hand van een voorbeeld: *“It always comes down as well to human decisions. Machine models will make you a proposition and then you would have to make a decision out of it. And this decision is how you're going to try out this material”.*

Respondent B gaat hierbij ook nog in, op relatie met een model dat extrapoleert: *“This could very well be as a function of extrapolation, right. And it's like, if you're far enough away from your space where you have good data density, you might just say yes, somebody should take a look at it before we make this.”*

Ook respondent A gaat in op de relatie tussen een review van een SME en een model dat extrapoleert: *“and as we work towards algorithms that are more designed for extrapolation, that's something we're gonna have to figure out is; do we always have to basically, do a hazard review, a subject matter thing where we list out the possible negative interactions and hard code against them.”*

Om te vermijden dat elke uitkomst van een model, door een SME gevalideerd zou moeten worden, zullen beperkingen ingebouwd moeten worden in het model.

Als bekend is welke factoren kunnen leiden tot een risico, bijvoorbeeld ingrediënten waarvan bekend is dat deze een gevaarlijke reactie kunnen opleveren, kan dit als een beperking ingebouwd worden in het model.

Respondent E noemt dit het inprogrammeren van de kennis van een SME: *“If you have a talented enough person who can program what that SME knows into the machine learning model to serve as a boundary then in time, yes, the SME’s presence won't need to be there.”*

Respondent C geeft echter aan, dat die stap naar de fysieke wereld uiteindelijk toch gemaakt moet worden: *“You have to go back to the physical world at some point. You have to measure the materials, you have to produce it or whatever, so you have to go back to the physical world.”*

Bepalend voor de accuraatheid van de voorspelling van het model, is de kwaliteit van de training- en test-data. Respondent B vindt dat als de test-set van voldoende kwaliteit is, dat deze het karakter van een experiment bezit, de fysieke validatie van de uitkomst van het model minder nodig is: *“The quality of the test set really determines on whether or not that should be acceptable. I mean, if you say it's like, hey, I have a test set that I hold out that essentially has the character of a new experiment. Then I think you should be able to go about it.”*

Echter, als de fysieke validatie niet tijdens het model-ontwikkelingsproject is gedaan, zal deze alsnog gedaan worden als het model eenmaal is uitgerold naar de eindgebruiker, zoals aangegeven door respondent A: *“if you didn't do the experimental validation, what would your next step be? It would be to deploy it and people would start using it by doing experiments and you would get validation information”.*

Ook respondent B geeft aan dat het zeer reëel is dat een materiaal nog getest wordt, voordat deze naar de klant gestuurd wordt. Met het verschil dat een chemicus een lager aantal experimenten moet uitvoeren, om het materiaal te valideren: *“if you have a model that recommends you a formulation or a material, before you send it to the customer, you probably are still going to check it at this time. The difference what it’s gonna be is, instead of doing 15 experiments you may be doing 2.”*

### 4.2.2 Deelvraag 4

Deelvraag vier is opgedeeld in verschillende onderwerpen, die zijn weergegeven in tabel 6. De resultaten worden gepresenteerd aan de hand van deze onderwerpen.

|  |  |
| --- | --- |
| **Deelvraag 4** | **Onderwerpen** |
| *Wat heeft Dow beschikbaar op het gebied van data, hardware en software, om de in de chemische sector gebruikte AI-/ML-technieken en -modellen, in de eigen organisatie te kunnen inzetten? Welke aanpassingen zijn er nodig in de bestaande IT-architectuur en -infrastructuur om AI/ML te kunnen inzetten binnen Dow, ter ondersteuning van de processen voor de ontwikkeling van nieuwe chemische producten?* | Data |
| Hardware |
| Software |
| Externe modellen |
| Externe data |
| Aanpassingen IT-infrastructuur en -architectuur |

*Tabel 6 – Onderwerpen van deelvraag 3*

* **Data**

Binnen R&D is data één van de grootse uitdagingen op het gebied van machine learning. Respondent A is hier heel duidelijk over: *“this is definitely, unanimously our biggest gap.”*

Deze uitdagingen bevinden zich in de verschillende stadia van het proces; vanaf het vastleggen van de data, tot aan het gereedmaken van data voor het trainen van een model.

Er zijn veel verschillende data-bronnen binnen R&D. De werkzaamheden van een R&D-wetenschapper zijn innovatief van aard. Het is niet te vergelijken met een proces-omgeving, waar er een zeer grote mate van standaardisatie is. De ontwikkelingen binnen R&D zijn kleiner van schaal, en elke wetenschapper legt gegevens vast, op een manier die voor hem of haar het beste werkt. Dit levert meteen al verschillen op, afhankelijk van de regio waar de desbetreffende wetenschapper zicht bevindt. Zo zal iemand in Noord-Amerika temperaturen vastleggen in Fahrenheit, in Europa wordt dit vastgelegd in Celsius.

Respondent A geeft hier een zeer uitgebreid voorbeeld van en geeft uiteindelijk ook aan, dat er op dit moment nog geen structurele oplossing voor is: *“we haven't come up with a good engineering solution to manage that problem yet, in terms of we can ask people to please write it down, right. There's an operational discipline kind of solution. … But we're really interested in infrastructure solutions that can manage all that”*

Een andere uitdaging is het verkrijgen van de data. Er is geen centrale database waar de data in wordt opgeslagen. Er bestaan centrale systemen waar laboratoriumgegevens in worden opgeslagen, echter speelt ook hier de aard van R&D-werkzaamheden een rol. Elke wetenschapper bouwt min of meer een eigen data-warehouse, op basis van het project of experiment waaraan gewerkt wordt. En is het dus aan de data-scientist om de data op te vragen bij verschillende personen en afdelingen, wanneer aan een machine learning project gewerkt wordt.

Zoals respondent C aangeeft met: *“A lot of the data are actually captured in documents. Like Word, PowerPoint, ELN [Electronic Laboratory Notebook], CRI’s... All kind of Office repositories, that are nice to create reports, but that are very hard to extract data from”.*

Er zijn initatieven gaande binnen Dow om te proberen deze tekortkomingen op te lossen, maar respondent A geeft duidelijk aan dat het meerdere jaren zal duren, voordat de resultaten hiervan zichtbaar zijn: *“There's a lot of activities going on at Dow to build infrastructure to make various pieces of that challenge fixed, but it's gonna…, I mean, it's five years minimum, probably more like 8 before we're gonna get in a good situation.”*

* **Hardware**

Op het gebied van hardware, lijken er op dit moment genoeg middelen voorhanden. Zo is er de beschikking over een high performance computing (HPC) cluster en kan er worden uitgeweken naar een cloud-omgeving, als er een noodzaak is voor extra middelen.

Vrijwel alle respondenten geven aan dat deze middelen voldoende zijn om machine learning modellen te ontwikkelen. Zo zegt respondent E hierover: *“if we exceed our hardware resources on premises, we can spill out into the cloud and buy compute-hours and GPU-hours, to satisfy all the needs that we would ever want with this company”.*

Er zijn echter wel enige restricties verbonden aan het gebruikmaken van de cloud-omgeving. Zo is dit afhankelijk van het budget dat beschikbaar is, om die middelen in te zetten. Ook de data waarmee gewerkt wordt, bepaalt of de cloud-omgeving gebruikt kan worden.

Zo is het bijvoorbeeld niet mogelijk om de cloud-omgeving te gebruiken als er met ‘crown jewel’-data gewerkt wordt. Deze data is van dusdanig gevoelige aard, dat deze absoluut niet buiten de organisatie gedeeld mag worden. Er wordt gewerkt aan mogelijkheden om dit in de toekomst wel mogelijk te maken, zoals respondent D aangeeft: *“There are a lot of controls associated with meeting the Crown jewel standards. And we don't have those in place today, but that is a capability we're looking to develop and continue to expand the data that is eligible to go out there.”*

Hardware-middelen worden niet beschouwd als limiterende factor, zoals respondent B zegt: “No, that's usually not where we run into the limitations. We are limited, especially in the material side, we are limited on the data first.”

* **Software**

Ook software lijkt geen limiterende factor te spelen. Respondent E geeft dit kort en bondig aan op de vraag of er voldoende software-middelen voorhanden zijn: *“Yes. Without a question, yes.”*.

Wat wordt bevestigd door de reactie van respondent B: *“Typically we don't have an issue with that. That's usually not the problems we run into.”*

Zo is er de beschikking over package managers voor Python en R, om nieuwe packages en releases te beheren. Ook is er een samenwerking met Microsoft, zodat gebruik gemaakt kan worden van de machine learning tools, die de Azure omgeving aanbiedt.

Een uitdaging hierbij is echter wel, dat de documentatie en kennis nog niet op het niveau is, om deze tools volledig te kunnen benutten. Zoals respondent C aangeeft met: *“But the reality is, very few people really know how to use all those tools. So, the limitation is not in the portfolio of the tools that we have. It’s in the documentation on how we can use them and who is the expert to help us.”*

* **Externe modellen**

Het gebruik van extern ontwikkelde modellen, onttrokken uit de literatuur, lijkt niet heel erg voor de hand liggend. Respondent E gaf aan dat hetgeen dat ontwikkeld is in de academica, in veel gevallen niet geschikt is voor een industriële omgeving: *“I wish I could say that academic level code is production ready for an industrial company, but it's not. I mean, this is a product of a student who's working 90 hours a week and got their code working just enough so that they can fill out that chapter in their dissertation and publish it. That's more often than not what’s happening.”*

Een visie die wordt gedeeld door respondent A die zegt: *“there's a big incentive for academia to generate new algorithms, when 90% of them probably aren't useful, really.”*

Ook respondent C acht de kans niet heel groot, dat een extern ontwikkeld model direct kan worden ingezet in de eigen organisatie.

Een voordeel kan wel zijn, dat het externe model een startpunt biedt om zelf iets te ontwikkelen: *“The opportunity to leverage a drop in model at Dow like, this model has been applied successfully for Bayer, it has been published and then we take it and we apply it for Dow, is unlikely. We probably would have to refactor the models and use the same methods. But at least you know where to start.”*

Ook respondenten B en E delen de mening, dat een extern model een goed uitgangspunt kan bieden om zelf een model te ontwikkelen. Immers, er is al min of meer bewezen dat de aanpak werkt.

Een tweede voordeel dat een extern model kan bieden, is dat het model is getraind op basis van data waar Dow zelf niet over beschikt. En dat wanneer de data een vergelijkbare context beslaat, waarvoor Dow het model wil inzetten, opnieuw trainen niet nodig is.

Zoals respondent C zegt: *“The advantage of using the external model will two-fold. The first will be they have used training data that we do not have. They have used a bigger training set than we have. OK, so I could make use of this well-trained model into my case, which is just a small set of data. And then I don't need to train it. I just need to test it.”*

Er zijn echter wel enkele voorwaarden aan het gebruik van een extern model. Zo moet de werking van het model en de stappen die zijn doorlopen om het model te ontwikkelen, voldoende zijn gedocumenteerd om te bepalen of het model bruikbaar is voor Dow.

Ook dient er te worden nagegaan of het model daadwerkelijk gebruikt mag worden in een commerciële omgeving. Er zal dus gekeken moeten worden of het gebruik niet is afgeschermd door patenten of licenties.

* **Externe data**

Het gebruik van externe data lijkt al meer ingeburgerd te zijn. Zowel respondent A als C gaven aan, dat dit al aan de orde is bij projecten die binnen Dow zijn uitgevoerd.

Respondent A zegt hierover: *“we've definitely done some things with external data. So I think we're looking at that. There's only a few areas that have really well curated datasets externally. So where those exist, I think we have pulled them in to assist”*

Deze opmerking geeft ook aan, dat het gegevensbeheer van die data op orde moet zijn. Iets wat door respondent C wordt bevestigd met de opmerking: *“Let's say for example atomic configurations of small molecules and things like that. That we can use as long as everything is documented, people are able to pull from those databases what they need.”*

Echter zijn de gebieden waar goed documenteerde en beheerde data beschikbaar is, nog zeer schaars. Wat volgens respondent A wel een aandachtspunt is in de chemische industrie: *“But it's pretty narrow the number of instances where that kind of data exists, it's something that the industry is working on.”*

* **Aanpassingen IT-infrastructuur en -architectuur**

Op de vraag welke aanpassingen aan de IT-infrastructuur en -architectuur, nodig zouden zijn om machine learning te ondersteunen, zijn veel verschillende gebieden genoemd.   
Er is specifiek de vraag gesteld, welke verbeteringen de respondenten vanuit hun visie zien. Echter, zijn gedurende het gehele interview, gebieden aangegeven waar verbeterpunten kunnen worden geïdentificeerd.

Respondent A gaf heel treffend aan, dat er eigenlijk op alle gebieden veranderingen nodig zijn en dat alleen de wiskunde en scheikunde hetzelfde blijft: *“So so many changes, yeah. Structural changes, work process changes, governance changes, skills changes, culture changes, it completely changes. So we're in the midst of it now … The math stays the same. The chemistry stays the same. Everything else differ”*

Een ander element dat door meerdere respondenten is aangegeven ter verbetering, betreft het uitrollen van een model aan de eindgebruikers. Op dit moment ontbreekt een gestandaardiseerd proces, voor het in productie nemen van een ontwikkeld model. Daarnaast is er een noodzaak om een proces te ontwikkelen voor ‘model lifecycle management’, waarbij het model geëvalueerd wordt nadat het een tijd in productie heeft gewerkt. Op deze manier wordt er een meer iteratief proces toegepast, om het model verder te ontwikkelen en te verbeteren, op basis van nieuwe data uit een productie-omgeving.

Volgens respondent B krijgt dit op het moment te weinig aandacht: *“The other big factor in that space is model maintenance and model lifetime management. That's something that I think doesn't get enough attention. So it's like anything that we put out there and that also has huge potential to be standardized. Anything that goes out there should have model performance metrics”*.

Respondent C gaf hierover aan dat er een initiatief gaande is, voor het opzetten van een Machine Learning Operations proces, om dit probleem aan te pakken: *“I'm contributing into as well the machine learning operations [MLOps/MLO]. We need to implement more control feedback OK, about how this model performs. How does it continue to perform, based on your data? At which point do we trust all the kind of reevaluation, retraining or remanufacturing of this models?”*

Ook met betrekking tot dataverzameling, lijken er nog veel mogelijkheden te zijn om te verbeteren. Te beginnen bij de instrumenten waarmee experimenten worden uitgevoerd. Er is op dat gebied een behoorlijke technische achterstand, waar het gaat om het moderniseren van die infrastructuur.

Er wordt aan gewerkt om deze achterstand in te halen, zoals respondent A aangeeft: *“Like there's a lot of technical debt to overcome in terms of modernizing the data generation instrumentation to allow it to securely be in a place where it can even convey that data to a centralized location.*

*So, they're doing hardware upgrades. They purchased like thousands of workstations to get everything onto a modern operating system. They're doing a whole bunch of networking stuff.”*

Dataverzameling is één onderwerp, voor de opslag van de data worden ook enkele mogelijkheden genoemd door de respondenten. Zowel respondent B en C gaven aan, dat de beschikking hebben over een datawarehouse en data-catalog, een flinke meerwaarde zou zijn voor het ontwikkelen van een model. Met de woorden van respondent B: *“having a data catalog and being like here is the data that's available and here's how you get it, would be awesome, right? I mean it's like one place and you can just look it up.”*

Er is ook nog ingegaan op andere elementen; zoals training, het hebben van de mensen met de juiste vaardigheden en een duidelijke opdeling van taken bij de verschillende rollen. Dit zijn echter meer organisatorische aanpassingen en worden hier dan ook verder niet benoemd.

## 4.3 Analyse

Het voornaamste risico dat binnen de materials science processen bestaat, zijn de chemische reacties die kunnen ontstaan als verschillende componenten worden samengevoegd. Het kan gebeuren dat componenten op een onverwachte en onbedoelde manier met elkaar reageren, wat kan leiden tot een risico met impact op mens en milieu.

Machine learning wordt in materials science onder andere gebruikt om voorspellingen te doen, over welke componenten samengevoegd moeten worden, om een materiaal met bepaalde eigenschappen te creëren. Of het wordt ingezet om de proces-condities voor te schrijven. In beide gevallen geldt, dat het machine learning model een voorspelling doet, die kan leiden tot een gevaarlijke reactie van de componenten.

Op de vraag wat er moet gebeuren, om machine learning op een veilige manier in te zetten in de materials science processen, kan een duidelijke conclusie worden getrokken.

Alle respondenten gaven aan, dat de risico’s zich voornamelijk in het gebied van extrapolatie bevinden. Dit houdt in dat de voorspelling zich buiten het bereik van de trainingsdata bevindt; er is nog geen eerdere kennis opgedaan over de variabelen waarin de voorspelling wordt gemaakt. Het tegenovergestelde, interpolatie, houdt in dat de voorspelling zich binnen het bereik van de trainingsdata bevindt, en dat de voorspelling dichter bij de bekende variabelen ligt.

Bij modellen die extrapoleren, is het raadzaam om de voorspelling van het model, te valideren met fysieke experimenten. Deze conclusie wordt door alle respondenten genoemd. Er kan echter ook worden geconcludeerd, dat dat experiment toch wel wordt uitgevoerd, voordat een materiaal op grotere schaal wordt geproduceerd. De voorspelling van het model, zal op zijn minst worden gevalideerd door de persoon die belast is met de taak om het materiaal te creëren. Dit is de overgang van de digitale wereld naar de fysieke wereld. Deze persoon zal op basis van eigen kennis en ervaring, een inschatting kunnen maken van de risico’s van de voorspelling.

Kijkend naar de middelen waar Dow over beschikt, om machine learning te kunnen inzetten, kan worden gesteld dat vooral data een probleemgebied vormt. Over de hardware- en software-middelen zijn alle respondenten het eens. Op die gebieden beschikt Dow over voldoende middelen, om machine learning in te zetten.

Voor software geldt dat de R&D-organisatie de beschikking heeft over de omgevingen die nodig zijn om machine learning modellen te ontwikkelen; R, Python en MATLAB. Ook is het over het algemeen mogelijk, om (open-source) packages te gebruiken, eventueel met behulp van een package manager.

Voor de uitrol van een model, is het afhankelijk van de schaal waarop het model gebruikt gaat worden. Deze valt grofweg in te delen in twee categorieën: R&D-niveau, en Business-niveau en groter.

De uitrol op R&D-niveau vindt meestal plaats binnen de omgeving, waarin het model is ontwikkeld zoals RStudio Connect, MATLAB Web App Server of Streamlit.

Voor de uitrol op grotere schaal, zal een webapplicatie ontwikkeld moeten worden, die door een grote user-base gebruikt kan worden. Hiervoor zal moeten worden samengewerkt met andere afdelingen binnen Dow.

Het is voor dit soort applicaties, dat er uitdagingen en mogelijkheden zijn geïdentificeerd. Zo is niet altijd de juiste kennis aanwezig of beschikbaar, om een dergelijke uitrol mogelijk te maken. Ook is er geen gestandaardiseerd proces voor de ontwikkeling en ondersteuning, van dit soort machine learning applicaties. Vier van de vijf respondenten hebben aangegeven, dat dit gebied voor verbetering vatbaar is.

Met betrekking tot hardware, waren de respondenten eensgezind. Alle respondenten gaven aan dat er voldoende middelen zijn, om machine learning modellen te ontwikkelen. Eén van de respondenten gaf aan dat als het gebruik van machine learning fors toeneemt, de huidige middelen misschien niet toereikend zullen zijn. Er is echter in veel gevallen de mogelijkheid om uit te wijken naar een cloud-omgeving, om de hardware-middelen op te schalen.

De grootste uitdaging wordt gevormd door de data. Alle respondenten die te maken hebben met het maken van modellen (vier van de vijf respondenten), geeft aan dat data dé limiterende factor is op het gebied van machine learning binnen/bij Dow. Er zijn dan ook verschillende aanpassingen nodig, om dat probleem aan te kunnen pakken.

Dit probleem betreft alle aspecten van data; genereren van data, vastleggen van data, en verzamelen en cureren van data. Dit aspect is zo omvangrijk, dat het niet eenvoudig is om hier een oplossing voor te vinden.

Er zijn dan ook al diverse initiatieven gestart binnen Dow, om deze problemen het hoofd te kunnen bieden. Echter, zoals zeer treffend aangegeven door één van de respondenten, zal dit nog vijf tot acht jaar duren, voordat de oplossingen daadwerkelijk geïmplementeerd zijn.

Over de mogelijkheden om gebruik te maken van extern ontwikkelde modellen, zijn de respondenten het met elkaar eens. Het is zeer onwaarschijnlijk dat een model direct toepasbaar is binnen Dow. Echter, een extern model kan wel als basis dienen voor een model dat Dow zelf kan ontwikkelen.

Voor het gebruik van externe modellen worden wel voorwaarden gesteld. Zo is het belangrijk dat het model, de gebruikte data en de ontwikkeling van het model, zeer gedetailleerd zijn gedocumenteerd. Dit is nodig om te kunnen bepalen of het überhaupt de moeite waard is, om te onderzoeken of het model van meerwaarde kan zijn voor Dow. En vervolgens om er een vergelijkbaar model van te ontwikkelen, binnen de eigen organisatie.

Er worden enkele kanttekeningen geplaatst, bij de modellen die uit de literatuur kunnen worden onttrokken. Het is dan ook de vraag, of deze modellen van voldoende kwaliteit zijn, om te gebruiken in een industriële omgeving.

Externe data daarentegen, kan wel degelijk een meerwaarde opleveren. Deze data kan worden gebruikt om de eigen datasets aan te vullen en te verrijken. Ook hier geldt, dat de data én de context van de data, gedetailleerd zijn gedocumenteerd. Open databases worden momenteel dan ook al gebruikt binnen Dow.

# 5. Conclusies en Aanbevelingen

## 5.1 Conclusies

## 5.2 Aanbevelingen

## 5.3 Reflectie

### 5.3.1 Methodologische en proces reflectie

### 5.3.2 Persoonlijke reflectie

# 6. Literatuurlijst

[Agra] Agrawal, A., & Choudhary, A. (2016b). Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Materials*, *4*(5), 053208. <https://doi.org/10.1063/1.4946894>

[Alm] De Almeida, A. F., Moreira, R., & Rodrigues, T. (2019). Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry*, *3*(10), 589–604. <https://doi.org/10.1038/s41570-019-0124-0>

[Alp] Alpaydin, E. (2016). *Machine Learning: The New AI (The MIT Press Essential Knowledge series)*. The MIT Press.

[Al-Sa] Al-Sahaf, H., Bi, Y., Chen, Q., Lensen, A., Mei, Y., Sun, Y., Tran, B., Xue, B., & Zhang, M. (2019). A survey on evolutionary machine learning. *Journal of the Royal Society of New Zealand*, *49*(2), 205–228. https://doi.org/10.1080/03036758.2019.1609052

[Ask] Askeland, D. R., Fulay, P. P., & Wright, W. J. (2010). *The Science and Engineering of Materials* (6de editie). CL Engineering.

[Att] Attig, N. (2004). *Computational Soft Matter: from Synthetic Polymers to Proteins*. NIC.

[Baum] Baum, Z. J., Yu, X., Ayala, P. Y., Zhao, Y., Watkins, S. P., & Zhou, Q. (2021). Artificial Intelligence in Chemistry: Current Trends and Future Directions. *Journal of Chemical Information and Modeling*, *61*(7), 3197–3212. <https://doi.org/10.1021/acs.jcim.1c00619>

[Bauer] Bauernhansl, T., Krüger, J., Reinhart, G., & Schuh, G. (2016). WGP-Standpunkt Industrie 4.0, *Wissenschaftliche Gesellschaft für Produktionstechnik WGP e. V.,* Darmstadt.

[Bur] Burke, K. (2012). Perspective on density functional theory. *The Journal of Chemical Physics*, *136*(15), 150901. <https://doi.org/10.1063/1.4704546>

[But] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2

[Cal] Jr., W. C. D., & Rethwisch, D. G. (2013). *Materials Science and Engineering: An Introduction* (9de editie). Wiley.

[Cram] Cramer, C. J., & Cramer, C. J. (2013). *Essentials of Computational Chemistry* (2de editie). Wiley.

[Das] Das, S., Pegu, H., Sahu, K. K., Nayak, A. K., Ramakrishna, S., Datta, D., & Swayamjyoti, S. (2020). Machine learning in materials modeling—fundamentals and the opportunities in 2D materials. In *Synthesis, Modeling, and Characterization of 2D Materials, and Their Heterostructures*(pp. 445-468). Elsevier.

[Dow] Dow.com *Purpose, ambition and values*. (z.d.). Dow Corporate. Geraadpleegd op 20 november 2021, van https://corporate.dow.com/en-us/about/company/ambition-and-values.html

[Geis] Geissbauer, R., Vedso, J. & Schrauf, S., 2016. *Industry 4.0: Building the Digital Enterprise, PwC, London*. Geraadpleegd op 24 augustus 2021, van <https://www.pwc.com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016.pdf>

[Guo] Guo, K., Yang, Z., Yu, C. H., & Buehler, M. J. (2021). Artificial intelligence and machine learning in design of mechanical materials. *Materials Horizons*, *8*(4), 1153–1172. https://doi.org/10.1039/d0mh01451f

[Jens] Jensen, F. (2017). *Introduction to Computational Chemistry* (3rd Edition). Wiley.

[Kit] Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. <https://doi.org/10.1177/2053951714528481>

[Kub] Kubat, M. (2015). *An Introduction to Machine Learning* (1st ed. 2015 ed.). Springer.

[Mehl] Mehlig, B. E. R. N. H. A. R. D. (2019). *Machine learning with neural networks.* arXiv preprint arXiv:1901.05639.

[Mor] Morgan, D., & Jacobs, R. (2020). Opportunities and Challenges for Machine Learning in Materials Science. *Annual Review of Materials Research*, *50*(1), 71–103. <https://doi.org/10.1146/annurev-matsci-070218-010015>

[Poul] Poulson, T., & Walter, L. (2012). *Introduction to Chemistry*. Van Haren Publishing.

[Racc] Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J., & Norquist, A. J. (2016). Machine-learning-assisted materials discovery using failed experiments. *Nature*, *533*(7601), 73–76. https://doi.org/10.1038/nature17439

[Sau] Saunders, M. N. K., Lewis, P., Thornhill, A., Arnoldy, M., & Smitt, P. (2019). *Methoden en technieken van onderzoek*. Pearson.

[Schled] Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M., & Fazzio, A. (2019). From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, *2*(3), 032001. <https://doi.org/10.1088/2515-7639/ab084b>

[Schw]Schwab, K. (2015, 12 december). *The Fourth Industrial Revolution: What It Means and How to Respond*. Foreign Affairs. Geraadpleegd op 25 augustus 2021, van https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution

[Seg] Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*(7698), 604–610. <https://doi.org/10.1038/nature25978>

[Set] Settles, B. (2009). Active learning literature survey.

[Sha] Sha, W., Guo, Y., Yuan, Q., Tang, S., Zhang, X., Lu, S., Guo, X., Cao, Y. C., & Cheng, S. (2020). Artificial Intelligence to Power the Future of Materials Science and Engineering. *Advanced Intelligent Systems*, *2*(4), 1900143. <https://doi.org/10.1002/aisy.201900143>

[Shalev] Shalev-Shwartz, S. (2014). *Understanding Machine Learning: From Theory to Algorithms* (1ste editie). Cambridge University Press.

[Shack] Shackelford, J. F. (2015). *Introduction to Materials Science for Engineers*. Pearson.

[Verhoev] Verhoeven, N. (2018). *Wat is onderzoek?* (6de editie). Boom Lemma.

[Weber] Weber, M. (2011) *Handboek ontwerpgericht wetenschappelijk onderzoek. Wetenschap met effect* (pp.177-202) Chapter: 11. Boom Lemma. Editors: Joan van Aken, Daan Andriessen

[Wei] Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., & Lei, M. (2019b). Machine learning in materials science. *InfoMat*, *1*(3), 338–358. <https://doi.org/10.1002/inf2.12028>

[Xi] Xi, Z. (2021). Functional Nanomaterials Design in the Workflow of Building Machine-Learning Models. *arXiv preprint arXiv:2108.13171.*

[Ye] Ye, W., Chen, C., Wang, Z., Chu, I. H., & Ong, S. P. (2018). Deep neural networks for accurate predictions of crystal stability. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06322-x

[Yuan] Yuan, R., Liu, Z., Balachandran, P. V., Xue, D., Zhou, Y., Ding, X., Sun, J., Xue, D., & Lookman, T. (2018). Accelerated Discovery of Large Electrostrains in BaTiO 3 ‐Based Piezoelectrics Using Active Learning. *Advanced Materials*, *30*(7), 1702884. <https://doi.org/10.1002/adma.201702884>

[Zhu] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/jproc.2020.3004555

[Zur] ŻUrański, A. M., Martinez Alvarado, J. I., Shields, B. J., & Doyle, A. G. (2021). Predicting Reaction Yields via Supervised Learning. *Accounts of Chemical Research*, *54*(8), 1856–1865. https://doi.org/10.1021/acs.accounts.0c00770

# 7. Bijlagen

Bijlage 1 – Onderzoeksopzet

Bijlage 2 - Checklist Systematisch Literatuuronderzoek DV1 - DV2

Bijlage 3 - Werkblad Deelvraag 1

Bijlage 4 – ML-technieken per toepassing

Bijlage 5 - Werkblad Deelvraag 2

Bijlage 6 – Vragenlijst Interviews

Bijlage 7 - Transcript A

Bijlage 8 - Transcript B

Bijlage 9 - Transcript C

Bijlage 10 - Transcript D

Bijlage 11 - Transcript E

Bijlage 12 - Codering transcripten