



Faculteit Bedrijf en Organisatie

Het gebruik van artificiële intelligentie bij de evalutie van bewegingsanalyses

Arjen Trinquet

Scriptie voorgedragen tot het bekomen van de graad van  
professionele bachelor in de toegepaste informatica

Promotor:  
Pieter Van Der Helst  
Co-promotoren:  
Roel De Ridder, Stijn Lievens

Instelling: —

Academiejaar: 2019-2020

Tweede examenperiode



Faculteit Bedrijf en Organisatie

Het gebruik van artificiële intelligentie bij de evalutie van bewegingsanalyses

Arjen Trinquet

Scriptie voorgedragen tot het bekomen van de graad van  
professionele bachelor in de toegepaste informatica

Promotor:  
Pieter Van Der Helst  
Co-promotoren:  
Roel De Ridder, Stijn Lievens

Instelling: —

Academiejaar: 2019-2020

Tweede examenperiode



## Woord vooraf

Dit onderzoek is geschreven vanuit een sterke persoonlijke interesse en vormt voor mij het ultieme sluitstuk van 8 jaar hard werk. In 2017 behaalde ik een masterdiploma in revalidatiewetenschappen en kinesitherapie. Tijdens die opleiding werd het echter al snel duidelijk dat het werkveld op vlak van technologische vooruitgang nog in zijn kinderschoenen staat. Als iemand met een brede technologische achtergrond was het daarom moeilijk om vrede te nemen met de huidige stand van zaken. Om bepaalde tekortkomingen aan te kunnen pakken, heb ik daarom besloten het werk als kinesitherapeut te combineren met de bacheloropleiding toegepaste informatica. In de drie daaropvolgende jaren, heb ik het gebrek aan technologische innovatie in de praktijk aan den lijve mogen ondervinden. Al die tijd zijn verschillende ideeën ontstaan over hoe informatica het werk als therapeut zou kunnen ondersteunen. Uiteindelijk heb ik één van deze ideeën gekozen als finale link tussen twee van mijn grootste interesses. Daarom was ik ook gemotiveerd om deze scriptie uit te werken tot een pioniersonderzoek. Hiermee hoop ik de deur open te zetten voor een verdere samenhang tussen de twee domeinen: meer doelgerichte interventies binnen de kinesitherapie door de technologische mogelijkheden van de informatica.

Hierbij wil ik graag Roel en Stijn bedanken voor de geweldige ondersteuning bij dit proces. Mede dankzij hun input is dit onderzoek geworden tot wat het nu is. Ook Cedric ben ik uitermate dankbaar. Zonder hem had ik nooit de fantastische data kunnen bemachtigen waarop dit onderzoek gebaseerd is. Pieter, mijn promotor die van in het begin geloofde in het idee, bedankt om zoveel geduld uit te oefenen. Vervolgens wil ik de verschillende mensen bedanken die zo vriendelijk geweest zijn om mijn schrijfkunsten te verbeteren. Mijn ouders, die mij blijven ondersteunen zijn ondanks al mijn gekke ideeën, bedankt. Tot slot, Nona, die mij ondertussen al 7 jaar aanzet om het beste uit mezelf te halen, zonder jou was ik hier nooit geraakt.



## **Abstract**

Meer en meer maakt men gebruik van preventieve screeningstesten om blessurerisico's in kaart te brengen. De lateral stepdown is één van die testen. Eén van de zaken waar men tijdens deze test naar kijkt, is de inwaartse verplaatsing van de knie, ook wel de knievalgus genoemd. Onderzoek heeft aangetoond dat onvoldoende lichamelijke controle over deze beweging kan leiden tot ernstige knieblessures zoals een voorstekruisbandscheur. De evaluatie van deze screening gebeurt in de meeste gevallen aan de hand van 2D-videoanalyse. Deze methode is echter traag en deels subjectief. Het doel van dit onderzoek is om te bekijken of men machine learning kan gebruiken om de kwaliteit van deze knievalgusbeweging tijdens de lateral stepdown te bepalen. Op deze manier zou men praktijkevaluaties sneller en objectiever kunnen laten verlopen.

Om dit te onderzoeken werd in eerste instantie een uitgebreid literatuuronderzoek gedaan over zowel de lateral stepdown als de mogelijke implementaties met machine learning. Vervolgens werd een experimenteel onderzoek uitgevoerd om te kijken in welke mate de resultaten van deze implementaties gelinkt kunnen worden aan de evaluaties van ervaren testers. Dit gebeurde aan de hand van een dataset van bijna 300 video's met uitvoeringen van de lateral stepdown. Voor elk van deze screenings werd de piekvalgushoek gestandaardiseerd geëvalueerd door getrainde onderzoekers. Als eerste stap werd op deze dataset een machine learning implementatie toegepast. Op basis van deze gegevens werden vervolgens de nodige piekvalgushoeken bepaald. Nadien werden met deze resultaten twee nieuwe evaluatiemethodes ontwikkeld. De eerste methode evaluateert door gebruik te maken van nieuwe criteria gebaseerd op de berekende piekvalgushoeken, de tweede op basis van een neuraal netwerk. Uiteindelijk werd er onderzocht in welke mate de originele en berekende gegevens verbanden vertonen.

De voorspelde piekvalgushoeken met de hulp van machine learning vertonen een positieve

Pearson correlatie van 0,69 met de originele evaluaties. De evaluatiemethode met nieuwe criteria behaalde een precisie van 0,44 op de originele dataset. De evaluaties met het neurale netwerk zijn inconsistent en ontoereikend.

Deze resultaten tonen aan dat er wel degelijk een verband bestaat tussen de valgushoeken die berekend worden met machine learning en de evaluaties van ervaren testers. De nieuwe evaluatiemethodes vertonen echter onvoldoende overeenkomst met de originele scores. Men kan zich de vraag stellen welke evaluatiemethode het meest representatief is om de kwaliteit van de valgusbeweging te onderzoeken, de originele scorecriteria bevatten namelijk een aantal subjectieve criteria. Programmatisch bepaalde hoeken zijn daarentegen wel volledig objectief en consistent reproduceerbaar. Het staat vast dat verder onderzoek noodzakelijk is om correcte evaluatiecriteria of neurale netwerken te ontwikkelen aan de hand van deze nieuwe methode. Dit onderzoek legt echter een solide basis om in de toekomst machine learning te gebruiken met als doel professionals in het werkveld te ondersteunen.

# Inhoudsopgave

<b>1</b>	<b>Inleiding</b>	<b>15</b>
1.1	Probleemstelling	16
1.2	Onderzoeks vraag	16
1.3	Onderzoeksdoelstelling	17
1.4	Opzet van deze bachelorproef	17
<b>2</b>	<b>Literatuurstudie</b>	<b>19</b>
2.1	Lateral stepdown	19
2.1.1	Testprotocol	19
2.1.2	Knievalgus en injury prediction	20
2.1.3	Valgusevaluatie en FPPA	21
2.1.4	Betrouwbaarheid	22
2.1.5	Evaluatiemogelijkheden	24

<b>2.2 Human pose estimation</b>	<b>25</b>
2.2.1 Gesuperviseerd vs. ongesuperviseerd leren . . . . .	25
2.2.2 Datasets . . . . .	27
2.2.3 Voorspellingsfout vs. voorspellingssnelheid . . . . .	27
2.2.4 Performantie beoordelen . . . . .	28
2.2.5 2D- vs. 3D-keypoint detectie . . . . .	28
<b>2.3 Evolutie van human pose estimation</b>	<b>29</b>
<b>3 Methodologie</b> .....	<b>35</b>
<b>3.1 Dataset</b>	<b>35</b>
3.1.1 Data exploratie . . . . .	36
3.1.2 Data cleaning . . . . .	37
<b>3.2 Keypoint detection</b>	<b>38</b>
<b>3.3 Piekvalgus bepaling</b>	<b>41</b>
3.3.1 Find peaks . . . . .	41
3.3.2 Moving averages . . . . .	42
3.3.3 Grondcontactbepaling . . . . .	43
3.3.4 Afstandsbepaling . . . . .	45
3.3.5 Starten of eindigen op de grond . . . . .	47
<b>3.4 Evaluatie</b>	<b>48</b>
<b>4 Resultaten</b> .....	<b>51</b>
<b>4.1 Dataverdeling</b>	<b>51</b>
4.1.1 Valgusevaluaties . . . . .	51
4.1.2 Piekvalgushoeken . . . . .	51

<b>4.2 Valgusscore bepaling</b>	<b>56</b>
4.2.1 Hypothetische evaluatiecriteria .....	56
4.2.2 Evaluatie adhv machine learning .....	56
<b>5 Discussie .....</b>	<b>59</b>
<b>6 Conclusie .....</b>	<b>63</b>
<b>A Onderzoeksvoorstel .....</b>	<b>65</b>
A.1 Introductie	65
A.2 State-of-the-art	66
A.3 Methodologie	67
A.4 Verwachte resultaten	68
A.5 Verwachte conclusies	68
<b>Bibliografie .....</b>	<b>69</b>



# Lijst van figuren

2.1	Lateral Stepdown .....	20
2.2	Normaal gealigneerde knie links, knievalgus rechts .....	21
2.3	Frontal Plane Projection Angle (Willson, Ireland & Davis, 2006, 5) ..	22
2.4	Valgusscore 0 .....	23
2.5	Valgusscore 1 .....	23
2.6	Valgusscore 2 of 3 .....	23
2.7	Keypoints tijdens de lateral stepdown 1 .....	26
2.8	Keypoints tijdens de lateral stepdown 2 .....	26
2.9	Keypoint heatmaps (Tompson, Goroshin, Jain, LeCun & Bregler, 2014) 28	
2.10	Weergave coördinaten op een afbeelding van 1920x1080 pixels	29
2.11	Keypoint coördinaten op een afbeelding van 1920x1080 pixels .	30
2.12	Onderlinge keypoint relaties adhv verenstructuur .....	31
2.13	Resultaten DeepPose (Toshev & Szegedy, 2013) .....	32
2.14	Stacked hourglass architectuur (Newell, Yang & Deng, 2016) ....	33
2.15	Resultaten stacked hourglass netwerk (Newell, Yang & Deng, 2016) 33	

3.1	Visualisatie van geïnverteerde FPPA berekend op basis van keypoint detection .....	40
3.2	Y-coördinaat van de enkel in functie van het framenummer .....	42
3.3	Pieken bepaald met de findPeaks functie op 3.2 zonder extra parameters .....	43
3.4	FindPeaks op een moving average van 3 frames .....	44
3.5	FindPeaks op een moving average van 5 frames .....	44
3.6	FindPeaks op een moving average van 10 frames .....	44
3.7	Pieken worden gemist wanneer de maximale y-waarde gebruikt wordt	
45		
3.8	Correcte pieken obv moving average van 50 frames .....	46
3.9	Slechts één piek per top adhv een minimale afstand .....	46
3.10	Voorbeeld van test waarbij gestart werd vanop de grond .....	47
3.11	Bepalen van de onderste pieken .....	48
3.12	Voorbeeld van een test met 5 piekmomenten waarvan 2 outliers	48
4.1	Verdeling van de valgusevaluaties .....	52
4.2	Poging met de kleinste minimale valgushoek .....	54
4.3	Poging met de grootste maximale valgushoek .....	54
4.4	Verdeling van de gemiddelde piekvalgushoek .....	54
4.5	Verdeling van de gemiddelde piekvalgushoek per score .....	55
4.6	Confusion matrix van de hypothetische evaluatiecriteria .....	57
A.1	Lateral stepdown .....	66
A.2	Valgushoek .....	67

# Lijst van tabellen

3.1	Data cleaning proces deel 1 .....	37
3.2	Resultaten van 4 voorgetrainde HRNet modellen op de COCO dataset 39	
3.3	Data cleaning proces deel 2 .....	41
4.1	Verdeling van de valgusevaluaties .....	52
4.2	Samenvatting berekende piekmomenten per video .....	53
4.3	Pearson correlatiecoëfficiënten van de verschillende piekvalgushoe- ken tov de originele score .....	55
4.4	Hypothetische valgusscore criteria .....	56



# 1. Inleiding

“An ounce of prevention is worth a pound of cure.”

— Benjamin Franklin

Iedereen kent het belang van preventieve maatregelen om grotere negatieve gevolgen te vermijden. In de praktijk is dit jammer genoeg iets wat regelmatig vergeten of genegeerd wordt. Vaak liggen tijd, geld of andere ongemakken aan de basis van de verwaarlozing van preventieve activiteiten. Recente technologische ontwikkelingen zoals artificiële intelligentie en computer vision hebben het echter mogelijk gemaakt om vervelende of repetitieve taken te automatiseren en verbeteren. De belangrijkste sector voor preventie is mogelijks de gezondheidszorg. Gezien de vaak strenge regelgeving en delicate ethische kwesties is dit eveneens de sector waar nog veel verbeteringen mogelijk zijn. Ook hier heeft artificiële intelligentie nog niet overal zijn weg gevonden om de praktijk te ondersteunen.

Een van de meest voorkomende en ernstige knieletsels bij contact- of hoge intensiteit-sporten is de voorste kruisbandscheur, of afgekort de VKB scheur. Sporters zoals skiërs of voetballers bij wie veel dynamische krachten inwerken op het kniegewricht zijn frequent het slachtoffer van deze letsels. Het gevolg is een lange revalidatie met een zware socio-emotionele impact. Tijdens het herstel zijn vele uren kinesitherapie noodzakelijk om het lichaam opnieuw klaar te stomen voor de sporthervervulling. Zo zijn intensieve kracht- en stabilisatietraining uitermate belangrijk om optimaal te kunnen hervatten en hervallen te voorkomen. Bij dit soort letsels wordt het duidelijk dat preventie zeer waardevol is, zowel op persoonlijk als socio-economisch vlak. Onnodige, vaak hoog oplopende ziekenhuis- en revalidatiekosten kunnen zo vermeden worden.

Bewegings- en krachttjesten worden vaak gebruikt als kostenefficiënte manieren om risicogroepen te screenen. De lateral stepdown of single leg squat is een van deze bewe-

gingstesten. Hierbij wordt onderzocht of zij een verhoogd risico hebben voor het krijgen van een VKB scheur of andere knieblessures. Tijdens de test worden verschillende evaluatiecriteria onderzocht. Aan de hand van deze factoren kunnen correcte interventies ondernomen worden om blessurerisico's te minimaliseren. Een belangrijk parameter die onderzocht wordt is het bewegen van de knie naar binnen, ook wel de knievalgus genoemd. Een overmatige en ongecontroleerde valgusbeweging is één van de oorzaken voor VKB scheuren waargenomen in wetenschappelijk onderzoek (Hewett e.a., 2005, 4). Gezien de dynamische uitvoering en beperkte verplaatsing van de knie, is het echter niet evident om deze beweging correct te beoordelen.

Er bestaan verschillende methodes om de hoeveelheid knievalgus te meten tijdens de test. 2D- of 3D-video opnames kunnen gebruikt worden om de beweging te vertragen en achteraf te evalueren. Het gebruik hiervan in de praktijk is echter beperkt. De grote investeringen die noodzakelijk zijn om 3D-videocamera's te voorzien en de tijd die de verwerking in beslag neemt, zorgen ervoor dat het gebruik bij deze testen eerder beperkt blijft. Als alternatief worden ze vaak met het blote oog geanalyseerd. Dit kan leiden tot een lage inter- en intratesterbetrouwbaarheid. In het slechtste geval zorgt dit ervoor dat de preventieve testen helemaal niet gebruikt worden. Het doel van deze bachelorproef is te onderzoeken of de mogelijkheid bestaat om de knievalgusbeweging objectief en betrouwbaar in kaart te brengen tijdens de lateral stepdown, met het gebruik van artificiële intelligentie en computer vision.

## 1.1 Probleemstelling

Het beperkte gebruik van preventieve bewegingstesten kan leiden tot een grotere incidentie van zware blessures. Personen uit de risicogroepen, patiënten, kinesitherapeuten die deze testen uitvoeren en de gezondheidszorg in het algemeen kunnen baat hebben bij het verbeteren van dit soort screenings.

In het werkveld vragen strikt objectieve metingen vaak veel tijd en energie. Om die reden worden vaak een aantal parameters subjectiever gemaakt om deze sneller te kunnen beoordelen. Indien men de tools nodig voor objectieve evaluaties toegankelijker kan maken, zouden clinici misschien meer geneigd zijn deze te gebruiken. Hierdoor kunnen de risicogevallen sneller geïdentificeerd worden en kan men preventief de nodige interventies uitvoeren.

## 1.2 Onderzoeksvraag

Om dit probleem te onderzoeken werden een onderzoeksvraag en een aantal deelvragen geformuleerd. Aan de hand van een literatuurstudie en een experimenteel onderzoek zal deze scriptie op deze vragen proberen te antwoorden.

- Kan machine learning gebruikt worden om de kwaliteit van de valgusbeweging tijdens de lateral stepdown te bepalen? Of op welke manier kan het evalueren

hiermee ondersteund worden?

- Op welke manieren kan de valgus tijdens de lateral stepdown momenteel geëvalueerd worden?
- Hoe kan men programmatisch een representatief beeld van de valgus voor de gehele test bekomen?
- Kunnen op basis van de eventuele resultaten correcte scorecriteria of neurale netwerken worden opgesteld?

## 1.3 Onderzoeksdoelstelling

Het voornaamste doel van deze scriptie is om te onderzoeken of er bestaande toepassingen van artificiële intelligentie gebruikt kunnen worden om de evaluatie van de lateral stepdown te ondersteunen. De focus hierbij ligt op de knievalgus tijdens deze test. Indien dit zo zou zijn, zal bekeken worden of er op basis van deze criteria objectieve testparameters kunnen worden opgesteld. Al deze doelen worden onderzocht aan de hand van een experimenteel onderzoek.

## 1.4 Opzet van deze bachelorproef

Het vervolg van deze scriptie is als volgt opgebouwd:

In Hoofdstuk 2 wordt eerst en vooral een globale context geschatst van de huidige stand van zaken. Hier worden zowel de lateral stepdown als de mogelijke toepassingen met artificiële intelligentie besproken.

In Hoofdstuk 3 wordt de gebruikte methodologie toegelicht. Hier worden de onderzoeks-technieken besproken waarmee er gepoogd is een antwoord te geven op de onderzoeks-vragen.

Hoofdstuk 4 wordt gebruikt om de resultaten van dit onderzoek toe te lichten.

Vervolgens worden in Hoofdstuk 5 deze resultaten besproken. De toepassingen samen met de mogelijke limitaties van het onderzoek worden overlopen.

Tot slot wordt in Hoofdstuk 6 de finale conclusie van dit onderzoek beschreven.



## **2. Literatuurstudie**

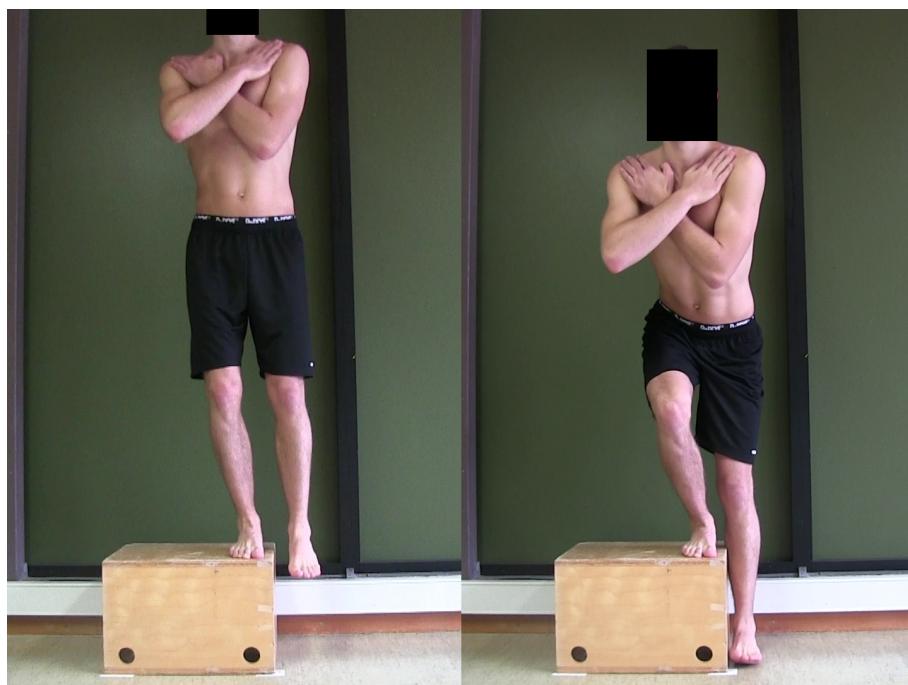
Zoals eerder vermeld focust dit onderzoek zich op de bewegingsanalyse van de lateral stepdown test. In dit gedeelte wordt eerst deze test en de relevantie hiervan uitvoerig besproken. Bovendien wordt er gekeken welke methodes er voor handen zijn om de kwaliteit van de valgusbeweging in kaart te brengen. Nadien worden er linken gelegd met machine learning en computer vision. Ook de huidige stand van zaken wat betreft deze technieken komt in dit onderdeel aan bod. Op basis hiervan werd een neuraal netwerk uitgekozen dat gebruikt werd in het experimenteel onderzoek.

### **2.1 Lateral stepdown**

#### **2.1.1 Testprotocol**

De lateral stepdown is een test die in een klinische setting vaak gebruikt wordt om bewegingspatronen van de onderste ledematen te onderzoeken. De Blaiser, De Ridder, Willems, Danneels en Roosen (2019, 6) hebben een testprotocol ontwikkeld op basis van bestaande protocols (Chmielewski e.a., 2007, 3; Piva e.a., 2006; Rabin & Kozol, 2010, 12), waarmee men op een consistente en betrouwbare wijze een evaluatie kan uitvoeren. De testpersoon wordt op één been op de rand van een box geplaatst met een hoogte van 30 centimeter. Wanneer de persoon kleiner is dan 170 centimeter wordt een box van 25 centimeter gebruikt. Belangrijk hierbij is dat dit gebeurt met de voet voorwaarts gericht. Tijdens het uitvoeren van de oefeningen wordt de tegenovergestelde knie volledig gestrekt en de heup aan deze zijde ook licht gebogen. De tester vraagt de persoon om de armen over de borst te kruisen en deze niet te gebruiken om het evenwicht te bewaren. Terwijl men voor zich uit blijft kijken plooit men de knie op een gecontroleerde manier tot de

hiel van het vrije been de grond raakt. Wanneer dit gebeurt is het belangrijk dat er geen gewichttransfer optreedt op de grond. Vervolgens keert men terug naar de startpositie en wordt de oefening herhaald (Figuur 2.1). Bijkomend worden er instructies gegeven om tijdens de gehele oefening de knie op dezelfde lijn te houden als de lijn die gevormd wordt tussen de heup en de voet. De testpersoon wordt gevraagd om een aantal herhalingen uit te voeren. Hierbij mag men zelf een comfortabele snelheid kiezen.



Figuur 2.1: Lateral Stepdown

### 2.1.2 Knievalgus en injury prediction

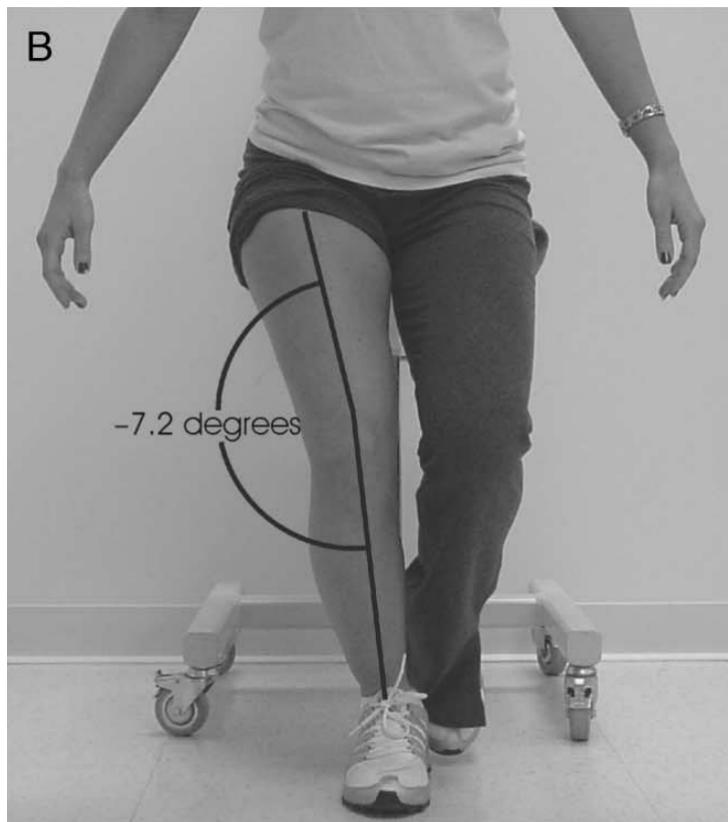
Deze test wordt voornamelijk gebruikt om te beoordelen of de verschillende segmenten van de onderste ledematen correct gealigneerd zijn. Hierbij wordt onder andere gekeken naar de mediale verplaatsing, of verplaatsing naar het midden toe, van de knie. Dit wordt ook wel de knievalgus genoemd (Figuur 2.2). Onderzoek heeft aangetoond dat de knievalgushoek gecorreleerd kan worden met verschillende knieletsels (Hewett e.a., 2005, 4; Ireland, 1999, 2; Souza & Powers, 2009, 1). Zo vonden Hewett e.a. (2005, 4) dat de knievalgusbeweging een voorspeller is van VKB letsels bij vrouwelijke atleten. Veel mensen vertonen onvoldoende controle over deze beweging. Russell, Palmieri, Zinder en Ingersoll (2006, 2) ontdekten bijvoorbeeld een grotere knievalgushoek tijdens een drop jump bij vrouwelijke atleten. Daarom worden testen zoals de lateral stepdown in de praktijk vaak gebruikt om deze beweging in kaart te brengen. Met gerichte interventies is het namelijk mogelijk om de musculaire controle te verbeteren en zo het risico op blessures te verkleinen (Bell, Oates, Clark & Padua, 2013, 4).



Figuur 2.2: Normaal gealigneerde knie links, knievalgus rechts

### 2.1.3 Valgusevaluatie en FPPA

Aangezien knievalgus een 3D-beweging is die bestaat uit externe knierotatie, adductie en interne rotatie van de heup wordt 3D-videoanalyse gezien als de gouden standaard wanneer het aankomt op het evalueren van de valgus. Zo zijn er verschillende onderzoeken die hiervan gebruik maken om de dynamische valgus te evalueren (Chappell, Yu, Kirkendall & Garrett, 2002, 2; Ford, Myer & Hewett, 2003, 10; Malinzak, Colby, Kirkendall, Yu & Garrett, 2001, 5; Scott G. McLean, Lipfert & van den Bogert, 2004, 6). Clinici die deze functionele bewegingen willen onderzoeken beschikken over het algemeen echter niet over dit soort apparatuur. 3D-videoapparatuur is duur, neemt veel plaats in en de verwerking van de resultaten neemt veel tijd in beslag. Het gebruik van 2D-videobeelden kan hierdoor een heel stuk interessanter zijn in de praktijk. Onderzoek van Willson, Ireland en Davis (2006, 5) heeft voor de analyse van 2D-beelden tijdens een single leg squat, een oefening die heel sterk aanleunt bij de lateral stepdown, de ‘frontal-plane projection angle’ (FPPA) geïntroduceerd (Figuur 2.3). Deze 2D-hoek is reeds in verschillende onderzoeken met variabele populaties gebruikt als testparameter om de valgus te onderzoeken (Herrington, 2014, 2; Munro, Herrington & Comfort, 2012, 4). Hij wordt gevormd door de lijn van het midden van de heup naar het kniegewricht en de lijn van het kniegewricht naar het midden van de enkel. Negatieve waarden wijzen op knievalgus terwijl positieve waarden een knievarus of lateralisatie van de knie aantonen. Verschillende studies hebben de validiteit van de FPPA van de knie vergeleken met bestaande 3D-technieken (S. G. McLean e.a., 2005, 6; Willson & Davis, 2008, 10). Resultaten toonden aan dat de 2D-FPPA sterk gecorreleerd is met zowel externe knierotatie als heupadductie, beiden onderdeel van een dynamische knievalgusbeweging. Uit het onderzoek blijkt dat 2D-videoanalyses geen



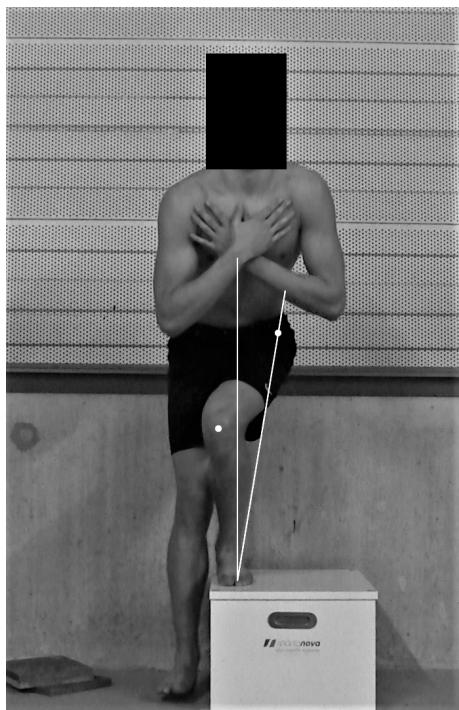
Figuur 2.3: Frontal Plane Projection Angle (Willson, Ireland & Davis, 2006, 5)

volwaardige vervanging zijn voor 3D-analyses omdat ze minder sterk zijn in het vastleggen van de rotatie component (Willson & Davis, 2008, 10). Ze zijn wel nuttig als screening tool voor het zoeken naar personen met een vergroot risico op bepaalde blessures (S. G. McLean e.a., 2005, 6). Onderzoek heeft bijvoorbeeld aangetoond dat patiënten met patellofemorale pijnklachten een significant grotere FPPA vertonen tijdens de single leg squat (Herrington, 2014, 2).

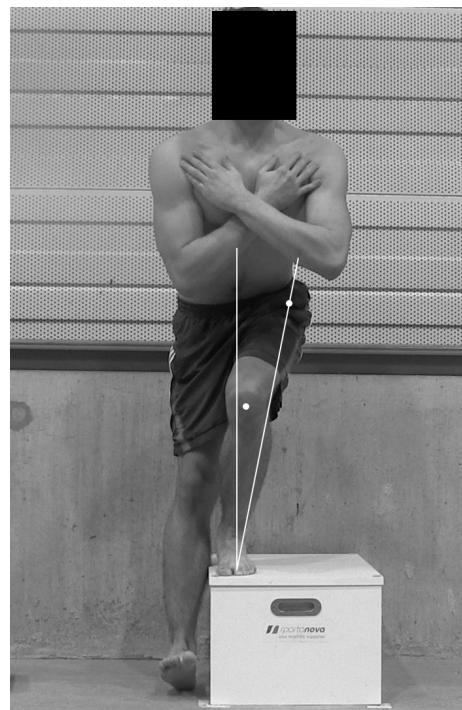
Naast het gebruik van de FPPA zijn er ook andere methoden ontwikkeld om de valgushoek te classificeren. Zo maken De Blaiser e.a. (2019, 6) gebruik van projectielijnen om de valgushoek een score te geven van 0 tot 3. Hoe hoger deze score hoe beter de uitvoering. Wanneer de knie zich op de lijn bevindt die gevormd wordt tussen de heup en de tweede teen (MT2) en stabiel blijft tijdens de uitvoering, krijgt de testpersoon een score 3. Wanneer de knie onstabiel is krijgt deze een score 2. Indien de knie zich recht boven de tweede teen bevindt, krijgt de testpersoon slechts een score 1. Als de knie ten slotte mediaal van de tweede teen gaat, wordt een score 0 gegeven (Figuren 2.4, 2.5 en 2.6). Al deze methodes zijn ontwikkeld om de grootte van de FPPA in verschillende gradaties in te delen.

#### 2.1.4 Betrouwbaarheid

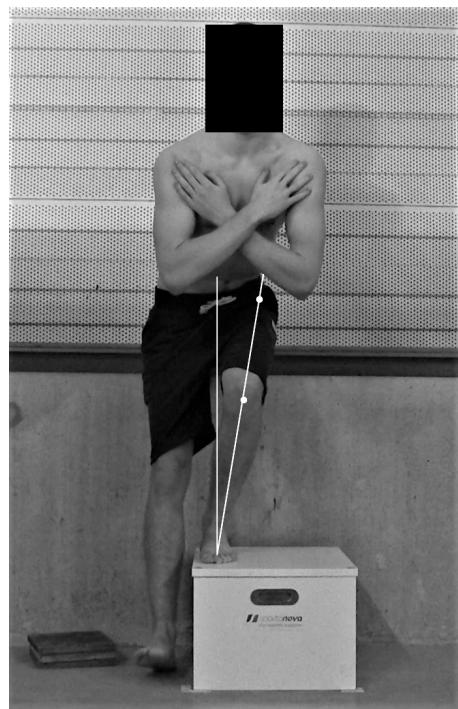
Een ander belangrijk aspect voor het gebruik van dit soort screenings is de test-hertest betrouwbaarheid. Dit wil zeggen dat de resultaten gelijkaardig blijven, wanneer een test



Figuur 2.4: Valgusscore 0



Figuur 2.5: Valgusscore 1



Figuur 2.6: Valgusscore 2 of 3

bij dezelfde persoon op een ander moment wordt uitgevoerd. Om dit in kaart te brengen wordt gebruik gemaakt van de kappa-waarde of de ICC (intraclasscorrelatiecoëfficiënt). Deze waarden geven weer in welke mate meerdere observaties van dezelfde onderzoeker overeenkomen, maar kunnen ook een beeld schetsen van de overeenstemming tussen twee of meer observatoren. Een uitkomst van 1 toont een perfecte overeenstemming aan. De waarde 0 geeft weer dat de observatie niet beter zou zijn dan wanneer deze volledig door toeval wordt bepaald. Volgens McHugh (2012) komen waarden minder dan 0,4 overeen met slechte correlatie, tussen 0,4 en 0,6 met gemiddelde correlatie, tussen 0,6 en 0,8 met goede correlatie en meer dan 0,8 met zeer sterke correlatie. Onderzoek heeft aangetoond dat de lateral stepdown over een goede inter- en intratesterbetrouwbaarheid beschikt (De Blaiser e.a., 2019, 6). Uit deze vaststellingen kan besloten worden dat valgusscreening tijdens een lateral stepdown aan de hand van 2D-beelden een betrouwbare manier is om preventief risicotatiënten te screenen.

### 2.1.5 Evaluatiemogelijkheden

Wanneer deze test in de praktijk gebeurt, wordt de beweging vaak gefilmd met een camera of een mobiele telefoon vanuit het frontale vlak. Vervolgens kan men de beelden visueel beoordelen of aan de hand van speciale software. Deze software wordt gebruikt om alle relevante herkenningspunten aan te duiden die nodig zijn om de FPPA, of de frontal-plane projection angle, te berekenen. Beide manieren hebben zowel voor- als nadelen. Ondanks de goede inter- en intratesterbetrouwbaarheid toont onderzoek aan dat de sensitiviteit (de mate waarin een beoordeling correct geïdentificeerd wordt) van valgusscreening soms beperkt is. Wanneer evaluaties gebeuren met het blote oog en op korte tijd, wordt een significant deel van de vergrote risico's gemist (Ekegren, Miller, Celebrini, Eng & Macintyre, 2009, 9). Dit kan men oplossen door het aanduiden van herkenningspunten met gesofisticeerde software. Het grote nadeel van deze manier van werken is de tijdinvestering. Het doel van de lateral stepdownscreening is om snel en zonder veel moeite de belangrijkste risico's te kunnen herkennen. Wanneer men na elke test verschillende videobewerkingen moet uitvoeren om een betrouwbaar resultaat te krijgen, gaat de snelheid verloren. Ook zijn niet alle clinici voldoende geschoold in het gebruik van dit soort software wat de toepassing ervan bemoeilijkt. Dit is waar machine learning een belangrijke rol kan spelen. Tot op de dag van vandaag is er nog geen onderzoek gedaan naar de haalbaarheid van het gebruik van machine learning in dit soort screenings. Nochtans zouden alle belangrijke keypoints of gewrichten samen met de FPPA mogelijk bepaald kunnen worden aan de hand van een speciaal getraind model. Dit kan eventueel leiden tot een significante toename in snelheid van de evaluatie zonder (de aankoop van) speciale apparatuur. Een eerste belangrijke stap in dit proces is het bepalen van alle relevante gewrichten of keypoints. Zoals eerder vermeld, heeft onderzoek aangetoond dat de FPPA betrouwbaar is om een beeld te krijgen van de valgushoek. Daarom zullen voor deze use case voornamelijk het heupgewricht, de knie en de enkel onderzocht moeten worden. In het volgende deel wordt besproken hoe machine learning kan helpen om dit probleem op te lossen.

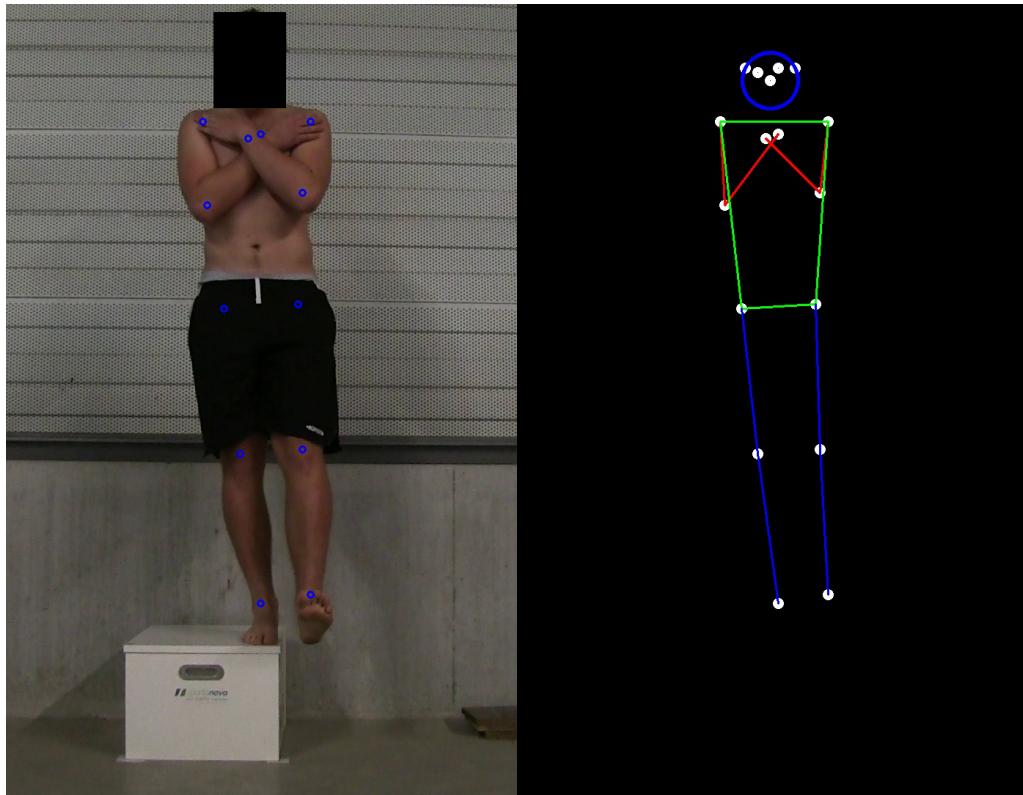
## 2.2 Human pose estimation

Human pose estimation is het proces om de configuratie van een lichaam te achterhalen aan de hand van een digitaal beeld of video. Dit gebeurt vaak aan de hand van keypoint detection ofwel de herkenning van belangrijke punten. In dit geval worden de grote gewrichten van een lichaam in kaart gebracht. Voor het onderste lidmaat zijn dit de heupen, knieën en enkels. Voor het bovenste lidmaat de schouders, ellebogen en polsen. Tot slot zullen er meestal ook een aantal herkenningspunten van het hoofd bepaald worden. Dit zijn bijvoorbeeld de neus, ogen en/of oren. Wanneer deze sleutelpunten op een correcte manier verbonden worden zullen deze in grote lijnen de vorm aannemen van een menselijk lichaam. Figuren 2.7 en 2.8 zijn voorbeelden van deze keypoints tijdens de lateral stepdown.

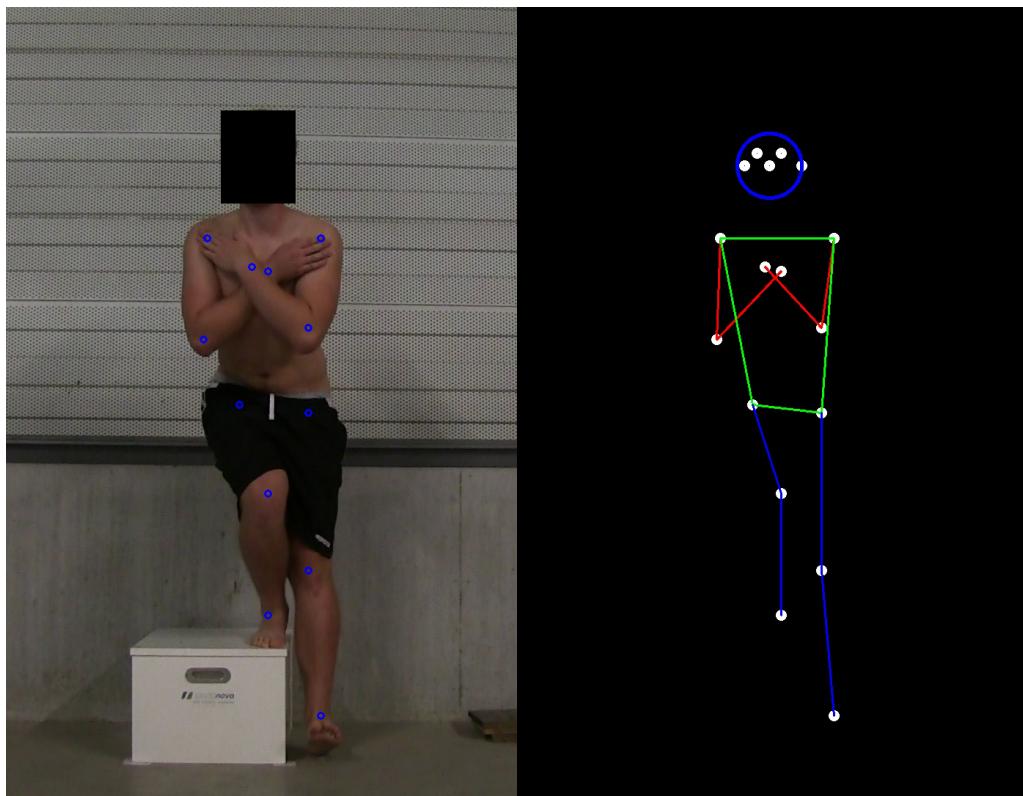
### 2.2.1 Gesuperviseerd vs. ongesuperviseerd leren

Keypoint detection en andere machine learning problemen worden opgelost aan de hand van neurale netwerken of getrainde modellen. In het domein van de machine learning kan er een onderscheid gemaakt worden tussen twee leermethodes, gesuperviseerd en ongesuperviseerd leren. Beide methodes kan men gebruiken om een netwerk te trainen en een probleem op te lossen. Welke methode het meest geschikt is voor het probleem hangt af van verschillende factoren. Bij de eerste manier, het gesuperviseerd leren, zijn er op voorhand voorbeelden beschikbaar met een correct resultaat, ook wel de ‘ground truth’ genoemd. Deze kunnen gebruikt worden om onder begeleiding het model te leren hoe het deze waarden kan voorspellen. Er is dus met andere woorden altijd een gelabelde dataset nodig. Wanneer dit toegepast wordt op het volgende probleem: het herkennen of een afbeelding een persoon bevat of niet, wil dit zeggen dat er een dataset nodig is die bestaat uit foto’s waarbij elke foto een label heeft: ‘persoon’ of ‘geen persoon’. Hierbij zal een deel van de dataset gebruikt worden om het model te leren welke afbeeldingen een persoon bevatten. Het andere deel wordt gebruikt om na elke leer-iteratie of epoch te controleren in welke mate de voorspellingen correct waren. Aan de hand van de trainingsdata stuurt het netwerk zichzelf bij om telkens een beter resultaat te krijgen. Een belangrijk aandachtspunt bij gesuperviseerd leren is de mate waarin een model afhankelijk wordt van de trainingsdata, ook wel de variantie genoemd. Wanneer men een netwerk niet lang genoeg traaint zal het mogelijks onvoldoende relaties herkennen en een lage variantie hebben. Dit fenomeen noemt men underfitting, veel zaken zullen dus niet herkend worden. Wanneer een model te lang wordt getraind of de dataset onvoldoende gevarieerd is zal er een hoge variantie zijn. Dit wil zeggen dat het model té specifiek getraind is op de trainingsdata en dus niet algemeen genoeg zal voorspellen. Dit noemt men overfitting. Hierbij zullen inputs in een andere context dan de trainingsdata slecht herkend worden.

Binnen deze categorie kan een verder onderscheid gemaakt worden tussen classificatieproblemen en regressieproblemen. Bij een regressieprobleem zal men voor elke input één output proberen bepalen. Een voorbeeld hiervan is het voorspellen van de marktwaarde van een woning aan de hand van zijn kenmerken. Een classificatieprobleem resulteert in twee of meer outputs voor elke input. Wanneer men slechts twee outputs heeft, spreekt men van een binair classificatieprobleem, zoals het eerder aangehaalde voorbeeld over



Figuur 2.7: Keypoints tijdens de lateral stepdown 1



Figuur 2.8: Keypoints tijdens de lateral stepdown 2

het herkennen van een persoon in een afbeelding. Hier zijn de outputs ‘persoon’ of ‘geen persoon’. Meer dan twee outputs noemt men multiclass classificatie.

Bij ongesuperviseerd leren beschikt men niet over een gelabelde dataset. Het netwerk heeft als het ware geen voorbeelden waaruit het kan leren. Het model moet dus zelf op zoek gaan naar een bepaalde structuur in de dataset. Voor problemen waarbij een gelabelde dataset beschikbaar is of waarbij deze te moeilijk is om samen te stellen, kunnen ongesuperviseerde leermethodes gebruikt worden. Een belangrijk nadeel is echter dat er geen specifieke manier is om de performantie van het netwerk te controleren. Deze methode wordt in de praktijk vaak gebruikt om een verkennende analyse uit te voeren en een hypothese te testen (Soni, 2018).

Aangezien het voor dit onderzoek noodzakelijk is dat de performantie van het netwerk kan worden nagegaan, is gesuperviseerd leren noodzakelijk. Doordat er bij keypoint detection meerdere outputs zijn, is dit dus een multiclass classificatieprobleem.

## 2.2.2 Datasets

Zoals eerder vermeld is het bij gesuperviseerd leren noodzakelijk om te beschikken over een gelabelde dataset. Op het moment van schrijven zijn er een aantal voornamme datasets ter beschikking met gelabelde keypoints. De twee belangrijkste zijn de COCO (Common Objects in Context) (Lin e.a., 2014) en MPII (Max Planck Institut Informatik) dataset (Andriluka, Pishchulin, Gehler & Schiele, 2014). COCO bestaat uit 330.000 afbeeldingen die in totaal meer dan 250.000 personen bevatten waarbij de keypoints aangeduid zijn. Hiertussen zitten dus ook veel extra afbeeldingen die minder relevant zijn voor human pose estimation. Zo hebben afbeeldingen van objecten of van dieren in dit geval minder nut. De COCO-dataset is zeer uitgebreid en kan onder meer nog gebruikt worden voor object detectie, segmentatie en herkenning binnen een context. In de MPII zijn er slechts 40.000 geannoteerde personen ter beschikking, verspreid over 25.000 afbeeldingen. Er wordt wel telkens een extra indeling gemaakt op basis van de activiteit die de persoon uitvoert. Deze dataset is specifiek samengesteld voor keypoint detection met computer vision. Beide datasets vormen dus een solide basis voor het trainen van netwerken voor human pose estimation. Verder onderzoek focust zich daarom ook op modellen die gebruik maken van gesuperviseerd leren met ten minste één van deze datasets.

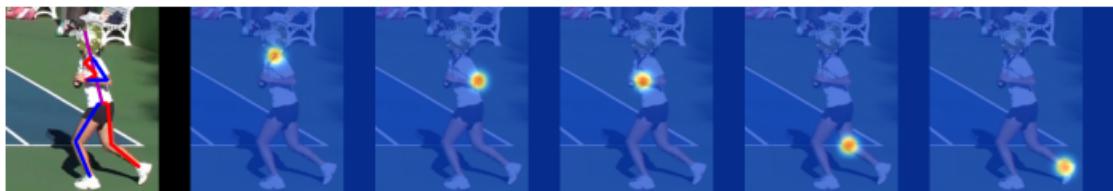
## 2.2.3 Voorspellingsfout vs. voorspellingssnelheid

Wanneer neurale netwerken getraind worden, moet men rekening houden met twee belangrijke doelen, de voorspellingsfout en voorspellingssnelheid. Het ultieme doel van een neurale netwerk is een lage voorspellingsfout met een hoge voorspellingssnelheid. Om deze doelen te beïnvloeden worden de design parameters aangepast. Deze design parameters zijn de bouwstenen van een neurale netwerk. Elk netwerk bestaat uit een input- en een outputlaag, een aantal verborgen lagen en een hoeveelheid neuronen. Door te experimenteren met deze parameters kunnen de snelheid en voorspellingsfout beïnvloed worden. Hoe meer neuronen en lagen een input zal moeten doorlopen, hoe trager een

netwerk zal voorspellen. Het doel hiervan is om te onderzoeken of human pose estimation kan gebruikt worden om menselijke evaluaties te vervangen. Hierbij is het dus uitermate belangrijk dat de precisie van de voorspelling zo groot mogelijk is. Een grote foutenmarge kan al snel exponentiële invloed hebben op de uiteindelijke berekening van de valgushoek. Hierdoor zouden resultaten compleet onbetrouwbaar zijn en geen volwaardige vervanging vormen voor manuele keypoint bepaling. Door het belang van de nauwkeurigheid, zal bij het vergelijken van verschillende bestaande netwerken de precisie doorslaggevend zijn.

## 2.2.4 Performantie beoordelen

Keypoint detection is een multiclass classificatieprobleem waar modellen getraind kunnen worden aan de hand van gesuperviseerd leren. In deze paragraaf zal besproken worden hoe een neurale netwerk deze keypoints in kaart brengt en hoe de performantie kan worden beoordeeld. De input voor deze use case zal steeds een afbeelding zijn. Bij elke afbeelding wordt per pixel of groep van pixels bekeken hoe groot de kans is dat deze regio een keypoint is. Dit resulteert in een heatmap (cfr. infrarood warmtebeelden), waarvan de kleur de waarschijnlijkheid weerspiegelt (Figuur 2.9). De plaats op de heatmap met de grootste



Figuur 2.9: Keypoint heatmaps (Tompson, Goroshin, Jain, LeCun & Bregler, 2014)

waarschijnlijkheid wordt vervolgens teruggegeven als een coördinaat. Elke afbeelding kan namelijk gezien worden als een rooster met een x- en een y-as waarbij iedere pixel een coördinaat op dit rooster is. De oorsprong van de assen is telkens in de linker-bovenhoek (Figuren 2.10 en 2.11). Om de correctheid van het resultaat te beoordelen, wordt frequent gebruik gemaakt van de AP (average precision) of AR (average recall). Precision of precisie is het aantal echt positieve (true positive) op het totale aantal positief bepaalde punten (echt positief + vals positief). De recall of sensitiviteit is het aantal echt positieve op het totaal werkelijk positieve (echt positief + vals negatief). Om een goede performantie van een model te hebben is het belangrijk dat beide waarden zo hoog mogelijk zijn. Wanneer bijvoorbeeld een netwerk een goede score heeft voor de AP maar een lage score heeft voor de AR, wil dit zeggen dat er veel positieve waarden incorrect als negatief geïdentificeerd werden. Omgekeerd geldt wanneer de AR hoog is en de AP laag dat er veel negatieve waarden als positief werden geïdentificeerd.

## 2.2.5 2D- vs. 3D-keypoint detectie

In de classificatie van netwerken voor keypoint detection kan ook een onderscheid gemaakt worden tussen 2D- en 3D-herkenning. 3D-modellen kunnen een inschatting maken van

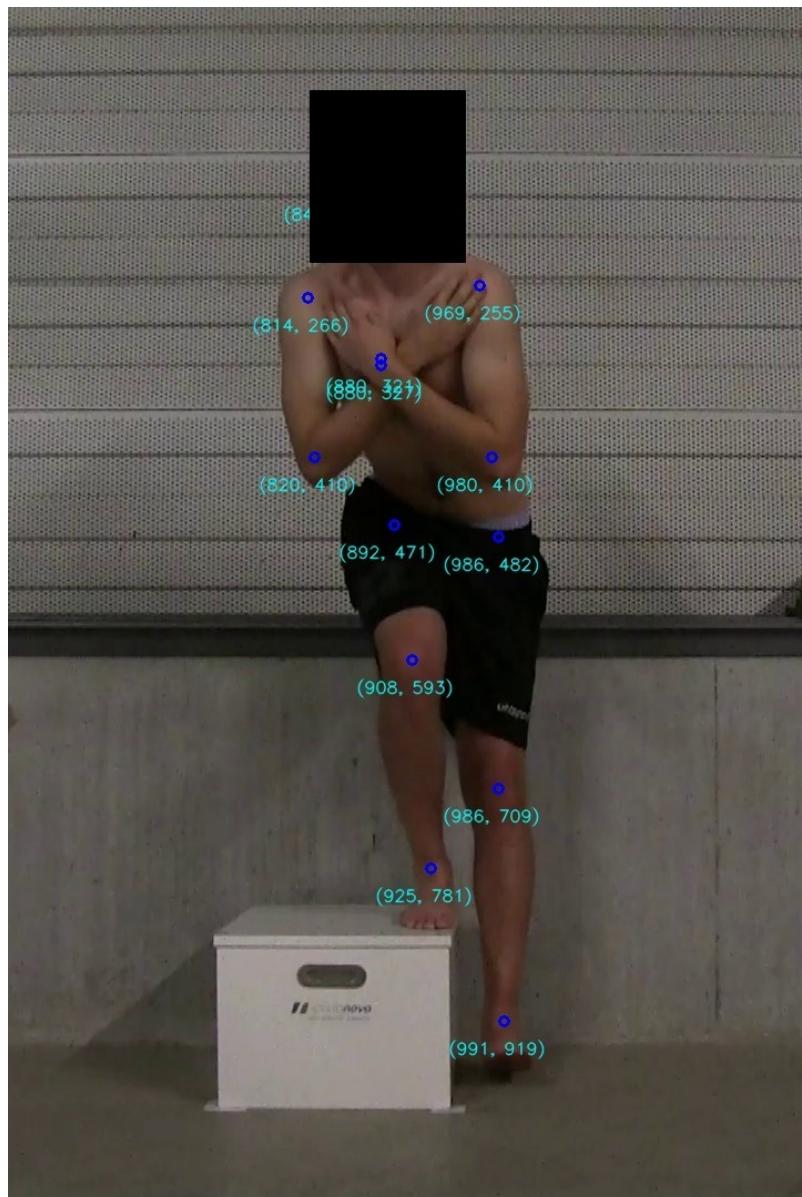


Figuur 2.10: Weergave coördinaten op een afbeelding van 1920x1080 pixels

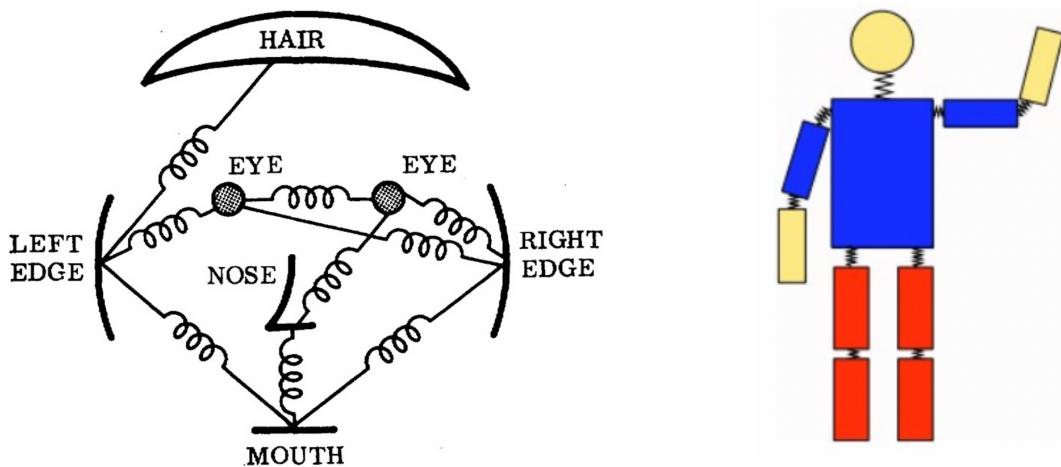
de relatieve diepte van een keypoint in de afbeelding. Hier wordt dus de coördinaat geprojecteerd op drie assen: de x-, y- en z-as. Bij 2D-detectie zal er geen herkenning van de diepte plaatsvinden en de coördinaat enkel op de x- en y-as geplaatst worden. Zoals eerder vermeld is de FPPA een betrouwbare 2D-projectie die een realistische inschatting geeft van de werkelijke valgushoek. Dit wil zeggen dat zowel modellen die gebruik maken van 2D- als van 3D-voorspellingen relevant zijn voor dit onderzoek. Belangrijk om te vermelden is dat zowel de COCO als de MPII dataset bij het annoteren van de data geen gebruik maken van de z-as voor de coördinaten van een keypoint. Er wordt dus enkel een 2D-coördinaat op de afbeelding geplaatst, zonder de absolute diepte op te lijsten. Wat wel wordt weergegeven is of een keypoint zichtbaar is in de afbeelding of dat het zich achter andere elementen bevindt. Zichtbare keypoints krijgen een waarde 1, onzichtbare de waarde 0. De vraag kan dus gesteld worden in welke mate 3D-coördinaten, voorspeld door 3D-netwerken, representatief zijn voor de werkelijkheid. De datasets die gebruikt worden voorzien namelijk geen trainingsdata waar het model kan van leren. Dit is de reden waarom 3D-modellen doorgaans geen gebruik maken van gesuperviseerd leren maar van semi-gesuperviseerde leermethodes. De 2D-annotaties van de betreffende datasets zullen gebruikt worden als ‘ground truth’, vervolgens zal via een ongesuperviseerde methode gezocht worden naar de diepte van de punten. Het spreekt voor zich dat de complexiteit van 3D pose estimation vele malen groter is dan 2D pose estimation en dus ook een heel stuk moeilijker op te lossen is.

## 2.3 Evolutie van human pose estimation

In het volgende onderdeel zullen een aantal bestaande modellen voor pose estimation en hun performantie besproken worden. Gezien de complexiteit van de 3D pose estimation



Figuur 2.11: Keypoint coördinaten op een afbeelding van 1920x1080 pixels



Figuur 2.12: Onderlinge keypoint relaties adhv verenstructuur

en het ontbreken van een ground truth zal gefocust worden op de 2D-varianten.

Al decennialang proberen onderzoekers het probleem van de human pose estimation op te lossen en de precisie van bestaande modellen te verbeteren. Oorspronkelijk werd een methode ontwikkeld die het lichaam voorstelde als verzameling van onderdelen. Al deze onderdelen werden verbonden met denkbeeldige veren. Het resultaat was een structuur waarbij onderlinge relaties tussen onderdelen werden afgebeeld (Figuur: 2.12). Een groot nadeel van deze manier van werken is dat het model geen rekening houdt met contextuele informatie uit de afbeelding. Door de opkomst van CNN's (convolutional neural networks) zijn er sinds 2013 steeds meer methodes ontwikkeld die de nauwkeurigheid van de voorspellingen significant verbeterden. Deep Pose was het eerste grote onderzoek dat deep learning heeft toegepast op human pose estimation (Toshev & Szegedy, 2013). Algemeen wordt aangenomen dat dit onderzoek de grondlegger is voor de huidige ontwikkelingen in human pose estimation. Validatie van de voorspellingen gebeurde bij Deep Pose aan de hand van de PCP of 'Percentage of Correct Parts'. Op het moment dat dit onderzoek gepubliceerd werd, waren de COCO of MPII datasets nog niet ontwikkeld of nog niet algemeen bekend. Hierdoor werd de performantie van dit onderzoek nog niet getest op deze datasets. Om deze reden is het moeilijk dit onderzoek te vergelijken met meer recente papers waar wel COCO of MPII datasets werden gebruikt in de evaluatie. In figuur 2.13 worden resultaten van Deep Pose weergegeven. De verschillende modellen slaagden erin om voor de bovenarm 56% van de delen correct te labelen. Bij de onderarm werd er slechts 38% correct geclassificeerd. Wanneer men kijkt naar de benen, werd voor het onderbeen 71% juist voorspeld en voor het bovenbeen 78%. Gemiddeld slaagde het beste model erin om 61% van de lichaamsdelen correct te voorspellen. Tot dan toe was er geen enkele andere methode die meer dan 60% van de keypoints juist kon indelen. Ondanks de ogenschijnlijk beperkte scores was dit onderzoek uitermate innovatief voor zijn tijd. Veel andere onderzoekers hebben nadien dit onderzoek gebruikt als springplank voor de ontwikkeling van nieuwere modellen.

Eind 2014 ontwikkelden Tompson, Goroshin, Jain, LeCun en Bregler (2014) een eerste

Method	Arm		Leg		Ave.
	Upper	Lower	Upper	Lower	
DeepPose	0.8	0.75	0.71	0.5	0.69
Pishchulin [18]	0.80	0.70	0.59	0.37	0.62
Johnson et al. [13]	0.75	0.67	0.67	0.46	0.64
Yang et al. [27]	0.69	0.64	0.55	0.35	0.56

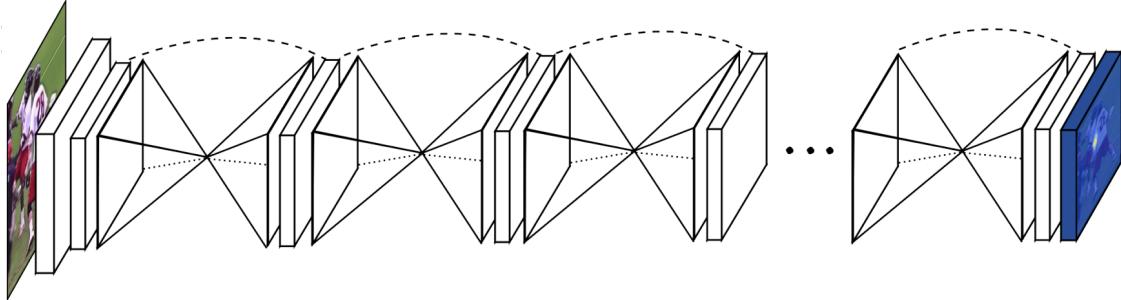
Table 2. Percentage of Correct Parts (PCP) at 0.5 on Image Parse dataset for DeepPose as well as two state-of-art approaches on Image Parse dataset. Results obtained from [18].

Figuur 2.13: Resultaten DeepPose (Toshev & Szegedy, 2013)

methode die gebruik maakte van heatmaps voor het voorspellen van gewrichten. Evaluatie van de performantie gebeurde hier wel aan de hand van de MPII dataset. Ook hier waren opnieuw grote sprongen in precisie zichtbaar. Wanneer men kijkt naar de AP waarden voor het voorspellen van het hoofd en de schouders werd er een percentage van meer dan 90% gehaald. Voor andere gewrichten zoals de knieën en de enkels bleven de resultaten echter wel steken rond de 60-70%. Deze grote verbeteringen waren te wijten aan het feit dat tot dan toe andere modellen steeds gebruik maakten van een regressiemethode om het pose estimation probleem op te lossen. De heatmaps die hier werden ontwikkeld zullen opnieuw door andere onderzoekers gebruikt worden als bouwsteen. Een van de beperkingen van dit model was echter wel dat er weinig tot geen rekening werd gehouden met biomechanische aspecten van het menselijk lichaam en gewrichten. Voorbeelden hiervan zijn het feit dat de enkel niet verbonden kan zijn aan de pols of dat de elleboog niet achterwaarts kan plooien. Rond de periode van dit onderzoek zijn verschillende andere methodes ontwikkeld die allemaal andere specifieke werkwijzen gebruikten om het human pose probleem op te lossen (Carreira, Agrawal, Fragkiadaki & Malik, 2015; Wei, Ramakrishna, Kanade & Sheikh, 2016). Geen enkele hiervan was in staat om significante vooruitgang te boeken ten opzichte van dit onderzoek.

Het volgende sleutelonderzoek werd geschreven door Newell, Yang en Deng (2016). Ook hier werd een nieuwe manier van verwerking ontwikkeld om tot betere resultaten te komen. In tegenstelling tot voordien werden hier wel alle andere onderzoeken duidelijk overtroffen. Om dit te bereiken werd een volledig nieuwe architectuur voor het voorspellingsmodel ontwikkeld. Deze architectuur noemde men een ‘stacked hourglass’ netwerk, een naam gekozen door de gelijkenissen met de wijze van verwerking. Door het principe van CNN’s wordt de resolutie van een afbeelding steeds kleiner wanneer deze door de verschillende lagen van een netwerk lopen. Dit noemt men een ‘top-down’ verwerking, vertrekken van een hoge resolutie en eindigen met een lage resolutie. Top-down verwerkingen zijn interessant voor het herkennen van lokale features zoals het herkennen van een pols. Bij pose estimation zijn echter niet enkel deze lokale features belangrijk maar ook de globale

features. Deze globale kenmerken zijn bijvoorbeeld de oriëntatie van de persoon, de opeenvolging van de ledematen of de relatie met aanliggende gewrichten. Wanneer men enkel gebruik maakt van top-down verwerking worden veel van deze features gemist. Er kan dus ook gewerkt worden van een lage resolutie naar een hoge resolutie, ofwel ‘bottom-up’ verwerking. Het stacked hourglass netwerk maakt gebruik van verschillende gestapelde zandlopers die zowel een top-down als een bottom-up verwerking gebruiken (Figuur 2.14). Door deze manier werd zowel de lokale als de globale informatie gebruikt,



**Fig. 1.** Our network for pose estimation consists of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference.

Figuur 2.14: Stacked hourglass architectuur (Newell, Yang & Deng, 2016)

met een grote vooruitgang als gevolg. De resultaten worden weergegeven in afbeelding 2.15. De anders moeilijk te herkennen knie- en enkelgewrichten werden hier met meer dan

**Fig. 7.** PCKh comparison on MPII

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Tompson et al. [16], CVPR’15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al. [19], CVPR’16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Pishchulin et al. [17], CVPR’16	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Hu et al. [27], CVPR’16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Wei et al. [18], CVPR’16	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Our model	<b>98.2</b>	<b>96.3</b>	<b>91.2</b>	<b>87.1</b>	<b>90.1</b>	<b>87.4</b>	<b>83.6</b>	<b>90.9</b>

**Table 2.** Results on MPII Human Pose (PCKh@0.5)

Figuur 2.15: Resultaten stacked hourglass netwerk (Newell, Yang & Deng, 2016)

80% nauwkeurigheid voorspeld in de MPII dataset. Ook werd voor het eerst de kaap van 90% overschreden voor de gemiddelde nauwkeurigheid van alle gewrichten.

Binnen een periode van drie jaar na de publicatie van de MPII dataset is men er in geslaagd om door middel van deep learning consistent hoge scores te behalen. Ook de scores op de complexere COCO dataset zijn op één jaar tijd van 60,5 mAP naar 72,1 mAP gegaan. In tegenstelling tot de MPII zijn hier dus wel nog significant verbeteringen mogelijk. Om die reden focussen de recentste onderzoeken zich voornamelijk op het verder verbeteren van de resultaten op de COCO dataset. Veel modellen van deze onderzoeken

hebben aanzienlijke verschillen in methodologie maar hebben slechts kleine verschillen in nauwkeurigheid. Hierdoor wordt het moeilijk om te onderscheiden welke aspecten nu net cruciaal zijn. Daarom is het onderzoek van Xiao, Wu en Wei (2018) teruggekeerd naar de basis. Men heeft onderzocht hoe goed een eenvoudig model kan presteren. In tegenstelling tot vele andere onderzoeken is de architectuur van het ontwikkelde netwerk een heel stuk eenvoudiger. Toch is men erin geslaagd om een score van 73,7 mAP te behalen op de COCO dataset. Gezien de eenvoud van het model is het verrassend om te zien dat de prestaties toch beter zijn dan veel netwerken met een complexere architectuur.

Zoals eerder vermeld bestaan er ondertussen meerdere modellen die gelijkaardige resultaten behalen op de COCO dataset. Een van de modellen die tot op de dag van vandaag nog steeds één van de hoogste scores heeft, is HRNet (High-Resolution Network) (Sun, Xiao, Liu & Wang, 2019). HRNet doet voorspellingen in twee grote fasen. In de eerste fase wordt de ‘bounding box’ of omsluitende rechthoek van alle personen die herkend worden in de input bepaald. Dit gebeurt door het bekende Resnet (He, Zhang, Ren & Sun, 2015), een model geoptimaliseerd voor het herkennen van personen en objecten. Nadien zullen binnen deze bounding box de keypoints worden bepaald. De gevonden coördinaten worden ten slotte terug geschaald ten opzichte van de volledige afbeelding. Klassieke netwerken verwerken inputs steeds van een hoge resolutie naar een lage resolutie om dan opnieuw te herleiden naar een hoge resolutie. Dit wordt nadien teruggegeven als output. HRNet is erin geslaagd om gedurende het gehele verwerkingsproces de afbeeldingen op een hoge resolutie te houden met een grotere nauwkeurigheid als gevolg. Dit zorgt ervoor dat ze in staat zijn veel van de andere huidige oplossingen te overtreffen. Hun beste model behaalde een mAP score van 77.0 op de COCO dataset.

## 3. Methodologie

In dit onderdeel zal de aanpak van het onderzoek besproken worden. Dit gaat van het zoeken naar een geschikte dataset tot de methodes die gebruikt zijn om een antwoord te krijgen op de opgestelde onderzoeks vragen. Het ultieme doel van dit onderzoek is om te onderzoeken of het mogelijk is om met de beschikbare informatie een proof-of-concept te ontwikkelen. Aan de hand van deze proof-of-concept zal gekeken worden of machine learning geschikt is om (een deel van) de manuele evaluatie van de lateral stepdown screening test uit te voeren.

### 3.1 Dataset

De eerste stap bij een onderzoek met machine learning is het opstellen van een geschikte dataset. Deze moet ten eerste relevant zijn voor het onderzoek. Op voorhand moet dus bepaald worden welke soort data nodig is en wat men uit deze data wil halen. Voorbeelden van soorten datasets zijn afbeeldingen, video's of numerieke data. Ten tweede is het belangrijk dat de dataset voldoende groot is. Een te kleine dataset kan er namelijk voor zorgen dat de resultaten geen goede steekproef zijn voor de werkelijkheid. Hoe groter de dataset, hoe representatiever deze is voor de situaties in de praktijk. Ten slotte moet men de data opruimen, ofwel 'data cleaning' uitvoeren. Dit is het proces waarbij alle incomplete data ofwel verwijderd of indien mogelijk correct aangevuld wordt. Hierdoor moet er uiteindelijk een uniforme dataset ontstaan.

Om deze dataset op te stellen werd er gekeken naar hoe een evaluatie van een lateral stepdown in de werkelijkheid gebeurt. Zoals eerder vermeld worden de uitvoeringen in de praktijk vaak gefilmd door middel van een camera. Hieruit werd besloten dat het

noodzakelijk was dat de dataset bestond uit video's van uitvoeringen van deze tests. Er moet gezorgd worden dat deze video's op een bepaalde wijze gestandaardiseerd zijn, om de uniformiteit te garanderen. In elke video moeten de camerastandpunten zo gelijk mogelijk zijn en moet de uitvoering van de test op dezelfde manier gebeuren. Om rekening te kunnen houden met zoveel mogelijk parameters werd er bijkomend gekozen om te zoeken naar data waarbij voor elke persoon de test zowel aan de linker- als de rechterzijde werd afgenoemt. Een dataset die op deze manier opgebouwd is zou een representatieve weergave moeten geven van praktijkrelevante situaties.

Naast een dataset van video's is er nood aan een manier om de nauwkeurigheid van de resultaten te evalueren. Om te kijken of resultaten die komen van een dataset verwerkt door machine learning nauwkeurig zijn, moet dit vergeleken worden met evaluaties die gebeurd zijn door ervaren testers. Aangezien dit onderzoek zich focust op de valgusbeweging van de knie zal elke video dus ook een beoordeling moeten bevatten voor de kwaliteit van deze beweging. Om te kunnen voldoen aan alle vooropgestelde voorwaarden werd contact opgenomen met de auteur van een onderzoek die aan de hand van videobeelden de lateral stepdown heeft beoordeeld. Na overleg, goedkeuring door het ethisch comité en het anoniem maken van de data kreeg deze bachelorproef de goedkeuring om de dataset van het onderzoek: „Reliability of two functional clinical tests to evaluate trunk and lumbopelvic neuromuscular control and proprioception in a healthy population.” van De Blaiser e.a. (2019, 6) te gebruiken.

### **3.1.1 Data exploratie**

In dit onderzoek is de lateral stepdown bij 139 personen geëvalueerd, zowel aan de linker- als de rechterzijde. Dit protocol verliep analoog aan de methode beschreven in hoofdstuk 2.1.1: Testprotocol. Bij dit onderzoek werd het aantal herhalingen echter wel specifiek meegedeeld. Elke persoon werd gevraagd de test 5 keer te herhalen. Voor de start kreeg elke testpersoon de mogelijkheid om 2 testpogingen uit te voeren, deze werden niet meegerekend in de finale score en werden ook niet gefilmd. Indien de hiel de grond niet raakte of de persoon van de box stapte, werd de poging als ongeldig beschouwd.

De dataset bestond in totaal uit 283 video's en een document met evaluaties van elke persoon. Elke video werd gefilmd vanop een afstand van 3 meter en een hoogte van 50 cm. De camera werd in lijn geplaatst met het steunbeen tijdens de test in het frontale vlak. Het document met de scores van de uitvoeringen was opgebouwd aan de hand van een Excel-bestand. Elke rij van dit bestand stelde een evaluatie van één van de video's voor. In totaal waren er 278 evaluaties. Per rij waren 8 kolommen ingevuld: een voor de naam van het videobestand waarmee de poging gelinkt was, de uitvoerende zijde (links of rechts), het aantal herhalingen waar rekening mee gehouden werd, eventuele opmerkingen en ten slotte 3 evaluatieparameters met een totaalscore voor de gehele oefening. Er werd een score gegeven op balans, bekkencontrole en knievalgus. Elk van deze elementen kreeg een score van 0 tot en met 3. De waarde 0 correleerde met een slechte score voor deze parameter, 1 met matig, 2 met goed en de score 3 met perfect. De totaalscore was hierdoor een score op 9 punten. Deze scores werden bepaald aan de hand

van een open source videobewerkingsprogramma, Kinovea<sup>1</sup>. Met dit programma werden referentiepunten geplaatst, segmenten door lijnen verbonden en ten slotte hoeken berekend. Knievalgus werd geëvalueerd op slechts één van de 5 pogingen. Hiervoor werd de meest representatieve poging van de hele test gekozen. De beoordelaar kreeg de instructies om alle pogingen na elkaar te bekijken en indien de kwaliteit van de herhalingen gelijkaardig was, de beste hiervan te evalueren. Wanneer er echter een grote variabiliteit in kwaliteit werd waargenomen, werden zowel de beste als de slechtste poging geëxcludeerd. Hierna werd de beste van de overgebleven pogingen geëvalueerd. Nadien werd de video van de test afgespeeld aan halve snelheid en gepauzeerd tijdens de gekozen poging op het moment dat de hiel de grond raakte. Op dit moment, ook wel de piekvalgus genoemd, werden aan de hand van Kinovea verschillende referentiepunten en lijnen geplaatst. Er werd een lijn getrokken van de heup naar de tweede teen en een verticale lijn vertrekende van de tweede teen. Wanneer het midden van de knieschijf op de lijn tussen de heup en de teen lag werd een score 2 of 3 gegeven. Het verschil tussen beide scores werd bepaald aan de hand van de videobeelden. Wanneer de knie niet stabiel was tijdens de uitvoering werd een score 2 gegeven, indien dit wel het geval was een score 3. Wanneer het midden van de knieschijf zich op de verticale lijn of lateraal hiervan bevond, kreeg de testpersoon een score 1. Indien de knieschijf zich aan de mediale zijde bevond, een score 0 (Figuren 2.4, 2.5 en 2.6).

### 3.1.2 Data cleaning

De ruwe dataset bestond uit in totaal 283 video's van de test uitgevoerd op 139 personen. Bij deze beelden waren er 2 slechte opnames. Deze werden niet geïncludeerd, wat het totaal op 281 video's met correcte uitvoeringen bracht. 4 video's waren opgenomen als 2de poging van 4 verschillende personen. Hiervan werden de originele pogingen niet opgenomen in de uiteindelijke dataset. De score was namelijk gelinkt aan de tweede poging. Bovendien was er 1 score waarvoor geen video voor handen was, deze werd ook niet geïncludeerd. Voor alle personen behalve één, was er dus zowel een rechter als linker uitvoering beschikbaar. De dataset die verder gebruikt werd bestond uit 277 video's met corresponderende score's voor 139 personen (Tabel 3.1).

Tabel 3.1: Data cleaning proces deel 1

	Aantal video's	Aantal score's
Originele dataset	283	278
Slechte video opnames	2	-
Aantal tweede pogingen	4	-
Aantal score's zonder video	-	1
Totaal	277	277

Ook in het document met de scoring werd er aan data cleaning gedaan. Aangezien er in dit onderzoek enkel wordt gefocust op de kneivalgus, werden de kolommen met de scores

<sup>1</sup><https://www.kinovea.org/>

voor de balans en bekkencontrole niet geïncludeerd. Als gevolg werd ook de totaalscore verworpen aangezien deze enkel relevant is wanneer alle drie de testparameters in rekening worden gebracht. In de kolom met opmerkingen werden zaken vermeld zoals ‘grond niet geraakt’, ‘gewichtstransfer’ of andere elementen die twijfel veroorzaakten bij de evaluatie. Deze kolom was niet gestandaardiseerd en werd vooral gebruikt als ondersteuning voor de onderzoekers. Omdat er naar de absolute scores van de knievalgus zal gekeken worden en er nood is aan gestandaardiseerde data werd ook deze kolom geëxcludeerd. De finale dataset van de scoring bestond uiteindelijk uit 278 elementen met 4 kolommen: videonummer, zijde, score knievalgus en herhalingen.

De bestandsnamen van de verschillende video’s waren initieel nog in de ruwe vorm. De naamgeving bestond uit een willekeurige cijfercombinatie gegenereerd door de videocamera. Om een consistente en duidelijke naamgeving te bekomen, werd een script geschreven dat alle bestanden op een uniforme wijze naamaf. De eerste video kreeg als naam ‘001’, elke video nadien werd incrementeel verhoogd. Uiteindelijk kregen alle video’s een naam lopende van 001 tot en met 278. In het document met de scoring werd ook de kolom met het videonummer geüpdatet om elke evaluatie te linken aan de correcte video.

## 3.2 Keypoint detection

Aan de hand van voorgaand literatuuronderzoek werd gekozen om het open source netwerk van HRNet te gebruiken voor de keypoint bepaling in de video’s. Hiervoor werd de bestaande repository (Sun e.a., 2019; Wang e.a., 2019; Xiao e.a., 2018) voor implementaties gebruikt en aangepast naar de noden van dit onderzoek. De originele code vereiste verschillende argumenten als inputparameters: de te verwerken video, de outputfolder, het aantal frames dat verwerkt moet worden en een configuratiebestand voor het netwerk. Vier getrainde modellen met hun respectievelijk configuratiebestand werden door de auteurs ter beschikking gesteld. Twee van deze modellen transformeerden de input naar een aspect ratio van 256x192 pixels, de andere twee naar 384x288 pixels. Per inputgroep bestond er ook telkens één model met een lager en één met een groter aantal parameters. De GFLOPs of ‘Giga Floating Point Operations / Second’ bepalen hoeveel berekeningen een neurale netwerk per seconde doet. De modellen met grotere inputformaten en meer parameters hebben een groter aantal GFLOPs en doen dus meer berekeningen per seconde. Als gevolg duurt het in dit geval langer om resultaten te bekomen maar zal er ook een betere nauwkeurigheid zijn. De resultaten op de COCO dataset zijn voor alle modellen rond 0,75 mAP. De modellen met de grootste input en de grotere complexiteit haalden iets betere resultaten. Aangezien precisie voor dit onderzoek de voorkeur krijgt op snelheid werd er gekozen voor het model met het meeste parameters en de grootste inputgrootte. In tabel 3.2 wordt een overzicht gegeven van de 4 beschikbare modellen. Het script uit de oorspronkelijke repository overloopt elk frame van de inputvideo. Indien een lagere framerate wordt opgegeven dan de inputvideo, zullen enkel deze frames verwerkt worden. Wanneer bijvoorbeeld de inputvideo een framerate heeft van 30 frames per seconde en het opgegeven aantal frames 15 is, zal slechts één op de twee frames verwerkt worden. Wanneer er meer frames dan de input worden opgegeven, wordt het script afgesloten en

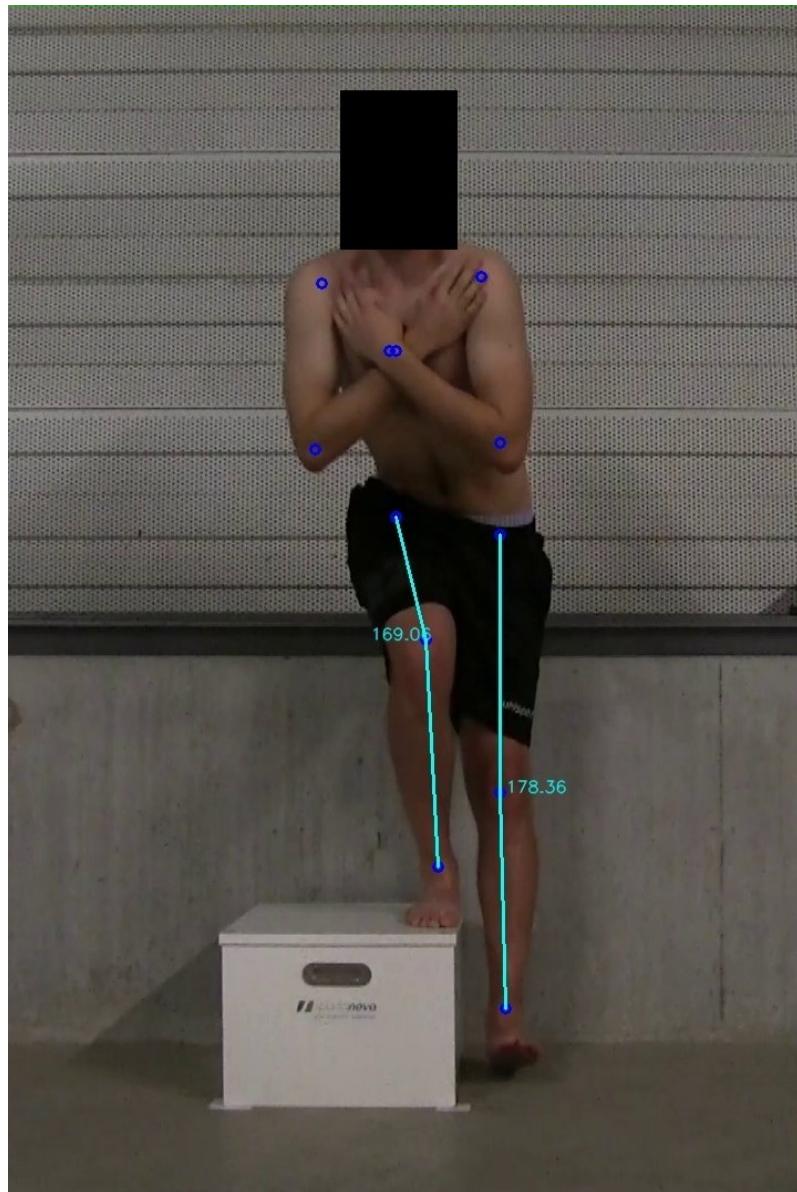
Tabel 3.2: Resultaten van 4 voorgetrainde HRNet modellen op de COCO dataset

Naam	Input formaat	#Params	GFLOPs	AP	Ap .5	AP .75	AR	AR .5	AR .75
pose_hrnet_w32	256x192	28.5M	7.1	0.744	0.905	0.819	0.798	0.942	0.865
pose_hrnet_w32	384x288	28.5M	16.0	0.758	0.906	0.825	0.809	0.943	0.869
pose_hrnet_w48	256x192	63.6M	14.6	0.751	0.906	0.822	0.804	0.943	0.867
pose_hrnet_w48	384x288	63.6M	32.9	0.763	0.908	0.829	0.812	0.942	0.871

een foutmelding gegeven. De video's in de dataset hebben een framerate van 25 frames per seconde. Er werd gekozen om elke video aan zijn originele framerate te verwerken om zeker geen relevante informatie te missen. Per frame dat verwerkt wordt, worden in eerste instantie alle personen gelokaliseerd door een omvattende rechthoek. Deze eerste voorspelling gebeurt door het bekende ResNet netwerk. Nadien wordt per rechthoek een tweede voorspelling gedaan door HRNet om de zichtbare keypoints te lokaliseren. Ten slotte worden alle keypoints in het frame opgeslagen als rij in een CSV-bestand. Per keypoint zijn er 2 kolommen, een voor de x-coördinaat en een voor de y-coördinaat. Zo heeft men met het oorspronkelijke script één CSV-bestand als output met 34 kolommen voor 17 keypoints. Elke rij komt overeen met alle gevonden keypoints in dat frame.

Aan deze oorspronkelijke implementatie werden verschillende aanpassingen gedaan voor dit onderzoek. Omdat het niet mogelijk was om een groot aantal video's in een keer te verwerken, werd de input van 1 video veranderd naar een input folder. In deze folder werden enkel videobestanden verwerkt die de eerder opgestelde naamgevingen volgden. Per video werd vervolgens ook een apart CSV-bestand opgeslagen met de gevonden keypoints. Hier werden een aantal kolommen aan toegevoegd. Als eerste, twee kolommen met algemene informatie van de video, één met het videonummer en één met het framenummer. Hiermee zou het in de toekomst mogelijk zijn om de keypoints opnieuw weer te geven op de correcte plaats in de video. Als tweede werden twee kolommen toegevoegd voor de FPPA. Met de x- en de y-coördinaten van de knie, enkel en heup is het mogelijk om deze hoek te berekenen. Voorgaand onderzoek heeft aangetoond dat deze hoek een goede inschatting geeft van de effectieve knievalgushoek. Met de vereiste goniometrie werden de FPPA's van beide knieën berekend. Aangezien het op dit moment nog niet mogelijk was om een onderscheid te maken tussen een linker- en een rechterevaluatie werden beide waarden per frame opgeslagen in een eigen kolom. Elke output bestond hierdoor uit 38 kolommen. Ten slotte werden ook verschillende lijnen code toegevoegd om visuele ondersteuning te bieden bij de resultaten. Hiermee was het mogelijk lijnen, punten of tekst te tekenen op de frames en deze op te slaan als foto of video. Een aantal voorbeelden hiervan zijn: de coördinaten plaatsen bij elk keypoint, lijnen tussen verschillende keypoints trekken of de FPPA weergeven (Figuren 2.11, 2.7 en 3.1).

Bij het inladen van de CSV-bestanden werd het duidelijk dat het niet mogelijk was om sommige rijen te importeren. Na onderzoek bleek dat in 11 video's een tweede persoon in beeld kwam. Hierdoor werden in dat frame extra keypoints bepaald en ook opgeslagen. Om de consistentie en standaardisatie van de dataset te garanderen werd besloten om enkel



Figuur 3.1: Visualisatie van geïnverteerde FPPA berekend op basis van keypoint detection

beelden waarin één testpersoon zichtbaar was in rekening te brengen. In twee van de video's was er gedurende de hele uitvoering een tweede persoon in beeld. Aangezien het niet mogelijk was een onderscheid te maken tussen welke coördinaten behoorden tot de testpersoon en welke niet, werden deze twee video's samen met hun evaluatie volledig geëxcludeerd. In de andere 9 video's kwam slechts in een deel van de video een extra persoon of deel van een andere persoon in beeld. In totaal ging het hier over een extra 289 frames. Aangezien het hier slechts om een klein deel van het totaal aantal frames ging, werd besloten om deze video's niet volledig te excluderen. Hier werden enkel de frames geëxcludeerd waar meer dan één persoon in beeld was. Exclusie gebeurde aan de hand van een script ontwikkeld voor dit onderzoek. De finale dataset bestond hierdoor uit 275 video's, waarvan er bij 9 een aantal frames niet in rekening werden gebracht. Tabel 3.3 geeft een overzicht van deze extra data cleaning.

Tabel 3.3: Data cleaning proces deel 2

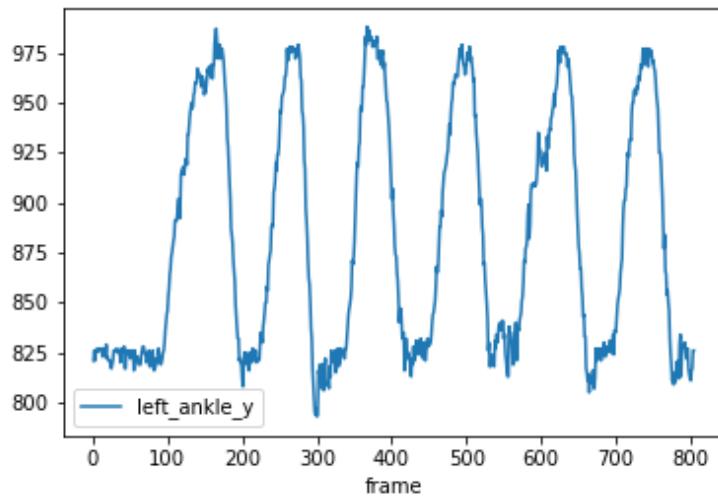
	Aantal video's	Totaal aantal frames
Dataset	277	198.799
Video's met 2de persoon hele tijd in beeld	2	1.200
Video's met 2de persoon deels in beeld	9	289
Totaal	275 (9 deels)	197.310

### 3.3 Piekvalgus bepaling

Zoals eerder beschreven gebeurde de scoring van de knievalgus in het originele onderzoek aan de hand van piekvalgus momenten, ofwel het moment waarop de hiel de grond raakt. Om te onderzoeken of er een correlatie is tussen de berekende hoek van de keypoint detection en de scores gegeven door menselijke testers, was het noodzakelijk om voor elke video de pieken te bepalen. Het doel van dit deel was dan ook om een representatieve FPPA te berekenen voor de verschillende piekvalgus momenten tijdens de test.

#### 3.3.1 Find peaks

Om de piekvalgus te bepalen werd gebruik gemaakt van de verplaatsingen van de y-coördinaat van de niet-uitvoerende enkel. Wanneer deze coördinaat in functie van de frames wordt uitgedrukt op een grafiek, geeft dit een sinusvormige functie (Figuur 3.2). De bovenste pieken op deze grafiek komen overeen met momenten waar de enkel van een neerwaartse beweging overgaat naar een opwaartse beweging. Voor de onderste pieken geeft dit de overgang van een opwaartse naar een neerwaartse beweging weer. Bij de bovenste pieken bevinden zich dus steeds de piekvalgus momenten, bij de onderste de momenten waar de enkel ter hoogte van de box is. Dit zijn belangrijke momenten om

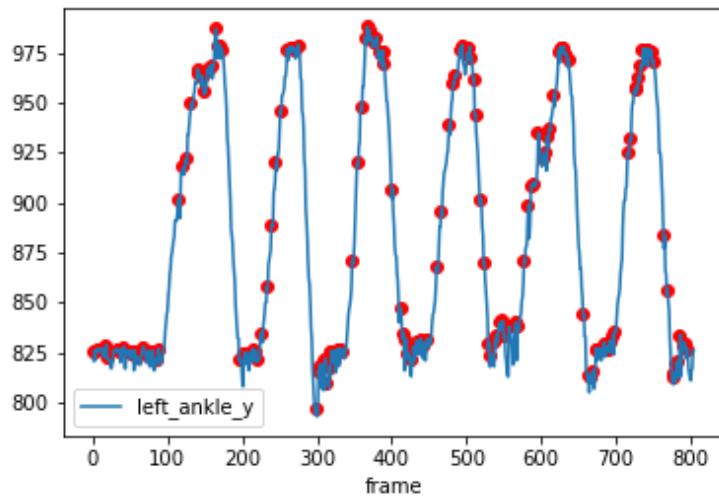


Figuur 3.2: Y-coördinaat van de enkel in functie van het framenummer

te bepalen voor de verdere verwerking. Niet elke piek komt echter overeen met een van deze twee momenten in een test. Indien men elke bovenste piek zou bepalen zouden er verschillende worden herkend die niet overeenkomen met het moment waarop de voet de grond raakt. Het moment waarop een persoon bijvoorbeeld tijdens een uitvoering even deels terugkeert voor de grond geraakt wordt, moet niet mee in rekening worden gebracht. Ook pieken die te snel op elkaar volgen kunnen wijzen op gewichtstransfer door evenwichtsverlies. Deze gevallen zijn voor dit onderzoek niet relevant en moeten er dus uit gefilterd worden. Om hier een onderscheid in te kunnen maken werd gebruik gemaakt van de functie ‘find\_peaks’ van de Python bibliotheek ‘Scipy’ (Scipy, 2019). Deze functie maakt het mogelijk om een minimale hoogte op de y-as en een minimale afstand tussen twee pieken op de x-as te specifiëren. Deze waarden mogen echter niet lukraak gekozen worden. Anders bestaat de mogelijkheid dat correcte pieken ontzettend worden geëxcludeerd. Om deze parameters zo correct mogelijk in te vullen werden een 15-tal unieke video’s geselecteerd. Hier ging het over 5 normale uitvoeringen en 10 uitvoeringen die niet perfect waren. Deze bevatten situaties zoals onvolledig contact met de grond, testpersoon die start vanop de grond of eindigt op de grond, evenwichtsverlies en het halverwege haperen en terugkeren. Voor elk van deze video’s werden de contactmomenten met de grond manueel door de onderzoeker bepaald. Deze werden nadien vergeleken met de berekende pieken om de correctheid te controleren.

### 3.3.2 Moving averages

Wanneer deze functie zonder extra parameters wordt toegepast op een grafiek, blijkt dat naast de piekvalgus momenten ook veel kleine pieken worden aangeduid (Figuur 3.3). Dit is te verklaren door de foutenmarge op de keypoint detection. Twee opeenvolgende frames hebben daarom, ondanks het feit dat de positie van de testpersoon quasi gelijk is, vaak licht verschillende coördinaten. Deze variaties veroorzaken kleine piekmomenten die worden

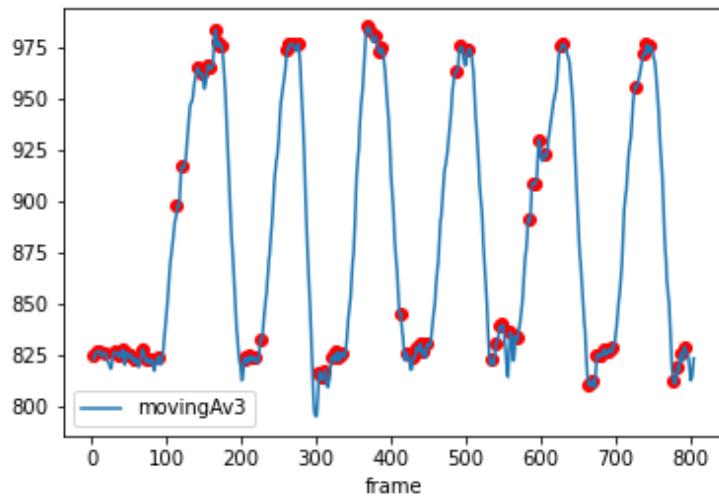


Figuur 3.3: Pieken bepaald met de findPeaks functie op 3.2 zonder extra parameters

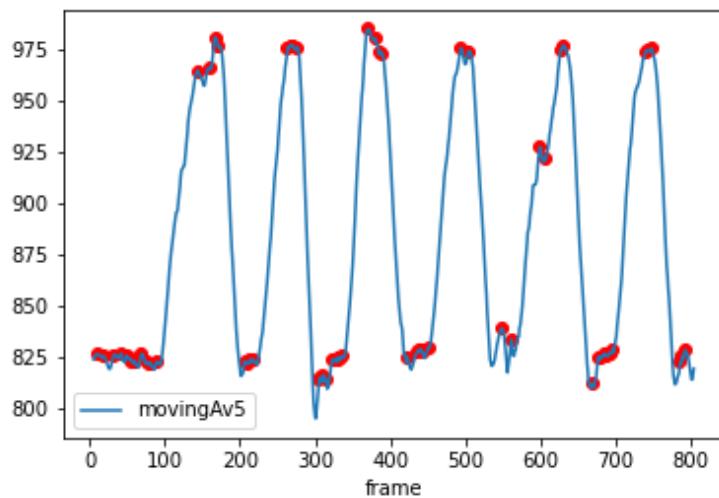
herkend door de functie. Om dit op te lossen werden de voortschrijdende gemiddelden, of ‘moving averages’, van de y-coördinaat berekend. Door per frame het gemiddelde te nemen van de coördinaten van een aantal voorgaande frames kan de curve egaler gemaakt worden. Piekmomenten door variaties in precisie van de voorspellingen worden hierdoor niet meer geïncludeerd. Elke video heeft een framerate van 25 frames per seconde. Belangrijk is dat de gekozen moving averages niet berekend worden over een te groot aantal frames. Anders zal er hierdoor een verschuiving van de grafiek plaatsvinden naar rechts. Wanneer de gevonden pieken daarna geprojecteerd worden op de originele video’s zullen deze later vallen dan het effectieve contact met de grond. Ook is het belangrijk dat de gemiddelden voldoende ruim worden genomen. Indien dit niet het geval is zullen er nog steeds oneffenheden aanwezig zijn. Bij nauwkeurige analyse van de videobeelden werd ontdekt dat de contactmomenten met de grond vaak langer dan 10 frames duurden. Om deze reden werden de moving averages voor 3, 5 en 10 frames onderzocht. In afbeeldingen 3.4, 3.5 en 3.6 kunnen de grafieken van deze gemiddelden op eenzelfde video vergeleken worden. Bij de moving averages van 3 frames waren in sommige gevallen nog steeds kleine piekmomenten aanwezig. Dit was veel minder het geval bij zowel de gemiddelden van 5 als van 10 frames. Wanneer de pieken gevonden bij 10 frames echter werden geprojecteerd op de originele video, vielen deze soms te laat. Hieruit werd geconcludeerd dat 5 frames de meest geschikte optie van de drie was.

### 3.3.3 Grondcontactbepaling

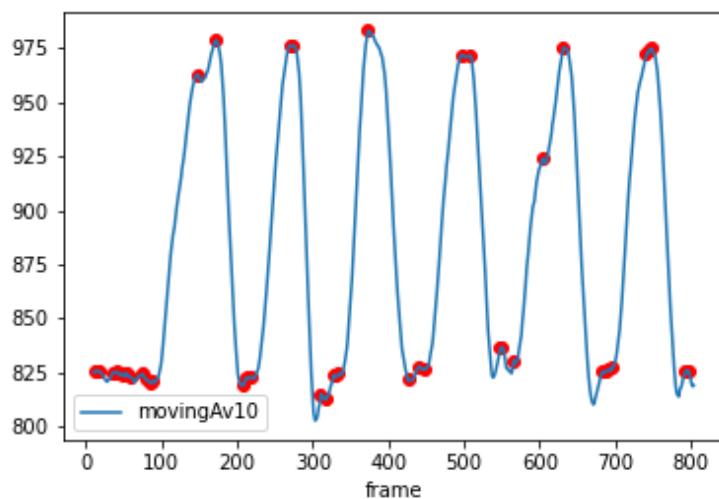
Zoals eerder vermeld ontstonden bij sommige video’s piekmomenten die niet voortkwamen uit de variatie in voorspellingen, maar uit het halverwege terugkeren tijdens een poging zonder contact met de grond. Deze gevallen veroorzaken wel duidelijke pieken in de grafiek maar zijn niet voldoende hoog. Om deze situaties er correct uit te filteren werd het gebruik van een minimale hoogte onderzocht. In eerste instantie werd de grootste



Figuur 3.4: FindPeaks op een moving average van 3 frames

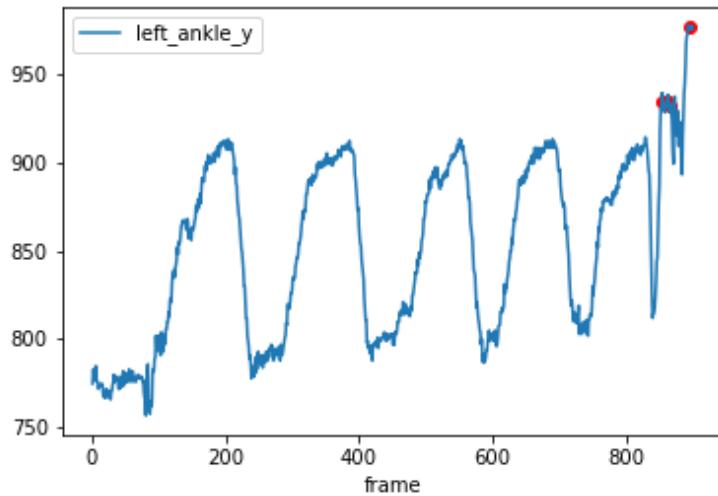


Figuur 3.5: FindPeaks op een moving average van 5 frames



Figuur 3.6: FindPeaks op een moving average van 10 frames

y-coördinaat tijdens de gehele test gebruikt. Om rekening te houden met kleine variaties die toch nog significant waren, werd de betrouwbaarheidsgrens van 0.95 van deze waarde genomen. Deze oplossing zorgde ervoor dat enkel de grootste pieken werden bepaald. In veel gevallen kwam dit ook overeen met de effectieve contactmomenten met de grond. In een aantal unieke situaties was dit echter niet het geval. Wanneer de grootste piek een situatie was zoals het afstappen van de box richting de camera, was deze maximale y-waarde te groot om de correcte pieken te bevatten (Figuur 3.7). Om dit op te lossen

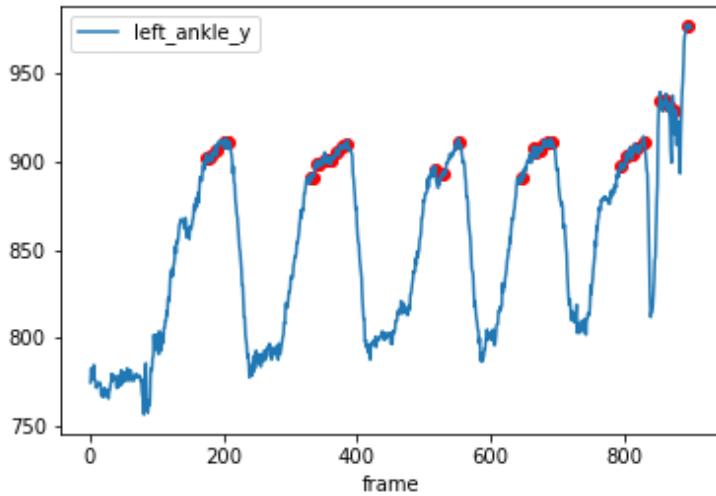


Figuur 3.7: Pieken worden gemist wanneer de maximale y-waarde gebruikt wordt

werd opnieuw gebruik gemaakt van de moving averages. Wanneer deze gemiddelden over een voldoende groot aantal frames wordt genomen, worden zeer extreme pieken sterk afgevlakt. Deze keer werden de moving averages van 25, 50 en 100 frames vergeleken. Deze kwamen overeen met 1, 2 en 4 seconden video. Ook hier werd 95% van de gevonden maximale grootte gebruikt om variaties op te vangen. Er werd gevonden dat 25 frames niet voldoende waren en dat 100 frames te veel waren. De moving averages van 50 frames zorgden ervoor dat alle correcte grondcontacten werden gevonden zonder te lage pieken te includeren (Figuur 3.8).

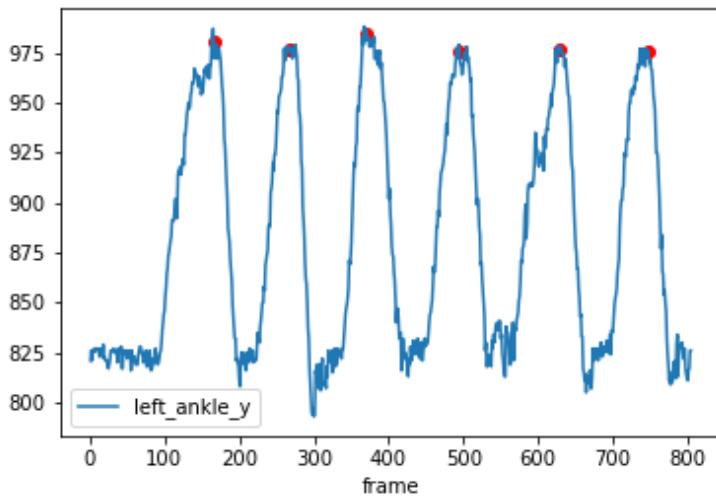
### 3.3.4 Afstandsbeleiding

Het volgende op te lossen probleem was het voorkomen van twee kort opeenvolgende pieken. Het is namelijk noodzakelijk dat van elke poging van de test slechts één piek wordt genomen. Bij slechte uitvoeringen kwam het soms voor dat er twee of meer pieken voor één poging waren. Dit kwam vaak door onvoldoende bewegingscontrole bij de testpersoon. Om hiermee rekening te houden was het noodzakelijk voldoende afstand te nemen tussen elke piek. De 'distance' parameter van de functie maakte dit mogelijk. In eerste instantie werd geëxperimenteerd met absolute waarden zoals 25, 50 of 100 frames. Aangezien iedere testpersoon een andere snelheid heeft waarmee de test wordt uitgevoerd, werd met elk van deze waarden ofwel te veel, ofwel te weinig afstand genomen. Bij elke video was



Figuur 3.8: Correcte pieken obv moving average van 50 frames

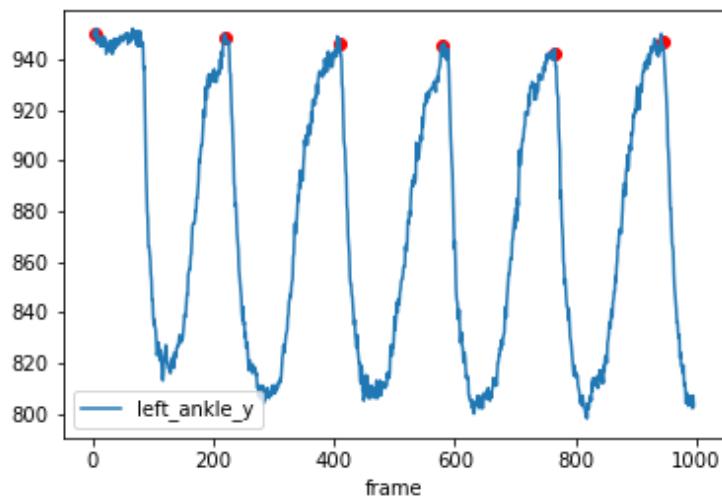
het echter wel zo dat enkel de test zelf werd gefilmd en dat per video gevraagd werd 5 pogingen te doen. Er was dus ook geen onnodige filmtijd voor of na de test. Bij personen die de test snel uitvoerden resulterde dit als gevolg in een lager totaal aantal frames dan personen met een trage uitvoering. Hierdoor was het mogelijk om een percentage te nemen van het totale aantal frames als minimale afstand. Door de afstand relatief te nemen ten opzichte van de videolengte werd er rekening gehouden met zowel snelle als trage uitvoeringen. Dit percentage werd bepaald aan de hand van trial en error. Bij elke video werd nadien gecontroleerd of de resulterende afstand voor een correcte spreiding van de pieken zorgde. Uiteindelijk bleek dat 10% van het totale aantal frames een goede verdeling was voor elk van de 15 video's (Figuur 3.9).



Figuur 3.9: Slechts één piek per top adhv een minimale afstand

### 3.3.5 Starten of eindigen op de grond

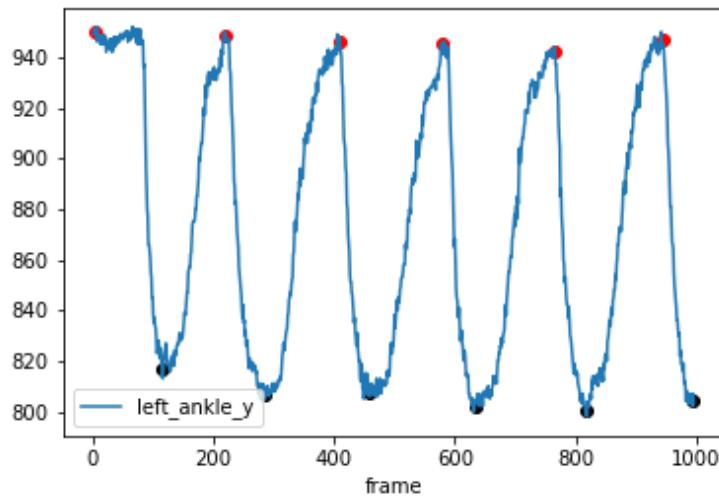
Het laatste probleem dat opgelost moest worden was het starten of eindigen van de test vanop de grond. Deze momenten vormen dezelfde pieken als het contact met de hiel op de grond, maar zijn geen onderdeel van de test. Afbeelding 3.10 is een voorbeeld van een uitvoering die begon op de grond en hierdoor een overbodig piekmoment geeft. Aangezien



Figuur 3.10: Voorbeeld van test waarbij gestart werd vanop de grond

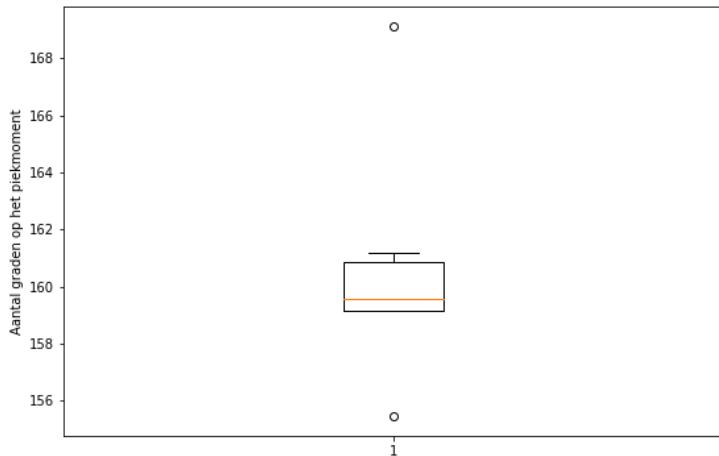
dit enkel bij de start of het einde van de video kan voorkomen, is het niet moeilijk om deze pieken te onderscheiden. De echte contactmomenten zullen namelijk altijd voorafgegaan en gevolgd worden door een geïnverteerde piek (wanneer de enkel ter hoogte van de box is). Wanneer dit niet het geval is zullen de pieken dus verwijderd worden. Om de omgekeerde toppen te bepalen werd de gehele dataset geïnverteerd. Nadien werden met de 'find\_peaks' functie de corresponderende frames bepaald op een analoge wijze. Het enige verschil was dat de y-coördinaat van de tegenovergestelde enkel in dit geval gebruikt werd om de minimale hoogte te bepalen. Deze bevindt zich namelijk steeds op de hoogte van de box. Afbeelding 3.11 geeft deze onderste pieken visueel weer. De bovenste en onderste toppen werden vervolgens samen in een lijst gestoken en zolang deze lijst startte of eindigde met een 'grond' piek werd deze uit de lijst verwijderd. Ten slotte werden deze waarden opgeslagen in een extra kolom in de CSV-bestanden. Met een paar kleine aanpassingen werd het hierdoor mogelijk om met het script voor de visualisatie deze piekmomenten uit een video te halen en op te slaan als foto.

Aan de hand van de correcte framenummers van alle contactmomenten met de grond was het mogelijk om de berekende hoeken op deze momenten te vergelijken met de score van de oorspronkelijke onderzoekers. Deze had als doel een zo representatief mogelijk beeld te schetsen van een gemiddelde poging. Daarom werden dus de beste en slechtste pogingen niet in acht genomen wanneer er een groot verschil in kwaliteit was tussen de verschillende herhalingen. Om hiermee rekening te houden bij de piekmomenten werden extreme waarden bepaald aan de hand van de interkwartielafstand. Wanneer een hoek



Figuur 3.11: Bepalen van de onderste pieken

zich niet in deze afstand bevond wou dit zeggen dat deze sterk verschillend was van de overige pogingen. De overgebleven hoeken geven als gevolg een representatief beeld van de gemiddelde prestaties van een testpersoon. In figuur 3.12 wordt een van deze testen met outliers visueel weergegeven in een boxplot.



Figuur 3.12: Voorbeeld van een test met 5 piekmomenten waarvan 2 outliers

### 3.4 Evaluatie

Eerst en vooral werd de verdeling van alle gegevens onderzocht. Nadien werd de correlatie van de gevonden FPPA's en de score op verschillende manieren vergeleken. De Pearson

correlatiecoëfficiënt werd bepaald voor het gemiddelde, minimum en maximum van de verschillende piekmomenten. Ook werd de verdeling van de gemiddelde hoeken ten opzichte van hun score geplot in een boxplot. Bovendien werden aan de hand van deze resultaten een aantal arbitraire classificatiecriteria opgesteld op basis van de gemiddelde valgushoeken. Deze criteria werden finaal met elkaar en met de oorspronkelijke scores vergeleken. Ook deze bevindingen werden aan de hand van verschillende grafieken en tabellen in kaart gebracht. Ten slotte werd gekeken of het mogelijk was een neuraal netwerk te ontwikkelen dat de originele scores kon voorspellen.



## 4. Resultaten

In dit onderdeel worden de gevonden resultaten van het onderzoek weergegeven en vergeleken. In eerste instantie wordt de verdeling van de scores bekijken. Nadien worden verschillende methodes gebruikt om de gevonden resultaten in beeld te brengen en de significantie te beoordelen. Ten slotte worden twee methodes bekijken om de valgusscore te bepalen.

### 4.1 Dataverdeling

#### 4.1.1 Valgusevaluaties

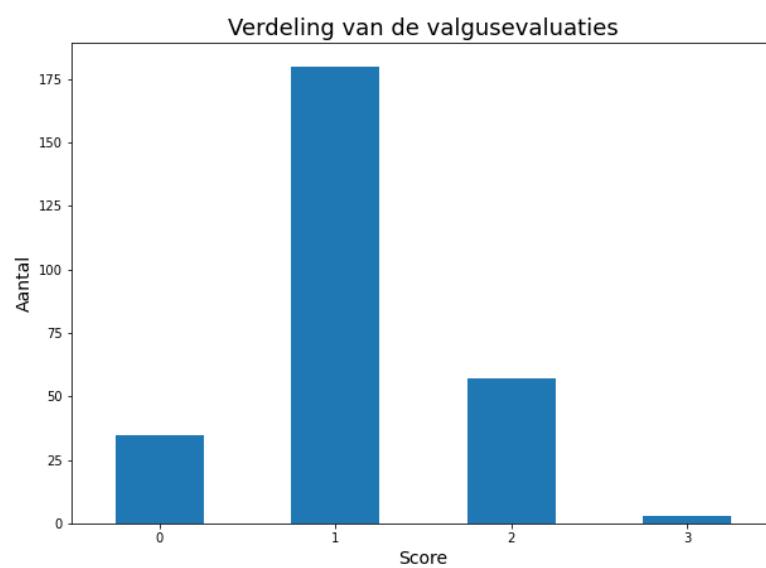
In tabel 4.1 wordt de verdeling van alle geïncludeerde scores weergegeven. Hierbij valt op dat ongeveer 2 op de 3 evaluaties overeenkomt met een score 1. Terwijl slechts 3 personen beoordeeld werden met de perfecte score 3. Dezelfde data wordt visueel voorgesteld in afbeelding 4.1.

#### 4.1.2 Piekvalgushoeken

Zoals in hoofdstuk 3 besproken wordt, zijn voor iedere video de piekvalgusmomenten dynamisch bepaald. Voor elk van deze momenten werd nadien de FPPA berekend aan de hand van de voorspelde keypoints. Er werd gekozen om het complement te nemen van de FPPA uitgelegd in hoofdstuk 2.1.3. Dit werd gedaan om de leesbaarheid te vergroten. Wanneer de hoek namelijk overeenkomt met de hoek die visueel geprojecteerd wordt, zorgt dit voor duidelijker interpretaties. Dit heeft echter geen invloed op de resultaten. Per test

Tabel 4.1: Verdeling van de valgusevaluaties

Aantal evaluaties	
Score 0	35
Score 1	180
Score 2	57
Score 3	3
Totaal	275



Figuur 4.1: Verdeling van de valgusevaluaties

werden vervolgens verschillende bewerkingen gedaan op de hoeken van de voorspelde piekmomenten. Voor elke video werden het totaal aantal piekmomenten, de gemiddelde valgushoek inclusief standaarddeviatie en variantie, minimale en maximale valgushoeken berekend. Een samenvatting hiervan wordt weergegeven in tabel 4.2.

Tabel 4.2: Samenvatting berekende piekmomenten per video

	Aantal piekmomenten	Minimale hoek	Maximale hoek	Gemiddelde hoek
Gemiddelde	4,66	159,45	172,17	165,54
SD	0,92	13,14	14,37	12,56
Min	2	115,20	120,28	118,32
25%	4	150,63	162,93	157,28
50%	5	159,15	171,38	165,11
75%	5	168,50	180,56	174,08
Max	7	198,15	220,01	200,57

Het gemiddelde aantal piekmomenten bepaald per video is 4,66. Dit ligt in lijn met de verwachtingen. Bij elke test werd namelijk gevraagd om 5 herhalingen te doen. Doordat de outliers van de pogingen niet geïncludeerd zijn, is het te verwachten dat dit gemiddelde iets lager ligt.

Wat opvalt aan deze data is dat de kleinste hoek van alle minimale pogingen in de hele dataset, ondanks het verwijderen van de outliers, 115,20 graden is. Ook de grootste hoek is met 220,01 graden, relatief gezien vrij groot. Gezien deze extreme waarden, is de correctheid hiervan nagegaan in de originele video. Voor de minimale waarde gaat het om een slechte uitvoering waarbij de romp niet helemaal recht gehouden werd tijdens de test (Figuur 4.2). De maximale waarde komt voor bij een test waarbij de testpersoon meerdere malen het evenwicht van de box verloor (Figuur 4.3). Beide pogingen zijn dus geen goede uitvoeringen van de test, maar vallen toch binnen de acceptable norm wanneer deze met de overige pogingen vergeleken worden.

Vervolgens is er gekeken naar de verdeling van de gemiddelde waarden van de piekvalgusmomenten. De waarschijnlijkheid hiervan is geplot op een histogram (Figuur 4.4). Op deze grafiek wordt duidelijk dat deze waarden een Gaussiaanse verdeling volgen. De gemiddelde representatieve piekvalgus van alle testen is 165,54 graden met een standaarddeviatie van 12,56 graden.

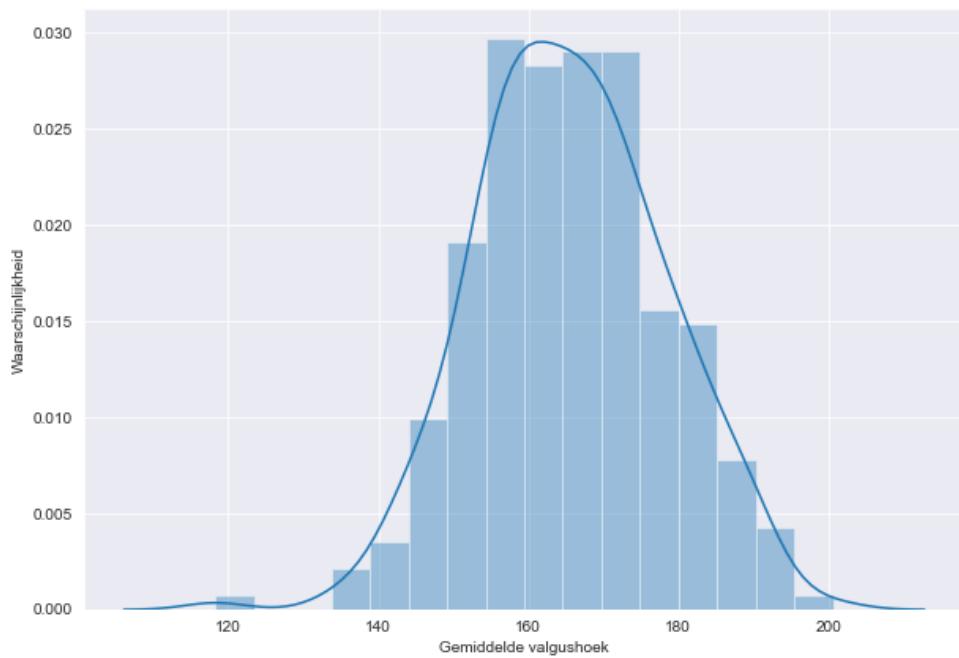
Ook de verdeling van de gemiddelde valgushoeken ten opzichte van de verkregen score werd grafisch bekeken (Figuur 4.5). Op deze grafiek kan men zien dat er een duidelijk verschil in spreiding is tussen de verschillende scores. Om na te gaan of deze waarden effectief gerelateerd zijn met de gegeven scores, wordt voor elke score de Pearson correlatiecoëfficiënt berekend. De resultaten hiervan, samen de correlatie met de minimale en maximale waarden worden in tabel 4.3 getoond.



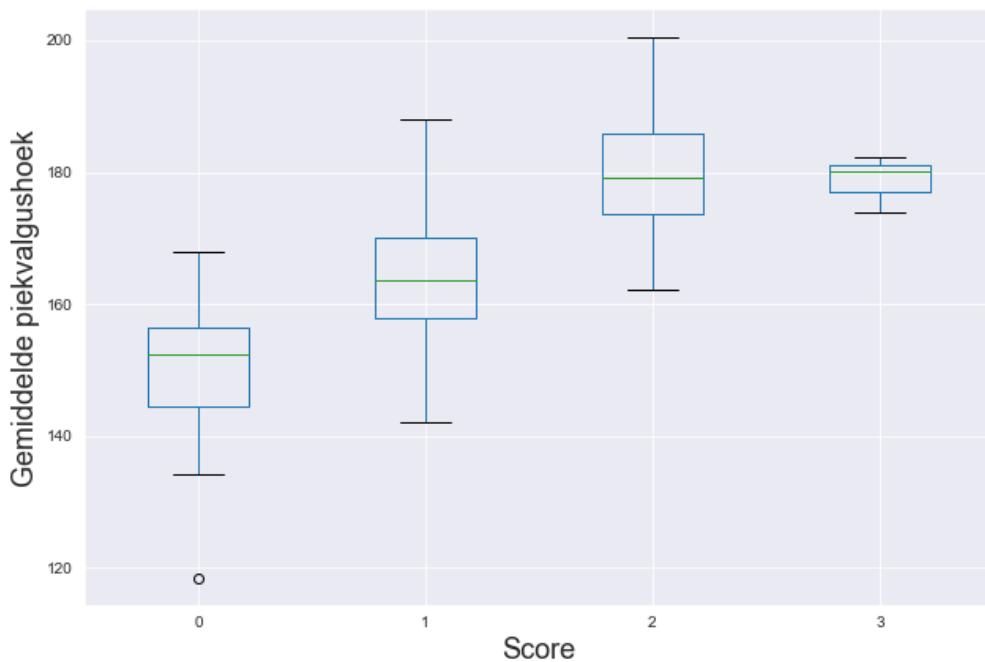
Figuur 4.2: Poging met de kleinste minimale valgushoek



Figuur 4.3: Poging met de grootste maximale valgushoek



Figuur 4.4: Verdeling van de gemiddelde piekvalgushoek



Figuur 4.5: Verdeling van de gemiddelde piekvalgushoek per score

Tabel 4.3: Peasson correlatiecoëfficiënten van de verschillende piekvalgushoeken tov de originele score

	Gemiddelde piekvalgushoek	Minimale piekvalgushoek	Maximale piekvalgushoek
Originele score	0,69	0,64	0,63

Deze correlaties tonen dat de gemiddelde piekvalgus effectief het beste correleert met de originele scores. De score van 0,69 wijs op een sterke correlatie tussen beide waarden. Hieruit kan besloten worden dat de gemiddelde piekvalgushoeken berekend in dit onderzoek wel degelijk een goede inschatting kunnen geven van de uiteindelijke valgusscore.

## 4.2 Valgusscore bepaling

### 4.2.1 Hypothetische evaluatiecriteria

Op basis van deze bevindingen is geprobeerd om voor elke score een aantal arbitraire grenzen op te stellen. Er is gekeken of de originele score op basis van deze grenzen voorspeld kan worden. Deze criteria zijn in grote lijnen gebaseerd op de interkwartielafstanden van elke score (Figuur 4.5). Tabel 4.4 geeft een overzicht van de opgestelde criteria. In de originele evaluatie werd het verschil tussen score 2 en 3 echter bepaald door de stabiliteit van de knie gedurende de hele test. Aangezien het niet mogelijk is om deze stabiliteit enkel met de piekvalgusmomenten te evalueren, is gekozen om als extra parameter de standaarddeviatie te gebruiken. Wanneer deze waarde klein is, wil dit zeggen dat er weinig verschil is tussen de groottes van de verschillende piekvalgusmomenten. Grote verschillen zullen dus mogelijks een indicatie geven dat de uitvoering minder stabiel was en dus eerder overeenkomen met een score 2.

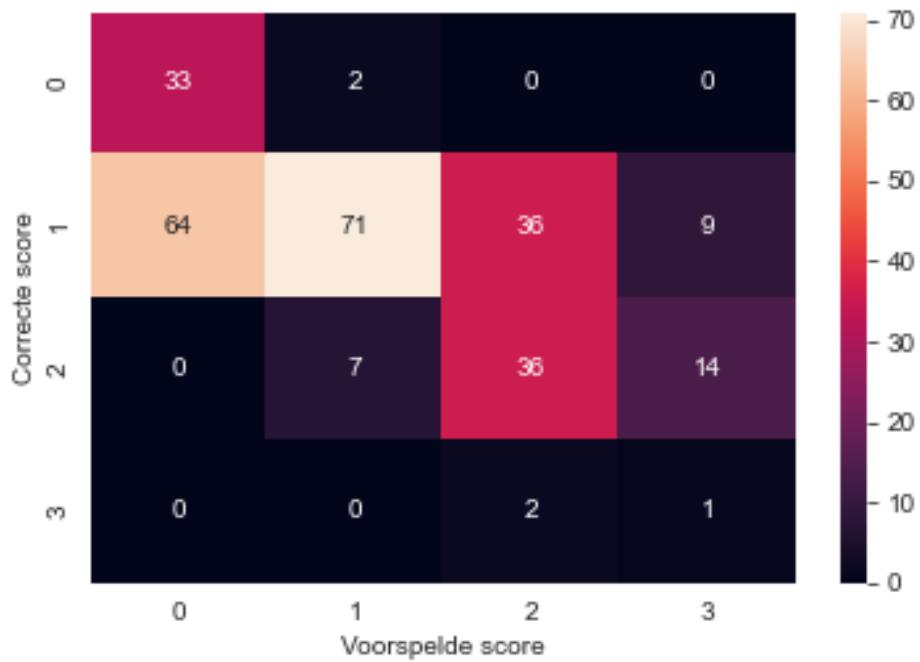
Tabel 4.4: Hypothetische valgusscore criteria

	Gemiddelde piekvalgushoek	Gemiddelde standaarddeviatie
Score 3	> 175	< 2,5
Score 2	> 170 en $\leq$ 175	$\geq$ 2,5
Score 1	> 160 en $\leq$ 170	-
Score 0	$\leq$ 160	-

De resultaten van dit hypothetisch model worden afgebeeld in de confusion matrix 4.6. De precisie van de resultaten van deze criteria is 0,44. Dit wil zeggen dat iets minder dan de helft van de scores overeen komen met de originele score. De vooropgestelde criteria zijn dus niet geschikt om de originele scores te voorspellen. Wat opvalt is dat bij de scores van het hypothetisch model de resultaten evenrediger verspreid zijn dan bij de originele scores. De vraag kan dus ook gesteld worden of de oorspronkelijke evaluatiecriteria voldoende verfijnd zijn om de valgushoek op te splitsen in gradaties.

### 4.2.2 Evaluatie adhv machine learning

Op basis van deze bevindingen werd besloten om te kijken of het mogelijk was om een neurale netwerk te ontwikkelen dat met deze hoeken de scores kon bepalen. De resultaten



Figuur 4.6: Confusion matrix van de hypothetische evaluatiecriteria

van dit netwerk waren echter niet goed. De vermoedelijke reden hiervoor is de scheve dataset. Het merendeel van de testpersonen kreeg een score 1, en slechts een klein aantal een score 3. Dit, gecombineerd met het beperkte aantal video's (275), is waarschijnlijk de oorzaak van deze slechte resultaten. Hiervoor is mogelijks een meer gevarieerde en uitgebreide dataset nodig.



## **5. Discussie**

Deze paper heeft onderzocht of machine learning en computer vision hun plaats hebben bij de evaluatie van bewegingstesten. De literatuurstudie maakte duidelijk dat op het moment van schrijven reeds meerdere modellen bestaan die menselijke gewrichten betrouwbaar kunnen bepalen. Deze gewrichten vormen de basis voor het beoordelen en classificeren van de analyses. Een van de meest nauwkeurige neurale netwerken werd gebruikt om dit te onderzoeken. De gevonden resultaten toonden aan dat de menselijke scores wel degelijk positief gecorreleerd zijn met de voorspelde gewrichten. Op basis hiervan werden een aantal mogelijke classificatiecriteria opgesteld om een evaluatie volledig automatisch te laten verlopen. De arbitraire grenzen opgesteld in dit onderzoek bleken echter niet voldoende specifiek te zijn om randgevallen te detecteren. Aangezien men bij het testen in de praktijk naast de piekmomenten ook gebruik maakt van een aantal dynamische criteria, was dit te verwachten. Men kijkt voor bepaalde scores namelijk ook naar de algemene stabiliteit van de knie tijdens de hele test. Dit is echter geen sluitend objectieve meting, eerder een subjectieve observatie. Veel van de huidige bewegingstesten maken gebruik van een aantal subjectieve criteria om de kwaliteit te beoordelen. Zo is er momenteel bijvoorbeeld geen objectieve afbakening tussen ‘veel’ en ‘weinig’ verplaatsing van de knie. Deze criteria kunnen dus ook door onafhankelijke testers verschillend worden geïnterpreteerd. Hierdoor kan de vraag gesteld worden of de huidige manier van beoordelen objectief genoeg is. De objectieve 3D-video analyses worden in de praktijk amper gebruikt door de grote kost en leercurve. Deze zijn dus ook geen geschikte oplossing voor dit probleem. De beschikbare alternatieven zijn 2D-videobewerkingen of het gebruik van subjectievere evaluatieparameters die eenvoudig te beoordelen zijn in de praktijk. De praktijk leert dat vooral onderzoekers gebruik maken van de 2D-videobewerkingen. Therapeuten in het werkveld hebben vaak niet de tijd of kennis om zulke bewerkingen te doen en gaan dus sneller overgaan tot visuele beoordelingen. Machine learning zou voor beide doelgroepen een brug kunnen zijn naar consistent objectieve evaluaties. Zo heeft dit onderzoek

gevonden dat keypoint detection het mogelijk maakt om de 2D-videobewerkingen grotendeels of volledig automatisch te laten verlopen. Ondanks het feit dat er nog steeds kleine foutenmarges op de voorspellingen zitten, kunnen therapeuten deze beelden wel gebruiken om hun eigen evaluaties op te baseren. Bovendien gebeuren evaluaties aan de hand van 2D-videobewerkingen in de praktijk zelden in éénzelfde behandeling. Aangezien het voor de therapeut zelf te veel tijd in beslag neemt om de video's te verwerken en te analyseren. Hierdoor kent de patiënt in veel gevallen pas later zijn resultaat. Wanneer de videobewerkingen automatisch en zonder manuele input gebeuren zou de patiënt op deze manier nog tijdens dezelfde behandeling zijn (objectieve) resultaten kunnen krijgen. Ook voor therapeuten met minder technische kennis kan dit een oplossing bieden. De vaak gesofisticeerde software is voor veel mensen afschrikwekkend. Indien deze stap overgeslagen wordt kunnen ook deze therapeuten gebruik maken van correcte markeringen op de videobeelden.

Ook kan dit onderzoek de basis leggen voor het ontwikkelen van specifieke neurale netwerken voor het beoordelen van bewegingstesten. Hiervoor zijn echter goede, betrouwbare en grote datasets nodig. Deze datasets zijn niet eenvoudig om op te stellen. Het is namelijk niet vanzelfsprekend om duizenden van deze testen gestandaardiseerd te filmen en te beoordelen. In het ideale scenario gebeuren de evaluaties ook door zo weinig mogelijk verschillende testers. Om dit te bereiken zouden andere onderzoeken zich kunnen focussen op het opstellen van een uniforme evaluatieschaal. Het ultieme doel van dit soort classificaties is om te voorspellen of personen in een bepaalde categorie behoren tot een risicogroep. Het is dus belangrijk dat deze categorieën rechtstreeks in verband kunnen gebracht worden met het voorkomen van sommige blessures. Het kan ook omgekeerd een positieve evolutie teweegbrengen. Wanneer een neuraal netwerk in staat is om automatisch een test te beoordelen, zou het mogelijk zijn om sneller op grote schaal testen uit te voeren. Wanneer er geen manuele verwerkingen meer hoeven te gebeuren, kunnen uitgebreide datasets op die manier veel efficiënter worden opgesteld. Bovendien is het toepassen van keypoint detection niet enkel gelimiteerd tot de knievalgus tijdens de lateral stepdown, de voorname focus van dit onderzoek. Ook andere parameters zoals balans en bekkenpositie kunnen eventueel in kaart gebracht worden door deze voorspellingen. Andere bewegingstesten zoals bijvoorbeeld de drop jump of tuck jump, waar alignatie van de gewrichten ook een cruciaal evaluatiecriterium is, kunnen baat hebben bij het gebruik van machine learning.

Er werden echter wel een aantal belangrijke beperkingen vastgesteld in dit onderzoek. Door de focus op nauwkeurigheid waren er duidelijk negatieve effecten zichtbaar op de uitvoersnelheid. Hierdoor is real-time verwerking van een video niet mogelijk aan een hoge framerate. Keypoint bepaling voor 30 frames of 1 seconde van een video nam minstens 9 seconden in beslag op een desktop met degelijke videokaart. Door video's te verwerken aan een lager aantal frames per seconde zou men dit eventueel kunnen versnellen. Hier moet evenwel opgepast worden dat belangrijke informatie niet gemist wordt. De voornaamste reden voor deze beperking is dat er telkens 2 netwerken een voorspelling moeten uitvoeren. In eerste instantie wordt de bounding box bepaald, nadat pas de keypoints. Door in accuraatheid in te leveren, kan er aan snelheid gewonnen worden. Modellen met minder complexiteit zouden in dit geval gebruikt kunnen worden als alternatief. Verder onderzoek zal moeten uitwijzen wat de minimaal vereiste precisie is van de voorspellingen voor objectieve evaluaties. Ook toepassingen op mobiele telefoon

lijken met de huidige ontwikkelingen nog niet aan de orde. De netwerken die ontworpen zijn om te werken op gsm's zijn een heel stuk minder accuraat. De voornaamste reden hiervoor is dat ze sterk moeten inboeten aan complexiteit. Om in de toekomst mobiele voorspellingen mogelijk te maken, zullen dus ofwel de mobiele telefoons krachtiger, ofwel de neurale netwerken minder complex moeten worden.



## **6. Conclusie**

Na uitvoerig onderzoek kan geconcludeerd worden dat machine learning wel degelijk een belangrijke meerwaarde kan hebben bij het evalueren van bewegingstesten. Verder onderzoek is echter noodzakelijk om sluitende evaluatiecriteria op te stellen. Hiermee moet het mogelijk zijn objectieve beoordelingen te bekomen op basis van de voorspelde gewrichten. Deze bevindingen kunnen de deur openzetten naar verdere ontwikkelingen in de kinesitherapie en het voorspellen van blessures. In veel gevallen zijn berekeningen uitgevoerd door computers namelijk veel objectiever dan observaties met het blote oog. Wanneer deze methodes op punt gesteld worden om de randgevallen te classificeren, is het mogelijk om op termijn deze subjectieve criteria verder te optimaliseren. Deze ontwikkelingen kunnen bovendien zorgen dat bestaande screenings op een grotere schaal onderzocht kunnen worden. Ook het afnemen van deze testen bij risicogroepen kan hierdoor grootschalig plaatsvinden. Indien men in staat is om efficiënt en gericht te screenen, kan dit voor een substantiële ontlasting van het gezondheidssysteem zorgen. Ook de positieve socio-economische impact hiervan is niet te onderschatten. Zowel voor therapeuten, onderzoekers, patiënten als de gezondheidszorg zijn technologische ontwikkelingen als deze meer dan welkom.



# A. Onderzoeksvoorstel

Het onderwerp van deze bachelorproef is gebaseerd op een onderzoeksvoorstel dat vooraf werd beoordeeld door de promotor. Dat voorstel is opgenomen in deze bijlage.

## A.1 Introductie

Sommige blessures die tijdens het sporten voorvallen kunnen leiden tot een lange revalidatie. Zo is een gemiddelde revalidatie na een voorste kruisbandreconstructie al snel 6-7 maanden vooraleer er opnieuw sportherstelling kan plaatsvinden. Vaak hebben zulke blessures bepaalde biomechanische kenmerken die het risico significant kunnen vergroten. Het herkennen en verhelpen van zulke risicofactoren is een belangrijke taak van de kinesitherapeut. Door de ernst van deze blessures is het belangrijk om voldoende aandacht aan preventie te besteden, voornamelijk bij de sporten met een groot risico. Bepaalde risicotvolle bewegingen kunnen snel en eenvoudig worden geëvalueerd aan de hand van een specifieke bewegingstest. Wanneer deze testen gebruikt worden is het echter moeilijk om met het blote oog deze objectief en consistent te evalueren. Het is mogelijk om de test te filmen en naderhand duidelijker de ernst in te schatten bij vertraagde beelden. Dit is echter zeer tijdrovend waardoor dit amper wordt gebruikt. Het meten via een 3D analyse is hiervoor een objectief alternatief. De grote kost en tijdsinvestering maken dit echter ook onbruikbaar in de praktijk. Door deze analyse te automatiseren en toegankelijker te maken zou dit kunnen zorgen voor een groter gebruik en meer objectiviteit van dit soort testen.

In dit onderzoek zullen volgende probleemstellingen worden onderzocht: *Is het mogelijk om aan de hand van AI (Artificial Intelligence) de evaluatie van de knievalgus tijdens een lateral stepdown te objectiveren en automatiseren?*

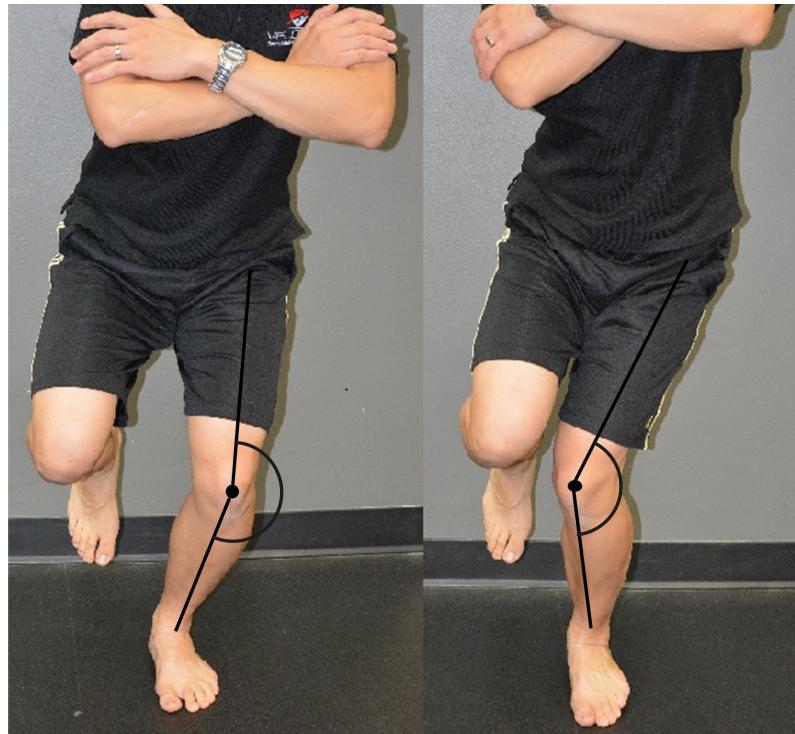


Figuur A.1: Lateral stepdown

- Welke mogelijkheden bestaan er om met AI videobeelden van bewegingen te analyseren?
- Op welke manieren is het mogelijk om de valgus te berekenen met AI?
- Hoe betrouwbaar is de evaluatie van de valgus met AI?

## A.2 State-of-the-art

Onderzoek toont aan dat een ongecontroleerde knievalgus belasting (Figuur A.2) een belangrijke risicofactor is voor een voorste kruisbandscheur (Hewett e.a., 2005, 4; Shultz, 2008). Bij sporten waar veel explosieve draaibewegingen worden uitgevoerd wordt dit risico nog een stuk groter (Besier, Lloyd, Cochrane & Ackland, 2000). Kinesitherapie heeft een belangrijke bijdrage bij het aanleren van een goede neuromusculaire controle over de knie. Via oefentherapie is het mogelijk om de kniecontrole tijdens excessieve valgusbewegingen te verbeteren (Bell e.a., 2013, 4; Ford e.a., 2003, 10). Hierbij is het noodzakelijk om een belangrijke nuance te maken. Het is niet noodzakelijk om deze bewegingen volledig te vermijden. Een bewuste controle over de knie tijdens deze bewegingen is voldoende om het risico op blessures te verlagen (Dischiavi, Wright, Hegedus & Bleakley, 2019). Om hier preventief te werk te kunnen gaan is het belangrijk te onderzoeken welke personen een verminderde neuromusculaire controle hebben over de knie. De lateral stepdown (Figuur A.1) of de single leg squat is aangetoond om hiervoor een geschikte en betrouwbare test te zijn (Ageberg e.a., 2010). Ook andere observationele screenings waaronder de drop-jump hebben bewezen dat ze over een grote validiteit beschikken voor het herkennen van risicotallen voor voorste kruisbandscheuren (Ekegren e.a., 2009, 9).



Figuur A.2: Valgushoek

### A.3 Methodologie

Voor dit onderzoek wordt gebruik gemaakt van 2D video analyses van de lateral stepdown in het frontale vlak. Via een uitgebreid literatuuronderzoek wordt onderzocht welke technieken momenteel beschikbaar zijn om aan de hand van AI de videobeelden te analyseren. Op basis van deze resultaten wordt bepaald welke techniek optimaal is voor dit onderzoek. Mogelijk gebeurt dit door aan de hand van bounding boxes het model te trainen om bepaalde lichaamsdelen zoals de knie en de enkel te leren herkennen. Ook of de test aan de linker- of rechterzijde wordt uitgevoerd zal hiermee worden bepaald. De valgushoek die ontstaat in de knie kan via deze boxes gemeten worden.

Het trainen van het model zal gebeuren met een dataset van 300 video's van uitvoeringen van de lateral stepdown. Elk van deze testen is gestandaardiseerd voor zowel de linker- als de rechter zijde. De mogelijke valgusbewegingen zijn hierbij telkens consistent gescoord. Een deel van deze video's zal gebruikt worden om het model te trainen, het andere deel zal gebruikt worden om het getrainde model te testen. De resultaten van dit deel van de video's zullen achteraf worden vergeleken met de oorspronkelijk uitgevoerde evaluatie. Zo zal de betrouwbaarheid van de AI evaluatie bekijken kunnen worden. Uiteindelijk zal een proof of concept ontwikkeld worden waarbij de focus ligt op de piekvalgus tijdens de beweging. Met dit model is het de bedoeling om een nieuwe uitvoering van een lateral stepdown te evalueren en automatisch de grootste waargenomen valgusbeweging te verwerken.

#### A.4 Verwachte resultaten

Er wordt verwacht dat een model zal kunnen ontwikkeld worden dat in staat is om te herkennen aan welke zijde een opgeladen test werd uitgevoerd. Ook zou het model de vereiste lichaamsdelen moeten kunnen herkennen voor het berekenen van de valgushoek. Door het gebruik van een gestandaardiseerde dataset wordt verwacht dat de resultaten van ongerelateerde uitvoeringen van de lateral stepdown met een vrij grote betrouwbaarheid zouden geëvalueerd kunnen worden. Door deze standaardisatie zal moeten onderzocht worden of het model in staat is om in een andere context zoals een ander standpunt van de camera of sterk verschillende achtergronden te werken. Ook uitvoeringen waarbij het been niet zichtbaar is door kledij zullen onderzocht moeten worden.

#### A.5 Verwachte conclusies

Aangezien het zorgvuldig trainen van een AI model een grote (10.000+) dataset nodig heeft is er een mogelijkheid dat ongerelateerde of minimaal afwijkende uitvoeringen niet correct worden geëvalueerd. Piekvalgus van testen in een gelijkaardige context zoals deze waarin het model getraind werd kunnen met een vrij grote betrouwbaarheid worden bepaald. Bedenkingen zullen echter gemaakt moeten worden over het aantal graden waarbij een valgusbeweging significant wordt beschouwd. In de toekomst kan hier nog een onderscheid en classificatie in worden bepaald aan de hand van het aantal graden. Aangezien alle opnames gemaakt zijn met 2D analyses en voor een grotere betrouwbaarheid de voorkeur wordt gegeven aan 3D analyse (Ho, Deaver, Nelson & Turner, 2019), is het mogelijk dat de berekende valgushoek geen volledig representatieve weergave geeft van de realiteit. Deze proof of concept zou echter een aanzet kunnen zijn naar verdere ontwikkelingen zoals automatiseren van 3D analyse in het domein van de biomechanica en injury prediction.

## Bibliografie

- Ageberg, E., Bennell, K., Hunt, M. A., Simic, M., Roos, E. M. & Creaby, M. W. (2010). Validity and inter-rater reliability of medio-lateral knee motion observed during a single-limb mini squat. *BMC Musculoskeletal Disorders*.
- Andriluka, M., Pishchulin, L., Gehler, P. & Schiele, B. (2014, juni 28). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, 28 juni 2014 (pp. 3686–3693). doi:10.1109/CVPR.2014.471
- Bell, D. R., Oates, D. C., Clark, M. A. & Padua, D. A. (2013). Two- and 3-dimensional knee valgus are reduced after an exercise intervention in young adults with demonstrable valgus during squatting. *Journal of athletic training*, 48, 442–449. doi:10.4085/1062-6050-48.3.16
- Besier, T., Lloyd, D., Cochrane, J. & Ackland, T. (2000). External loading of the knee joint during running and cutting maneuvers. *Medicine & Science in Sports & Exercise*.
- Carreira, J., Agrawal, P., Fragkiadaki, K. & Malik, J. (2015, juli 23). Human Pose Estimation with Iterative Error Feedback. arXiv: 1507.06550v3 [cs.CV]
- Chappell, J. D., Yu, B., Kirkendall, D. T. & Garrett, W. E. (2002). A comparison of knee kinetics between male and female recreational athletes in stop-jump tasks. *The American journal of sports medicine*, 30, 261–267. doi:10.1177/03635465020300021901
- Chmielewski, T. L., Hodges, M. J., Horodyski, M., Bishop, M. D., Conrad, B. P. & Tillman, S. M. (2007). Investigation of clinician agreement in evaluating movement quality during unilateral lower extremity functional tasks: a comparison of 2 rating methods. *The Journal of orthopaedic and sports physical therapy*, 37, 122–129. doi:10.2519/jospt.2007.2457
- De Blaizer, C., De Ridder, R., Willems, T., Danneels, L. & Roosen, P. (2019). Reliability of two functional clinical tests to evaluate trunk and lumbopelvic neuromuscular control and proprioception in a healthy population. *Brazilian journal of physical therapy*, 23, 541–548. doi:10.1016/j.bjpt.2018.10.014

- Dischiavi, S. L., Wright, A. A., Hegedus, E. J. & Bleakley, C. M. (2019). Rethinking Dynamic Knee Valgus and Its Relation to Knee Injury: Normal Movement Requiring Control, Not Avoidance. *Journal of Orthopaedic & Sports Physical Therapy*, 49(4), 216–218. doi:10.2519/jospt.2019.0606
- Ekegren, C. L., Miller, W. C., Celebrini, R. G., Eng, J. J. & Macintyre, D. L. (2009). Reliability and validity of observational risk screening in evaluating dynamic knee valgus. *The Journal of orthopaedic and sports physical therapy*, 39, 665–674. doi:10.2519/jospt.2009.3004
- Ford, K. R., Myer, G. D. & Hewett, T. E. (2003). Valgus knee motion during landing in high school female and male basketball players. *Medicine and science in sports and exercise*, 35, 1745–1750. doi:10.1249/01.MSS.0000089346.85744.D9
- He, K., Zhang, X., Ren, S. & Sun, J. (2015, december 10). Deep Residual Learning for Image Recognition. arXiv: 1512.03385v1 [cs.CV]
- Herrington, L. (2014). Knee valgus angle during single leg squat and landing in patellofemoral pain patients and controls. *The Knee*, 21, 514–517. doi:10.1016/j.knee.2013.11.011
- Hewett, T. E., Myer, G. D., Ford, K. R., Heidt, R. S., Colosimo, A. J., McLean, S. G., ... Succop, P. (2005). Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a prospective study. *The American journal of sports medicine*, 33, 492–501. doi:10.1177/0363546504269591
- Ho, K.-Y., Deaver, B. B., Nelson, T. & Turner, C. (2019). Using a Mobile Application to Assess Knee Valgus in Healthy and Post-Anterior Cruciate Ligament Reconstruction Participants. *Journal of Sport Rehabilitation*, 28(5), 532–535. doi:10.1123/jsr.2018-0278
- Ireland, M. L. (1999). Anterior cruciate ligament injury in female athletes: epidemiology. *Journal of athletic training*, 34, 150–154.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2014, mei 1). Microsoft COCO: Common Objects in Context. arXiv: 1405.0312v3 [cs.CV]
- Malinzak, R. A., Colby, S. M., Kirkendall, D. T., Yu, B. & Garrett, W. E. (2001). A comparison of knee joint motion patterns between men and women in selected athletic tasks. *Clinical biomechanics (Bristol, Avon)*, 16, 438–445. doi:10.1016/s0268-0033(01)00019-5
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochimia medica: Biochimia medica*, 22(3), 276–282.
- McLean, S. G. [S. G.], Walker, K., Ford, K. R., Myer, G. D., Hewett, T. E. & van den Bogert, A. J. (2005). Evaluation of a two dimensional analysis method as a screening and evaluation tool for anterior cruciate ligament injury. *British journal of sports medicine*, 39, 355–362. doi:10.1136/bjsm.2005.018598
- McLean, S. G. [Scott G.], Lipfert, S. W. & van den Bogert, A. J. (2004). Effect of gender and defensive opponent on the biomechanics of sidestep cutting. *Medicine and science in sports and exercise*, 36, 1008–1016. doi:10.1249/01.mss.0000128180.51443.83
- Munro, A., Herrington, L. & Comfort, P. (2012). Comparison of landing knee valgus angle between female basketball and football athletes: possible implications for anterior cruciate ligament and patellofemoral joint injury rates. *Physical therapy in sport* :

- official journal of the Association of Chartered Physiotherapists in Sports Medicine*, 13, 259–264. doi:10.1016/j.ptsp.2012.01.005
- Newell, A., Yang, K. & Deng, J. (2016, maart 22). Stacked Hourglass Networks for Human Pose Estimation. arXiv: 1603.06937v2 [cs.CV]
- Piva, S. R., Fitzgerald, K., Irrgang, J. J., Jones, S., Hando, B. R., Browder, D. A. & Childs, J. D. (2006). Reliability of measures of impairments associated with patellofemoral pain syndrome. *BMC musculoskeletal disorders*, 7, 33. doi:10.1186/1471-2474-7-33
- Rabin, A. & Kozol, Z. (2010). Measures of range of motion and strength among healthy women with differing quality of lower extremity movement during the lateral step-down test. *The Journal of orthopaedic and sports physical therapy*, 40, 792–800. doi:10.2519/jospt.2010.3424
- Russell, K. A., Palmieri, R. M., Zinder, S. M. & Ingersoll, C. D. (2006). Sex differences in valgus knee angle during a single-leg drop jump. *Journal of athletic training*, 41, 166–171.
- Scipy. (2019, december 19). *scipy.signal.findpeaks*. Verkregen 24 april 2020, van [https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find\\_peaks.html](https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html)
- Shultz, Y. S. S. (2008). Mechanisms of Noncontact Anterior Cruciate Ligament Injury. *Journal of Athletic Training*, 396–408.
- Soni, D. (2018, maart 22). Supervised vs. Unsupervised Learning. Verkregen 16 april 2020, van <https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d>
- Souza, R. B. & Powers, C. M. (2009). Differences in hip kinematics, muscle strength, and muscle activation between subjects with and without patellofemoral pain. *The Journal of orthopaedic and sports physical therapy*, 39, 12–19. doi:10.2519/jospt.2009.2885
- Sun, K., Xiao, B., Liu, D. & Wang, J. (2019, februari 25). Deep High-Resolution Representation Learning for Human Pose Estimation. arXiv: 1902.09212v1 [cs.CV]
- Tompson, J., Goroshin, R., Jain, A., LeCun, Y. & Bregler, C. (2014, november 16). Efficient Object Localization Using Convolutional Networks. arXiv: 1411.4280v3 [cs.CV]
- Toshev, A. & Szegedy, C. (2013, december 17). DeepPose: Human Pose Estimation via Deep Neural Networks. doi:10.1109/CVPR.2014.214. arXiv: 1312.4659v3 [cs.CV]
- Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... Xiao, B. (2019, augustus 20). Deep High-Resolution Representation Learning for Visual Recognition. arXiv: 1908.07919v2 [cs.CV]
- Wei, S.-E., Ramakrishna, V., Kanade, T. & Sheikh, Y. (2016, januari 30). Convolutional Pose Machines. arXiv: 1602.00134v4 [cs.CV]
- Willson, J. D. & Davis, I. S. (2008). Utility of the frontal plane projection angle in females with patellofemoral pain. *The Journal of orthopaedic and sports physical therapy*, 38, 606–615. doi:10.2519/jospt.2008.2706
- Willson, J. D., Ireland, M. L. & Davis, I. (2006). Core strength and lower extremity alignment during single leg squats. *Medicine and science in sports and exercise*, 38, 945–952. doi:10.1249/01.mss.0000218140.05074.fa
- Xiao, B., Wu, H. & Wei, Y. (2018, april 17). Simple Baselines for Human Pose Estimation and Tracking. arXiv: 1804.06208v2 [cs.CV]