Ontwerpen van een machine learning algoritme om hemodynamische verslechtering te voorspellen bij volwassen IC-patiënten

Technisch Geneeskundige Opdracht

TGO groep 26

Anoniem 1
Anoniem 2
Anoniem 3

UNIVERSITY OF TWENTE.

Opdrachtgevers

Technische Universiteit Eindhoven (TU/e)
Catharina Ziekenhuis Eindhoven (CZE)

Begeleiders

Anoniem 4
Anoniem 5
Anoniem 6

Datum

Maandag 26 juni 2023

1 Abstract

Introductie De Hemodynamic Deterioration Index (HDI) is een machine learning algoritme ontwikkeld om hemodynamische instabiliteit één uur van tevoren te voorspellen bij intensive care patiënten. Hiermee kan de informatiestress op de IC worden weggenomen door één voorspellende waarde. Artsen kunnen daarop proactief handelen, wat gezondheidsvoordelen oplevert voor patiënten. Telkens wordt een terugkoppeling gemaakt naar de implementatie van de HDI op de IC van het Catharina Ziekenhuis Eindhoven (CZE), waar veel cardio-thoracale chirurgische patiënten liggen.

Methoden De classificatiemodellen Logistische Regressie, Adaptive Boosting, Gradient Boosting en Random Forest werden getraind op de Medical Information Mart for Intensive Care-III (MIMIC-III) database. Na voorbewerking zijn 10.401 patiënten geïncludeerd met 207 variabelen, waaronder extra berekende parameters als shockindices en statistische (trend)parameters zoals gemiddelden en standaarddeviaties. Hyperparameter optimalisatie gebeurde in gestratificeerde 6-fold kruisvalidatie. De modellen werden hertraind na een verkennende gegevensanalyse waarin onder andere insignificant correlerende parameters werden verwijderd. Hier zijn 116 parameters overgehouden. Ook werd hertraind op een subset van alleen cardiale patiënten, wat relevant is voor het CZE. Uitkomstmaten bestonden onder andere uit area under receiver operator characteristic curve (AUROC), sensitiviteit, specificiteit en belangrijkheid van parameters.

Resultaten Gradient Boosting blijkt de beste classificator voor het voorspellen van HI, met een AUROC van 0,895 voor de voorbewerkte MIMIC-dataset (*se. 0,755, sp. 0,891*), een AUROC van 0,855 na aanvullende bewerkingsstappen in een verkennende gegevensanalyse (*se. 0,709, sp. 0,859*) en een AUROC van 0,861 voor de cardiale patiëntengroep (*se. 0,811, sp. 0,760*). Recent gemeten systolische bloeddrukken, standaarddeviaties van de hartslag en gemiddelde zuurstofsaturaties blijken belangrijke parameters voor de voorspelling van HI.

Conclusie De totstandkoming van de HDI is uitgebreid en klinisch onderbouwd, tevens volgens de TRIPOD-richtlijn, wat zorgt voor een goed interpreteerbaar en reproduceerbaar model. Gradient Boosting (AUROC=0,895) is het best presterende classificatie ML-model voor de voorspelling van HI en werkt beter dan vergelijkbare modellen voor HI. Daarnaast zijn de prestaties van de HDI goed bij de subset van cardiale patiënten, wat bewijs zou kunnen zijn dat de HDI ook goed zou kunnen werken op de intensive care van het CZE. De belangrijkste parameters voor de voorspelling van HI waren de systolische bloeddrukken, standaarddeviaties van de hartslag en gemiddelde zuurstofsaturaties.

Trefwoorden: *Hemodynamische Instabiliteit, Hemodynamic Deterioration Index, HDI, Machine Learning, Multivariabel Predictiemodel, Intensive Care.*

Inhoudsopgave

2 Introductie

Hemodynamische instabiliteit (HI) is slecht gedefinieerd [1, 2] maar een belangrijk klinisch begrip om aan te geven dat een patiënt perfusiefalen ontwikkelt, wat kan leiden tot (symptomen van) circulatoire shock en/of gevorderde hartfalen. Het wordt ook gebruikt wanneer een of meer metingen al dan niet pathologisch buiten normaalwaarden vallen [3]. Tot wel één op de drie patiënten op de intensive care (IC) raakt op enig moment in circulatoire shock [4] en wordt dus hemodynamisch instabiel. Het komt dus veel voor en heeft tevens een hoge mortaliteit van 38,3% [4, 5]. Daarom vormt het monitoren van de hemodynamica een van de belangrijkste onderdelen van de zorg op de IC, waarbij gebruik wordt gemaakt van heel veel klinische variabelen zoals vitale kenmerken, hematologische en chemische labwaarden en variabelen verkregen uit lichamelijk onderzoek [1, 6, 7]. Omdat vroegtijdige interventies gezondheidsvoordelen kunnen opleveren [8] is het nodig dat al deze parameters tijdig en frequent worden gemeten en geïnterpreteerd.

De overweldigende kwantiteit aan klinische variabelen zorgt voor informatiestress (*data overload*) bij artsen op de IC [9]. Dat maakt het moeilijk om de hemodynamica van de patiënt in de gaten te houden, laat staan te voorspellen. Hierdoor kunnen artsen de benodigde interventie pas uitvoeren op het moment dat er sprake is van een instabiel incident [6]. Idealiter wordt er ingegrepen voordat de patiënt instabiel wordt. Zo kunnen HI en gevorderde complicaties voorkomen worden. Onderzoek toont aan dat verlate toediening van vasopressoren en noradrenaline geassocieerd wordt met een verhoogde mortaliteit en vertraagd herstel van de gemiddelde arteriële bloeddruk (*mean arterial pressure*, MAP) in patiënten met septische shock [8, 10].

Er zijn nieuwe parameters ontwikkeld die een hogere correlatie hebben met HI dan losse klinische variabelen. Deze combineren meerdere klinische variabelen, zoals hartslag en MAP, in een poging de informatiestress te verminderen. Voorbeelden daarvan zijn *shock index* (SI), *rate pressure product* (RPP) en de ROX index. Ze zijn van toegevoegde waarde bewezen in medische alarmeringssystemen, maar blijken nog steeds geen goede prognosemaat voor HI, vooral vanwege de hoge maat van vals-positieve alarmen en omdat deze gericht zijn op detectie en nog niet op voorspellen [11, 12].

Om de behandelaar vroegtijdig te kunnen waarschuwen over de verslechtering van de klinische situatie van een patiënt, worden steeds vaker *early warning systems* (EWS), ofwel vroegtijdige waarschuwingssystemen, geïntroduceerd op basis van (supervised) machine learning (ML) [13, 14]. Een dergelijk ML-model wordt getraind met grote gelabelde (patiënt)datasets, waarbij het model met behulp van verschillende (klinische) parameters probeert te voorspellen of de gelabelde gebeurtenis zal plaatsvinden [13, 15]. Deze voorspellende ML-modellen worden reeds ingezet in de kliniek [13], zoals bij het voorspellen van (acuut) nierfalen [16, 17], sepsis [18, 19] en bij het voorspellen van intra-operatieve hypotensie op basis van de arteriële drukgolfvorm [20].

Een bestaand ML-model dat hemodynamische instabiliteit bij IC-patiënten voorspelt is de Hemodynamic Stability Index (HSI) [6, 11, 21]. Deze index is ontwikkeld op basis van retrospectieve IC-patiëntdata uit de eICU *Collaborative Research Database* [22] en uit de *Medical Information Mart for Intensive Care III*-database (MIMIC-III) [23]. Het ML-model geeft, op basis van realtime gemeten klinische parameters, een enkele predictieve waarde voor de kans op hemodynamische instabiliteit. Een bijkomend voordeel van het gebruik van een enkele predictieve waarde is dat de informatiestress voor een behandelend arts sterk wordt verminderd [6]. De methoden voor het opstellen van de HSI worden echter onvoldoende (klinisch) onderbouwd. Hierbij mist er vooral transparantie over keuzes die zijn gemaakt in het modelleringsproces en duiding over welke specifieke klinische meetgegevens worden meegenomen in het ML-model. Een duidelijk voorbeeld hiervan is dat het niet helder is of de gemeten bloedgassen arterieel of veneus zijn afgenomen, wat van invloed kan zijn op het ML-model en de resultaten ervan, omdat deze twee soorten bloedgassen significant van elkaar verschillen [24].

Onduidelijke en niet-transparante rapportage van het modelleringsproces van voorspellende ML-modellen is een breder probleem en maakt het voor clinici en onderzoekers moeilijk om ze te interpreteren, reproduceren en valideren [25]. Daarnaast worden verschillende ML-technieken onjuist en inconsistent gebruikt, waardoor soms onjuiste conclusies worden getrokken [26]. Hierdoor neemt de betrouwbaarheid van predictieve ML-modellen, zoals de HSI, af en zullen clinici minder snel geneigd zijn om

resultaten van dergelijke modellen te gebruiken bij hun klinische besluitvorming [26]. Daarom is het belangrijk dat clinici worden geïnformeerd over de (klinische) afwegingen die in voorspellende ML-algoritmes worden gemaakt.

Het doel van dit onderzoek is daarom het ontwerpen van een klinisch onderbouwd ML-model dat hemodynamische instabiliteit kan voorspellen bij IC-patiënten. Hiervoor wordt de Hemodyamic Deterioration Index (HDI) voorgesteld, waarmee de informatiestress wordt verholpen door één voorspellende waarde weer te geven dat HI één uur van tevoren kan voorspellen. In die tijd kunnen clinici proactief handelen, wat gezondheidsvoordelen voor patiënten zou kunnen opleveren. Het algoritme zal worden getraind op *Medical Information Mart for Intensive Care III* (MIMIC-III) data [23] waarbij alle keuzes en aannames in het modelleringsproces klinisch (en technisch) worden onderbouwd en transparant worden gerapporteerd, in lijn met de *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis* (TRIPOD)-richtlijn [27]. Deze transparantie draagt bij aan de validatie en interpretatie van het ML-algoritme zodat het betrouwbaarder geïmplementeerd kan worden bij klinische besluitvorming van hemodynamische instabiele IC-patiënten. De transparantie komt ook de reproduceerbaarheid ten goede, zodat de HDI gepersonaliseerd en geoptimaliseerd toegepast kan worden op de IC's van verschillende ziekenhuizen. Uiteindelijk is de onderzoeksvraag opgesteld:

'Hoe kan hemodynamische verslechtering bij volwassen intensive care patiënten voorspeld worden door middel van een machine learning algoritme?'

Hiervoor werden voorbewerkingsstappen ondernomen om hemodynamische instabiliteit te herkennen in de data. Er werd onderzocht welke (afleidbare) parameters klinisch significant en belangrijk zijn in het voorspellen van HI. Tot slot zijn meerdere algoritmen getraind en zijn er subsetanalyses uitgevoerd, waarvan de prestaties met elkaar werden vergeleken. De modellen werden bijvoorbeeld opnieuw getraind voor de subset cardiale patiënten om te onderzoeken hoe goed de HDI theoretisch zou kunnen werken op de IC van het CZE, waar veel cardio-thoracale chirurgie patiënten verblijven. Er zijn hypothesen opgesteld aan de hand van het vooronderzoek in sectie 3. Deze hypothesen zijn aan het einde van het theoretisch kader opgenomen.

3 Theoretisch kader

In dit hoofdstuk zal meer context en toelichting worden gegeven over de onderwerpen hemodynamische instabiliteit en machine learning algoritmes. Ten eerste zal de etiologie en definitie van hemodynamische instabiliteit worden uitgelicht en wordt het management van HI beschreven. Ten tweede worden er verschillende ML-modellen uitgelegd aan de hand van hun eigenschappen en achterliggende mechanismen. Tot slot wordt aan de hand van voorgaande theoretische uiteenzetting hypothesen opgesteld

3.1 Hemodynamische instabiliteit

Zoals in de inleiding beschreven, wordt hemodynamische instabiliteit gedefinieerd als een of twee van de volgende kenmerken: 1) perfusiefalen wat kan leiden tot (symptomen van) circulatoire shock en/of gevorderde hartfalen en 2) wanneer één of meer metingen, al dan niet pathologisch, buiten normaalwaarden vallen [3]. Wat deze twee kenmerken precies inhouden wordt in dit hoofdstuk beschreven in de volgorde van deze twee kenmerken van HI.

3.1.1 Perfusiefalen

Perfusiefalen kan ontstaan door een verlaagde *preload*, contractiliteit en/of *afterload*, wat in een vergevorderd stadium respectievelijk overeenkomt met hypovolemische, cardiogene en/of distributieve shock. Een verhoogde afterload kan in een vergevorderd stadium zorgen voor een obstructieve shock. Deze vier pathologieën zijn vormen van circulatoire shock, wat beschreven kan worden als een levensbedreigende, gegeneraliseerde vorm van acuut circulatoir falen, geassocieerd met onvoldoende zuurstofaanbod aan cellen [5, 28]. Algemene symptomen van HI door perfusiefalen zijn fatigue, dyspnoe, een verminderde capillaire flow en hiermee koudere extremiteiten, oligurie van <0,5 mL/kg/h en een veranderde mentale staat waarbij verwarring, desoriëntatie of een delirium te herkennen zijn [5]. Onderliggende (patho)fysiologische symptomen kunnen het best beschreven worden aan de hand van de gevorderde stadia van HI, de typen shock, zoals onderstaand. Hypovolemische shock wordt gekenmerkt door een verminderde preload van het hart, dit is het volume waarmee het ventrikels worden gevuld. Een verlaagde preload wordt veroorzaakt door een laag veneus terugkerend volume, wat komt door verlies aan circulerend volume, bijvoorbeeld door uitdroging of een bloeding. Het gevolg is een lager hartminuutvolume, ofwel de *cardiac output* (CO), met hiermee een lagere orgaanperfusie wat kan leiden tot shock en orgaanfalen. Cardiogene shock kan optreden wanneer de contractiliteit van het hart dusdanig verminderd is door hartfalen bijvoorbeeld door een infarct, myocarditis, ernstige aritmie of wanneer de hartkleppen ernstig zijn aangedaan [29]. De pathologieën verslechteren de CO wat hypotensie en verlaagde orgaanperfusie als gevolg heeft. Distributieve shock kan optreden wanneer de vaattonus systemisch ernstig zijn aangedaan, waarbij de bloedvaten verwijden (perifere vasodilatatie) en ze een hogere permeabiliteit krijgen. Dit kan worden veroorzaakt door complementactivatie of door een verminderde autonome neurologische innervatie (zoals bij een dwarslaesie). Complementactivatie kan optreden als gevolg van een stressreactie van het lichaam, zoals na of tijdens een operatie, bij een acute ontsteking, bij ischemie of tijdens een infectie zoals bij sepsis het geval is. De hiervoor genoemde symptomen zijn veelvoorkomend bij patiënten die acuut worden opgenomen op de IC. Deze overmatige complement-activatie wordt systemisch inflammatoir respons syndroom genoemd (SIRS) [30, 31]. Dit fysiologische proces zorgt voor een verlaagde afterload. De afterload geeft de weerstand aan die de vaten geven op het gepompte volume door het hart Verlaging van de afterload betekent dus een lagere bloeddruk en kan leiden tot perfusiefalen. In verdere stadia geeft dit een distributieve shock, bijvoorbeeld in sepsis (septische shock), waarbij orgaanfalen ontstaat [32]. De laatste vorm van circulatoire shock is de obstructieve shock. Deze wordt veroorzaakt door een obstructie in de vaten of van buiten op de vaten, zoals een trombus bij longembolie, maar ook door een harttamponade of een spanningspneumothorax. Dit kan leiden tot een verhoogde afterload en dus perfusiefalen.

3.1.2 Metingen buiten normaalwaarden

Een andere kenmerk dat kan zorgen voor HI is wanneer één of meer metingen, al dan niet pathologisch, buiten de normaal waarden vallen. Wanneer deze parameters buiten de normaalwaarden vallen is dit bewijs van verlaagde perfusie en/of hypovolemie. Deze normaalwaarden kunnen zelfstandige parameters zijn in de vorm

van de bloeddruk, hartminuutvolume of de gemengd veneuze zuurstofsaturatie (SvO_2). Daarnaast kunnen deze normaalwaarden ook dynamische parameters zijn om hypovolemie vast te stellen. Dit is mogelijk door te kijken naar het effect van een vloeistoftoediening of door de fysiologische respons hiervan te voorspellen aan de hand van bloeddruk veranderingen door ademhaling, met een autologe vloeistof-responstest waarbij de benen passief verhoogd worden of met echo-onderzoek. Bij een volumeverhoging wordt een relatief kleine bolus met een hoge stroomsnelheid geïnjecteerd wat resulteert in een verhoging van de preload. Vervolgens kijkt men of een van de hiervoor genoemde zelfstandige parameters positieve reactie vertoont wat wijst op hypovolemie. Als tweede kan men de polsdrukvariatie (PPV) onderzoeken tussen in- en uitademing. Een hoge variatie in druk komt overeen met een lage preload. Met een echo onderzoek is het mogelijk om de 'instorting' van de vena cava in beeld te brengen tijdens verschillende fasen van de ademhaling. Een hoge mate van instorting van de vena cava wordt geassocieerd met een laag veneus terugkerend volume en hiermee hypovolemie. Als laatst kan met behulp van de passieve beenheffingstest een autologe vloeistofresponsiviteit test uitvoeren waarmee wordt getoetst of het verhoogd terugkerend veneus volume, afkomstig van de benen, de CO bevordert. Positieve bevinding bij deze onderzoeken wijst op vloeistof responsiviteit en dus hypovolemie [33].

3.2 Management van HI

De hemodynamiek van HI-patiënten kan gereguleerd worden met medicatie, vloeistoftherapie (volumeresuscitatie), bloedtransfusies en in het uiterste geval met mechanische hartondersteuning. De medicatietoedieningen bestaan uit vasopressoren en inotropica. De vloeistoftherapieën bestaan uit crystalloïde en colloïde oplossingen.

3.2.1 Vasopressoren

De typen medicatie die onder vasopressoren vallen worden gebruikt voor vasoconstrictie. Deze interventie wordt veel ingezet bij patiënten met hypotensie. Vasopressoren bestaan vooral vooral uit alfa-agonisten die perifere arteriële vasoconstrictie stimuleren wat zorgt voor een verhoogde systemische vasculaire weerstand (SVR). Als de SVR, dus de afterload toeneemt, zal de gemiddelde arteriële bloeddruk (*Mean Arterial Pressure*, MAP) ook toenemen. Verder bestaat het voornamelijk uit contractiliteit en hiermee CO verhogende beta-1 agonisten. Vasopressine heeft hier bovenop een anti-diuretisch effect wat hypovolemie tegengaat. Veel gebruikte vasopressoren zijn: fenylefrine, norepinephrine, epinephrine, dopamine en vasopressine [4, 34].

3.2.2 Inotropica

Inotrope medicatie heeft een werking op de hartspier en wordt veelal gebruikt bij patiënten met cardiogene shock. Dobutamine, een voorbeeld hiervan heeft in tegenstelling tot vasopressoren meer beta dan alfa activiteit en dus sterkere bevordering van de hartspier-contractiliteit. Milrinon, een andere veelgebruikte medicatie, verhoogt cyclische AMP niveaus die aangrijpen op het hartspierweefsel. Beide medicijnen leiden tot directe verhoging van CO en stabilisatie van de MAP. Zoals genoemd zijn dobutamine en milrinon veel gebruikte voorbeelden van inotrope medicatie [34].

3.2.3 Vloeistof toediening

Patiënten die een vloeistoftoediening krijgen komen bijna altijd in aanraking met crystalloïden, colloïden of subtypen hiervan. Het wordt veel gebruikt bij patiënten met hypovolemische shock. In de literatuur bestaat een aanhoudende discussie welke typen van vloeistoftoedieningen beter werkt, maar in essentie hebben deze hetzelfde effect. Beide verhogen namelijk de preload door het algehele circulatoire volume te verhogen [35].

3.2.4 Bloedtransfusies

Packed Red Blood Cells (PRBC's) worden gemaakt door rode bloedcellen van het bloedplasma te scheiden door middel van centrifuge. Ze hebben een typisch volume van 250 tot 300 mL en bevatten 65% tot 80% hematocriet. Ze worden typisch gebruikt om shock te voorkomen bij interne bloedingen of bij bloedarmoede. Effecten van deze bloedtransfusie zijn een significante verhoging van de MAP en SVR en een verlaging van de CO en hartslag [36].

3.3 Machine learning

Om hemodynamische instabiliteit te voorspellen met behulp van een machine learning (ML) model, is het van belang de typen modellen te identificeren die hiervoor het meest geschikt

zijn. Machine learning is een vorm van kunstmatige intelligentie (AI) die autonoom patronen kan ontdekken in (grote) datasets. ML kan primair worden opgedeeld in twee soorten probleemoplossende methoden, namelijk *unsupervised* en *supervised* ML.

Bij unsupervised ML wordt gebruik gemaakt van een ongelabelde dataset en probeert het model zelf overeenkomsten of verbanden vast te stellen. Bij supervised ML maakt men gebruik van een dataset met een bekende uitkomstmaat (continu of categoriaal), waar het ML-model op getraind moet worden. Het model probeert de belangrijkste kenmerken binnen elk datapunt te herkennen, om zo zelf de uitkomstmaat te kunnen voorspellen. Met een regressiemodel kan een continue uitkomstmaat worden voorspeld, terwijl met een classificatiemodel een categoriale (ofwel dichotome) uitkomstmaat, zoals HI, kan worden voorspeld. Voor het classificatiemodel zou de dichotome uitkomstmaat hemodynamische instabiliteit dus als label kunnen worden genomen. Hiermee zou het ML-model, na training met een (retrospectief) HI-gelabelde patiëntdataset, aan de hand van nieuwe, onbekende patiëntgegevens kunnen voorspellen of de patiënt hemodynamisch instabiel wordt. In dit onderzoek wordt hierom gebruik gemaakt van een supervised ML-model met gelabelde patiëntdata. Er bestaan veel verschillende categorieën supervised classificatie ML-modellen, waarbij in dit onderzoek de focus zal liggen op *lineaire* en *beslisboom* modellen.

3.3.1 Lineaire classificatiemodellen

Lineaire classificatie ML-modellen gaan uit van een lineaire relatie tussen de kenmerken in de datapunten en de uitkomstmaat. Hierbij probeert het model het gewicht van ieder kenmerk te leren, dat de mate van invloed van het kenmerk op het voorspellen van de categorische uitkomstmaat aanduidt. Logistische regressie is veel gebruikte lineaire classificator.

Logistische regressie
Logistische regressie is dus een lineair classificatie model, waarbij een dichotome uitkomstvariabele kan worden voorspeld op basis van een of meerdere verklarende variabelen [37]. Hierbij wordt de uitkomstvariabele getransformeerd zodat een benadering van lineaire regressie mogelijk is. De uitkomst ($\overline{y}_i$) van een logistisch regressie model wordt als volgt beschreven:

$$ln(odds) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \dots b_nx_n$$

Met x_n als de verschillende verklarende variabelen en b_n als de gewichten die door het ML-model worden bepaald, die bepalen hoeveel de verklarende variabelen bijdragen aan het voorspellen van de uitkomstmaat. De *odds* is de verhouding van de kans (p) op het optreden van instabiliteit ten opzichte van de kans $(1-p)$ op het niet optreden van instabiliteit. Waarbij ln(*odds*) de *log odds* of *logit* wordt genoemd en als volgt kan worden beschreven:

$$odds = \frac{p(label = instabiel)}{1 - p(label = instabiel)}$$

Dit geheel kan worden omgeschreven naar de kans (p) voor het optreden van instabiliteit, wat erg lijkt op een sigmoïdefunctie en dus altijd tussen 0 en 1 ligt:

$$p = \frac{1}{1 + e^{-(b_0+b_1x_1+b_2x_2+b_3x_3 \dots b_nx_n)}}$$

Voordat het model kan worden getraind, moet eerst normalisatie of standaardisatie worden toegepast. De parameters verschillen onderling in eenheden en bereik. Dat maakt het heel moeilijk voor logistische regressie om bijvoorbeeld zuurstofsaturatie (±95-100%) direct te vergelijken met pH (±7,4) of lichaamstemperatuur (±37^{o}C). Daarom worden alle waarden van de parameters geschaald naar eenzelfde bereik, bijvoorbeeld tussen 0 en 1 [38].

Om vervolgens de gewichten te berekenen wordt door het logistische ML-model slim gebruik gemaakt van een *verliesfunctie*. Deze functie geeft de fout, ofwel het kwadratische verschil tussen de voorspelde uitkomst (tussen 0 of 1, $\overline{y}_i$) en de daadwerkelijke uitkomst (enkel 0 of 1, y_i). Het ML-model probeert de uitkomst van deze verliesfunctie te minimaliseren door de gewichten van de verschillende verklarende variabelen te wijzigen, waardoor een beter voorspellend model ontstaat.

$$L = \sum_{i=1}^{n} (y_i - \overline{y}_i)^2$$

Vaak wordt, voordat het model wordt getraind met behulp van een verliesfunctie, nog regularisatie toegepast. Regularisatie zorgt voor vermindering van overfitting en voor het verbeteren van prestaties en generaliseerbaarheid van het model. Er zijn twee belangrijke vormen van regularisatie, namelijk L1-regularisatie en L2-regularisatie. Deze twee regularisatieparameters (λ), worden als strafterm aan de verliesfuntie toegevoegd. Hierdoor moeten de gewichten kleiner worden, omdat het ML-model de uitkomst van de verliesfunctie probeert te minimaliseren.

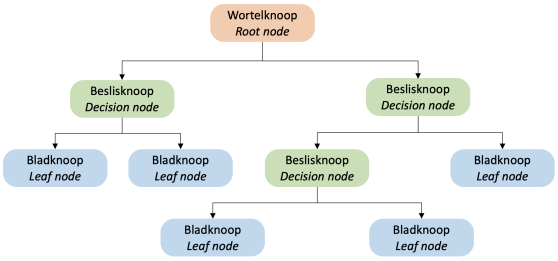
- L1-regularisatie, ofwel Lasso regressie, zorgt ervoor dat sommige gewichten minder tot géén waarde meer krijgen. Hierdoor zal het logistische regressiemodel deze verklarende variabelen niet meer meenemen in de modelvorming. Dit zorgt voor een selectie van variabelen en dus voor het vereenvoudigen van het model, wat zorgt voor verminderende kans op overfitting. Een bijkomend voordeel van deze methode is dat het minder rekenkracht vereist en dat het model beter is te interpreteren, omdat er minder verklarende variabelen zijn die gewicht aan het model geven.

- L2-regularisatie, ofwel Ridge regressie, zorgt voor minder waarde van sommige gewichten en zorgt dus ook voor een verminderde kans op overfitting. Met deze regressiemethode worden wel alle variabelen meegenomen in het model en zorgt dus niet voor vereenvoudiging. Daarnaast helpt L2-regularisatie bij het verminderen van multicollineariteit. Bij (multi)collineariteit zijn de verklarende variabelen afhankelijk van elkaar. Collineariteit ontstaat wanneer er een (sterke) correlatie bestaat tussen twee (onafhankelijke) variabelen, waardoor het lastiger is om de individuele gewichten te bepalen. De mate van collineariteit tussen de verklarende variabelen, moet dus worden beperkt bij toepassing van een logistisch regressie model.

De regularisatieparameter is een voorbeeld van een zogenaamde hyperparameter. Dit wordt gebruikt om de instellingen van het ML-model aan te passen (*tuning*).

3.3.2 Beslisboom classificatiemodellen

Een ander type classificatoren maakt gebruik van beslisbomen (Engels: *decision trees*). Een beslisboom is een manier om verbanden te leggen in een grote dataset op basis van klassen, de labeling in de data. In dit onderzoek zijn de labels 'stabiel' en 'instabiel'. Bij elke interne knoop wordt de dataset gesplitst op één variabele die de klassen kan voorspellen.



Figuur 1: Een schematische weergave van een beslisboom.

Zie afbeelding 1 hierboven voor een schematische weergave van een beslisboom. Een beslisboom begint (bovenaan) bij een wortelknoop (Engels: *root node*), dat is de variabele die de dataset het beste kan splitsen. Daarna komen de beslisknopen (Engels: *decision nodes*), die wederom de dataset splitsten op één variabele. Wanneer een terminaal onderscheid gemaakt kan worden tussen de klassen, wordt dat gerepresenteerd door externe knopen, zogenaamde bladknopen (Engels: *leaf nodes*). De splitsingen creëren aftakkingen in de beslisboom, die worden *branches* genoemd. Algoritmen kunnen op verschillende manieren gebruik maken van beslisbomen, door bijvoorbeeld andere criteria te stellen aan de splitsingen, overbodige takken weg te 'snoeien' (Engels: *pruning*) of meerdere beslisbomen op een bepaalde manier te gebruiken in de voorspellingen. Dit laatstgenoemde heet *ensemble learning*, waarvan hierna een paar voorbeelden worden uitgewerkt.

- **Adaptive boosting**
  Adaptive boosting, beter bekend als AdaBoost, is zo een ensemble ML-model. Allereerst wordt een beslisboom gevormd door de parameter met de laagste onzuiverheid, bepaald door de Gini-index, als wortelknoop te nemen. De Gini-index bepaalt de kans dat een willekeurige sample verkeerd wordt geklassificeerd. Dus, hoe lager de Gini-index, hoe nauwkeuriger de splistingen in de boom en de voorspellingen van het model. Aanvankelijk krijgen alle datapunten hetzelfde gewicht en vervolgens krijgen foutieve voorspellingen gedurende het trainen een hoger gewicht. AdaBoost blijft het model uitbreiden met een nieuwe beslisboom gebaseerd op de foute voorspellingen van de vorige, om deze vervolgens te verbeteren. Het sequentieel combineren van deze beslisbomen wordt *boosting* genoemd en houdt aan tot de hoogste prestatie is bereikt. AdaBoost is tevens een algoritme dat al ge-

bruikt wordt in predictiemodellen voor HI, onder andere door Rahman et al. [21, 39].

- **Gradient boosting**
  Gebaseerd op AdaBoost is later gradient boosting ontwikkeld. De twee algoritmen zijn hierom erg vergelijkbaar maar ze identificeren de fouten in het model op andere manieren. Bij AdaBoost wordt dit vooral gedaan door een foutieve voorspelling een hoger gewicht toe te wijzen. Bij gradient boosting wordt dit gedaan door de residuen in het ensemble te bepalen en later te minimaliseren en hiermee de prestatie van classificatoren te identificeren. Dit maakt gradient boosting minder gevoelig voor uitschieters in vergelijking met AdaBoost.

- **Random forest**
  Random forest (RF) is weer een ander type model dan *boosting*, maar maakt ook gebruik van beslisbomen. Er worden telkens nieuwe beslisbomen gemaakt parallel en onafhankelijk van elkaar, op basis van maar een subset van de beschikbare data. Dit verschilt met boosting, waar de losse beslisbomen sequentieel en adaptief worden gevormd. Elke beslisboom wordt dus getraind op weer een nieuwe, unieke subset van de data. De subset bevat dus niet alle rijen (patiënten) en niet alle kolommen (variabelen). Het wordt geselecteerd volgens de *bootstrap aggregating* (*bagging*) methode, waarbij willekeurige samples worden genomen van de database. Dat gebeurt 'met vervanging', wat betekent dat samples meerdere malen kunnen worden geselecteerd. Ook is er data dat helemaal niet zal worden geselecteerd, dat zijn de zogenaamde *out of bag samples*.

  Het aantal variabelen van een bootstrap sample wordt aangegeven met de letter m. Een standaard grootte is $\mathrm{m} = \sqrt{\mathrm{p}}$ met p het aantal variabelen van de oorspronkelijke dataset. De bagging methode zorgt voor minder correlatie en dus minder variantie tussen individuele beslisbomen. Hoe kleiner m, hoe minder variantie, maar hoe meer risico op overfitting. Daartegenover staat dat als m te klein wordt gekozen, er meer kans is op underfitting.

3.3.3 Hyperparameter optimalisatie en validatie

Een machine learning algoritme wordt getraind op een groot deel van de data, de trainingdata, en gevalideerd op een kleiner deel, de testdata. Voor de beste prestaties moeten eerst de hyperparameters, de 'instellingen' van het algoritme, worden geoptimaliseerd. Deze optimalisatie vindt plaats in een zogenaamde *inner loop*. Binnen de traindata worden een k-aantal splitstingen gemaakt (*folds*), namelijk k-1 training folds en één test-fold.

Voor elke hyperparameter zijn een paar van tevoren gespecificeerde mogelijkheden. Bij een *random search* wordt uit alle mogelijkheden een van tevoren bepaald aantal combinaties (iteraties) genomen. Met die combinaties van hyperparameters worden de algoritmes getraind op de traindata en gevalideerd op de testdata in de inner loop. Bij een *gridsearch* worden de algoritmes met álle combinaties uit van tevoren gespecificeerde hyperparameters getraind om de beste combinatie te achterhalen. De gridsearch kost meer tijd en rekenkundige kracht, maar verkent alle mogelijke combinaties en is dus nauwkeuriger dan een random search. Dit proces wordt vervolgens herhaald waarbij een andere fold gebruikt wordt voor validatie en de rest weer voor training. Elke fold is dus één keer een test-fold en k-1 keer een train-fold, deze procedure heet *k-fold cross validation*. Uiteindelijk wordt een geoptimaliseerde combinatie aan waarden van hyperparameters verkregen die na de kruisvalidatie de beste prestaties geven.

Als kruisvalidatie niet wordt toegepast neemt de kans op overfitting enorm toe, omdat een bepaalde combinatie hyperparameters heel goed is afgesteld op de volledige traindata maar vervolgens niet meer goed werkt op de validatiedata. Voor nog minder overfitting en geoptimaliseerde validatie van de dataset kan de k-fold cross validation ook nog een aantal keer herhaald worden. Met r het aantal herhalingen van de k-fold cross validation heet dat *r times repeated k-fold cross-validation*. Een nadeel van (r-times repeated) k-fold cross validation is de grote benodigde rekenkundige kracht [40]. Het gebruik van k=10 wordt vaak gebruikt en ook aangeraden [41]. Voor enorme datasets van 5.000 tot 100.000 instances, wordt aangeraden om 5-6 folds te gebruiken, zonder herhalingen (r=1) [40].

Tegelijk met kruisvalidatie wordt stratificatie toegepast. Het aantal patiënten met een label, dus stabiel of instabiel, blijft dan gelijk verdeeld in alle folds wat voor minder *imbalance*

bias zorgt. Wanneer hier geen rekening mee wordt gehouden een inaccurate evaluatie van het model ontstaan [41]. Stratificatie is vooral belangrijk als gewerkt wordt met kleine datasets en/of wanneer een van de labels in grote minderheid voorkomt maar zou toch ook nog positief kunnen bijdragen aan de performance voor grote datasets [42].

3.4 Hypothesen

Aan de hand van het hiervoor uitgewerkte theoretische kader kunnen de onderstaande hypothesen op de onderzoeksvraag worden opgesteld. Perfusiefalen ligt ten grondslag aan HI, daarom wordt vermoed dat de sterke indicatoren hiervan ook belangrijk zullen zijn in de voorspellende ML-modellen. Dit bevat vooral de klinische parameters zoals bloeddrukken, hartslag, zuurstofsaturatie en combinaties hiervan zoals verschillende typen shock indices. Ook labwaarden die (perfusiefalen-geïnduceerde) ischemie aantonen, zoals lactaat, pH, pO2 en pCO2, worden verwacht belangrijke parameters te zijn in het voorspellen van HI.

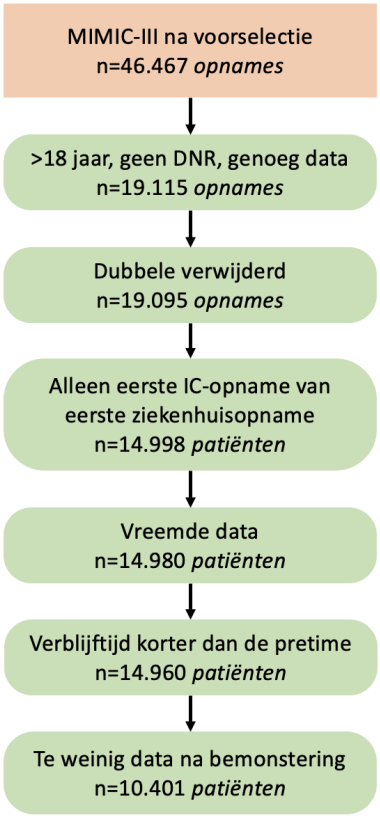
In het algemeen staan beslisboom modellen bekend sterker te zijn in classificatie ten opzichte van regressie modellen, vanwege de gemakkelijke accurate omgang met niet-lineaire relaties, uitschieters en complex samenhangende parameters. Binnen de beslisbomen kunnen bagging-modellen overfitting goed voorkomen en boosting-modellen bias en variantie sterk verlagen. AdaBoost, zoals gebruikt in soortgelijke onderzoeken, levert goede resultaten en zal vermoedelijk in dit onderzoek ook hoge prestaties leveren [21, 39].

4 Methoden

Om het management van hemodynamische instabiliteit bij IC-patiënten meer proactief te maken is er een machine learning model ontwikkeld, die met een enkele predictieve waarde een uur in de toekomst hemodynamische instabiliteit kan aankondigen. Het model is getraind op MIMIC-III IC-patiëntdata. Deze data is voorbewerkt, waarbij klinische data zijn geselecteerd en gefilterd, hemodynamische stabiele en instabiele patiënten zijn gelabeld en aanvullende (klinische) parameters zijn toegevoegd. Hierna is de correlatie bepaald tussen de mogelijke voorspellende parameters en het instabiliteitslabel. Tot slot is het ML-model getraind op basis van de voorbewerkte en gelabelde patiëntdata.

4.1 Databron

Het HDI algoritme werd getraind en getest met de retrospectieve 'Medical Information Mart for Intensive Care'-III (MIMIC-III) database. Voor de totstandkoming en totale inhoud van de database wordt verwezen naar [23]. De data is geanonimiseerd in overeenstemming met HIPAA-regelgeving door persoonsgegevens zoals naam en adres te verwijderen en de tijdgerelateerde gegevens willekeurig te verschuiven naar de toekomst. Wel bleven de verblijftijd op de IC, tijdstippen op de dag en dag van de week relatief hetzelfde. Aanvullend voor identiteitsbescherming werden geboortedata van patiënten ouder dan 89 jaar ook willekeurig verschoven, waardoor de leeftijd van die patiënten boven 300 jaar uitkwam en is dus niet meer accuraat [23].



Figuur 2: Een stroomdiagram voor alle voorbewerkingsstappen waarbij patiënten werden geëxcludeerd. 'n *opnames*' wordt gebruikt om aan te geven dat het om bewerking van bepaalde ziekenhuisopnames gaat – patiënten kunnen namelijk meerdere keren op de IC hebben gelegen. Na het excluderen van de meerdere IC-verblijven representeert n het aantal patiënten dat overblijft na de bewerkingsstappen.

4.2 Voorbewerking

In deze sectie wordt beschreven welke stappen in het selecteren en verwerken van de data zijn gedaan om uiteindelijk een dataset te creëren waarmee het ML-algoritme wordt getraind en getest. Deze database moet vanzelfsprekend kloppend zijn en klinisch significante parameters bevatten. Verdere verwerking voor het bouwen van het model bestond uit het definiëren van HI en daarmee het maken van een label voor de data. Ook het bemonsteren van de data in de tijd rondom een hemodynamisch (in)stabiel event wordt hier beschreven. Alle voorbewerkingsstappen die resulteerden in het verwijderen van patiëntdata worden samengevat in een stroomdiagram, zie figuur 2.

4.2.1 Voorselectie

Eerst werden de methoden van Rahman et al. [21] gevolgd, dus alleen patiënten >18 jaar en zonder een wilsverklaring tot niet-reanimeren (Engels: *do-not-resuscitate order*, kortweg DNR) werden geïncludeerd. De keuze voor de exclusie van patiënten jonger dan 18 jaar – hier dus de pasgeborenen en kinderen tussen 16 en 18 jaar die wel in de MIMIC-III data stonden – is gemaakt om het model zo goed mogelijk af te stemmen op het gebruik in het CZE, waar op de intensive care alleen maar volwassenen liggen. Bovendien verschilt de hemodynamica van kinderen simpelweg van die van volwassenen, dus zouden de gebruikte klinische parameters moeten worden aangepast [43].

Ook patiënten met een DNR werden geëxcludeerd. De patiënten die hier gebruik van maken zijn doorgaans van hogere leeftijd, hebben een grotere comorbiditeit en zijn ernstiger ziek ten opzichte van patiënten zonder DNR [44, 45]. De aanwezigheid van een DNR kan het ook handelen van medisch personeel beïnvloeden. Patiënten met een DNR zouden misschien minder snel naar een IC worden overgeplaatst [46] of krijgen wellicht minder interventies als het toedienen van bloed of vasoactieve middelen [46, 47]. De eerder genoemde patiëntkarakteristieken en de mogelijke invloed van het DNR op de interventies zorgen voor bias in het model, wat wordt voorkomen door exclusie.

Verder worden alleen patiënten meegenomen waarvan genoeg en dus betrouwbare data is verzameld. Dat is, tevens in lijn met [21], wanneer er per patiënt per dag beschikbaar is: $\geq$7 registraties van medicatie via infuus, $\geq 0,75$ ventilatie en luchtwegregistraties in het behandelplan en $\geq$10 registraties in de respiratoire registratietabel. Uiteindelijk is hier een dataset met 19.115 IC-opnames uitgekomen waarmee de rest van de voorbewerking werd gedaan. Een beschrijving van de patiëntkarakteristieken wordt gegeven in sectie 1. Tot slot worden alleen klinische parameters uit de MIMIC-III dataset meegenomen die in de afgelopen twee jaar ook op de IC in het Catharina Ziekenhuis zijn gemeten. Daarmee zijn alle parameters waarmee het model wordt getraind daadwerkelijk relevant voor de toepassing binnen het CZE. Een overzicht van alle klinische variabelen in de dataset wordt gegeven in appendix A.

4.2.2 Verdere selectie

De dataset van 19.115 IC-opnames onderging meer voorbewerking stappen. Allereerst worden de dubbele IC-opnames die per ongeluk nog in de dataset zitten verwijderd. Vervolgens wordt alleen gekeken naar de eerste IC-opname tijdens de eerste ziekenhuisopname van een patiënt. Meerdere verblijven in het ziekenhuis en op de IC hebben invloed op de gezondheidstoestand van de patiënt en vormen een risico op selectiebias. Vervolgens worden patiënten niet-kloppende data ook verwijderd uit de dataset. Dat zijn patiënten die, volgens de data, eerder uit het ziekenhuis zijn weggegaan of eerder zijn overleden dan dat ze op de intensive care terecht kwamen. Doordat er wel gemeten data van deze patiënten beschikbaar zijn (ook na overlijden relatief aan IC-opname), zou men kunnen kiezen om wel data mee te kunnen nemen ter voorkoming van conformatie bias en er zo een grotere hoeveelheid data beschikbaar is. Echter, de data van deze patiënten zijn onbetrouwbaar, omdat niet met zekerheid gezegd kan worden of de data gemeten zijn tijdens IC-opname of daarvoor. Dit zijn weinig patiënten en leidt dus niet tot groot dataverlies.

4.2.3 Pre-time

IC-opnames hebben veel verschillende redenen, waarbij patiënten vaak ernstig ziek zijn en per definitie intensieve zorg nodig hebben. Deze patiënten worden dus reeds instabiel, al dan niet hemodynamisch instabiel, opgenomen. Dit zorgt voor problemen bij het voorspellen van hemodynamische verslechtering op de IC. Met de HDI wordt namelijk getracht om hemodynamische verslechtering (ofwel instabiliteit) te voorspellen en niet voor de reeds aanwezige hemodynamische instabiliteit te alarmeren. Dit kan zorgen voor overbodige alarmeringen, wat juist zorgt voor extra informatiestress. Clinici zullen vooral aan het begin van de IC-opname trachten de hemodynamisch verslechterende of instabiele patiënt, stabiel te krijgen met therapie en daarna de rest van de opname stabiel te houden. Dit laatste is waar de HDI tot zijn recht komt, waarbij na het stabiel krijgen van de patiënt gealarmeerd kan worden voor een aankomend (nog onbekend) instabiel moment. Daarnaast zijn artsen en verpleegkundigen in het begin van de IC-opname wellicht nog bezig met het aansluiten van medische apparatuur, zoals sensoren en infusen. Deze omstandigheden maken de eerste uren aan data van de patiënt onbetrouwbaar, wat kan zorgen voor selectiebias. Hieruit volgend is besloten dat de eerste zes uur aan data van elke patiënt niet wordt meegenomen, wat is gebaseerd op het onderzoek van Rahman et al. [21]. Ook na een korte observatie van de data werd vernomen dat veel patiënten al in de eerste zes uur instabiel werden, na therapie weer stabiel werden binnen deze zes uur (volgens de criteria die worden genoemd in sectie 4.2.4) en pas latere tijd weer een instabiele periode doormaakten. Na deze zogenaamde *pre-time* is de data dus betrouwbaarder om hemodynamische verslechtering te kunnen voorspellen. Als gevolg van het hanteren van de pre-time, worden patiënten die een IC-opnameduur hebben die korter is dan de pre-time geëxcludeerd, zie ook het stroomdiagram (figuur 2).

4.2.4 (In)stabiliteit labelen

Zoals eerder beschreven is hemodynamische instabiliteit een slecht gedefinieerd begrip. Om een (supervised) ML-model te trainen op het voorspellen van hemodynamische instabiliteit, moet er echter wel een eenduidige klinische definitie beschikbaar zijn. Er wordt gekozen voor de definitie dat een patiënt *hemodynamisch instabiel* is als deze één of meer HI-gerelateerde interventies kreeg en *hemodynamisch stabiel* is als de patiënt deze interventies niet kreeg. De hemodynamische interventies bestaan uit medicatietoedieningen, vochttoedieningen en/of bloedtransfusies.

De toedieningscriteria om hemodynamische instabiliteit te labelen, zijn gebaseerd op de criteria beschreven door Rahman et al. [21], waarin tot een consensus wordt gekomen op basis van een panel van medische experts. De grondslag van deze bedachte gouden standaard is te lezen in een bijlage door Rahman et al. (Additional file 1). Hierin zijn de (tijdgebonden) volume criteria klinisch onderbouwd op basis van de etiologie van hemodynamische instabiliteit, zoals uiteengezet in sectie 3.1.

Allereerst wordt iedere toegediende hoeveelheid van verscheidene inotrope (contractiliteit vergrotende) en vasopressor (vaatvernauwende) medicatie gezien als een (significante) hemodynamische interventie. De medicatietoedieningen die als hemodynamische interventie worden gelabeld zijn:

- Fenylefrine (vasopressor)
- Noradrenaline (vasopressor)
- Dopamine (vasopressor met inotrope eigenschappen)
- Epinefrine (vasopressor)
- Vasopressine (vasopressor)
- Milrinon (inotroop)
- Dobutamine (inotroop)

Daarnaast worden vochttoedieningen en bloedtransfusies met verschillende doseringen binnen een bepaald tijdvenster als hemodynamische interventie gelabeld. De voorwaarden van een significante hemodynamische vloeistofinterventie, bestaande uit zowel kristalloide als colloïde (volumeresuscitatie) oplossingen zijn:

- 700 mL in één uur
- 1500 mL in vier uur
- 2400 mL in acht uur
- 3000 mL in twaalf uur

De voorwaarden van bloedtransfusies bestaande uit *packed red blood cells* (PRBC's) zijn:

- 500 mL in twee uur
- 800 mL in 24 uur

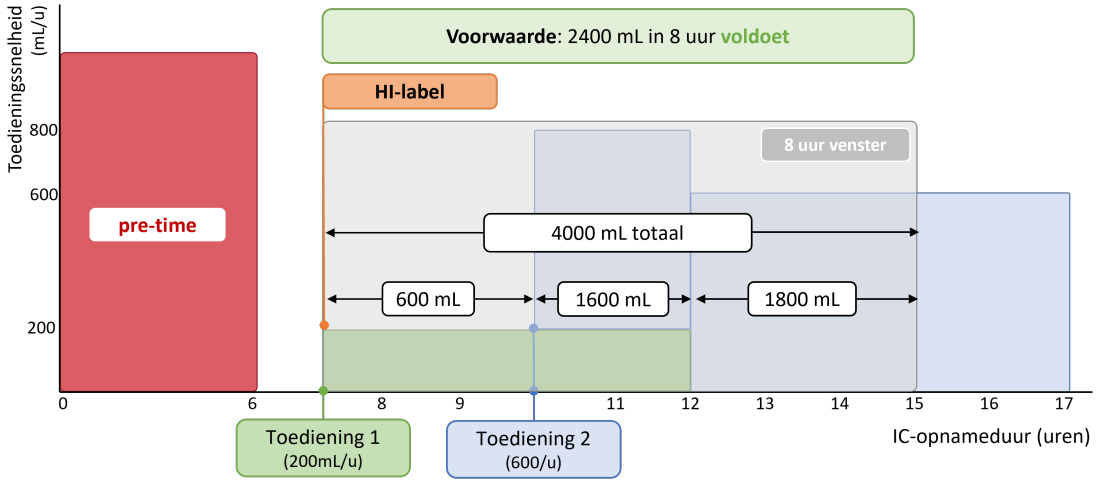
Anders dan door Rahman et al. wordt er niet gekeken naar of een bloedtransfusie wel of niet wordt gevolgd door vloeistoftherapie. Er wordt geacht dat een dergelijke bloedtransfusie hoe dan ook significant is, ongeacht eventuele vochttoediening in de periode daarna. Bovendien is dit criterium makkelijker te interpreteren, omdat het beschreven aanvullende criterium: 'niet gevolgd door vochttoediening binnen 24 uur' ingewikkeld lijkt samen te hangen met het aanvullende criterium daarvoor: 'wel gevolgd door vochttoediening binnen 12 uur'. Tevens wordt het onderscheid tussen beide additionele criteria niet toereikend klinisch onderbouwd door Rahman et al.

Patiënten worden als hemodynamisch instabiel gelabeld op het moment dat zij voor het eerst een significante hemodynamische interventie kregen binnen een bepaald tijdvenster, bekeken vanaf de pre-time. Dit kan een moment zijn van medicatietoediening of op basis van vloeistoftherapieën of bloedtransfusies die voldoen aan de criteria. Deze significante vloeistoftherapieën en bloedtransfusies worden gelabeld over de tijd op basis van een op een tijdas bemonsterde infuussnelheid, met een bemonsteringstijd van één minuut. Door een dergelijke aanpak kan op ieder moment over de bemonstering, precies het totale toegediende vloeistofvolume berekend worden, ook bij meerdere gelijktijdige toedieningen. Vervolgens worden er tijdvensters geschoven over de bemonsterde tijd en wordt er berekend, aan de hand van de sommatie van toegediende volumes, of er binnen het venster voldaan is aan een van de criteria (zie figuur 3). Deze methode komt voort uit diverse afwegingen die worden uiteengezet in appendix B.

Het instabiliteitslabel zal vervolgens aan het begin van het venster worden geplaatst waarin een significante interventie plaatsvond. Dit is dus het moment waarop de arts de hemodynamische behandeling startte voor een instabiele patiënt, getoetst op basis van de genoemde instabiliteitscriteria. Echter heeft het ML-model trainen op basis van deze tijd van instabiliteit weinig nut. Het model zal een arts dan namelijk waarschuwen op het moment dat de patiënt instabiel is en dus wanneer een arts uit de dataset het nodig achtte om een hemodynamische behandeling te starten. Een dergelijk model zou handig kunnen zijn voor een arts die twijfelt of een HI-behandeling noodzakelijk is, maar er is dan geen tijd meer om HI te voorkomen. Door

het HI-label nog een periode van een uur terug in de tijd te schuiven, kan het model vroegtijdig een hemodynamisch instabiel moment voorspellen. Deze tijd wordt de *predictietijd* genoemd. Door een periode van een uur te nemen (gebaseerd op Rahman et al.), heeft de arts tijd tot anticiperen op de aankomende instabiliteit, waardoor er genoeg tijd is om eventueel extra onderzoek te doen ter conformatie. Maar ook om een hemodynamische therapie te starten en deze in te laten werken, om zo HI te voorkomen en het risico op overlijden te verlagen [8]. Daarnaast is er een vertraging aanwezig tussen het feitelijke begin van een instabiel moment en het moment van het geplaatste instabiele label. Dit komt doordat het moment van plaatsen van het instabiliteitslabel gebaseerd is op de interventie van een arts. Een arts zal namelijk vaak later reageren dan het eigenlijke begin van het hemodynamische instabiele moment door de tijd die nodig is om HI te herkennen, wat lastiger wordt door de benoemde informatiestress.

Patiënten die niet hemodynamisch instabiel werden bevonden tijdens hun IC-verblijf, worden als hemodynamisch stabiel gelabeld. Bij deze patiënten wordt een willekeurig moment tijdens hun IC-verblijf, na de pre-time, geselecteerd als stabiel moment. Na het bepalen van een (in)stabiel moment per patiënt, zullen de klinische parameters retrospectief worden geëxtraheerd vanaf de predictietijd. Deze data zal eerst nog gefilterd worden met een plausibiliteitsfilter.



Figuur 3: Hemodynamische instabilteit labelen op basis van tijdgebonden volumecriteria.

4.2.5 Plausibiliteitsfilter

Bij het trainen en testen van een machine learning algoritme is het van belang dat de data eerst wordt opgeschoond. Enorme uitschieters en meetfout, hierna samengenomen en 'uitschieters' genoemd, kunnen zorgen voor onbetrouwbare voorspellingen en slechtere performance van het algoritme [48]. Van een meetfout is tevens niet herleidbaar of de waarde, als het correct gemeten was, hoog of laag zou zijn. Daarom kunnen uitschieters niet worden afgekapt op bepaalde waarden, maar moeten ze dus worden verwijderd uit de dataset, door middel van een zogenaamd plausibiliteitsfilter.

De meest betrouwbare manier om uitschieters te herkennen en te filteren is het definiëren van een fysiologisch of technologisch mogelijk bereik. Alle waarden die onder of boven het gespecificeerde bereik vallen kunnen worden verwijderd ('laten vallen', *dropping*). Echter is het niet altijd mogelijk, of erg moeilijk, om voor elke parameter een dergelijk mogelijk bereik te definiëren. Bovendien is het niet makkelijk reproduceerbaar, omdat andere instellingen of laboratoria dezelfde parameters op een andere manier of in andere eenheden meten. Een voorbeeld hiervan is de chemische labwaarde C-reactief eiwit (*C-reactive protein*, CRP) gemeten in mg/dL. Extreem hoge waarden van CRP, >50 mg/dL, wijzen op een hele heftige infectie, maar zijn dus wel (patho)fysiologisch mogelijk [49]. In appendix C.1 worden meerdere methoden getest om om te gaan met uitschieters van deze parameter. Uiteindelijk blijkt het toepassen van een robuuste Z-score een goed alternatief als er geen duidelijk fysiologisch bereik gedefinieerd kan worden. Dus, waar mogelijk wordt een fysiologisch bereik gedefinieerd, an-

ders wordt een robuuste Z-score toegepast.

Het laten vallen van bepaalde metingen kost data waar op zichzelf wel klinische waarde in zit. Namelijk, het feit dat er wel gemeten is, komt waarschijnlijk omdat een clinicus dat nodig achtte. Om deze indicatie wel mee te nemen werd van bepaalde gefilterde parameters nog een parameter aangemaakt met een waar/onwaarwaarde om aan te geven dat er wel naar gemeten is. Dit werd niet gedaan voor alle parameters die niet continu werden gemeten (zoals invasieve bloeddrukken), maar wel voor parameters die niet zo vaak zijn gemeten zoals labwaarden en non-invasieve bloeddrukken.

De locatie van bloedafname heeft invloed op het resultaat van een bloedgasmeting. In de dataset na voorselectie werd onderscheid gemaakt tussen bloedgasmetingen met arterieel, veneus, gemixt en centraal veneus bloed. In de data is te zien dat waarden tussen gemixt, centraal veneus en veneus bloed verschillen heel weinig [50] en werden samengenomen als 'veneus'. Ook waren er onbekende afnamelocaties, welke als nog zijn gelabeld als arterieel of veneus volgens de methode in appendix C.2. Een overzicht van alle parameters met waar mogelijk hun fysiologisch bereik is ook opgenomen in appendix A. Wanneer de data is gefilterd bevat het geen uitschieters en foutieve metingen meer, dus kan het worden bemonsterd per patiënt afhankelijk van de tijd van (in)stabiliteit.

4.2.6 Data bemonstering

Allereerst worden van alle patiënten de demografische gegevens meegenomen, bestaande uit onder andere geslacht, gewicht, leeftijd en lengte. In de MIMIC-III database werden leeftijden van patiënten van 89 jaar of ouder gemaskeerd en verschoven naar een leeftijd van 300 jaar of ouder. Om het uiteindelijk wel correct weer te geven werden leeftijden ingedeeld in groepen van tien jaar, beginnend bij 18-25 jaar en eindigend met 85-plus.

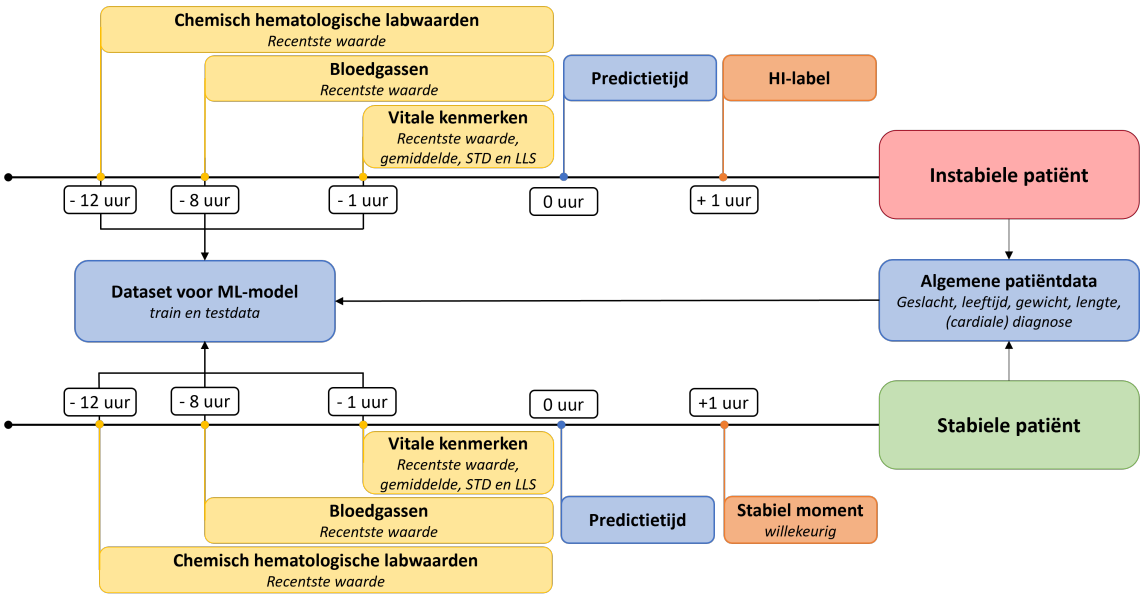
Na het bepalen van een (in)stabiel moment per patiënt, worden de gefilterde klinische parameters retrospectief bemonsterd vanaf de predictietijd, zie figuur 4. Het tijdvenster voor de predictietijd waarin de gegevens worden bemonsterd is anders voor de verschillende datasets en is gebaseerd op de duur waarvoor een arts de gemeten parameter nog vertrouwt. Dit was niet vindbaar in de literatuur, dus deze vensters zijn afgeleid uit de data door te kijken naar de gemiddelde tijdregistraties tussen niet-continu gemeten parameters voor elke dataset. Deze gemiddelden staan beschreven in appendix D (tabel 11, waaronder ook het 75e percentiel van die tijden tussen de tijdregistraties. De uiteindelijke lengte van de tijdvensters waarin bemonsterd werd, is iets ruimer genomen dan dit 75e percentiel. Een ruim tijdvenster waardoor veel data bemonsterd kan worden moet opwegen tegen de duur waarin een arts de meting nog beschouwt als valide, waarin voor dit 75e percentiel is gekozen als een goed middelpunt in deze afweging. Hieronder worden hieruit volgende tijdvenster beschreven per dataset, gebaseerd op tabel 11 in appendix D.

- De vitale kenmerken worden tot één uur voor de predictietijd bemonsterd. Het 75e percentiel van de gemiddelde intervallen tussen alle metingen was 51 minuten (afgerond op minuten).
- De hematologische chemische labwaarden worden tot twaalf uur voor de predictietijd bemonsterd. Het 75e percentiel van de gemiddelde intervallen tussen alle metingen was 11 uur en 52 minuten (afgerond op minuten).
- De bloedgassen worden tot acht uur voor de predictietijd bemonsterd. Het 75e percentiel van de gemiddelde intervallen tussen alle metingen was 6 uur en 56 minuten (afgerond op minuten).

Binnen het tijdvenster wordt voor elke dataset per gemeten variabele de meest recente waarde overgenomen als losse parameter. Hiernaast worden van de vitale kenmerken alle metingen binnen het tijdvenster meegenomen in een statistische (trend)analyse met onder andere een gemiddelde en de variantie, zie sectie 4.2.8. Dit wordt alleen gedaan voor vitale kenmerken omdat hier naar verwachting genoeg data beschikbaar voor is, aangezien deze zeer frequent of zelfs continu gemeten zijn. Bepaalde bloedgassen, namelijk pCO_2, pO_2, Base Excess (base overmaat) en pH, zijn ook continu gemeten en net als de vitale kenmerken doorgaans elke vijf minuten geregistreerd. Hiervan worden géén trendanalyses uitgevoerd, omdat binnen het CZE deze variabelen niet continu worden gemeten. Deze trendanalyses hebben dus niet veel betekenis wanneer dit model zal worden toegepast binnen het CZE en zorgen voor een onnodige uitbreiding van de dataset.

Bij de bemonstering van data wordt tevens rekening gehouden met de pre-time. Data dat binnen de pre-time van zes uur is gemeten wordt

niet meegenomen in de dataset. Mocht het tijdvenster gedeeltelijk over de pre-time vallen, dan worden alleen de metingen meegenomen die later dan de pre-time zijn gemeten (wederom door de onbetrouwbaarheid van deze metingen). Als blijkt dat er uiteindelijk niet genoeg data bemonsterd kan worden, worden de respectievelijke patiënten helemaal niet meegenomen in de dataset. Dit geldt alleen wanneer er voor een patiënt geen enkel vitaal kenmerk en/of labwaarde bemonsterd kon worden. Deze stap in de voorbewerking is de laatste aanvulling in het stroomdiagram, figuur 2.



Figuur 4: Data bemonstering van vitale kenmerken, bloedgassen en chemisch en hematologische labwaarden bij hemodynamisch (in)stabiele patiënten (tijd relatief aan predictietijd)

4.2.7 Extra fysiologische parameters

Parameters die meerdere variabelen combineren kunnen in sommige gevallen een hogere correlatie geven met de onderzochte uitkomst dan individuele variabelen [11]. Daarom worden een enkele extra fysiologische variabelen meegenomen die hieronder worden beschreven.

- **(Systolische) Shock Index**
  De verhouding tussen hartslag en systolische bloeddruk (SBP) wordt de shock index (SI) genoemd. Pre-intubatieve SI heeft aangetoond post-intubatieve hypotensie goed te kunnen voorspellen [51, 52]. Om onderscheid te maken met de hierna geïntroduceerde diastolische shock index, wordt deze parameter dus 'systolische shock index' (SSI) genoemd.

$$\text{SSI} = \frac{\text{HR}}{\text{SBP}}$$

- **Diastolische Shock Index**
  De klassieke SI maakt gebruik van de systolische bloeddruk. Recentelijk onderzoek heeft een diastolische shock index (DSI) bedacht, waarin de systolische bloeddruk wordt vervangen door de diastolische bloeddruk (DBP). Deze DBP heeft namelijk een direct verband met de vasculaire tonus en een omgekeerd verband met de duur van een hartcyclus. Dit maakt het ook een goede hemodynamische index [53].

$$\text{DSI} = \frac{\text{HR}}{\text{DBP}}$$

- **Gemodificeerde Shock Index**
  Naast de klassieke SI bestaat er ook een gemodificeerde shock index (*Modified Shock Index*, MSI). Deze geeft de verhouding tussen hartslag en gemiddelde arteriële bloeddruk (*Mean Arterial Pressure*, MAP) [54]. MSI heeft aangetoond een significante correlatie te hebben met mortaliteit in IC-patiënten [55].

$$\text{MSI} = \frac{\text{HR}}{\text{MAP}}$$

- **Rate Pressure Product**
  Door HR te vermenigvuldigen met SBP verkrijgt men de *Rate Pressure Product* (RPP). RPP geeft een indicatie van de

mate van stress die op de hartspier ontstaat [6].

$$\text{RPP} = \text{HR} \cdot \text{SBP}$$

- **Oxygenatie index**
  Weefsel voorzien van zuurstof is het primaire doel van het circulatoire systeem [56]. De oxigenatie index (OI) is een goede indicator voor acuut hypoxisch respiratoir falen (AHRF) en geeft een inzicht over de mate van extractie van zuurstof door het lichaam in de weefsels [57]. Het kan berekend worden door middel van de gemiddelde luchtwegdruk (*Mean Airway Pressure*, P_{mean}), arteriële zuurstofspanning (PaO_2) en de fractie van ingeademde zuurstof (FiO_2).

$$\text{OI} = \frac{\text{P}_{\text{mean}} \cdot \text{FiO}_2}{\text{PaO}_2}$$

- **Zuurstofsaturatie index**
  Naast OI is de zuurstofsaturatie index (*Oxygen Saturation Index*, OSI) ook een goede indicator voor zuurstofvoorziening. In plaats van de arteriële zuurstofspanning wordt de zuurstofsaturatie (SpO_2) gebruikt.

$$\text{OSI} = \frac{\text{P}_{\text{mean}} \cdot \text{FiO}_2}{\text{SpO}_2}$$

- **ROX index**
  Ook een indicator van hypoxisch respiratoir falen is de ROX index. Deze maakt gebruik van de arteriële zuurstofspanning (PaO_2), de fractie van ingeademde zuurstof (FiO_2) en de ademhalingsfrequentie (*Respiratory Rate*, RR). Het wordt gebruikt als een voorspellende indicator voor het succes van *high flow* zuurstoftherapie [58, 59].

$$\text{ROX} = \frac{\text{SpO}_2/\text{FiO}_2}{\text{RR}}$$

- **Body Mass Index**
  Ondanks de implementatie van een te hoog Body Mass Index (BMI) in de Elixhauser Comorbidity Index, die direct afleidbaar is uit de database, is het interessant om BMI apart te bekijken. Door BMI alleen in te delen op wél of niet obesitas verliest men namelijk informatie. Een hogere Body Mass Index (BMI) wordt vaak geassocieerd met verscheidene hart pathologieën. Een hogere BMI van een patiënt zou daarom een hoger standaard risico niveau kunnen betekenen voor HI. [60].

$$\text{BMI} = \frac{\text{Gewicht}}{\text{Lengte}^2}$$

- **CK-MB/CK ratio**
  Verhoogde creatine kinase (CK) in het bloed indiceert spierschade en heeft een hoge sentiviteit voor acuut myocardinfarct. Het iso-enzym CK-MB is de zogenaamde *muscle brain* creatine kinase en is een specifieke marker voor schade aan de hartspier. De ratio tussen deze twee biedt inzicht in de locatie van eventuele spierschade en kan dus onderscheid maken tussen cardiale en niet-cardiale spierschade [61].

- **Cardiale diagnose**
  Tot slot worden patiënten die een cardiale diagnose kregen tijdens hun ziekenhuisopname, gelabeld als cardiale patiënt. Dit zijn onder andere patiënten met diagnose labels zoals myocardinfarct, endocarditis, coronair, atriaal, mitraal, (congestief) hartfalen, pericardeffusie en andere hartproblematiek.

4.2.8 Statistische (trend)variabelen

Voor de frequent gemeten vitale kenmerken worden, naast het selecteren van de meest recente waarde per patiënt, ook andere variabelen berekend op basis van de gemeten waarden die binnen het vitale tijdvenster van één uur vallen (zie sectie 4.2.6). Ten eerste worden het gemiddelde en de standaarddeviatie berekend voor de gemeten vitale kenmerken in het tijdvenster. De standaarddeviatie geeft de mate van spreiding van de vitale kenmerken aan binnen het tijdvenster en dus vóór het moment van instabiliteit. Het zegt wat over de veranderlijkheid in waarde van de gemeten vitale kenmerken. Ten tweede wordt, om de veranderlijkheid van de gemeten vitale kenmerken over de tijd te duiden, een *ordinary least squares* regressie methode (ofwel kleinste-kwadratenmethode) toegepast. Met deze enkelvoudige lineaire regressie methode kan een regressielijn worden berekend tussen de gemeten waarden, die deze waarden zo goed mogelijk voorspelt over de tijd door de kwadratische afwijking tussen de meetpunten en de regressielijn te minimaliseren. De helling van deze regressielijn geeft de richting en snelheid van de verandering aan, ofwel de trend, van de gemeten vitale kenmerken voorafgaand aan een instabiel moment.

4.3 Verkennende gegevensanalyse

Na afronding van de voorbewerking is de dataset bijna gereed voor training van het ML-model. Voorafgaand hieraan is het nodig om de correlaties en multicollineariteit te onderzoeken om zo overwegingen te kunnen maken tussen mogelijke parameters. Dit wordt correlatie gebaseerde parameterselectie genoemd. Zo worden namelijk alleen parameters meegenomen die genoeg toegevoegde waarde hebben zonder dat de informatie overlapt met een andere. In andere woorden, de ideale parameter heeft een hoge correlatie met het label maar een lage correlatie met andere parameters. Niet alleen is dit noodzakelijk voor regressie modellen, het voorkomt ruis, verlaagt de benodigde rekenkracht en bevordert de uitlegbaarheid [62].

4.3.1 Ontbrekendheidsgraad parameters

Te veel missende waarden in de dataset kunnen leiden tot een lagere precisie en hogere bias van een predictiemodel [63]. Het is daarom belangrijk goed om te gaan met missende waarden, waarvoor veel technieken zijn ontwikkeld [64, 65]. Deze kunnen worden toegepast voor het trainen van het ML algoritme, zie daarvoor sectie 4.4.3. Eerst worden variabelen die hoe dan ook te veel missende waarden bevatten verwijderd uit de dataset. De literatuur verschaft geen duidelijke afkapwaarde voor een verhouding tussen missende en beschikbare data waarvoor de variabele gedropt zou moeten worden, dus is dit bepaald aan de hand van een korte data-analyse. Hierbij is het doel om een balans te vinden tussen het aantal geëxcludeerde parameters en het maximale percentage missende data per parameter, ofwel de NaN-ratio. Voor verschillende afkapwaarden van NaN-ratio's en bij dezelfde afkapwaarden voor absolute correlatie en multicollineariteit is het aantal parameters bepaald waarna een weloverwogen afkapwaarde wordt gekozen.

4.3.2 Correlatie van continue variabelen

De punt-biseriële correlatiecoëfficiënt, welke gebaseerd is op Pearson's r, kan gebruikt worden om de correlatie tussen een continue variabele en een dichotome variabele meetbaar te maken [66]. De continue variabelen zijn alle parameters, dus de vitale kenmerken en labwaarden, die in de pre-processing zijn bemonsterd. Ook van demografische gegevens wordt de correlatie bepaald. De dichotome uitkomst is hier het label, stabiel/instabiel. In tegenstelling tot Spearman's monotone associatie-maat geldt bij punt-biseriële correlatie de voorwaarde dat de data normaal is verdeeld. Dit komt doordat het gebruik maakt van lineaire correlatie, hierbij is de meerwaarde van deze methode de betrouwbaarheid [67]. In dit onderzoek werd, vanwege de grootte van de dataset (aantal observaties per parameter $\gg 10$), aangenomen dat de data normaal is verdeeld [68].

Vervolgens kan punt-biseriële correlatie toegepast worden. Dit wordt gedaan door de significantie te berekenen van de correlatie tussen parameter en label, welke wordt weergegeven in de vorm van een p-waarde. Er is een afkapwaarde van $\alpha = 0{,}05$ gehanteerd, zoals gebruikelijk in medische predictiemodellen [69]. Vervolgens zullen alle parameters met p-waarde $> \alpha$, ergo, zonder significante correlatie, geëxcludeerd worden.

4.3.3 Correlatie van dichotome variabelen

De dataset bestaat ook uit dichotome variabelen. Dat zijn de parameters gender en meetgraad (zie sectie 4.2.5). Bij de analyse van de parameters gender en meetgraad maakt men gebruik van een χ^2-toets. De toets geeft de significantie weer in de vorm van een p-waarde. Hier wordt wederom dezelfde afkapwaarde van significantie gebruikt: $\alpha = 0{,}05$.

4.3.4 Multicollineariteit tussen parameters

Naast de hiervoor beschreven methode voor de bepaling van associatie tussen parameter en uitkomst van stabiliteit is het van toegevoegde waarde om de parameters met behulp van Pearson's correlatiecoëfficiënt onderling te vergelijken. Deze methodiek is alleen noodzakelijk op het logistische regressie ML-model, maar zal voor uniformiteit van de dataset ook toegepast worden op de beslisboom modellen. In tegenstelling tot beslisboom modellen zijn regressie modellen namelijk erg gevoelig voor multicollineariteit van onafhankelijke parameters. Multicollineariteit betekent dat parameters sterk van elkaar afhankelijk zijn. De multicollineaire parameters verkrijgen in dit geval onjuiste gewichten waardoor instabiele schattingen en inaccurate varianties ontstaan. Wanneer men dit observeert is het belangrijk na te gaan waar dit afkomstig van is en vervolgens te overwegen hiervan slechts één parameter mee te nemen in het ML-model. Naast de hiervoor genoemde noodzakelijkheid is het een groot bijkomend voordeel

dat parameterselectie met eliminatie van multicollineariteit zorgt voor het voorkomen van ruis, verlagen van de benodigde rekenkracht en bevordering van de uitlegbaarheid. Er worden namelijk alleen parameters meegenomen die unieke informatie bevatten.

Allereerst wordt gedefinieerd wat wordt gezien als een sterkte correlatie. Op het gebied van medische parameters wordt een Pearson's correlatiecoefficient boven 0,8 gezien als sterke correlatie [70]. Onder deze parameters die multicollineariteit vertonen is de parameter behouden die het meest significant correleert met het label. Middels deze methodiek is de verdere parameterselectie toegepast om logistische regressie toe te kunnen passen en betrouwbare modellen te kunnen genereren [71, 72].

4.4 Machine Learning

Na de voorbewerking volgens sectie 4.2 is een database geconstrueerd. Met deze data in het geheel en met verder bewerkte delen van de dataset worden de modellen getraind. Deze trainingen, de evaluaties daarvan en aanvullende ondernomen bewerkingsstappen worden hieronder uiteengezet.

4.4.1 Trainingen

De machine learning algoritmen werden getraind op verschillende delen (*subsets*) van de data na de voorbewerking, waarna de prestaties met elkaar werden vergeleken. Allereerst werden de algoritmen getraind op de volledige dataset voor én na de verkennende gegevensanalyse. De bewerkingsstappen van sectie 4.3 kunnen immers de generaliseerbaardheid en uitlegbaarheid van het model verbeteren, maar kunnen de prestaties ook beïnvloeden. Vervolgens werd een subsetanalyse uitgevoerd op de subgroep hartpatiënten. Goede prestaties voor die specifieke patiëntgroep zou bewijs zijn dat de HDI ook goed zou kunnen werken op de intensive care van het CZE, waar veel (post-operatieve) hartpatiënten liggen. Deze cardiale subset komt voort uit de door de verkennende gegevensanalyse verkregen dataset. Er werd gebruik gemaakt van de *scikit-learn package* [73] in Python (versie 3.11).

4.4.2 Classificatiemodellen

Voor het kiezen van een gepast classificatiemodel zijn veel opties en er bestaan geen duidelijke richtlijnen voor [74]. In klinische predictiemodellen blijken logistische regressie en beslisboom ML-modellen erg effectief en worden vaak toegepast [21, 39, 74, 75]. Beslisboom modellen kunnen, in tegenstelling tot logistische regressie modellen, complexere verbanden in de data ontdekken en geven een meer intuïtieve weergave van de data en klinische besluitvorming wat aantrekkelijk is voor clinici [76].

Er zijn uiteindelijk vier classificatie ML-algoritmen geselecteerd die binnen regressie- en belisboommodellen het meest gangbaar zijn. Dat is logistische regressie (*Logistic Regression*) en voor de belisboommodellen *Adaptive Boosting*, *Gradient Boosting* en *Random Forest*. In appendix F wordt voor elk model hun (unieke) hyperparameters beschreven.

4.4.3 Missende waarden

Zoals beschreven in sectie 4.3.1 is het belangrijk om goed om te gaan met missende waarden, waarvoor eerst alle parameters werden verwijderd die meer dan 60% missende waarden hadden. Vervolgens worden binnen de overgebleven parameters alle missende waarden simpel geïmputeerd met het gemiddelde van die parameter.

4.4.4 Normalisatie

Voor logistische regressie moet eerst normalisatie of standaardisatie toegepast worden. Er wordt een schaling toegepast op basis van minimale en maximale waarden [38]:

$$X_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Omdat het gebaseerd is op extreme waarden is het wel erg gevoelig voor uitschieters. Aangezien deze al zijn gefilterd door een plausibiliteitsfilter (sectie 4.2.5) vormt dat geen probleem. Beslisboom modellen zijn ongevoelig voor de schaling van de parameters, omdat deze elke parameter apart behandelen [38]. Desalniettemin wordt de schaling ook voor deze modellen toegepast om uiteindelijk de prestaties goed te kunnen vergelijken met die van logistische regressie.

4.4.5 Modelprestatie

Om de prestatie van de vier ML-modellen te beoordelen, werden de modellen getest voor iedere subset, met behulp van de volgende prestatiematen [77]:

- Nauwkeurigheid en precisie
- Sensitiviteit en specificiteit

- ROC-curve (*Receiver Operating Characteristic*) en hieruit de *Area Under the Curve* (AUC)
- F1-score

De nauwkeurigheid van een ML-model omschrijft de verhouding correct HI en stabiel geclassificeerde patiënten ten opzichte van het totale aantal patiënten. De precisie geeft daarentegen de verhouding tussen het aantal correct geclassificeerde HI-patiënten en het totale aantal geclassificeerde HI-patiënten (correct en incorrect). Een hoge precisie geeft dus weinig incorrecte HI geclassificeerde patiënten. Deze twee prestatiematen zijn echter minder goed voor een dataset waar de prevalentie van het label (HI) niet gelijk is.

De sensitiviteit, ook wel *recall* genoemd, en specificiteit hebben geen last van een scheve verdeling tussen klassen. De sensitiviteit beschrijft het aantal correct voorspelde HI-patiënten ten opzichte van alle HI-patiënten, terwijl de specificiteit het aantal correct voorspelde stabiele patiënten ten opzichte van alle stabiele patiënten geeft. Een goed voorspellend ML-model heeft een goede sensitiviteit en specificiteit, dit wordt samengenomen in de ROC-curve.

De ROC-curve zet de sensitiviteit uit tegen de aspecificiteit (1-specificiteit, ofwel de incorrecte HI-geclassificeerden) en heeft dus ook geen last van een scheve verhouding tussen de klassen. Het laat de relatie tussen de sensitiviteit en specificiteit en waar nog verbetering mogelijk is duidelijk zien. De oppervlakte onder deze curve (AUC) omvat deze curve in één gemakkelijk interpreteerbaar getal (*Area Under the Receiver Operating Characteristic-curve*, AUROC) en maakt het een goede validator voor de kruisvalidatie. Hierom is gekozen de AUROC als evaluatiemaat te gebruiken bij de optimalisatie van de ML-modellen.

nDe F-score wordt berekend als een gewogen harmonisch gemiddelde, de eerdergenoemde precisie (*p)* en sensitiviteit (recall, *r*) volgens onderstaande formule. Het geeft een waarde tussen 0 en 1, waarbij 1 een perfecte sensitiviteit en precisie zou aangeven. De β geeft er een bepaalde weging aan, waarbij de sensitiviteit β-keer zoveel gewogen is als de precisie. Hier wordt gekozen voor β=1, dus een F1-score [78].

$$\mathrm{F}_\beta = (1+\beta)\frac{\mathrm{pr}}{\mathrm{r}+\beta^2\mathrm{p}},\ \mathrm{F}_1 = \frac{2\mathrm{pr}}{\mathrm{r}+\mathrm{p}}$$

4.4.6 Hyperparameter optimalisatie en kruisvalidatie

De dataset werd gestratificeerd gesplitst in een trainingsset van 80% van de data en een validatieset van 20% van de data. Deze splitsing werd gedaan met een *random state*, wat ervoor zorgt dat dezelfde splitsing van de data reproduceerbaar is. Binnen de trainingsdata werden eerst de hyperparameters geoptimaliseerd onder toepassing van gestratificeerde *6-fold cross-validation* zonder herhalingen (r=1). Vanwege gelimiteerde tijd en rekenkundige kracht zal eerst een *random search* worden uitgevoerd met n=50 iteraties. Hiervoor zijn aannemelijke waarden van hyperparameters van tevoren gespecificeerd, zie appendix F.1. De beste paar waarden gebaseerd op de modelprestaties, of een klein bereik daaromheen, worden vervolgens gebruikt in een gridsearch.

4.4.7 Belangrijkheid van parameters

Om inzicht te krijgen in welke parameters de grootste voorspellende waarde hebben voor HI werd na voltooiing van elke training voor elk model de (relatieve) belangrijkheid van iedere parameter bepaald. Voor de logistische regressie classificatie ML-modellen werd relatieve belangrijkheid bepaald aan de hand van de coëfficiënten (relatieve gewichten) van de parameters. Grotere (absolute) coëfficiënten representeren relatief belangrijkere parameters in het voorspellen van HI.

Beslisboom modellen bieden andere manieren om inzicht in de belangrijkheid van variabelen, waarvan er twee in dit onderzoek werden gebruikt. Een van die methoden is gebaseerd op de Gini-index, ofwel de *Gini-impurity*. Voor elke splitsing in een beslisboom op een bepaalde parameter, wordt het verschil in onzuiverheid tussen de voorgaande knoop en de onzuiverheid van de twee 'dochterknopen' bepaald. Deze verschillen worden voor elke parameter bepaald, waarbij een relatief groter verschil een relatief belangrijke parameter indiceert. Bij bagging modellen, hier dus Random Forest, wordt deze score voor elke parameter voor elke boom bepaald en vervolgens gemiddeld. Een tweede manier is gebaseerd op permutatie van de parameters. Bij parameter wordt de associatie met het label (instabiliteit) verstoord door de op een willekeurige manier de volgorde van de waarden te schudden. Vervolgens wordt per parameter de daling in prestatie (nauwkeurigheid) gemeten, wat als maatstaf dient voor de belangrijkheid van die parameter.

5 Resultaten

5.1 Beschrijving data

5.1.1 Patiëntkarakteristieken

In de tabel hieronder wordt de onderzoekspopulatie na alle voorbewerkingsstappen beschreven. Data van deze patiënten werden dus meegenomen in het trainen en testen van de ML-modellen.

Tabel 1: Patiëntkarakteristieken

Eigenschap	**n=10.401**
Geslacht, man (%)	60,0
Label instabiel (%)	41,9
Hartpatiënten (%)	34,3
Leeftijd (jaren)	55-65 [45-55, 65-75]
Gewicht (kg)	82,7 ±22,6
Lengte (cm)	170,2 ±10,6
BMI (kg/m^2)	28,7 ±7,1
Elixhauser index	0,72 ±2,6

5.2 Verkennende gegevensanalyse

Het aantal (klinische en demografische) parameters dat na voorbewerking is verkregen is 210 en vormt de primaire subset: 'Vóór verkennende gegevensanalyse (voor VGA)'. Deze parameters zijn verder geselecteerd op basis van ontbrekendheidsgraad, correlatie en multicollineariteit, wat de subset 'Na verkennende gegevensanalyse (na VGA)' en de basis van de hartpatiënten subset vormt.

5.2.1 Ontbrekendheidsgraad parameters

In de onderstaande tabel is weergegeven hoe de afkapwaarde van de ontbrekendheidsgraad, ofwel NaN-ratio, het aantal parameters beïnvloed.

Tabel 2: Analyse NaN-ratio afkapwaarden

Afkapwaarden	**50%**	**60%**	**75%**	**90%**
Zonder exclusie	210	210	210	210
Na NaN-ratio	97	116	129	152
Na correlatie	67	83	90	104

De NaN-ratio (ontbrekendheidsgraad) per parameter is weergegeven in appendix E in de tabel E.1. Voor inclusie van een parameter is een afkapwaarde van NaN-ratio $< 0{,}6$ (maximaal 60% missende data) gekozen. In geval van een NaN-ratio hoger dan 0,6 wordt de parameter gedropt en gaat hij dus niet mee naar de 'Na VGA'-subset. Bij een dergelijke afkapwaarde werden 94 parameters geëxcludeerd als gevolg van te hoge NaN-ratio.

5.2.2 Correlatie

Significantie (p-waarde) van correlatie tussen parameter en het HI-label is weergegeven in de tabel in E.1. In geval van en insignificante correlatie (p-waarde $> 0{,}05$) werd de parameter gedropt. 22 parameters werden geëxcludeerd ten gevolge van een insignificante correlatie met het HI-label.

5.2.3 Multicollineariteit

De parameters die zijn gedropt als gevolg van multicollineariteit zijn weergegeven in appendix E in de tabel E.1. Een volledig overzicht van de multicollineariteit tussen alle continue variabelen is weergegeven in de vorm van een heatmap, zie bijlage E figuur 11. Er werden 23 parameters geëxcludeerd ten gevolge van multicollineariteit.

5.3 Machine learning resultaten

5.3.1 Optimale hyperparameters

De eindresultaten van de *Grid Search (GS) cross-validation* hyperparameter optimalisatie bij de subsets: vóór de verkennende gegevensanalyse (VGA), na de verkennende gegevensanalyse en de cardiale patiënten, zijn respectievelijk te vinden in tabel 3, tabel 4 en tabel 5. De tussenresultaten van de daarvoor uitgevoerde *Random Search (RS) cross-validation* hyperparameter optimalisatie en de daaruit afgeleide inputs voor de GS zijn te vinden in de appendix sectie F.2 (vóór de VGA), appendix sectie F.3 (na de VGA) en in appendix sectie F.4 (hartpatiënten).

5.3.2 Modelprestaties

Ieder geoptimaliseerde model werd getest met een validatieset per subset met behulp van verschillende prestatiematen. De resultaten hiervan zijn te vinden in tabel 6. De modellen laten allemaal een afname van de prestatiematen zien na de bewerkingsstappen door de verkennende gegevensanalyse. Verder laten de modellen allemaal een stijging zien van de AUROC en sensitiviteit bij de hartpatiënten subset, terwijl de specificiteit bij deze groep patiënten juist afneemt. Random Forest laat bij deze groep een toename

zien in AUROC, sensitiviteit en F1, maar een afname van de nauwkeurigheid, precisie en specificiteit.

Op basis van alle prestatiematen, met uitzondering van de specificiteit, presteerde Gradient Boosting het beste voor alle subsets. Hierna volgenden Adaptive Boosting en Random Forest, waarbij Adaptive Boosting over het algemeen beter presteerde voor de cardiale patiënten subset en Random Forest beter presteerde voor de 'na VGA'-subset. Logistische regressie presteerde het slechts van alle modellen, maar maakte wel een grote prestatieverbetering door na het selecteren van de cardiale patiënten. Hierbij steeg de AUROC, nauwkeurigheid, precisie, sensitiviteit en F1 flink, maar daalde de specificiteit. Tot slot laat figuur 5 de ROC-curven zien voor de vier verschillende modellen per subset. Ook hier blijkt Gradient Boosting het beste te presteren van de classificatiemodellen, terwijl logistische regressie wederom het slechts blijkt te presteren voor alle subsets.

5.3.3 Belangrijkheid van parameters

Om de belangrijkheid van de parameters voor de logistische regressie classificatoren te bepalen werden de relatieve gewichten bepaald voor iedere parameter per subset. Daarnaast werd voor de beslisboom classificatoren de *Gini-impurity* en de permutatie belangrijkheid bepaald. De permutatie parameter belangrijkheid werd met behulp van de validatieset bepaald. Voor beide soorten classificatoren zijn de resultaten hiervan te vinden in appendix G. Hieruit volgend werd voor het best presterende model de vijf belangrijkste kenmerken bepaald per subset, zie tabel 7. Het valt op dat hierbij vaak de recenste (rec.) gemeten (systolische) bloeddruken, standaarddeviatie (std.) van de hartslag en gemiddelde (gem.) zuurstofsaturatie de grootste (relatieve) belangrijkheid hebben.

Tabel 3: Meest optimale hyperparameters (*subset: vóór verkennende gegevensanalyse*)

Logistische Regressie		**Adaboost**		**GradientBoost**		**RandomForest**	
Parameter	*Waarde*	*Parameter*	*Waarde*	*Parameter*	*Waarde*	*Parameter*	*Waarde*
C	0.81	n_estimators	200	n_estimators	3000	n_estimators	1000
pentalty	l2	learning_rate	0.1	learning_rate	0.005	min_samples_leaf	0.0001
solver	liblinear	min_samples_leaf	0.0092	min_samples_leaf	0.0043	max_leaf_nodes	None
max_iter	1000	max_leaf_nodes	20	max_leaf_nodes	50	max_depth	128
		algorithm	SAMME	max_depth	16	max_features	log2
				max_features	log2	criterion	entropy
				loss	log_loss		

Tabel 4: Meest optimale hyperparameters (*subset: na verkennende gegevensanalyse*)

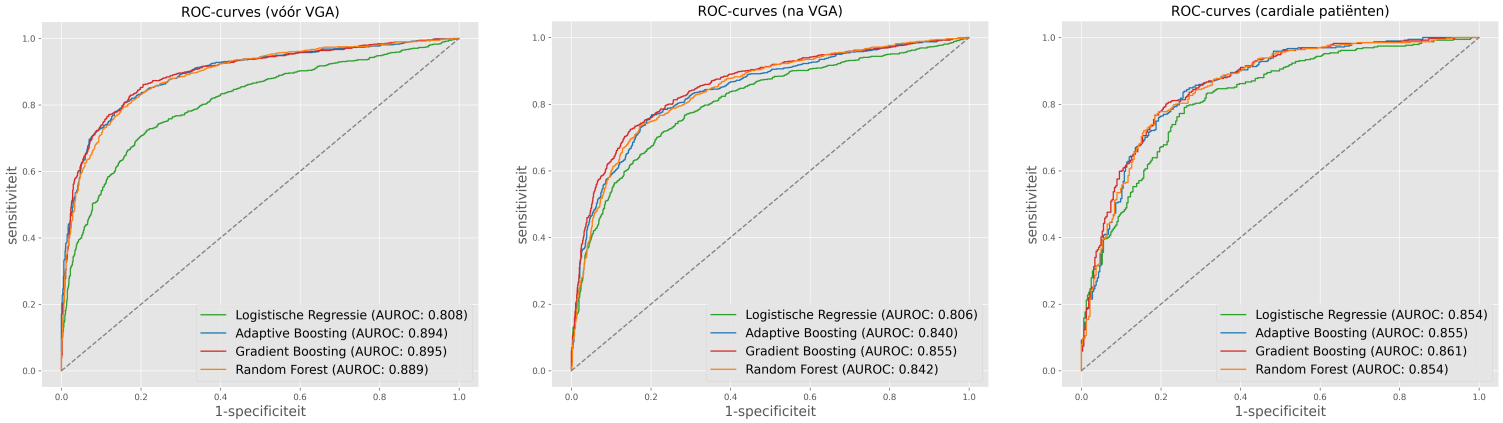
Logistische Regressie		**Adaboost**		**GradientBoost**		**RandomForest**	
Parameter	*Waarde*	*Parameter*	*Waarde*	*Parameter*	*Waarde*	*Parameter*	*Waarde*
C	1.88	n_estimators	200	n_estimators	1000	n_estimators	1000
pentalty	l1	learning_rate	0.5	learning_rate	0.01	min_samples_leaf	0.001
solver	saga	min_samples_leaf	0.0092	min_samples_leaf	0.0043	max_leaf_nodes	500
max_iter	2000	max_leaf_nodes	10	max_leaf_nodes	50	max_depth	64
		algorithm	SAMME	max_depth	16	max_features	None
				max_features	log2	criterion	gini
				loss	log_loss		

Tabel 5: Meest optimale hyperparameters (*subset: cardiale patiënten*)

Logistische Regressie		**Adaboost**		**GradientBoost**		**RandomForest**	
Parameter	*Waarde*	*Parameter*	*Waarde*	*Parameter*	*Waarde*	*Parameter*	*Waarde*
C	0.81	n_estimators	200	n_estimators	1000	n_estimators	1000
pentalty	l1	learning_rate	0.05	learning_rate	0.005	min_samples_leaf	0.001
solver	liblinear	min_samples_leaf	0.001	min_samples_leaf	0.002636	max_leaf_nodes	500
max_iter	100	max_leaf_nodes	20	max_leaf_nodes	50	max_depth	64
		algorithm	SAMME	max_depth	16	max_features	log2
				max_features	log2	criterion	entropy
				loss	log_loss		

Tabel 6: Modelprestaties van Logistische regressie, AdaBoost, GradientBoost en RandomForest voor alle subsets

	AUROC	Nauwkeurigheid	Precisie	Sensitiviteit	Specificiteit	F1
Logistische Regressie						
Vóór VGA	0.808	0.759	0.744	0.649	0.838	0.694
Na VGA	0.806	0.751	0.731	0.645	0.828	0.685
Cardiale patiënten	0.854	0.772	0.785	0.803	0.735	0.794
Adaptive Boosting						
Vóór VGA	0.894	0.826	0.825	0.742	0.886	0.782
Na VGA	0.840	0.776	0.762	0.678	0.846	0.718
Cardiaal	0.855	0.788	0.795	0.824	0.745	0.809
Gradient Boosting						
Vóór VGA	0.895	0.834	0.834	0.755	0.891	0.792
Na VGA	0.855	0.796	0.785	0.709	0.859	0.744
Cardiaal	0.861	0.788	0.803	0.811	0.760	0.807
Random Forest						
Vóór VGA	0.889	0.821	0.819	0.739	0.881	0.777
Na VGA	0.842	0.785	0.786	0.669	0.868	0.723
Cardiaal	0.854	0.772	0.785	0.803	0.735	0.794



(a) Vóór VGA (b) Na VGA (c) Cardiale patiënten

Figuur 5: ROC-curves van alle geoptimaliseerde ML-modellen getraind op de verschillende subsets

Tabel 7: Top vijf belangrijkste parameters GradientBoost per subset

	Gini-impurity	Permutatie
Vóór VGA		
1	Systolische Arteriële Bloeddruk (rec.)	Systolische Arteriële Bloeddruk (rec.)
2	Systolische Arteriële Bloeddruk (gem.)	Leeftijd
3	Hartslag (std.)	Systolische Arteriële Bloeddruk (gem.)
4	SI (rec.)	Systolische NI Arteriële Bloeddruk (rec.)
5	Hartslag (lls.)	Hartslag (std.)
Na VGA		
1	Systolische Arteriële Bloeddruk (rec.)	Systolische Arteriële Bloeddruk (rec.)
2	SI (rec.)	Systolische NI Arteriële Bloeddruk (rec.)
3	Systolische NI Arteriële Bloeddruk (rec.)	Systolische Arteriële Bloeddruk (gem.)
4	Gemiddelde Arteriële Bloeddruk (rec.)	Zuurstofsaturatie (gem.)
5	Zuurstofsaturatie (gem.)	Leeftijd
Cardiale patiënten		
1	Systolische Arteriële Bloeddruk (rec.)	Zuurstofsaturatie (gem.)
2	Zuurstofsaturatie (gem.)	Systolische Arteriële Bloeddruk (rec.)
3	SI (rec.)	Ademhalingsfrequentie (gem.)
4	Systolische NI Arteriële Bloeddruk (rec.)	SI (rec.)
5	RPP (rec.)	Systolische NI Arteriële Bloeddruk (rec.)

6 Discussie

6.1 Hoofdlijnen

Dit onderzoek betreft het ontwerpen van een machine learning algoritme dat hemodynamische instabiliteit (HI) kan voorspellen bij intensive care (IC) patiënten. Hiervoor wordt de Hemodyamic Deterioration Index (HDI) voorgesteld, waarmee getracht wordt informatiestress weg te nemen bij IC-clinici door één voorspellende waarde weer te geven. De HDI voorspelt HI één uur van tevoren zodat artsen proactief kunnen handelen, wat gezondheidsvoordelen voor patiënten zou kunnen opleveren.

Er zijn vier classificatiemodellen getraind op de MIMIC-III data, namelijk Logistische Regressie, Adaptive Boosting, Gradient Boosting en Random Forest. Er is een uitgebreide voorbewerking op de data uitgevoerd, waarin veel aandacht is besteed aan onder andere data labeling en toevoeging van extra berekende parameters en trendvariabelen. De vier modellen zijn getraind op de primaire voorbewerkte MIMIC-data en na extra bewerkingsstappen in een verkennende gegevensanalyse (VGA) om de effecten van de VGA te bepalen. Tot slot zijn de modellen hertraind op alleen cardiale patiënten om te onderzoeken hoe goed de HDI theoretisch zou kunnen werken op de IC van het CZE, waar veel cardio-thoracale chirurgie patiënten verblijven. Voor elke training zijn de hyperparameters, de modelprestaties en de belangrijkheid van parameter als uitkomsten bepaald.

Omdat perfusiefalen ten grondslag ligt aan HI, werd verwacht dat parameters als zuurstofsaturatie, hartslag, bloeddrukken en trends of combinaties hiervan zoals verschillende typen shockindices, belangrijk zullen zijn in het voorspellen van HI. Ook labwaarden die perfusiefalen kunnen aantonen, zoals lactaat, pH en pCO2, werden verwacht sterke voorspellers te zijn. Daarnaast werden betere prestaties verwacht bij beslisboom modellen ten opzichte van logistische regressie, omdat deze goed overweg kunnen met niet-lineaire relaties en onderlinge collineariteit tussen parameters. AdaBoost, zoals gebruikt in soortgelijke onderzoeken [21, 39], levert goede resultaten en zal vermoedelijk in dit onderzoek ook hoge prestaties leveren.

Uit de resultaten blijkt GradientBoosting de beste classificator voor het voorspellen van HI, met een AUROC van 0,895 voor de voorbewerkte MIMIC-dataset (se. 0,755, sp. 0,891), een AUROC van 0,855 na aanvullende bewerkingsstappen in een verkennende gegevensanalyse (se. 0,709, sp. 0,859) en een AUROC van 0,861 voor de cardiale patiëntgroep (se. 0,811, sp. 0,760). Hierbij blijken recent gemeten systolische bloeddrukken, standaarddeviaties van de hartslag en gemiddelde zuurstofsaturaties belangrijke parameters voor de voorspelling van HI.

6.2 Bevindingen

Goede datakwaliteit is essentieel voor het maken van een accuraat predictiemodel, omgang met missende waarden is hier een onderdeel van. In dit onderzoek is gekozen om parameters boven een bepaalde ontbrekendheidsgraad te excluderen van het onderzoek. Opvallend in tabel 2 is dat met een afkapwaarde van $< 50\%$ veel parameters worden gedropt, maar dat tussen 60%, 75% en 90% minder verschil zit vergeleken met het verschil in NaN-ratio tolerantie. Een afkapwaarde van 60% lijkt hierom een goede keuze om niet te veel parameters te verwijderen maar tegelijkertijd wel zekerheid te bieden dat elke kolom minimaal 40% data bevat. Deze keus is arbitrair bepaald op basis van een afkapwaarde die passend is voor deze specifieke dataset. In het geval dat het model opnieuw getraind zal worden op CZE-data zou deze waarde dus heroverwogen kunnen worden. Idealiter zou een uniforme procedure gebruikt worden om tot deze situatiespecifieke waarde te komen. Een waarde op basis van andere onderzoeken en de dimensies van de betreffende dataset zou de resultaten van dit onderzoek makkelijker te vergelijken maken met soortgelijke onderzoeken. Verder zijn enkele uitzonderingen gemaakt op deze NaN-ratio tolerantie. De extra berekende parameters OI, OSI en ratio CK-MB/CK zijn hoe dan ook geïncludeerd om een benadering te kunnen maken van de toegevoegde waarde in het model, de lage aanwezigheid van data maakt de benadering van deze uitkomsten echter minder representatief.

Met het doel om slechts parameters mee te nemen met klinische waarde voor een HI predictiemodel, zijn de parameters met HI gecorreleerd met de χ^2-toets en punt-biseriële correlatie. Zoals in de hypothese genoemd blijken parameters die geassocieerd worden met perfusie en de hemodynamiek ook sterk te correleren volgens de VGA. Zoals in hypothese is beschreven zijn de laatste en gemiddelde systolische arteriële bloeddrukken in nagenoeg elk model de sterkste voor-

spellers. Bij respiratoire kenmerken als ademfrequentie en zuurstofsaturatie worden hierbij ook de trendanalyses als belangrijk bevonden. Dit is vermoedelijk omdat de standaard ademfrequentie erg patiëntafhankelijk is en slechts een snelle verandering hierin een indicator kan zijn voor hemodynamische verslechtering. De standaarddeviatie en regressie van zuurstof-saturatie blijkt een erg goede predictie-maat te zijn, slechte perfusie staat dan ook gelijk aan een instabiele of verlaagde zuurstoftoevoer. Parameters afkomstig van arteriële labwaarden op het gebied van base excess, pH en lactaat zijn erg sterk in alle modellen, deze geven immers indicatie van acidose en hiermee perfusiefalen aan. Om de toegevoegde waarde hiervan verder te brengen zouden base excess en lactaat frequenter gemeten worden. De standaarddeviatie van de hartslag valt daarnaast ook erg hoog uit. Dit is terug te leiden naar verscheidene onderzoeken die de veranderlijkheid van de hartslag, beter bekend als *heart rate variability*, als indicator gebruiken voor stabiliteit van de patiënt [79]. Als laatst was opvallend dat de leeftijd van de patiënt belangrijker dan verwacht werd bevonden door een aantal modellen.

In totaal zijn 22 verschillende parameters uit de dataset gefilterd (p-waarde < 0.05) vanwege insignificante correlatie met HI. Dit werd voornamelijk gedaan voor labwaarden zoals hartziekte indicatoren en ontstekingswaarden, die dus geassocieerd worden met andere ziektebeelden.

In de VGA is de voorspellende waarde van elke parameter benaderd aan de hand van significantie (p-waarde). Voor continue waarden is dit met punt-biseriële en voor dichotome waarden met χ^2 methode berekend. Na het trainen van de modellen is de ranglijst van significantie vergeleken met het (relatief) parameterbelang in de ML-modellen, door de Gini-impurity of permutatie test bij de beslisboom modellen en gewichten bij het logistische regressie model. Direct opvallend is dat de dichotome variabelen van gemetenheid in de VGA erg hoog uitkomen, terwijl dit in de modellen niet het geval is. Dit is enerzijds mogelijk door verschil tussen significantie uit de punt-biseriële en χ^2 methode. De p-waarden afkomstig van de χ^2 worden misschien erg overschat omdat het een tetrachorische correlatie betreft. Anderzijds komt dit mogelijk door de verschillende aard van de twee soorten bivariabele toetsen. χ^2 benadert correlatie en significantie, terwijl Gini en permutatie het relatieve parameterbelang aangeven voor de classificatie van HI. Dit laatste zou ook kunnen worden ingezet voor de parameterselectie, ter vervanging van de statistische correlatietesten. De parameters met het laagste parameterbelang zouden bij een dergelijke methode iteratief kunnen worden verwijderd, waardoor enkel de parameters overblijven die het belangrijkst zijn voor de voorspelling van HI.

Multicollineariteit is in dit onderzoek aangetoond met de χ^2 en Pearson's r methode. Deze geven coëfficiënten van correlatie, wat in specifieke gevallen voor collineariteit kan zorgen. Parameters die als gevolg van deze selectie verwijderd werden, bestonden uit parameters die in hele sterke mate overeen kwamen met andere parameters en dus in feite dezelfde informatie bevatten. Dit bestond deels uit bloeddrukken en onderlinge aspecten hiervan zoals gemiddelde en de recentste waarde, omdat deze natuurlijkerwijs sterk samenhangen. Verder komen erg veel gemetenheid parameters sterk met elkaar overeen, dit is te verklaren doordat bij aanvraag van bloedonderzoek gelijk een bepaalde groep labwaarden onderzocht wordt. Op het gebied van (multi)collineariteit is correlatie een goede indicator, echter blijkt uit de literatuur VIF een sterkere indicator te zijn van collineariteit, omdat correlatie kan duiden op collineariteit, maar het niet altijd hoeft te bewijzen. VIF is een speciaal ontworpen indicator om (multi)collineariteit aan te tonen en zou in volgende onderzoeken overwogen moeten worden.

Bij training en evaluatie is gefocust op AUROC, dit is tevens de primaire uitkomstmaat van dit onderzoek. Op maximale AUROC gesorteerd heeft Gradient Boosting (AUROC=0,895; vóór VGA) de beste prestaties, in het eerdere HSI-onderzoek van de studenten BMT behaalde dit model met een AUROC van 0,806. AdaBoosting (AUROC=0,894; vóór VGA), in het onderzoek van Philips, door Rahman et al., behaalde dit model een AUROC van 0,82. Random Forest (AUROC=0,889; vóór VGA) behaalde in het HSI-onderzoek van de studenten BMT een AUROC van 0,771. Als laatst behaalde logistische regressie de slechtste resultaten (AUROC=0,854; cardiale patiënten), hier bestaat in de klinische literatuur nog geen vergelijkbaar onderzoek van. De verhoogde prestaties ten opzichte van soortgelijke onderzoeken zou te verklaren kunnen zijn door de nieuwe methode van labeling, maar overfitting zou een mogelijke reden kunnen zijn. Hierom zou het model extern gevalideerd moeten worden.

Zoals verwacht zijn beslisbomen dus sterkere classificatiemodellen dan het logistische regressie modellen. Ook op het gebied van uitlegbaarheid

en interpreteerbaarheid zijn deze sterker en kunnen deze dus meer toegevoegde waarde hebben in de kliniek. De beslisboommodellen verschillen onderling slechts in geringe mate van elkaar. De scores van secundaire uitkomstmaten nauwkeurigheid, precisie, sensitiviteit, specificiteit en F1 hebben nagenoeg dezelfde verhoudingen tussen de vier modellen. Enkele secundaire uitkomstmaten zijn opvallend, de sensitiviteit is namelijk beduidend lager dan specificiteit voor alle modellen, met uitzondering op de subgroep met alleen cardiale patiënten.

Na parameterselectie, gebaseerd op de VGA, verslechtert de AUROC van elk model relatief sterk. Slechts het logistische regressie model is nagenoeg gelijk gebleven in prestatie. Zoals verwacht zijn de beslisboommodellen erg goed in staat om de complexe relaties van de vele parameters vóór uitvoering van VGA te verwerken om zo tot een accuraat predictiemodel te komen. Deze complexe relatie en afhankelijkheid zorgen in het logistische regressiemodel duidelijk toch voor enige multicollineariteit. Wel is te zien bij logistische regressie voor de subset vóór VGA dat L2 als optimale hyperparameter naar voren komt, terwijl bij de subset na VGA L1 regularisatie als optimale hyperparameter naar voren komt. Naar alle waarschijnlijkheid komt dit door het feit dat L2 beter geschikt is voor multicollineariteit en wanneer dit (deels) wordt genomen L1 betere prestaties geeft.

Echter, de eliminatie van multicollineariteit heeft niet opgewogen tegen de eliminatie van voorspellende parameters en geeft dus een verlaging van modelprestaties. Toch is het opvallend dat alle modellen vóór VGA de beste prestaties behalen, ná VGA erg verslechteren en als laatst bij cardiaal, dus na VGA met enkel cardiale patiënten, de prestaties opeens beduidend verbeteren. Dit is vooral op basis van de F1-score die in het algemeen hoog is bij uniformiteit van data binnen een groep. Hierop gebaseerd kan men de hypothese stellen dat de reden van opname in een dergelijk model veel van invloed is op het predictiemodel. Een mogelijke manier om dit te verwerken zou zijn om de modellen apart te trainen per subgroep van reden van opname en dus bij binnenkomst van patiënt op IC te moeten instellen welk model toegepast moet worden.

6.3 Limitaties

Dit onderzoek haalt vooral zijn meerwaarde ten opzichte van andere hemodynamische predictiemodellen uit meerdere aspecten. De voorbewerkingsstappen zijn uitgebreid beschreven en zoveel mogelijk klinisch onderbouwd, wat ten goede komt aan de interpretatie en reproduceerbaarheid van de HDI. Het voldoet tevens voldoet aan de TRIPOD-richtlijn, wat gedetailleerd is besproken in appendix H. Een voorbeeld daarvan is de data labeling, waarvoor meerdere opties worden besproken en ondersteund met duidelijke figuren. Het theoretische kader biedt een goede instap in het onderwerp en machine learning, wat voor clinici met weinig voorkennis de mogelijkheid biedt de HDI geheel te begrijpen. De modellen zijn getraind op een grote dataset die nog is uitgebreid met ontbrekendheidsgraden, extra berekende parameters zoals shock en oxygenatie indices en met trendanalyses. Tijdens het ontwerpen van de HDI is telkens een terugkoppeling gemaakt naar de implementatie ervan op de intensive care van het Catharina Ziekenhuis Eindhoven (CZE), een tertiar ziekenhuis met veel cardiothoracale chirurgische patiënten. Toch zijn er nog veel aspecten aan het onderzoek die geoptimaliseerd of verbeterd kunnen worden.

De HDI is getraind op de Amerikaanse *Medical Information Mart for Intensive Care III*-database (MIMIC-III). Deze data verschilt van typische Nederlandse IC-data. De patiëntengroepen verschillen, alleen al in demografische gegevens. Ook de hemodynamische interventies kunnen op andere manieren en/of tijdstippen worden gegeven, wat invloed heeft op de labeling van de data. Daarnaast is de database opgezet tussen 2001 and 2012 [23], wat ook kan betekenen dat inmiddels de standaard kliniek is veranderd en dus interventies ook op andere manieren of tijdstippen gegeven kunnen worden. De labeling van de data is dus specifiek voor deze dataset en niet direct toepasbaar op andere data.

In de voorbewerking zijn meerdere exclusies gemaakt van data en patiënten om een betrouwbaardere dataset te krijgen, wat wel ten koste gaat van de inzetbaarheid van de HDI. Allereerst zijn de methoden van Rahman et al. [21] aangehouden. Zo is er alleen data meegenomen van patiënten zonder DNR. Daarnaast wordt van elke patiënt in de database alleen het eerste moment van instabiliteit van de eerste IC-opname en de eerste ziekenhuisopname meegenomen. Meerdere ziekenhuis- en IC verblijven en momenten van instabiliteit hebben invloed op de gezondheidstoestand van de patiënt en vormen een risico op selectiebias. Dat houdt wel in dat het algoritme niet (intern) gevalideerd is voor patiënten met een DNR en gevallen van meerdere IC-opnames.

Daarnaast is er rekening gehouden met een bepaalde *pretime* van zes uur. Data binnen de eerste zes uur van een IC-opname wordt als onbetrouwbaar beschouwd omdat patiënten vaak al instabiel zijn en dus verschillende therapieën (vloeistof en medicatie) krijgen. Bovendien is bij IC-patiënten bij binnenkomst vaak sprake van SIRS, een lichamelijke stressreactie op bijvoorbeeld sepsis, operatie of mentale stress. Na de pretime is de data betrouwbaarder om de algoritmen te trainen en voorspellingen op te doen. Toch is deze exclusie van de eerste zes uur aan data lichtelijk tegenintuïtief, omdat eerder beschreven is dat meerdere momenten van instabiliteit invloed hebben op elkaar en een risico vormen op selectiebias. Toch is het de bedoeling van de HDI om pas alarm te geven wanneer het nodig is. Als de artsen dus al medicaties aan het toedienen zijn, is een alarm overbodig. De HDI is van toegevoegde waarde wanneer het voor stabiele patiënten kan voorspellen of ze instabiel worden. Dit is dus een weloverwogen concessie, wederom gebaseerd op Rahman et al. [21]. De implementatie van de pretime binnen dit onderzoek houdt in dat alle therapieën die begonnen zijn in de eerste zes uur van opname niet worden meegenomen. Toch zijn er therapieën zoals vloeistoftoedieningen die voor langere tijd kunnen duren, zelfs meerdere dagen, die nu niet worden meegenomen in de labeling van de data. Een andere manier zou kunnen inhouden dat alleen de eerste zes uur van deze in de pretime gestarte therapieën niet worden meegenomen, maar dat ze dus worden afgekapt op zes uur.

Voor hemodynamische instabiliteit is geen harde definitie uitgedrukt in klinische metingen, zoals bloeddruk of zuurstofsaturatie. Daarom is de definitie aangehouden die Rahman et al. [21] ook hebben gebruikt, namelijk wanneer een patiënt een significante hemodynamische interventie krijgt. Omdat deze definitie gebruikt wordt als label in het predictiemodel, wordt in werkelijkheid niet voorspeld wanneer een patiënt hemodynamisch instabiel wordt, maar eigenlijk wanneer een patiënt een daaraan gerelateerde interventie krijgt. Het ware moment van instabiliteit bevindt zich daarvoor, waarop een arts zal reageren. Met deze systematische onnauwkeurigheid wordt rekening gehouden in de zogenaamde predictietijd, hier is dat één uur gekozen. De data wordt dus vanaf een uur voor het label retrospectief geëxtraheerd, zodat het model een uur voor het moment van interventie HI kan voorspellen. Eén uur is gekozen als ruime tijd waarin de arts nog proactief kan handelen.

Wanneer een interventie precies significant wordt, hangt af van de toediening. Elke toedoening van inotropica of vasopressieve medicatie wordt als significant beschouwd. Nou kan het voorkomen dat een patiënt deze medicatie toegediend krijgt vanwege het klaarmaken voor bijvoorbeeld een operatie. De patiënt wordt dan wel als instabiel wordt gelabeld, ook al is dat niet per definitie het geval. In dit onderzoek wordt geen rekening gehouden met deze situatie. Deze gevallen zouden geïdentificeerd kunnen worden wanneer er data van vitale kenmerken mist voor langere periode na toediening van deze medicatie, omdat de patiënt dan van de IC weg is. Dit incorrecte 'instabiele' moment zou vervolgens geëxcludeerd kunnen worden.

Daarnaast worden bepaalde hoeveelheden toegediende vloeistof en/of rode bloedcellen binnen een specifiek tijdsvenster ook als significant beschouwd. Binnen de data wordt deze significantie bepaald door het tijdsvenster per minuut te verschuiven over een aangemaakte tijdas en daarin de vocht- en bloedtoedieningen te sommeren. Wanneer de sommatie boven de drempelwaarde uitkwam voor dat tijdsvenster, werd het label 'instabiel' geplaatst aan het begin van het venster. Zie ook appendix B voor duidelijke figuren.

Idealiter zou het HI-label niet aan het begin van het tijdvenster worden gezet, maar op het moment dat de arts een significante interventie inzet binnen datzelfde tijdvenster. Ter verduidelijking, het tijdsvenster dat voldoet aan het criterium blijft dus hetzelfde, echter wordt de plaatsing van het HI-label binnen dat venster dus accurater. Volgens de definitie van Rahman et al. wordt hemodynamische instabiliteit namelijk gedefinieerd als een periode waarin een significante hemodynamische interventie plaatsvindt. Het meest wenselijk zou dan ook zijn om het label precies te plaatsen waar deze significante interventie start. Ter illustratie, een aannemelijke situatie is een patiënt die een relatief 'stabiele', dus langdurige en langzame vochttoediening krijgt als een soort onderhoudsdosis. Als het slecht gaat met de patiënt, zal een arts een extra toediening geven. Het is juist deze verhoging, bovenop de stabiele dosis, waar het label accurater geplaatst kan worden. Om dit te realiseren zou de afgeleide van de infuussnelheid berekend kunnen worden en het label te plaatsen bij een significante verhoging.

Wanneer de labeling wordt bepaald aan de hand van infuussnelheden is het belangrijk rekening te houden met het lichaamsgewicht van de patiënt. Eenzelfde infuussnelheid zou significanter zijn voor een patiënt met een lager li-

chaamsgewicht ten opzichte van een patiënt met een hoger lichaamsgewicht. Er zou bijvoorbeeld gerekend kunnen worden met een infuusnelheid per kilogram, mL/kg/uur, zoals [39]. Daarnaast kwamen in de MIMIC data enorm hoge infuussnelheden voor binnen hele korte tijd, waarin liters vloeistof in één korte interventie werden toegediend. De vraag is of dit een meetfout is of een uitschieter, maar een labeling gebaseerd op infuussnelheden kan wellicht gebaat zijn bij het filteren op plausibiliteit.

Om op dezelfde manier meetfouten en uitschieters te filteren in de overige data worden twee manieren gebruikt. Als er voor een parameter gemakkelijk een bepaald bereik gedefinieerd kan worden, waartussen de parameter nog aannemelijke waarden heeft, kunnen alle waarden buiten dat bereik direct worden verwijderd. Deze plausibele bereiken kunnen worden bepaald worden aan de hand van bijvoorbeeld literatuur. Als dat niet lukt, bijvoorbeeld omdat er geen duidelijke plausibele afkapwaarden te definiëren zijn, werden uitschieters herkend en gefilterd door middel van een robuuste Z-score die, in tegenstelling tot een normale Z-score, resistent is tegen enorme uitschieters. In dit onderzoek zijn plausibele bereiken opgesteld aan de hand van voornamelijk visuele inspectie van de data. Wanneer parameters geen enorme uitschieters bevatten, kon dus niet makkelijk een bereik gedefinieerd worden en is een robuuste Z-score gebruikt. De beste methode blijft het definiëren van een plausibel bereik op basis van literatuur of eigen inzicht, omdat met een Z-score nooit met zekerheid kan worden gezegd of alle uitschieters zijn verwijderd en of er geen goede data overbodig wordt verwijderd.

(Normaal)waarden van bloedgassen zijn afhankelijk van de afnamelocatie. Binnen MIMIC-III wordt onderscheid gemaakt tussen arteriële, veneuze, centraal veneuze, gemengd veneuze of een onbekende afnamelocatie. Centraal en gemengd veneuze bloedmonsters zijn samengenomen met veneuze monsters omdat deze waarden dicht bij elkaar lagen en op die manier de data te versimpelen. Onbekende afnamelocaties kwamen vaak voor ($\pm$14,8%) en werden overgeschreven naar arterieel of veneus op basis van de hoogste waarschijnlijkheid zoals gespecificeerd in appendix C.2. Deze overschrijving is nooit perfect en fouten kunnen zorgen voor ruis in de data, wat de prestaties negatief zou kunnen beïnvloeden. Toch wordt geschat dat de overschrijvingen wel accuraat genoeg zijn om geen last te hebben van significante prestatievermindering. Binnen de plausibele bereiken van deze labwaarden werd geen onderscheid gemaakt tussen arterieel en veneus omdat, hoewel dat deze parameters van elkaar verschillen, ze wel binnen dezelfde ordegrootte vallen.

Het bemonsteren van de data, dus van hematologische chemie, bloedgassen en vitale kenmerken gebeurt binnen een bepaald venster vóór de predictietijd. Dit tijdvenster is anders voor de verschillende datasets en is gebaseerd op de duur waarvoor een arts de gemeten parameter nog vertrouwt. Deze vensters zijn afgeleid uit de data door te kijken naar de gemiddelde tijdregistraties tussen niet-continu gemeten parameters voor elke dataset. Hoewel de vitale kenmerken en sommige bloedgassen (o.a. pH, pCO2, pO2) doorgaans elke vijf minuten waren geregistreerd, werden dus alleen de variabelen meegenomen in het gemiddelde die minder frequent zijn gemeten. Op die manier worden de vensters bepaald zodat ook minder frequent gemeten variabelen genoeg worden bemonsterd. Daarnaast kan het voorkomen dat een patiënt de IC tijdelijk verlaat, bijvoorbeeld voor een operatie, waardoor alle metingen tijdelijk stoppen. De tijd tussen de laatste meting voor het verlaten en de eerste bij terugkeer is niet representatief voor de normale frequentie van metingen maar wordt wel meegenomen in dit gemiddelde. Uiteindelijk is het 75e percentiel als richtlijn voor de lengte van het tijdvenster genomen zodat van de parameters veel data meegenomen kan worden maar nog steeds representatief zijn.

In de data-analyse naar correlatie en multicollineariteit is gebruik gemaakt van Pearson's correlatiecoëfficiënt en een specificatie hiervan, namelijk punt-biseriële correlatie. Voor beide methoden geldt voorwaarde van een normale verdeling in de onderzochte data. Voor grote datasets zoals in dit onderzoek betreffend zou dit visueel onderzocht kunnen worden door middel van QQ-plots of een staafdiagram. Omwille van de tijd en de grootte van de dataset is uitgegaan van een normale verdeling. Indien dit echter niet het geval is zouden de uit VGA verkregen resultaten onbetrouwbaar kunnen zijn.

Voor het trainen van de ML algorimtes werd eerst de dataset gesplitst in 80% training- en 20%-testdata, maar had ook op andere manieren gekund zoals 60/40-, 70/30- of 80/20% [42]. Voor grote datasets zou ook een 90/10%-split een goede optie kunnen zijn [42], Rahman et al. gebruiken een 80-20%-split. Als gevolg van deze splitsing kan overfitting ontstaan, wanneer de training subset heel erg verschilt van de validatie-subset. Dan zal het model minder goed voorspellingen kunnen doen op de

validatie-subset en zal de performance dus slechter uitvallen [40, 41].

Tijdens het optimaliseren van hyperparameters werd binnen de traindata textit(r times repeated k-fold cross validation toegepast om overfitting te voorkomen. Vaak wordt k=10 gebruikt en ook aangeraden [41, 80], maar vanwege de grote dataset en de daarvoor grote benodigde rekenkundige kracht is gekozen voor k=6 en r=1 gebaseerd op [40]. Theoretisch zijn meer folds en repeats altijd beter. Kruisvalidatie kan ook toegepast worden op de oorspronkelijke 80/20%-split in zogenaamde *nested cross validation* (nCV). Dit resulteert in enorm veel iteraties, namelijk $k_{outer} \cdot k_{inner} \cdot n_{iter}$ per klassificatiemodel. Bovendien wordt voor elke outer fold een nieuw model gecreëerd, elk met eigen geoptimaliseerde hyperparameters specifiek voor de outer folding. Van elk model wordt vervolgens de performance geëvalueerd en gemiddeld in de nCV-score. Deze nCV-score is een goede maat voor het selecteren van een klassificatiemodel, maar is dus niet bedoeld voor het selecteren van een set geoptimaliseerde hyperparameters binnen één classificator. Vanwege gelimiteerde tijd en rekenkundige kracht werd voor hyperparameter optimalisatie eerst een *random search* uitgevoerd met n=50 iteraties. Hiervoor zijn aannemelijke waarden van hyperparameters van tevoren gespecificeerd, zie appendix F.1. De beste paar waarden gebaseerd op de modelprestaties, of een klein bereik daaromheen, worden vervolgens gebruikt in een *gridsearch*. In een ideale situatie wordt er enkel een gridsearch uitgevoerd, omdat er dan modellen worden getraind op alle mogelijke combinaties van waarden van hyperparameters, waaruit de beste combinatie kan worden geselecteerd. Dit laatste is nauwkeuriger dan de random search met een gelimiteerd aantal iteraties.

Eerder is beschreven dat er variabelen met een bepaalde ratio van missende op beschikbare data worden verwijderd uit de dataset. Aanvullend moeten in de overgebleven kolommen de onbekende waarden nog worden verholpen. Dit is gedaan met simpele imputatie met het gemiddelde. Dit gebeurde na het plausibiliteitsfilter dus wordt het gemiddelde niet beïnvloedt door uitschieters of meetfouten. Het gebruik van het gemiddelde zou een onnauwkeurige weergave zijn van de waarden die er eigenlijk hadden moeten staan. Er zijn veel andere manieren om de missende waarden in te vullen, zoals regressie, hot deck (ook meerdere vormen van) of expectation maximisation. Ook zijn er verschillende machine learning algoritmes die hiervoor zijn ontwikkeld zoals *K nearest neighbours* of ook beslisboom modellen. Het is interessant om verder te onderzoeken wat al deze verschillende manieren voor effect hebben op de prestaties van de modellen, bijvoorbeeld door de verschillende AUROCs te vergelijken.

Voor logistische regressie classificatie modellen moet eerst voor elke parameter normalisatie worden toegepast. Dat is gedaan door middel van een standaard schaler tussen 0 en 1 die gebruik maakt van minimale en maximale waarden van de parameter. Een probleem treedt hier op, omdat de dataset zowel dichotome als normale continue variabelen bevat. De dichotome parameters worden gecodeerd als 0 en 1, en het verschil daartussen is per definitie groter dan bijvoorbeeld het geschaalde verschil tussen bijvoorbeeld zuurstofsaturatie of gewicht. Dit kan leiden tot minder nauwkeurige voorspellingen. Daarnaast is het zo dat de normalisatie uitsluitend gebaseerd moet zijn op de (minimale en maximale waarden van) de traindata. Als de testdata wel wordt meegenomen in de normalisatie zal het model worden getraind op een deel informatie van de testdata, wat voor overfitting kan zorgen. Dit is helaas foutief toegepast in dit onderzoek.

6.4 Conclusie

De totstandkoming van de HDI is uitgebreid en klinisch onderbouwd, tevens volgens de TRIPOD-richtlijn, wat zorgt voor een goed interpreteerbaar en reproduceerbaar model. Gradient Boosting (AUROC=0,895) is het best presterende classificatie ML-model voor de voorspelling van HI. Dit model werkt beter dan vergelijkbare modellen voor HI. Er werd verwacht dat Adaptive Boosting het beste resultaat zou leveren. Hoewel de hypothese incorrect is, presteert een vergelijkbare boosting methode wel het beste. Daarnaast zijn de prestaties van de HDI goed bij de subset van cardiale patiënten, wat bewijs zou kunnen zijn dat de HDI ook goed zou kunnen werken op de intensive care van het CZE. De belangrijkste parameters voor de voorspelling van HI waren, in lijn met wat verwacht werd, de systolische bloeddrukken, standaarddeviaties van de hartslag en gemiddelde zuurstofsaturaties.

6.5 Klinische implicaties en aanbevelingen

De HDI kan de informatiestress op de intensive care verhelpen en tegelijkertijd de arts de mo-

gelijkheid geven proactief te handelen bij hemodynamische verslechtering. De totstandkoming van de HDI is erg uitgebreid gedocumenteerd, waardoor clinici de HDI goed kunnen begrijpen en toepassen. Ook is de HDI reproduceerbaar, zodat het kan worden gepersonaliseerd en geoptimaliseerd voor implementatie op IC's van zowel het CZE als andere ziekenhuizen. Vóór de implementatie kan nog veel onderzoek gedaan worden naar de HDI. Allereerst wordt aangeraden om de hiervoor beschreven limitaties te optimaliseren door te experimenteren met:

- parameter inclusie van variabelen binnen MIMIC-III die nog niet worden gemeten op de IC maar wel hoog correleren met HI,
- labeling op basis van significante infuussnelheden,
- verschillende tijdsvensters voor data bemonstering,
- herhaalde kruisvalidatie met meer folds,
- *nested cross validation* voor klassificatiemodel selectie,
- een volledige gridsearch voor hyperparameter optimalisatie.

Daarnaast zouden de invloeden van predictietijd onderzocht kunnen worden, waarbij de modelprestaties kunnen worden vergeleken op verschillende tijden vooraf aan het label. Verder is een externe validatie nodig op een DNR-patiëntgroep om te controleren hoe goed de HDI voor die groep werkt. Uiteindelijk is het de bedoeling dat de transparante en uitgebreide documentatie van de HDI toepassingen binnen andere ziekenhuizen dan het CZE mogelijk maakt. Daarvoor is het belangrijk dat alle keuzes in de voorbewerking heroverwogen kunnen worden met inachtneming van ziekenhuis-specifieke klinische praktijk en belangen. Zo kunnen bijvoorbeeld invoervariabelen zoals base excess en lactaat frequenter gemeten worden en pH zelfs continu. Parameters van andere ziekenhuizen moeten vergeleken worden met die op de eigen IC van belang zijn, zoals in dit onderzoek gedaan is voor het CZE. Er kan een subsetanalyse worden uitgevoerd op een bepaalde patiëntengroep, zoals in dit geval hartpatiënten, om een betere schatting te krijgen van de prestaties van de HDI bij specifieke patiënten. Ook keuzes voor pretime en predictijd zijn onderhevig aan verandering als daar behoefte aan is. Zo zou bij een desgewenste kleinere pretime al eerder in de IC-opname de HDI worden ingezet maar met kans dat het al snel instabiliteit zal aangeven.

Referenties

1. Marik PE, Monnet X en Teboul JL. Hemodynamic parameters to guide fluid therapy. Annals of Intensive Care 2011 Mar; 1:1–9. DOI: 10.1186/2110-5820-1-1/FIGURES/4
2. Tod Brindle C, Malhotra R, O'rourke S, Currie L, Chadwik D, Falls P, Adams C, Swenson J, Tuason D, Watson S en Creehan S. Turning and repositioning the critically ill patient with hemodynamic instability: A literature review and consensus recommendations. Journal of Wound, Ostomy and Continence Nursing 2013 May; 40:254–67. DOI: 10.1097/WON.0B013E318290448F
3. Weil MH. Defining Hemodynamic Instability. *Functional Hemodynamic Monitoring*. Red. door Pinsky MR en Payen D. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005 :9–17. DOI: 10.1007/3-540-26900-2{_}2
4. Sakr Y, Reinhart K, Vincent JL, Sprung CL, Moreno R, Ranieri VM, De Backer D en Payen D. Does dopamine administration in shock influence outcome? Results of the Sepsis Occurrence in Acutely Ill Patients (SOAP) Study. Critical Care Medicine 2006 Mar; 34:589–97. DOI: 10.1097/01.CCM.0000201896.45809.E3
5. Vincent JL, Ince C en Bakker J. Clinical review: Circulatory shock - an update: a tribute to Professor Max Harry Weil. Critical Care 2012 Nov; 16:1–5. DOI: 10.1186/CC11510/FIGURES/4
6. Hanqing Cao, Eshelman L, Chbat N, Nielsen L, Gross B en Saeed M. Predicting ICU hemodynamic instability using continuous multiparameter trends. *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008 Aug :3803–6. DOI: 10.1109/IEMBS.2008.4650037
7. Sevransky J. Clinical assessment of hemodynamically unstable patients. Current Opinion in Critical Care 2009 Jun; 15:234–8. DOI: 10.1097/MCC.0B013E32832B70E5
8. Colon Hidalgo D, Patel J, Masic D, Park D en Rech MA. Delayed vasopressor initiation is associated with increased mortality in patients with septic shock. Journal of Critical Care 2020; 55:145–8. DOI: https://doi.org/10.1016/j.jcrc.2019.11.004
9. Joon Lee, Scott DJ, Villarroel M, Clifford GD, Saeed M en Mark RG. Open-access MIMIC-II database for intensive care research. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011 Aug :8315–8. DOI: 10.1109/IEMBS.2011.6092050
10. Li Y, Li H en Zhang D. Timing of norepinephrine initiation in patients with septic shock: a systematic review and meta-analysis. Critical care (London, England) 2020 Aug; 24:488. DOI: 10.1186/s13054-020-03204-x
11. Cao H, Eshelman LJ, Nielsen L, Gross BD, Saeed M en Frassica JJ. Hemodynamic Instability Prediction Through Continuous Multiparameter Monitoring in ICU. Journal of Healthcare Engineering 2010; 1:509–34. DOI: 10.1260/2040-2295.1.4.509
12. Chambrin MC. Alarms in the intensive care unit: How can the number of false alarms be reduced? Critical Care 2001 May; 5:184–8. DOI: 10.1186/CC1021/TABLES/1
13. Adlung L, Cohen Y, Mor U en Elinav E. Machine learning in clinical decision making. Med 2021 Jun; 2:642–65. DOI: 10.1016/j.medj.2021.04.006
14. Rajkomar A, Dean J en Kohane I. Machine Learning in Medicine. New England Journal of Medicine 2019 Apr; 380:1347–58. DOI: 10.1056/NEJMra1814259
15. Liu Q en Wu Y. Supervised Learning. *Encyclopedia of the Sciences of Learning*. Boston, MA: Springer US, 2012 :3243–5. DOI: 10.1007/978-1-4419-1428-6{_}451
16. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, Mottram A, Meyer C, Ravuri S, Protsyuk I, Connell A, Hughes CO, Karthikesalingam A, Cornebise J, Montgomery H, Rees G, Laing C, Baker CR, Peterson K, Reeves R, Hassabis D, King D, Suleyman M, Back T, Nielson C, Ledsam JR en Mohamed S. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019 Aug; 572:116–9. DOI: 10.1038/s41586-019-1390-1

17. Koyner JL, Carey KA, Edelson DP en Churpek MM. The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model. Critical Care Medicine 2018 Jul; 46:1070–7. DOI: 10.1097/CCM.0000000000003123

18. Peiffer-Smadja N, Rawson T, Ahmad R, Buchard A, Georgiou P, Lescure FX, Birgand G en Holmes A. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clinical Microbiology and Infection 2020 May; 26:584–95. DOI: 10.1016/j.cmi.2019.09.009

19. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, Fuchs BD, Meadows L, Lynch M, Donnelly PJ, Pavan K, Fishman NO, Hanson CW en Umscheid CA. A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock. Critical Care Medicine 2019 Nov; 47:1485–92. DOI: 10.1097/CCM.0000000000003891

20. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, Rinehart J en Cannesson M. Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. Anesthesiology 2018 Oct; 129:663–74. DOI: 10.1097/ALN.0000000000002300

21. Rahman A, Chang Y, Dong J, Conroy B, Natarajan A, Kinoshita T, Vicario F, Frassica J en Xu-Wilson M. Early prediction of hemodynamic interventions in the intensive care unit using machine learning. Critical Care 2021 Dec; 25:1–9. DOI: 10.1186/S13054-021-03808-X/FIGURES/4

22. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG en Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data 2018 Sep; 5:180178. DOI: 10.1038/sdata.2018.178

23. Johnson AE, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L en Mark RG. MIMIC-III, a freely accessible critical care database. Scientific Data 2016 3:1 2016 May; 3:1–9. DOI: 10.1038/sdata.2016.35

24. Chong WH, Saha BK en Medarov BI. Comparing Central Venous Blood Gas to Arterial Blood Gas and Determining Its Utility in Critically Ill Patients: Narrative Review. Anesthesia & Analgesia 2021 Aug; 133:374–8. DOI: 10.1213/ANE.0000000000005501

25. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KGM en Hooft L. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Medical Research Methodology 2022 Dec; 22:12. DOI: 10.1186/s12874-021-01469-6

26. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S en Berk M. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. Journal of Medical Internet Research 2016 Dec; 18:e323. DOI: 10.2196/jmir.5870

27. Collins GS, Reitsma JB, Altman DG en Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Medicine 2015; 13:1. DOI: 10.1186/s12916-014-0241-z

28. Cecconi M, De Backer D, Antonelli M, Beale R, Bakker J, Hofer C, Jaeschke R, Mebazaa A, Pinsky MR, Teboul JL, Vincent JL en Rhodes A. Consensus on circulatory shock and hemodynamic monitoring. Task force of the European Society of Intensive Care Medicine. Intensive Care Medicine 2014 Nov; 40:1795. DOI: 10.1007/S00134-014-3525-Z

29. Parlow S, Weng W, Di Santo P, Jung RG, Lepage-Ratte MF, Motazedian P, Prosperi-Porta G, Abdel-Razek O, Simard T, Chan V, Labinaz M, Froeschl M, Mathew R en Hibbert B. Significant Valvular Dysfunction and Outcomes in Cardiogenic Shock: Insights From the Randomized DOREMI Trial. The Canadian journal of cardiology 2022 Aug; 38:1211–9. DOI: 10.1016/J.CJCA.2022.04.004. Available from: https://pubmed.ncbi.nlm.nih.gov/35430192/

30. Mulder MP, Broomé M, Donker DW en Westerhof BE. Distinct morphologies of arterial waveforms reveal preload-, contractility-, and afterload-deficient hemodynamic instability: An in silico simulation study. Physiological Reports 2022 Apr; 10:e15242. DOI: 10.14814/PHY2.15242

31. Chakraborty RK en Burns B. Systemic Inflammatory Response Syndrome. StatPearls 2023 May. Available from: https://www.ncbi.nlm.nih.gov/books/NBK547669/

32. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, Poll T van der, Vincent JL en Angus DC. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb; 315:801. DOI: 10.1001/jama.2016.0287

33. Pinsky MR. Functional hemodynamic monitoring. Critical care clinics 2015 Jan; 31:89–111. DOI: 10.1016/J.CCC.2014.08.005

34. Saric L, Prkic I en Karanovic N. Inotropes and Vasopressors. Signa Vitae 2023 Feb; 13:46–52. DOI: 10.22514/SV131.032017.6

35. Lewis SR, Pritchard MW, Evans DJ, Butler AR, Alderson P, Smith AF en Roberts I. Colloids versus crystalloids for fluid resuscitation in critically ill people. The Cochrane Database of Systematic Reviews 2018 Aug; 2018. DOI: 10.1002/14651858.CD000567.PUB7

36. Saugel B, Klein M, Hapfelmeier A, Phillip V, Schultheiss C, Meidert AS, Messer M, Schmid RM en Huber W. Effects of red blood cell transfusion on hemodynamic parameters: a prospective study in intensive care unit patients. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 2013 Dec; 21:21. DOI: 10.1186/1757-7241-21-21

37. Rebala G, Ravi A en Churiwala S. Regressions. *An Introduction to Machine Learning*. Cham: Springer International Publishing, 2019 :25–40. DOI: 10.1007/978-3-030-15729-6{_}3

38. El Morr C, Jammal M, Ali-Hassan H en El-Hallak W. Logistic Regression. *Machine Learning for Practical Decision Making. International Series in Operations Research & Management Science*. Deel 334. Springer, 2022. Hfdstk. Logistic Regression:231–49. DOI: 10.1007/978-3-031-16990-8{_}7/FIGURES/14

39. Potes C, Conroy B, Xu-Wilson M, Newth C, Inwald D en Frassica J. A clinical prediction model to identify patients at high risk of hemodynamic instability in the pediatric intensive care unit. Critical Care 2017 Nov; 21:1–8. DOI: 10.1186/S13054-017-1874-Z/FIGURES/3

40. Yadav S en Shukla S. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. Proceedings - 6th International Advanced Computing Conference, IACC 2016 2016 Aug :78–83. DOI: 10.1109/IACC.2016.25

41. Berrar D. Cross-Validation Call for Papers for Machine Learning journal: Machine Learning for Soccer View project Cross-validation. DOI: 10.1016/B978-0-12-809633-8.20349-X. Available from: https://doi.org/10.1016/B978-0-12-809633-8.20349-X

42. Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. 2018

43. Zhang X, Haneishi H en Liu H. Multiscale modeling of the cardiovascular system for infants, children, and adolescents: Age-related alterations in cardiovascular parameters and hemodynamics. Computers in Biology and Medicine 2019 May; 108:200–12. DOI: 10.1016/J.COMPBIOMED.2019.03.021

44. Jayes RL, Zimmerman JE, Wagner DP, Draper EA, Knaus WA, Chernow B, Dantzker D, Leiken J, Parrillo JE, Sibbald WJ en Vincent JL. Do-Not-Resuscitate Orders in Intensive Care Units: Current Practices and Recent Changes. JAMA 1993 Nov; 270:2213–7. DOI: 10.1001/JAMA.1993.03510180083039

45. Cohen RI, Lisker GN, Eichorn A, Multz AS en Silver A. The impact of do-not-resuscitate order on triage decisions to a medical intensive care unit. Journal of Critical Care 2009; 24:311–5. DOI: 10.1016/j.jcrc.2008.01.007

46. Beach MC en Sean Morrison R. The Effect of Do-Not-Resuscitate Orders on Physician Decision-Making. Journal of the American Geriatrics Society 2002 Dec; 50:2057–61. DOI: 10.1046/J.1532-5415.2002.50620.X

47. Driggers KE, Dishman SE, Chung KK, Olsen CH, Ryan AB, McLawhorn MM en Johnson LS. Perceptions of care following initiation of do-not-resuscitate orders. Journal of Critical Care 2022 Jun; 69. DOI: 10.1016/J.JCRC.2022.154008

48. Li W, Mo W, Zhang X, Squiers JJ, Lu Y, Sellke EW, Fan W, DiMaio JM en Thatcher JE. Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging. https://doi.org/10.1117/1.JBO.20.12.121305 2015 Aug; 20:121305. DOI: 10.1117/1.JBO.20.12.121305

49. Vanderschueren S, Deeren D, Knockaert DC, Bobbaers H, Bossuyt X en Peetermans W. Extremely elevated C-reactive protein. European Journal of Internal Medicine 2006 Oct; 17:430–3. DOI: 10.1016/j.ejim.2006.02.025

50. Chen K, Kong W, Liao C, Liang Y, Ding J, Zhu X en Yang K. Comparison of laboratory results between central venous access devices and venipuncture: A systematic review and meta-analysis. Journal of Vascular Access 2023 Feb. DOI: 10.1177/11297298231155522/ASSET/IMAGES/LARGE/10.1177{_}11297298231155522-FIG4.JPEG

51. Trivedi S, Demirci O, Arteaga G, Kashyap R en Smischney NJ. Evaluation of preintubation shock index and modified shock index as predictors of postintubation hypotension and other short-term outcomes. Journal of Critical Care 2015 Aug; 30:1–861. DOI: 10.1016/J.JCRC.2015.04.013

52. Kamikawa Y en Hayashi H. Equivalency between the shock index and subtracting the systolic blood pressure from the heart rate: an observational cohort study. BMC Emergency Medicine 2020 Dec; 20:1–8. DOI: 10.1186/S12873-020-00383-2/FIGURES/3

53. Dalmau R. The diastolic shock index works... but, what is it? Annals of Intensive Care 2020 Dec; 10:103. DOI: 10.1186/S13613-020-00720-5

54. Althunayyan SM, Alsofayan YM en Khan AA. Shock index and modified shock index as triage screening tools for sepsis. Journal of Infection and Public Health 2019 Nov; 12:822–6. DOI: 10.1016/J.JIPH.2019.05.002

55. Liu Yc, Liu Jh, Fang ZA, Shan Gl, Xu J, Qi Zw, Zhu Hd, Wang Z en Yu Xz. Modified shock index and mortality rate of emergency patients. World Journal of Emergency Medicine 2012; 3:114. DOI: 10.5847/WJEM.J.ISSN.1920-8642.2012.02.006

56. Elsayed Y en Abdul Wahab MG. A new physiologic-based integrated algorithm in the management of neonatal hemodynamic instability. European Journal of Pediatrics 2022 Mar; 181:1277–91. DOI: 10.1007/S00431-021-04307-5/FIGURES/3

57. Dechert RE, Park PK en Bartlett RH. Evaluation of the oxygenation index in adult respiratory failure. Journal of Trauma and Acute Care Surgery 2014 Feb; 76:469–73. DOI: 10.1097/TA.0B013E3182AB0D27

58. Roca O, Messika J, Caralt B, García-de-Acilu M, Sztrymf B, Ricard JD en Masclans JR. Predicting success of high-flow nasal cannula in pneumonia patients with hypoxemic respiratory failure: The utility of the ROX index. Journal of critical care 2016 Oct; 35:200–5. DOI: 10.1016/J.JCRC.2016.05.022

59. Roca O, Caralt B, Messika J, Samper M, Sztrymf B, Hernández G, García-De-Acilu M, Frat JP, Masclans JR en Ricard JD. An index combining respiratory rate and oxygenation to predict outcome of nasal high-flow therapy. American Journal of Respiratory and Critical Care Medicine 2019; 199:1368–76. DOI: 10.1164/rccm.201803-0589OC

60. Bai S, Wu B, Yao Z, Zhu X, Jiang Y en Wang H. Development and validation of a clinical model to predict intraoperative hemodynamic instability in patients with pheochromocytomas surgery. Endocrine Journal 2020; 67:81–9. DOI: 10.1507/ENDOCRJ.EJ19-0278

61. Al-Hadi HA en Fox KA. Cardiac Markers in the Early Diagnosis and Management of Patients with Acute Coronary Syndrome. Sultan Qaboos University Medical Journal 2009 Dec; 9:231

62. Kubat M. An Introduction to Machine Learning. An Introduction to Machine Learning 2017 Sep :1–348. DOI: 10.1007/978-3-319-63913-0/COVER

63. Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E en Lix LM. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health and quality of life outcomes 2019 Jun; 17. DOI: 10.1186/S12955-019-1181-2. Available from: https://pubmed.ncbi.nlm.nih.gov/31221151/

64. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B en Tabona O. A survey on missing data in machine learning. Journal of Big Data 2021 8:1 2021 Oct; 8:1–37. DOI: 10.1186/S40537-021-00516-9

65. Hastie T, Tibshirani R en Friedman J. The Elements of Statistical Learning. 2de ed. Springer Series in Statistics. New York: Springer New York, 2009. DOI: 10.1007/978-0-387-84858-7

66. Kornbrot D. Point Biserial Correlation. Wiley StatsRef: Statistics Reference Online 2014 Apr. DOI: 10.1002/9781118445112.STAT06227

67. Hauke J en Kossowski T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae 2011 Jun; 30:87–93. DOI: 10.2478/V10117-011-0021-1

68. Schmidt AF en Finan C. Linear regression and the normality assumption. Journal of Clinical Epidemiology 2018 Jun; 98:146–51. DOI: 10.1016/J.JCLINEPI.2017.12.006

69. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW en Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. Journal of Clinical Epidemiology 2016 Mar; 71:76–85. DOI: 10.1016/J.JCLINEPI.2015.10.002

70. Akoglu H. User's guide to correlation coefficients. Turkish Journal of Emergency Medicine 2018 Sep; 18:91–3. DOI: 10.1016/J.TJEM.2018.08.001

71. M R Senaviratna NA, J A Cooray TM en Alberto Ferreira MM. Diagnosing Multicollinearity of Logistic Regression Model. Asian Journal of Probability and Statistics 2019 Oct; 5:1–9. DOI: 10.9734/AJPAS/2019/V5I230132. Available from: https://journalajpas.com/index.php/AJPAS/article/view/96

72. Chandrashekar G en Sahin F. A survey on feature selection methods. Computers & Electrical Engineering 2014 Jan; 40:16–28. DOI: 10.1016/J.COMPELECENG.2013.11.024

73. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, Vanderplas J, Cournapeau D, Pedregosa F, Varoquaux G, Gramfort A, Thirion B, Grisel O, Dubourg V, Passos A, Brucher M, Perrot M en Duchesnay É. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12:2825–30

74. Shipe ME, Deppen SA, Farjah F en Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. Journal of Thoracic Disease 2019; 11:574–84. DOI: 10.21037/JTD.2019.01.25

75. Schober P en Vetter TR. Logistic Regression in Medical Research. Anesthesia & Analgesia 2021 Feb; 132:365–6. DOI: 10.1213/ANE.0000000000005247

76. Banerjee M, Reynolds E, Andersson HB en Nallamothu BK. Tree-Based Analysis: A Practical Approach to Create Clinical Decision Making Tools. Circulation. Cardiovascular quality and outcomes 2019 May; 12. DOI: 10.1161/CIRCOUTCOMES.118.004879

77. Bouter L, Dongen M van, Zielhuis G en Zeegers M. Leerboek epidemiologie. 7de ed. Houten: Bohn Stafleu van Loghum, 2016 :189–220. DOI: 10.1007/978-90-368-0562-9

78. Flach PA en Kull M. Precision-Recall-Gain Curves: PR Analysis Done Right. Advances in Neural Information Processing Systems 2015; 28

79. Koko KR, McCauley BD, Gaughan JP, Fromer MW, Nolan RS, Hagaman AL, Brown SA en Hazelton JP. Spectral analysis of heart rate variability predicts mortality and instability from vascular injury. Journal of Surgical Research 2018 Apr; 224:64–71. DOI: 10.1016/J.JSS.2017.11.029

80. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995. Available from: http://robotics.stanford.edu/~ronnyk

81. Ley C, Klein O, Bernard P en Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology 2013 Jul; 49:764–6. DOI: 10.1016/J.JESP.2013.03.013

A Metadata MIMIC-III

Hier volgt een beschrijving van de inhoud van de MIMIC-III data na voorselectie. Tevens is een kolom opgenomen met 'Plausibiliteit' die beschrijft of er volgens sectie 4.2.5 een robuuste Z-score of een bepaald fysiologisch bereik wordt toegepast om uitschieters te filteren. Twee parameters, Sodium (mEq/dL) en EGFR, bevatten door een fout geen inhoud en werden dus niet gebruikt voor dit onderzoek. Deze zijn gelabeld als 'leeg' onder de kolom 'Plausibiliteit'.

Tabel 8: Overzicht van MIMIC-III data na voorselectie

Database Onderdeel	Data type	Eenheid	Plausibiliteit	Beschrijving
general				
subject-id	num	-	-	Unique identifier which specifies an individual patient. Links all tables back to icu-stays table.
hadm-id	num	-	-	Unique identifier which represents a single patient's admission to the hospital. Links all tables back to icu-stays table.
icustay-id	num	-	-	Unique identifier which represents a patient's ICU stay.
charttime-rel	deltatime	days hh:mm:ss	-	Time of measurement/sampling relative to icu-intime
starttime-rel	deltatime	days hh:mm:ss	-	Start time of medication/fluid relative to icu-intime
endtime-rel	deltatime	days hh:mm:ss	-	End time of medication/fluid relative to icu-intime
icu-stays				
demographics				
gender	str	-	-	The genotypical sex of the patient
height	num	cm	100-250	Height of the patient at the start of its icu stay
weight	num	kg	30-300	Weight of the patient at the start of its icu stay
age	num	years	-	Age of a patient at the start of its icu stay

Gaat door op volgende pagina

deathtime	deltatime	days hh:mm:ss	-	Time of death relative to icu-intime
icu-death	bin	-	-	Death within current icu stay
hosp-death	bin	-	-	Death within current hosp stay
dbsource	str	-	-	Database source (CareVue or MetaVision)
data-start	datetime	dd-mm-yyyy hh:mm:ss	-	timepoint from which datapoints are included in the dataset, either icu-intime - 12 hours if first icu stay in 24 hours, or halfway between previous icu endtime and icu-intime if there was a previous stay in ¡24 hours
data-end	datetime	dd-mm-yyyy hh:mm:ss	-	timepoint until which datapoints are included in the dataset, either icu-intime + 12 hours if last icu stay in 24 hours, or halfway between next icu-intime and icu endtime if there was a next stay in ¡24 hours
icu-intime	datetime	dd-mm-yyyy hh:mm:ss	-	Starttime of the icu stay
icu-los	deltatime	days hh:mm:ss	-	Length of stay of current icu stay
icu-stayorder	num	-	-	# icu stay of the current icu stay within one hospital stay
first-careunit	str	-	-	First intensive care unit type within current icu stay
last-careunit	str	-	-	Last intensive care unit type within current icu stay
hosp-intime	deltatime	days hh:mm:ss	-	Time of hospital admission relative to icu-intime
hosp-outtime	deltatime	days hh:mm:ss	-	Time of hospital release relative to icu-intime

Gaat door op volgende pagina

hosp-los	deltatime	days hh:mm:ss	-	Length of stay of current hospital stay
hosp-stayorder	num	-	-	# hospital stay for current patient
admit-type	str	-	-	Hospital admission type
admit-loc	str	-	-	Hospital admission location
disch-loc	str	-	-	Discharge location after hospital stay
ed-intime	deltatime	days hh:mm:ss	-	Time of emergency department admission relative to icu-intime
ed-outtime	deltatime	days hh:mm:ss	-	Time of emergency department release relative to icu-intime
diagnosis	str	-	-	Diagnosis upon admission to hospital
dnr	bin	-	-	Do not resuscitate code status
comorbidities				
congestive-heart-failure	bin	-	-	heart muscle not pumping optimally
cardiac-arrhythmias	bin	-	-	irregular heartbeat
valvular-disease	bin	-	-	damage within any of the heart valves
pulmonary-circulation	bin	-	-	any complication with the pulmonary circulation system
peripheral-vascular	bin	-	-	any complication with the peripheral vascular system
hypertension	bin	-	-	high blood pressure
paralysis	bin	-	-	loss of the ability to move
other-neurological	bin	-	-	any other neurological complication
chronic-pulmonary	bin	-	-	chronic pulmonary (breathing) conditions
diabetes-uncomplicated	bin	-	-	Diabetis without end organ damage

Gaat door op volgende pagina

diabetes-complicated	bin	-	-	Complicated diabetes is defined as diabetes associated with end organ damage such as peripheral neuropathy, nephropathy and/or PAD
hypothyroidism	bin	-	-	Unveractive thyroid gland, resulting in tiredness, weight gain, depression
renal-failure	bin	-	-	Kidney dysfunction
liver-disease	bin	-	-	Liver dysfunction
peptic-ulcer	bin	-	-	Stomach, esophagus or small intestine ulcer
aids	bin	-	-	Acquired Immunodeficiency Syndrome
lymphoma	bin	-	-	Lymphatic system cancer
metastatic-cancer	bin	-	-	Cancer spreading from the original site to a distant site in the body
solid-tumor	bin	-	-	Non-metastatic tumor
rheumatoid-arthritis	bin	-	-	autoimmune disease causing pain, stiffness, swelling in the joints
coagulopathy	bin	-	-	impaired blood clotting
obesity	bin	-	-	overweight
weight-loss	bin	-	-	excessive weight loss
fluid-electrolyte	bin	-	-	electrolyte imbalance/mineral imbalance
blood-loss-anemia	bin	-	-	Excessive blood loss
deficiency-anemias	bin	-	-	Anemia due to any type of blood disorder
alcohol-abuse	bin	-	-	Excessive alcohol use
drug-abuse	bin	-	-	Excessive drug use

Gaat door op volgende pagina

psychoses	bin	-	-	Psychotic disorders, losing contact with reality, hallucinations
depression	bin	-	-	Mental disorder, persistant sadness or lack of interest/pleasure/joy
elixhauser-sid30	num	-	-	An integer comorbidity risk score relating comorbid burden to mortality
inputevents-fluids				
label	str	-	-	crystalloids, colloids or packed rbcs
fluid-rate	num	mL/h	-	the average rate of fluid administration during the given time period
fluid-amount	num	mL	-	amount of fluid administered in the given time period
duration-hours	num	hours	-	duration of the given rate, timeframe in which amount is administered
inputevents-medication				
label	str	-	-	dobutamine, dopamine, epinephrine, norepinephrine, phenylephrine, vasopressin, or milrinone
vaso-rate	num	mcg/(kg*min), (U/min for vasopressin)	-	the average rate of medication administration during the given time period
vaso-amount	num	mg, (U for vasopressin)	-	amount of medication administered in the given time period

Gaat door op volgende pagina

duration-hours	num	hours	-	duration of the given rate, time-frame in which amount is administered
labs-blood-gases				
specimen	str	-	-	ART, VEN, MIX, CENTRAL VENOUS (arterial blood, venous blood, mixed venous blood, central venous blood)
blood gas				
Temperature (C)	num	C	30-45	Blood temperature
pH	num	-	6-11	Blood acidity
pCO2 (mmHg)	num	mmHg	Z-score	Partial pressure of carbon dioxide in blood
Base Excess (mEq/L)	num	mEq/L	-55-+55	Defines metabolic alkalosis
pO2 (mmHg)	num	mmHg	Z-score	Partial pressure of oxygen in blood
FiO2 (%)	num	%	21-100	Fraction of inspired oxygen
SO2 (%)	num	%	20-100	Oxygen saturation
Carboxyhemoglobin (%)	num	%	0-10	level of hemoglobin bound to carbon monoxide
Methemoglobin (%)	num	%	0-10	level of oxidized hemoglobin
whole blood				
Hemoglobin (g/dL)	num	g/dL	3-21	Oxygen carrier
Glucose (mg/dL)	num	mg/dL	Z-score	Blood sugar
Potassium (mEq/L)	num	mEq/L	0-20	Potassium blood level
Sodium (mEq/L)	num	mEq/L	100-180	Sodium blood level
Chloride (mEq/L)	num	mEq/L	65-150	Chloride blood level
Free Calcium (mmol/L)	num	mmol/L	0-11	Ionized calcium not attached to proteins

Gaat door op volgende pagina

Lactate (mmol/L)	num	mmol/L	Z-score	Measures amount of lactic acid produced by muscles and rbcs
labs-hematology-chemistry				
hematology				
Hemoglobin (g/dL)	num	g/dL	Z-score	Oxygen carrier
Hematrocrit (%)	num	%	Z-score	Percentage of red blood cells
Platelets (K/uL)	num	K/uL	Z-score	Thrombocytes that help blood clotting
White Bloodcell Count (K/uL)	num	K/uL	Z-score	Leukocyte test
Neutrophils (%)	num	%	Z-score	Most abundant granulocyte
Lymphocytes (%)	num	%	Z-score	B and T cells, antibody production and targeted cell killing, adaptive immune system
Monocytes (%)	num	%	Z-score	innate immune system
Eosinophils (%)	num	%	Z-score	Combating multicellular parasites
Basophils (%)	num	%	Z-score	Immune surveillance, similar cells to mast cells
chemistry				
Sodium (mEq/dL)	num	mEq/L	leeg	Sodium blood level
Potassium (mEq/dL)	num	mEq/L	Z-score	Potassium blood level
Chloride (mEq/dL)	num	mEq/L	Z-score	Chloride blood level
Calcium (mg/dL)	num	mg/dL	Z-score	Calcium blood level
Phosphate (mg/dL)	num	mg/dL	Z-score	Phosphate blood level
Magnesium (mg/dL)	num	mg/dL	Z-score	Magnesium blood level
BUN (mg/dL)	num	mg/dL	Z-score	Blood urea nitrogen, indicates kidney urea clearance, urea is a protein breakdown product

Gaat door op volgende pagina

Creatinine (mg/dL)	num	mg/dL	Z-score	Energy production chemical leftover compound, indicates kidney clearance
Bilirubin Total (mg/dL)	num	mg/dL	Z-score	RBCs breakdown product, total conjugated and non-conjugated bilirubin
Bilirubin Direct (mg/dL)	num	mg/dL	Z-score	RBCs breakdown product, conjugated bilirubin, attached to glucoronic acid
Alkaline Phosphate (IU/L)	num	IU/L	Z-score	Liver damage or bone disorder marker
GGT (IU/L)	num	IU/L	Z-score	Gamma Glutamyltransferase, liver or bile duct disease marker
ASAT (IU/L)	num	IU/L	Z-score	Aspartate aminotransferase, cardiac damage marker
ALAT (IU/L)	num	IU/L	Z-score	Alanine aminotransferase, liver damage marker
LD (IU/L)	num	IU/L	Z-score	Lactate dehydrogenase, tissue damage marker
CK (IU/L)	num	IU/L	Z-score	Creatine Kinase, skeletal muscle breakdown product
Cholesterol (mg/dL)	num	mg/dL	Z-score	Blood lipids
Triglyceride (mg/dL)	num	mg/dL	Z-score	Blood lipids
Glucose (mg/dL)	num	mg/dL	Z-score	Blood sugar
CRP (mg/dL)	num	mg/dL	Z-score	C-reactive protein, inflammation marker
Albumin (g/dL)	num	g/dL	Z-score	Carrier protein
Lipase (IU/L)	num	IU/L	Z-score	Pancreatitis marker
EGFR	num	-	leeg	Estimated glomerular filtration rate, indicates kidney clearance

Gaat door op volgende pagina

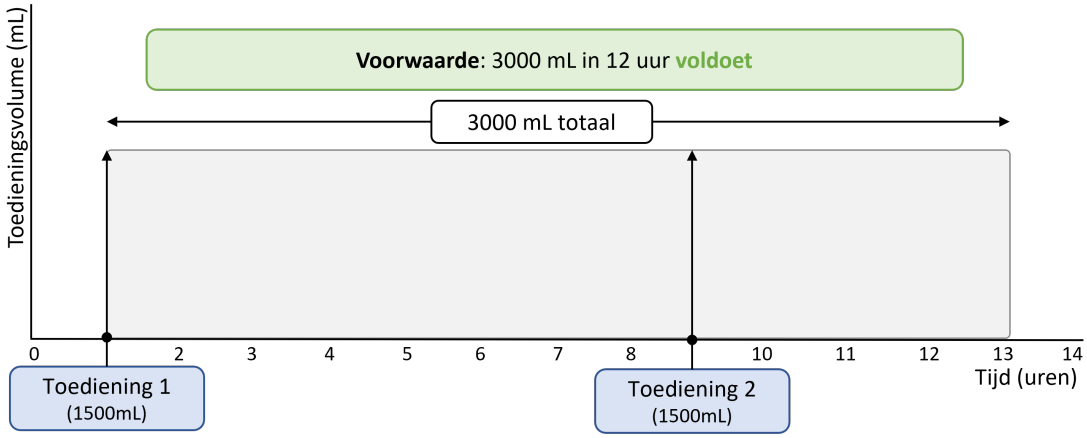
CK-MB (ng/mL)	num	ng/mL	Z-score	Creatine Kinase-MB, heart muscle breakdown product
Troponin-T (ng/mL)	num	ng/mL	Z-score	Cardiac damage marker
Bicarbonate (mEq/dL)	num	mEq/L	Z-score	Blood base
Anion Gap (mEq/dL)	num	mEq/L	Z-score	Blood charge
vital-signs				
Heart Rate (bpm)	num	bpm	20-220	Amount of heartbeats per minute
Systolic Arterial Blood Pressure (mmHg)	num	mmHg	20-250	Invasive blood pressure in the arteries when the heart contracts
Systolic NI Blood Pressure (mmHg)	num	mmHg	20-250	Non-invasive and manual blood pressure when the heart contracts
Diastolic Arterial Blood Pressure (mmHg)	num	mmHg	10-250	Invasive blood pressure in the arteries when the heart rests
Diastolic NI Blood Pressure (mmHg)	num	mmHg	10-250	Non-invasive and manual blood pressure when the heart rests
Mean Arterial Blood Pressure (mmHg)	num	mmHg	10-350	Invasive average blood pressure in the arteries
Mean NI Blood Pressure (mmHg)	num	mmHg	10-350	Non-invasive and manual average blood pressure
Respiratory Rate (resp/min)	num	resp/min	6-120	Amount of respirations per minute
Temperature (C)	num	C	30-46	Body temperature, can be Axillary or Rectal
Central Venous Pressure (mmHg)	num	mmHg	0-50	Blood pressure in the vena cava
Mean Airway Pressure (cmH2O)	num	cmH2O	2-45	Pmean, refers to the mean pressure applied during positive-pressure mechanical ventilation

Gaat door op volgende pagina

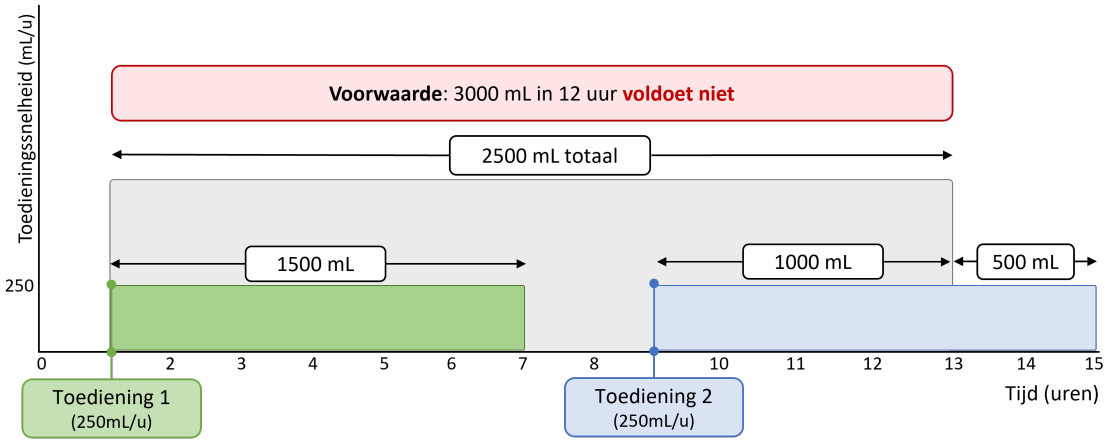
Peak Inspiratory Pressure (cmH2O)	num	cmH2O	5-90	PIP, highest pressure measured during the respiratory cycle. Can be set for patients receiving mechanical ventilation.
Postive End-Expiratory Pressure (cmH2O)	num	cmH2O	5-50	PEEP, the pressure in the lungs above atmospheric pressure that exists at the end of expiration. Can be set for patients receiving mechanical ventilation.
Oxygen Saturation (%)	num	%	70-100	Measure of how much hemoglobin is bound to oxygen compared to unbound hemoglobin, measured using pulse oxymetry
Fingerstick Glucose (mg/dL)	num	mg/dL	Z-score	Blood sugar level measured using a finger-stick device

B Labeling van hemodynamisch instabiele IC-patiënten

Het labelen van de hemodynamische instabiele patiënten is een zeer belangrijke stap in de voorbewerking van de data. Het bepaalt wat het model gaat voorspellen en bepaalt dus de volledige validiteit van het ontwikkelde machine learning model. Het model kan een goede performance hebben, maar als het niet de juiste uitkomst voorspelt omdat de uitkomstmaat niet goed is gelabeld, zal men niets aan het model hebben. Er werden verschillende methoden gevonden om de HI-patiënten zo accuraat mogelijk te labelen op basis van de door Rahman et al. opgestelde interventie criteria [21] elk met hun eigen concessies en voordelen.
Omdat dit onderzoek voortborduurt op eerdere onderzoeken naar de HDI in opdracht van TU/e en CZE, werd eerst bekeken welke methode reeds werd gehanteerd bij de labeling van HI. Hier werd uit geconcludeerd dat HI foutief werd gelabeld in deze onderzoeken. De labeling van patiënten aan de hand van vochttoediening en bloedtransfusie criteria werd voorheen namelijk gedaan door slechts te kijken naar de starttijden van toedieningen binnen het gedefinieerde tijdvenster en de totale hoeveelheid van de toediening. Toedieningen verschillen echter ook veel van elkaar in de infuussnelheid (ofwel toedieningssnelheid) en de duur van toediening. Het probleem dat hieruit voortkomt is dat patiënten in bepaalde gevallen als instabiel werden gelabeld, terwijl zij dit niet waren en andersom.
Als voorbeeld wordt het criterium van 3000 mL volumeresuscitatie (kristalloid of colloïd) in twaalf uur genomen (op basis van Rahman et al.). In figuur 6 is te zien dat er twee vochttoedieningen van elk 1500 mL totaal worden gestart binnen een periode van twaalf uur. Voorgaande onderzoeken sommeerde deze totale vochttoedieningen en concludeerden dat deze patiënt hemodynamisch instabiel was, omdat aan de voorwaarde van 3000 mL binnen twaalf uur werd voldaan. In figuur 7 wordt de werkelijke situatie van een vochttoediening weergegeven. Deze verkrijgt men als ook de infuussnelheid en toedieningsduur worden meegenomen in het model. Beide toedieningen hebben in dit voorbeeld een stroomsnelheid van 250 mL/u. Dit betekent dat tweede toediening niet volledig binnen het tijdvenster van twaalf uur valt en dus ook niet in de volume sommatie mag worden meegenomen. Wanneer dan het totale toedieningsvolume wordt berekend, resulteert dit in een totaal volume van 2500 mL en dat de patiënt dus geen hemodynamisch instabiel label krijgt. Concluderend zorgt het niet hanteren van infuussnelheden voor vals positieve gelabelde HI-patiënten.

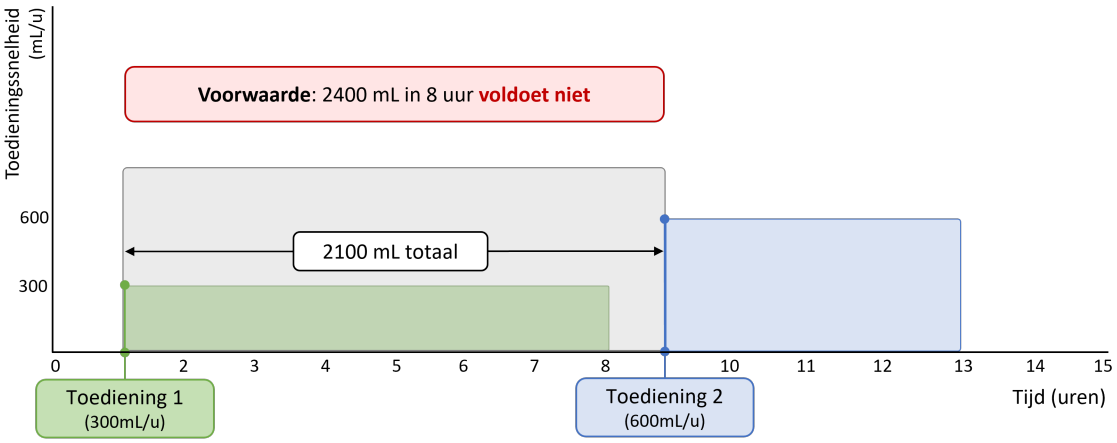


Figuur 6: Foutieve methode van HI labeling in voorgaande onderzoeken met impuls van toedieningen

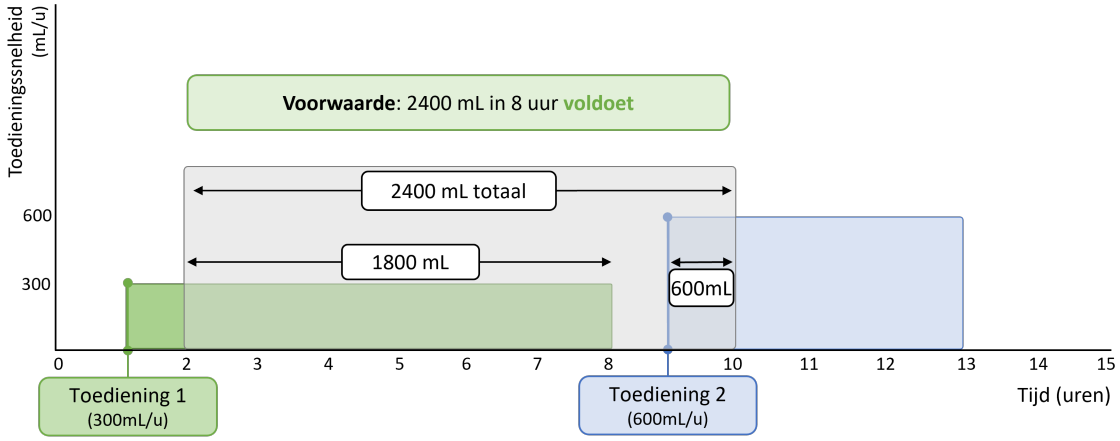


Figuur 7: Correcte methode van HI labeling met infuussnelheid (mL/u) en toedieningsduur

Daarnaast werd er in voorgaande onderzoeken de sommatie van toegediende vloeistofvolumes bepaald door het tijdvenster te itereren over de starttijd van de toedieningen en niet over een discrete bemonsteringstijd. Dit zorgt voor een onderbemonstering van de vloeistoftoediening en zorgt dus voor het missen van instabiele patiënten. Zoals te zien is in figuur 8 wordt deze patiënt niet als hemodynamisch instabiel gelabeld, omdat de sommatie van de vochttoedieningen niet boven de voorwaarde van 2400 mL binnen acht uur komt. Om deze onderbemonstering op te lossen werd ervoor gekozen om de vloeistof en PRBC's toedieningssnelheden over een discrete bemonsteringstijd van één minuut te bepalen en vervolgens het (criterium) tijdvenster over deze bemonsterde tijd te schuiven in plaats van over de starttijden van toedieningen. Voor iedere patiënt werd dus per minuut bekeken hoeveel mL/u toediening er plaatsvond, hierbij werd het dus ook mogelijk om toedieningen die tegelijkertijd plaatsvonden te sommeren. In figuur 9 is te zien wat het voordeel is van het bemonsteren van de toedieningen over een discreet tijdsinterval. In figuur 8 werd de patiënt niet als HI gelabeld, maar wanneer het tijdvenster een uur vooruit wordt geschoven (zie figuur 9) voldoet de patiënt wel aan het criterium en wordt dus als HI gelabeld. In werkelijkheid wordt het tijdvenster dus iedere keer een minuut opgeschoven en gecontroleerd of er aan het criterium wordt voldaan. Vervolgens wordt het label HI aan het begin van het tijdvenster gezet.

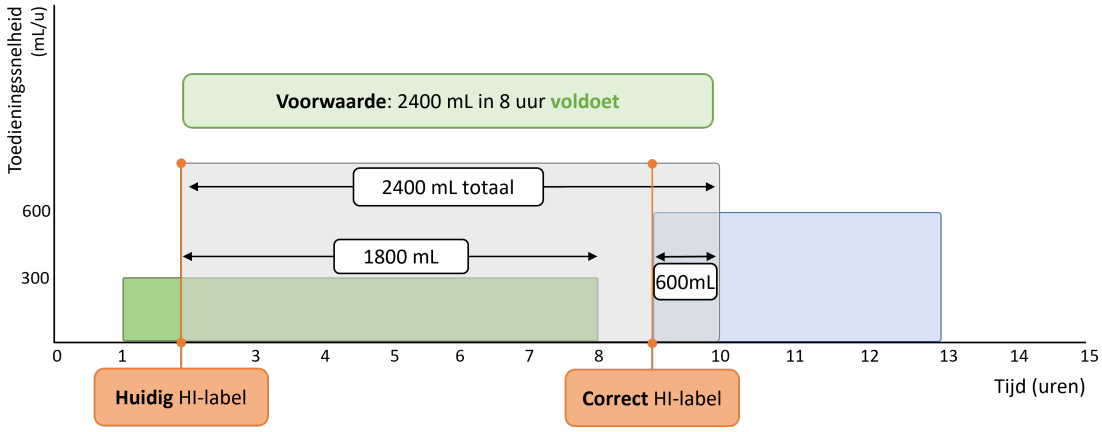


Figuur 8: Tijdvenster iteratie over de starttijden van toedieningen zorgt voor onderbemonstering



Figuur 9: Tijdvenster iteratie over een discreet tijdinterval zorgt voor een nauwkeurigere manier van HI-labeling

Idealiter zou het HI-label niet aan het begin van het tijdvenster worden gezet, maar op het moment dat de arts een significante interventie inzet binnen datzelfde tijdvenster. Ter verduidelijking, het tijdvenster dat voldoet aan het criterium blijft dus hetzelfde, echter wordt de plaatsing van het HI-label binnen dat venster dus accurater. Volgens de definitie van Rahman et al. wordt hemodynamische instabiliteit namelijk gedefinieerd als een periode waarin een significante hemodynamische interventie plaatsvindt. Het meest wenselijk zou dan ook zijn om het label precies te plaatsen waar deze significante interventie start. Voor het voorbeeld in figuur 9 zou kunnen worden gesteld dat er pas vanaf de tweede toediening een significante ophoging start en dat de eerste toediening een soortement onderhoudsdosering is, zie figuur 10. De eerste toediening draagt dus wel bij aan het voldoen van het criterium van 2400 mL in acht uur, maar is niet de significante interventie en dus ook niet het moment dat de patiënt daadwerkelijk hemodynamisch instabiel werd. Een aanpak zou kunnen zijn dat de afgeleide van de infuussnelheid wordt berekend en vervolgens het HI-label te plaatsen waar deze afgeleide een significante elevatie heeft binnen het tijdvenster (eventueel maximale elevatie). Deze methode blijft ook arbitrair, maar is wel een accuratere manier van HI-labeling.



Figuur 10: Huidige en accuratere labelingsmethode van hemodynamische instabiliteit

C Plausibiliteit

C.1 Plausibiliteitsfilter

Er zijn verschillende manieren om de uitschieters te herkennen, waarna ze kunnen worden verwijderd. Zo wordt data bij *trimming* verwijderd die buiten een gespecificeerd percentiel valt, maar resulteert in het verwijderen van data ongeacht of het uitschieters zijn. Uitschieters kunnen ook herkend worden met een robuuste Z-score berekening, wat gebruik maakt van de mediaan en de standaarddeviatie rondom de mediaan (zie de vergelijking hieronder), in plaats van het gemiddelde en de standaarddeviatie rondom het gemiddelde van een normale Z-score. Deze verschillen zorgen ervoor dat de robuuste Z-score veel minder wordt beïnvloedt door enorme uitschieters [81]. De robuuste Z-score wordt ook nog geschaald, zodat de niet-uitschieters ongeveer dezelfde Z-score krijgen als ze hadden gekregen zonder uitschieters in de data. Al met al is beste methode is generaliseerbaar op alle parameters en zorgt niet voor onnodig dataverlies.

$$Z_{\text{robuust}} = 0,7418\frac{x_i - \text{mediaan}}{\text{MAD}}$$

$$\text{MAD} = \text{median}(|x_i - \text{median}|)$$

Als van een parameter een duidelijk bereik te definiëren is, zoals bij parameters met procentuele eenheden, worden waarden buiten dat bereik verwijderd, ofwel 'laten vallen, *dropping*'. Dat zal altijd de meest eenvoudige manier blijven om zeker te weten dat er geen uitschieters meer aanwezig zijn. Echter is het niet altijd mogelijk, of erg moeilijk, om voor elke parameter een dergelijk mogelijk bereik te definiëren. Ook is het niet makkelijk reproduceerbaar, omdat andere instellingen of laboratoria dezelfde parameters op een andere manier of in andere eenheden meten.
Ter illustratie wordt een parameter geanalyseerd waarvan geen duidelijke afkapwaarden bestaan. Er wordt wel een fysiologisch bereik opgesteld, maar dat blijft erg arbitrair. Eerst wordt de parameter geanalyseerd zonder filtering, dan na *dropping*, na een robuuste Z-score en na *dopping* in combinatie met een normale Z-score. Trimming wordt niet toegepast wegens eerder genoemde redenen. Daarna wordt er een conclusie getrokken en zal beschreven worden welke methode zal worden toegepast. Een parameter waarbij moeilijker duidelijke afkapwaarden kan worden gedefinieerd is bijvoorbeeld de C-reactief eiwit (*C-reactive protein*, CRP), gemeten in mg/dL. Extreem hoge waarden van CRP, >50 mg/dL, wijzen op een hele heftige infectie, maar zijn dus wel (patho)fysiologisch mogelijk [49]. Ter illustratie wordt hier voor CRP een mogelijk bereik tussen 0 en 150 mg/dL gedefinieerd.

Tabel 10: Drie methoden voor het filteren van uitschieters

Analyse	Aantal	Gem. ±SD [min; max]
Ongefilterd	588.000	85,7 ±80,1 [0,12; 299,0]
Rob. Z-score	560.000	76,2 ±69,5 [0,12; 250,3]
Dropping	460.000	50,3 ±44,1 [0,12; 150,0]

Het valt al op dat CRP geen extreme uitschieters heeft maar wel tot hoog oploopt. De robuuste Z-score heeft best veel data gekost maar levert een beter bereik op tussen 0,12 en 250,3 mg/dL. In conclusie, als er geen duidelijk plausibel bereik kan worden gedefinieerd werkt de robuuste Z-score erg goed. Toch blijft de beste optie om, als het (goed) mogelijk is, afkapwaarden van een plausibel bereik te definiëren. In de tabel in appendix A is een overzicht gegeven van alle parameters waarin wordt gespecificeerd of er met een plausibel bereik, en zo ja welk bereik, of met een robuuste Z-score wordt gewerkt.

C.2 Labelling onbekende afnamelocaties

Binnen de dataset van bloedgasmetingen waren veel afnamelocaties onbekend, welke als nog zijn gelabeld als arterieel of veneus volgens een stappenplan dat hieronder is beschreven.

1. Als eerst is voor elke parameter een gemiddelde uitgerekend voor zowel arterieel als veneus label.
2. Er is een lijst gemaakt met de absolute verschillen tussen die gemiddelden, op volgorde van groot naar klein. Dit is dus een lijst waar op volgorde de parameters het meest verschillen tussen arterieel en veneus.
3. Per onbekende afnamelocatie werd in dezelfde rij in de dataset gekeken naar de parameters op volgorde van de lijst van 2.
4. Als de eerste parameter, met dus het grootste verschil tussen arterieel en veneus, gemeten is in dezelfde rij, wordt gekeken naar die gemeten waarde.
5. Als die waarde dichter bij het arteriële gemiddelde van die parameter zit, wordt het label overgeschreven naar arterieel, anders naar veneus.
6. Als die waarde niet gemeten is, wordt gekeken naar de volgende parameter in de lijst van 2. Dit is dus de parameter met het een na grootste verschil in gemiddelde tussen arterieel en veneus.
7. Hetzelfde geldt voor die waarde als bij 5 en 6, totdat alle labels zijn overgeschreven.

D Bemonsteringsinterval

Het bemonsteren van de data, dus van hematologische chemie, bloedgassen en vitale kenmerken gebeurt binnen een bepaald venster vóór de predictietijd. Dit tijdvenster is anders voor de verschillende datasets en is gebaseerd op de duur waarvoor een arts de gemeten parameter nog vertrouwt. Dit was slecht vindbaar in literatuur, dus deze vensters zijn afgeleid uit de data door te kijken naar de gemiddelde tijdregistraties tussen niet-continu gemeten parameters voor elke dataset. Deze gemiddelden staan hieronder in de tabel uitgewerkt. Uiteindelijk is het 75e percentiel als richtlijn voor de lengte van het tijdvenster genomen.

Tabel 11: Vitale kenmerken

Datset	Aantal	Beschrijving (gem. [25%; 75%])
Vitale kenmerken	n=14.984	00:45:00 [00:39:10; 00:51:00]
Chemie	n=14.819	09:34:39 [06:28:33; 11:52:31]
Bloedgassen	n=13.427	06:17:12 [01:49:53; 06:56:01]

E Resultaten verkennende gegevensanalyse

E.1 Resultaten verkennende gegevensanalyse

Onderstaande tabel bevat de resultaten van het parameter selectieproces tijdens de verkennende gegevensanalyse. De eerste kolom bevat de gemeten (klinische of demografische) parameter. De tweede kolom noemt het afgeleide aspect van deze parameter. Vitale kenmerken bevatten: last (recentste niet-NAN waarde in het venster), mean (gemiddelde waarde in venster), std (standaarddeviatie van waarden in venster) en lls (kleinste-kwadratenmethode regressie lijn van waarden in venster). Labwaarden bevatten: last (recentste waarde in venster) en gemetenheid (binaire indicator van aanwezigheid van een waarde in venster (ook NaN), ofwel de indicator óf deze parameter überhaupt is gemeten). Achter deze aspecten staat de bijbehorende NaN-ratio (x100) (percentage missende data). Hierachter de bijbehorende significantie (p-waarde) van correlatie met het label. In de laatste kolom, Drop, wordt benoemd of de parameter wel of niet is geëxcludeerd voor ML-model training: *nee*: parameter is meegenomen, *ja*: parameter is geëxcludeerd vanwege NaN-ratio of significantie,*ja (mc)* = parameter is geëxcludeerd vanwege multicollineariteit (mc).

Parameter	**Aspect**	**NaN-ratio**	**Significantie**	**Drop**
OI		0.967	0.000	nee
OSI		0.968	0.000	ja (mc)
DSI		0.347	0.000	nee
SI		0.347	0.000	nee
BMI		0.472	0.000	nee
MSI		0.349	0.000	ja (mc)
RPP		0.347	0.000	nee
CK-MB/CK		0.942	0.099	nee
age		0.000	0.000	nee
gender		0.000	0.341	ja
weight		0.143	0.086	ja
height		0.470	0.024	nee
Elixhauser index		0.000	0.000	nee
cardiac		0.000	0.000	nee
Heart Rate	last	0.015	0.000	ja (mc)
	mean	0.015	0.000	nee
	std	0.745	0.007	ja
	lls	0.745	0.155	ja
Systolic Arterial Blood Pressure	last	0.345	0.000	nee
	mean	0.345	0.000	ja (mc)
	std	0.784	0.001	ja
	lls	0.784	0.049	ja
Systolic NI Blood Pressure	last	0.598	0.000	nee
	mean	0.598	0.000	ja (mc)
	std	0.936	0.000	ja
	lls	0.936	0.324	ja
Mean NI Blood Pressure	last	0.601	0.000	ja

Gaat door op volgende pagina

	mean	0.601	0.000	ja
	std	0.937	0.000	ja
	lls	0.937	0.807	ja
Mean Arterial Blood Pressure	last	0.347	0.000	nee
	mean	0.347	0.000	ja (mc)
	std	0.788	0.030	ja
	lls	0.788	0.219	ja
Peak Inspiratory Pressure	last	0.864	0.000	ja
	mean	0.864	0.000	ja
	std	0.997	0.233	ja
	lls	0.997	0.319	ja
Diastolic NI Blood Pressure	last	0.598	0.000	nee
	mean	0.598	0.000	ja (mc)
	std	0.937	0.000	ja
	lls	0.937	0.323	ja
Diastolic Arterial Blood Pressure	last	0.345	0.000	ja (mc)
	mean	0.345	0.000	ja (mc)
	std	0.784	0.023	ja
	lls	0.784	0.147	ja
Central Venous Pressure	last	0.617	0.000	ja
	mean	0.617	0.000	ja
	std	0.860	0.071	ja
	lls	0.860	0.655	ja
Mean Airway Pressure	last	0.837	0.000	ja
	mean	0.837	0.000	ja
	std	0.996	0.055	ja
	lls	0.996	0.356	ja
Respiratory Rate	last	0.029	0.000	nee
	mean	0.029	0.000	nee
	std	0.725	0.684	ja
	lls	0.725	0.066	ja
Oxygen Saturation	last	0.041	0.000	nee
	mean	0.041	0.000	nee
	std	0.739	0.241	ja
	lls	0.739	0.792	ja
Positive End-Expiratory Pressure	last	0.821	0.000	ja
	mean	0.821	0.000	ja
	std	0.992	0.047	ja
	lls	0.992	0.348	ja
Fingerstick Glucose	last	0.860	0.379	ja
	mean	0.860	0.380	ja
	std	0.992	0.947	ja
	lls	0.992	0.341	ja

Gaat door op volgende pagina

Temperature	last	0.566	0.954	ja
	mean	0.566	0.581	ja
	std	0.911	0.962	ja
	lls	0.911	0.014	ja
Arterial lactate	last	0.869	0.000	ja
	gemetenheid		0.000	nee
Venous lactate	last	0.976	0.000	ja
	gemetenheid		0.000	nee
Arterial base excess	last	0.587	0.000	nee
	gemetenheid		0.000	ja (mc)
Venous base excess	last	0.947	0.000	ja
	gemetenheid		0.000	nee
Bicarbonate	last	0.513	0.000	ja (mc)
	gemetenheid		0.000	ja (mc)
Anion gap	last	0.533	0.000	nee
	gemetenheid		0.000	ja (mc)
Arterial pH	last	0.558	0.000	nee
	gemetenheid		0.000	ja (mc)
Venous pH	last	0.924	0.000	ja
	gemetenheid		0.000	ja (mc)
Arterial pO2	last	0.598	0.000	nee
	gemetenheid		0.000	ja (mc)
Venous pO2	last	0.547	0.000	nee
	gemetenheid		0.000	ja (mc)
Arterial pCO2	last	0.597	0.000	nee
	gemetenheid		0.000	nee
Venous pCO2	last	0.949	0.333	ja
	gemetenheid		0.000	ja (mc)
Glucose	last	0.535	0.000	nee
	gemetenheid		0.000	nee
Arterial glucose	last	0.808	0.000	ja
	gemetenheid		0.000	nee
Venous glucose	last	0.972	0.232	ja
	gemetenheid		0.587	ja
Potassium	last	0.487	0.000	nee
	gemetenheid		0.000	ja (mc)
Arterial potassium	last	0.812	0.052	ja
	gemetenheid		0.000	nee
Venous potassium	last	0.975	0.547	ja
	gemetenheid		0.178	ja
Chloride	last	0.496	0.000	nee
	gemetenheid		0.000	ja (mc)

Gaat door op volgende pagina

Arterial chloride	last	0.973	0.104	ja
	gemetenheid		0.000	nee
Venous chloride	last	0.998	0.828	ja
	gemetenheid		0.360	ja
Calcium	last	0.610	0.000	ja
	gemetenheid		0.000	nee
Arterial free calcium	last	0.731	0.085	ja
	gemetenheid		0.000	nee
Venous free calcium	last	0.959	0.234	ja
	gemetenheid		0.000	nee
Phosphate	last	0.608	0.000	ja
	gemetenheid		0.000	ja (mc)
Magnesium	last	0.523	0.282	ja
	gemetenheid		0.000	nee
Arterial sodium	last	0.951	0.420	ja
	gemetenheid		0.000	nee
Venous sodium	last	0.997	0.815	ja
	gemetenheid		0.446	ja
Hemoglobin	last	0.539	0.000	nee
	gemetenheid		0.000	nee
Arterial hemoglobin	last	0.937	0.003	ja
	gemetenheid		0.000	nee
Venous hemoglobin	last	0.990	0.030	ja
	gemetenheid		0.000	nee
Hematocrit	last	0.434	0.003	nee
	gemetenheid		0.000	nee
Platelets	last	0.529	0.000	nee
	gemetenheid		0.000	nee
Arterial methemoglobin	last	0.999	0.414	ja
	gemetenheid		0.716	ja
Arterial carboxyhemoglobin	last	0.999	0.030	ja
	gemetenheid		0.009	nee
Venous carboxyhemoglobin	last	0.999	0.818	ja
	gemetenheid		0.999	ja
Arterial temperature	last	0.874	0.934	ja
	gemetenheid		0.002	nee
Venous temperature	last	0.984	0.630	ja
	gemetenheid		0.000	nee
Arterial sO2	last	0.841	0.361	ja
	gemetenheid		0.000	nee
Venous sO2	last	0.939	0.000	ja
	gemetenheid		0.000	nee
CK-MB	last	0.931	0.000	ja

Gaat door op volgende pagina

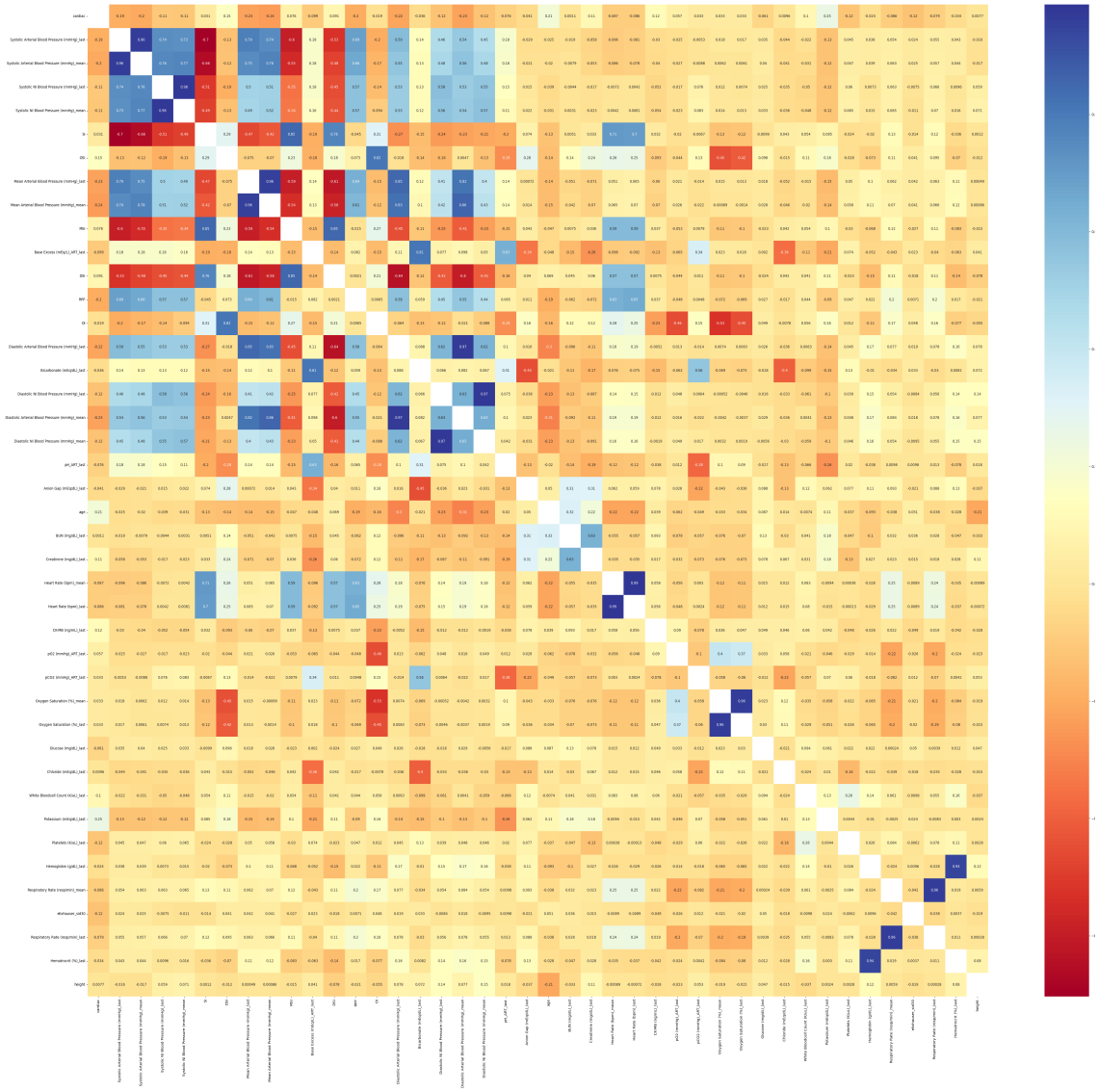
	gemetenheid		0.004	nee
CK	last	0.914	0.156	ja
	gemetenheid		0.060	ja
Arterial FiO2	last	0.884	0.000	ja
	gemetenheid		0.000	nee
Venous FiO2	last	0.988	0.000	ja
	gemetenheid		0.000	nee
BUN	last	0.527	0.000	nee
	gemetenheid		0.000	ja (mc)
Creatinine	last	0.551	0.000	nee
	gemetenheid		0.000	ja (mc)
White Bloodcell count	last	0.544	0.000	nee
	gemetenheid		0.000	nee
Eosinophils	last	0.955	0.001	ja
	gemetenheid		0.594	ja
Neutrophils	last	0.954	0.002	ja
	gemetenheid		0.594	ja
Lymphocytes	last	0.954	0.002	ja
	gemetenheid		0.594	ja
Basophils	last	0.957	0.025	ja
	gemetenheid		0.594	ja
Monocytes	last	0.954	0.061	ja
	gemetenheid		0.594	ja
Bilirubine direct	last	0.988	0.221	ja
	gemetenheid		0.210	ja
Bilirubine	last	0.988	0.089	ja
	gemetenheid		0.047	nee
Troponin-T	last	0.958	0.254	ja
	gemetenheid		0.000	nee
ASAT	last	0.923	0.486	ja
	gemetenheid		0.081	ja
ALAT	last	0.924	0.811	ja
	gemetenheid		0.076	ja
Albumine	last	0.928	0.000	ja
	gemetenheid		0.097	ja
Alkaline phosphate	last	0.909	0.198	ja
	gemetenheid		0.072	ja
Triglyceride	last	0.989	0.662	ja
	gemetenheid		0.341	ja
Lipase	last	0.967	0.766	ja
	gemetenheid		0.593	ja
Cholesterol	last	0.992	0.911	ja
	gemetenheid		0.803	ja

Gaat door op volgende pagina

GGT	last	0.999	0.992	ja
	gemetenheid		0.409	ja
LD	last	0.950	0.787	ja
	gemetenheid		0.488	ja
CRP	last	0.998	0.897	ja
	gemetenheid		0.516	ja

E.2 Heatmap (collineariteit continue variabelen)

Disclaimer: door in te zoomen op dit figuur (digitale versie) zijn alle gegevens te zien.



Figuur 11: Een heatmap met de collineariteit van alle continue variabelen.

F Resultaten hyperparameter optimalisatie

In deze appendix zullen eerst kort de initiële keuzes voor de waarden van hyperparameter worden uiteengezet voor de verschillende ML-modellen, zie sectie F.1. Hierna wordt voor elk ML-model en voor iedere subset beschreven: inputs (hyperparameters) van random search (RS), resultaten van RS, inputs van de gridsearch (GS) en resultaten van de GS. Voor de hyperparameter optimalisatie vóór de verkennende gegevensanalyse zie sectie F.2, ná de verkennende gegevensanalyse zie sectie F.3 en voor de cardiale patiëntgroep zie sectie F.4. De selectie van de parameters per subset is te zien in appendix E in de tabel E.1.

F.1 Keuzes voor hyperparameters

De initiële keuzes voor de belangrijkste waarden van hyperparameters voor de random search, specifiek voor ieder gebruikt ML-model.

- **Logistische regressie**
  - *C*: De inverse van de regularisatiesterkte. Hoe hoger dit getal hoe minder de regularisatiefactor wordt meegenomen in de verliesfunctie en dus hoe minder het gewicht van de kenmerken wordt geminimaliseerd. Normaal 1 en gekozen voor een logistisch bereik van 0,1 tot 10, door de grootte van onze dataset.
  - *penalty*: De gekozen regularisatie methode met de opties: L1, L2, elastisch (L1&L2) en geen regularisatie methode. Normaal L2 en gekozen voor alle mogelijke opties.
  - *solver*: Het type algoritme waarmee de logistische regressie methode wordt uitgevoerd, met de opties: 'lbfgs', 'liblinear', 'newton-cg', 'newton-cholesky', 'sag', 'saga'. De solvers werken niet voor elk type penalty. Gekozen om alles mee te nemen, omdat de computatie snelheid van logistische regressie dit toeliet. Schaling, ofwel normalisatie zijn voor 'sag' en 'saga' beide erg belangrijk. Normalisatie is in dit onderzoek toegepast en beide solvers kunnen dus worden gebruikt.
  - *max_iter*: Maximaal aantal iteraties dat de solver mag uitvoeren. Wanneer deze te laag wordt gekozen kan het betekenen dat convergentie niet wordt bereikt (een foutieve voorspelling). Normaal ligt deze waarde rond 100, maar er kan gekozen worden voor getallen tussen de 500 en 10.000, omdat de dataset heeft erg veel features heeft.
- **AdaBoost**
  - *n_estimators*: Het aantal beslisbomen dat wordt gecreëerd om de uitkomst te voorspellen. Een hogere waarde zorgt voor een complexer model en dus betere voorspellingen, maar ook zorgen voor overfitting. Normaalwaarden zijn tussen 50 en 200. Grotere datasets hebben meer variabelen en kunnen dus baat hebben bij een complexer model, waarvoor bijvoorbeeld 500 estimators gebruikt kan worden.
  - *learning_rate*: Het gewicht dat wordt gegeven aan de voorgaande beslisboom, waarop een nieuwe wordt gebaseerd. Een hogere learning rate geeft meer gewicht aan voorgaande beslisbomen, wat het model sneller laat convergeren tot de voorspelling. Er is dan wel kans op overfitting, omdat te veel waarde gehecht wordt aan de traindata en een minder complex model wordt gevormd dat beter kan presteren bij ongeziene data. De learning rate ligt typisch tussen 0,1 en 1, maar voor grote datasets is een complexer model nodig en eventueel dus nog een lagere learning rate.
  - *min_samples_leaf*: Het minimumaantal samples dat nodig is om een bladknoop (*leaf node*) te vormen. Een hogere waarde zorgt dat de boom met een hoger aantal samples kan eindigen in een terminale knoop, waarmee het model minder complex wordt. De waarde wordt bepaald door een bepaald percentage te nemen van het aantal samples. Hier zijn dat patiënten en wordt een percentage gegenereerd met: np.linspace(0.001, 0.01, 12).
  - *max_leaf_nodes*: Dit limiteert het aantal bladknopen van de uiteindelijke beslisboom. Voor een hogere waarde kan een complexer model worden gevormd dat eventueel nauwkeurigere voorspellingen kan doen maar wel gevoeliger is voor overfitting. Normale waarden hiervoor liggen tussen 10-100, maar kan voor grotere datasets oplopen.
  - *algorithm*: Nieuwe beslisbomen worden gegenereerd op basis van de fouten van de voorgaande beslisbomen. De gewichten die aan die fouten worden gegeven, waarmee de volgende beslisbomen rekening houden, worden bepaald aan de hand van deze algoritmen. Twee verschillende algoritmes zijn beschikbaar: 'SAMME.R' kan goed overweg met meerdere klassen, 'SAMME' maar met twee.
- **Gradient boosting**

- *n_estimators*
- *learning_rate*
- *min_samples_leaf*
- *max_leaf_nodes*
- *max_depth*: Bepaald de complexiteit (diepte) van de beslisbomen in het model. Een hogere waarde kan leiden tot een complexer model met nauwkeurigere voorspellingen maar heeft ook kans op overfitting. Typische waarden zitten tussen ongeveer tussen 1 en 100, waarbij grotere waarden vaak worden gebruikt bij grotere datasets.
- *max_features*: Geeft het aantal parameters aan dat wordt gebruikt voor het creëren van de beslisbomen. Een hogere waarde betekent dus een complexer model met meer kans op overfitting, maar kan wel nauwkeurigere resultaten geven. Opties zijn 'sqrt', de wortel van het aantal parameters, 'log2', het logaritme van het aantal parameter of 'None', wanneer geen maximum wordt opgegeven. Voor grote datasets wordt meestal wel een maximum gebruikt.
- *loss*: De verliesfunctie (*loss function*) geeft de verschillen tussen voorspelde en daadwerkelijke uitkomsten aan, die geminimaliseerd kunnen worden op twee manieren. 'log-loss' komt overeen met de verliesfunctie van logistische regressie, wat de log-likelihood van het model maximaliseert. De andere optie is 'exponential', wat beter geschikt is wanneer de verdelingen van de data een exponentieel karakter hebben. Over het algemeen wordt 'log-loss' gebruikt.

- **Random forest**
  - *n_estimators*
  - *min_samples_leaf*
  - *max_leaf_nodes*
  - *max_depth*
  - *max_features*
  - *criterion*: Dit zijn de methoden die een maat van onzuiverheid weergeven, waarop de splitsingen in een beslisboom worden gebaseerd. Hierbij wordt de onzuiverheid geminimaliseert op basis van de Gini-index ('gini') of entropie ('entropy'), die wiskundig erg op elkaar lijken en vaak dezelfde resultaten geven. Er is dus niet echt een richtlijn welke van de twee te gebruiken.

F.2 Optimalisatie vóór verkennende gegevensanalyse

Tabel 12: Hyperparameter optimalisatie Logistische Regressie (vóór verkennende gegevensanalyse)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
C	np.logspace(-1, 1, 12)	0.5336699	np.logspace(-1, 1, 12)	0.8111308
penalty	['l1', 'l2']	l2	['l1', 'l2']	l2
solver	['liblinear', 'newton-cg', 'lbfgs', 'sag', 'saga', None]	saga	['liblinear', 'newton-cg', 'lbfgs', 'sag', 'saga', None]	liblinear
max_iter	[500, 1000, 3000, 10000]	3000	[1000, 3000, 5000, 10000]	1000
AUROC trainset		0.819		0.819
AUROC testset		0.809		0.808

Tabel 13: Hyperparameter optimalisatie AdaBoost (vóór verkennende gegevensanalyse)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	200	[100, 200, 1000]	200
learning_rate	[0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 10]	0.25	[0.1, 0.25, 0.5, 1]	0.1
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.0083636	[0.0075455, 0.0083636, 0.0091818]	0.0091818
max_leaf_nodes	[2, 10, 100, None]	10	[10, 20]	20
algorithm	['SAMME', 'SAMME.R']	SAMME	SAMME	SAMME
AUROC trainset		0.896		0.898
AUROC testset		0.891		0.894

Tabel 14: Hyperparameter optimalisatie GradientBoost (vóór verkennende gegevensanalyse)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	1000	[500, 1000, 3000]	3000
learning_rate	[0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 10]	0.005	[0.001, 0.005, 0.01]	0.005
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.0034545	[0.0026364, 0.0034546, 0.0042727	0.0042727
max_leaf_nodes	[2, 10, 100, None]	100	[50, 100, 150]	50
max_depth	[1, 2, 4, 8, 16, 32, 64, 128]	16	[8, 16, 32]	16
max_features	['log2', 'sqrt', None]	log2	log2	log2
loss	['log_loss', 'exponential']	log_loss	log_loss	log_loss
AUROC trainset		0.899		0.902
AUROC testset		0.894		0.895

Tabel 15: Hyperparameter optimalisatie RandomForest (vóór verkennende gegevensanalyse)

Hyperparameter	**Input RS**	**Resultaat RS**	**Input GS**	**Resultaat GS**
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	200	[100, 200, 1000]	1000
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.001	[0.0001, 0.001, 0.0018182]	0.0001
max_leaf_nodes	[2, 10, 100, None]	None	[500, None]	None
max_depth	[1, 2, 4, 8, 16, 32, 64, 128]	128	[64, 128, 256]	128
max_features	['log2', 'sqrt', None]	log2	log2	log2
criterion	['gini', 'entropy', 'log_loss']	entropy	entropy	entropy
AUROC trainset		0.889		0.894
AUROC testset		0.880		0.889

F.3 Optimalisatie na verkennende gegevensanalyse

Tabel 16: Hyperparameter optimalisatie Logistische Regressie (na verkennende gegevensanalyse)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
C	np.logspace(-1, 1, 12)	1.8738174	[1.2328467, 1.8738174, 2.8480359]	1.8738174
penalty	['l1', 'l2']	l1	l1	l1
solver	['liblinear', 'newton-cg', 'lbfgs', 'sag', 'saga', None]	saga	['liblinear', 'saga']	saga
max_iter	[500, 1000, 3000, 10000]	3000	[2000, 3000, 5000]	2000
AUROC trainset		0.894		0.810
AUROC testset		0.806		0.806

Tabel 17: Hyperparameter optimalisatie AdaBoost (na verkennende gegevensanalyse)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	200	[100, 200, 1000]	200
learning_rate	[0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 10]	0.25	[0.1, 0.25, 0.5, 1]	0.5
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.0083636	[0.0075455, 0.0083636, 0.0091818]	0.0091818
max_leaf_nodes	[2, 10, 100, None]	10	[10, 20]	10
algorithm	['SAMME', 'SAMME.R']	SAMME	SAMME	SAMME
AUROC trainset		0.846		0.848
AUROC testset		0.844		0.840

Tabel 18: Hyperparameter optimalisatie GradientBoost (na verkennende gegevensanalyse)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	1000	[500, 1000, 3000]	1000
learning_rate	[0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 10]	0.005	[0.001, 0.005, 0.01]	0.01
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.0034545	[0.0026364, 0.0034546, 0.0042727]	0.0042727
max_leaf_nodes	[2, 10, 100, None]	100	[50, 100, 150]	50
max_depth	[1, 2, 4, 8, 16, 32, 64, 128]	16	[8, 16, 32]	16
max_features	['log2', 'sqrt', None]	log2	log2	log2
loss	['log_loss', 'exponential']	log_loss	log_loss	log_loss
AUROC trainset		0.855		0.856
AUROC testset		0.854		0.855

Tabel 19: Hyperparameter optimalisatie RandomForest (na verkennende gegevensanalyse)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	200	[100, 200, 1000]	1000
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.001	[0.0001, 0.001, 0.0018182]	0.001
max_leaf_nodes	[2, 10, 100, None]	None	[500, None]	500
max_depth	[1, 2, 4, 8, 16, 32, 64, 128]	128	[64, 128, 256]	64
max_features	['log2', 'sqrt', None]	None	None	None
criterion	['gini', 'entropy', 'log_loss']	gini	gini	gini
AUROC trainset		0.843		0.843
AUROC testset		0.842		0.842

F.4 Optimalisatie cardiale patiënten

Tabel 20: Hyperparameter optimalisatie Logistische Regressie (cardiale patiënten)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
C	np.logspace(-1, 1, 12)	1.8738174	[1.2328467, 1.8738174, 2.8480359]	0.8111308
penalty	['l1', 'l2']	l1	l1	l1
solver	['liblinear', 'newton-cg', 'lbfgs', 'sag', 'saga', None]	liblinear	['liblinear', 'saga']	liblinear
max_iter	[500, 1000, 3000, 10000]	500	[100, 500, 1000]	100
AUROC trainset		0.832		0.834
AUROC testset		0.822		0.822

Tabel 21: Hyperparameter optimalisatie AdaBoost (cardiale patiënten)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	100	[64, 100, 200]	200
learning_rate	[0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 10]	0.1	[0.05, 0.1, 0.25, 0.5]	0.05
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.0091818	[0.0083636, 0.0091818, 0.01]	0.01
max_leaf_nodes	[2, 10, 100, None]	10	[10, 20]	20
algorithm	['SAMME', 'SAMME.R']	SAMME	SAMME	SAMME
AUROC trainset		0.863		0.864
AUROC testset		0.850		0.855

Tabel 22: Hyperparameter optimalisatie GradientBoost (cardiale patiënten)

Hyperparameter	Input RS	Resultaat RS	Input GS	Resultaat GS
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	1000	[500, 1000, 3000]	1000
learning_rate	[0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 10]	0.005	[0.001, 0.005, 0.01]	0.005
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.0034545	[0.0026364, 0.0034546, 0.0042727]	0.0026364
max_leaf_nodes	[2, 10, 100, None]	100	[50, 100, 150]	50
max_depth	[1, 2, 4, 8, 16, 32, 64, 128]	16	[8, 16, 32]	16
max_features	['log2', 'sqrt', None]	log2	log2	log2
loss	['log_loss', 'exponential']	log_loss	log_loss	log_loss
AUROC trainset		0.871		0.872
AUROC testset		0.860		0.861

Tabel 23: Hyperparameter optimalisatie RandomForest (cardiale patiënten)

Hyperparameter	**Input RS**	**Resultaat RS**	**Input GS**	**Resultaat GS**
n_estimators	[1, 8, 16, 64, 100, 200, 1000]	200	[100, 200, 1000]	1000
min_samples_leaf	np.linspace(0.001, 0.01, 12)	0.001	[0.0001, 0.001, 0.0018182]	0.001
max_leaf_nodes	[2, 10, 100, None]	None	[500, None]	500
max_depth	[1, 2, 4, 8, 16, 32, 64, 128]	128	[64, 128, 256]	64
max_features	['log2', 'sqrt', None]	log2	log2	log2
criterion	['gini', 'entropy', 'log_loss']	entropy	entropy	entropy
AUROC trainset		0.863		0.865
AUROC testset		0.853		0.854

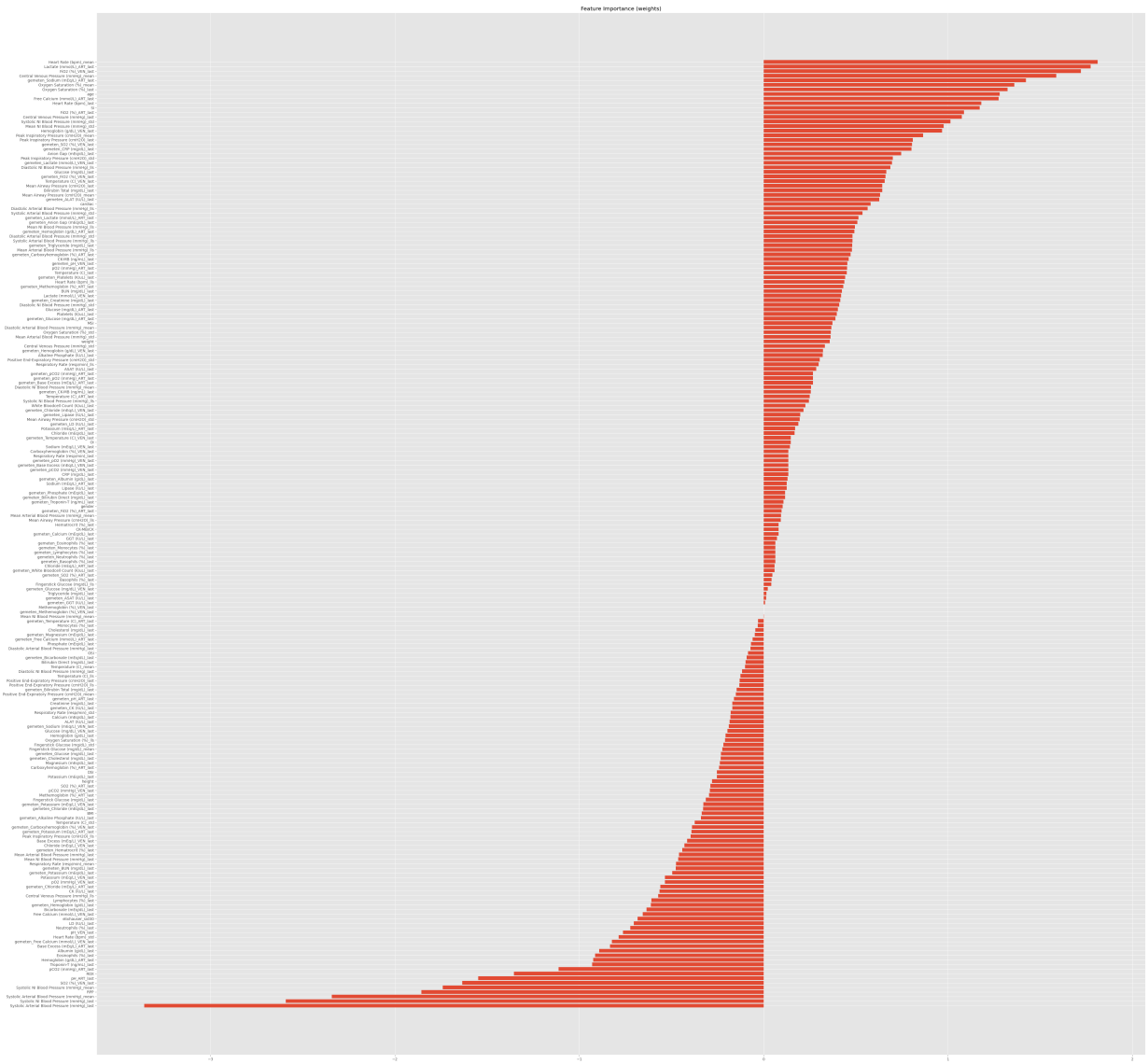
G Belangrijkheid van parameters

In deze appendix zullen de belangrijkheid van iedere meegenomen parameter, voor ieder model, voor iedere subset worden bekeken. Voor de belangrijkheid van parameters voor de subset: vóór de verkennende gegevensanalyse zie sectie G.1, ná de verkennende gegevensanalyse zie sectie G.2 en voor de cardiale patiëntgroep zie sectie G.3. De selectie van de parameters per subset is te zien in appendix E in de tabel E.1.

Disclaimer: door in te zoomen op de figuren (digitale versie) zijn alle gegevens in deze figuren te zien.

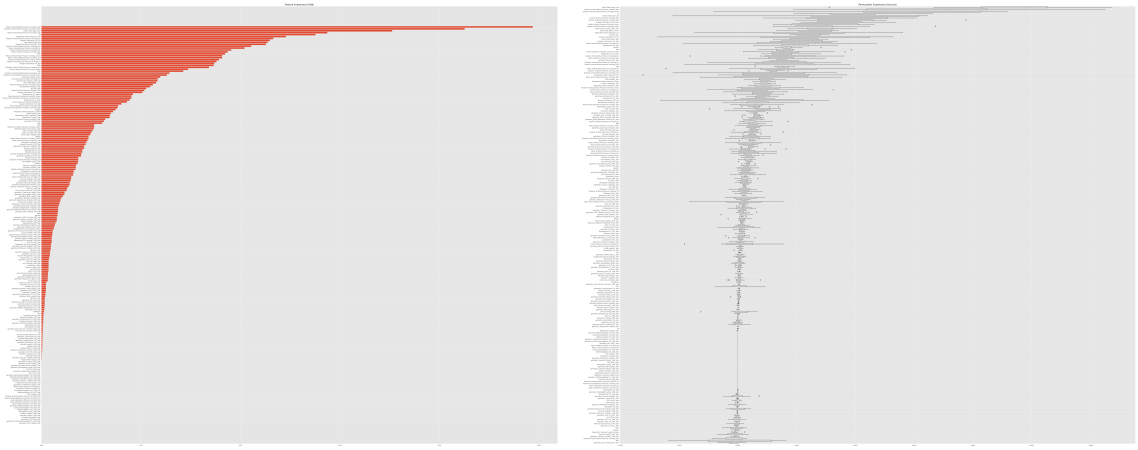
G.1 Parameter belangrijkheid vóór verkennende gegevensanalyse

G.1.1 Logistische regressie



Figuur 12: Belangrijkheid van parameters (relatieve gewichten): Logistische regressie (*subset: vóór VGA*)

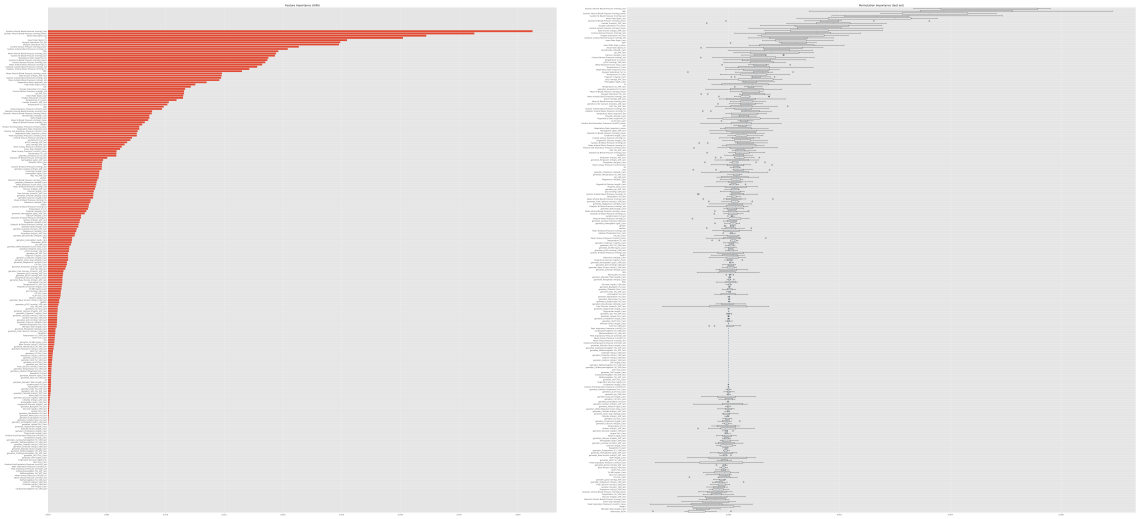
G.1.2 Adaptive Boosting



(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 14: Belangrijkheid van parameters: Adaboost (*subset: vóór VGA*)

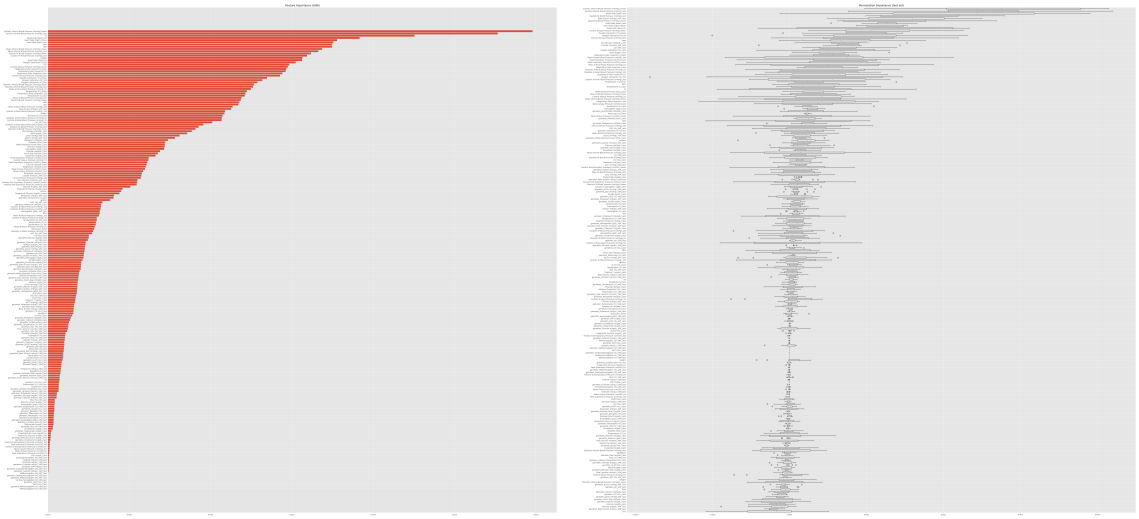
G.1.3 Gradient Boosting



(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 16: Belangrijkheid van parameters: GradientBoost (*subset: vóór VGA*)

G.1.4 Random Forest

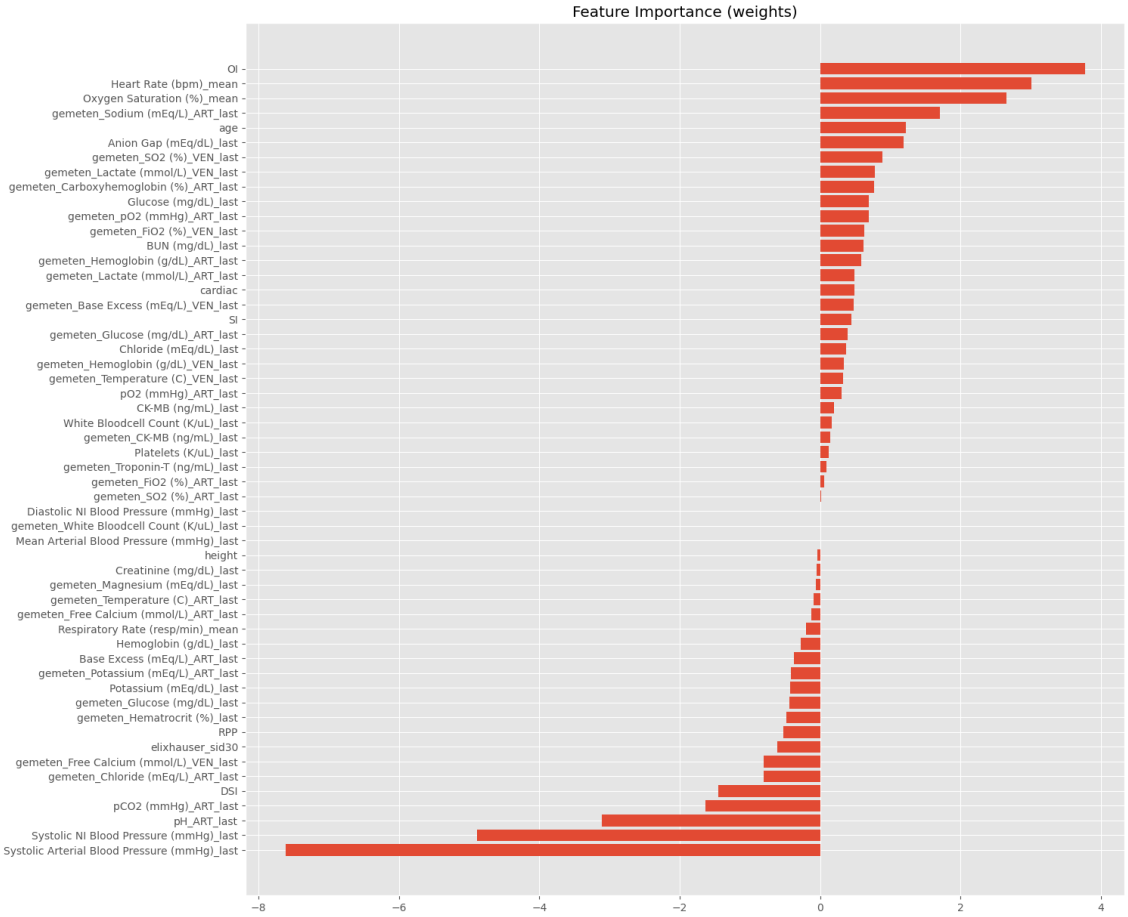


(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 18: Belangrijkheid van parameters: RandomForest (*subset: vóór VGA*)

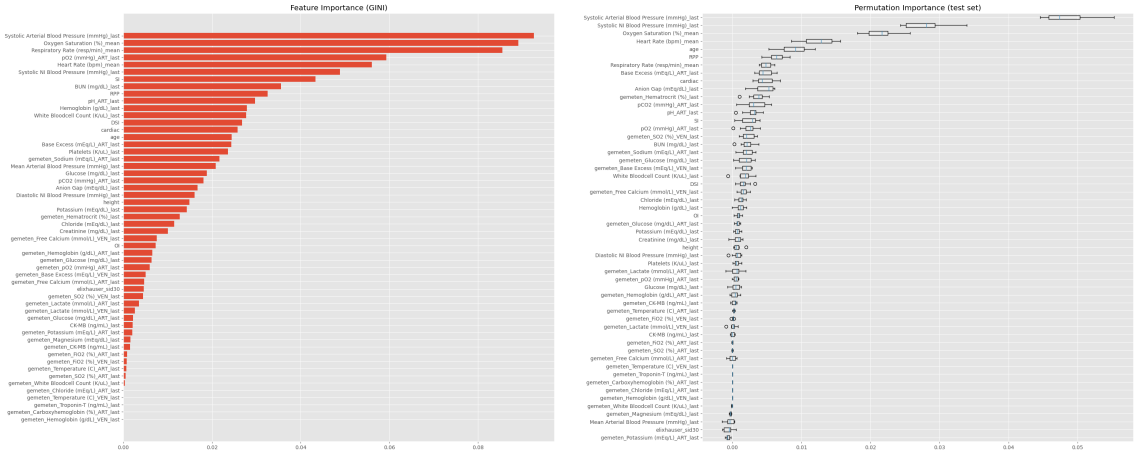
G.2 Parameter belangrijkheid na verkennende gegevensanalyse

G.2.1 Logistische regressie



Figuur 19: Belangrijkheid van parameters (relatieve gewichten): Logistische regressie (*subset: na VGA*)

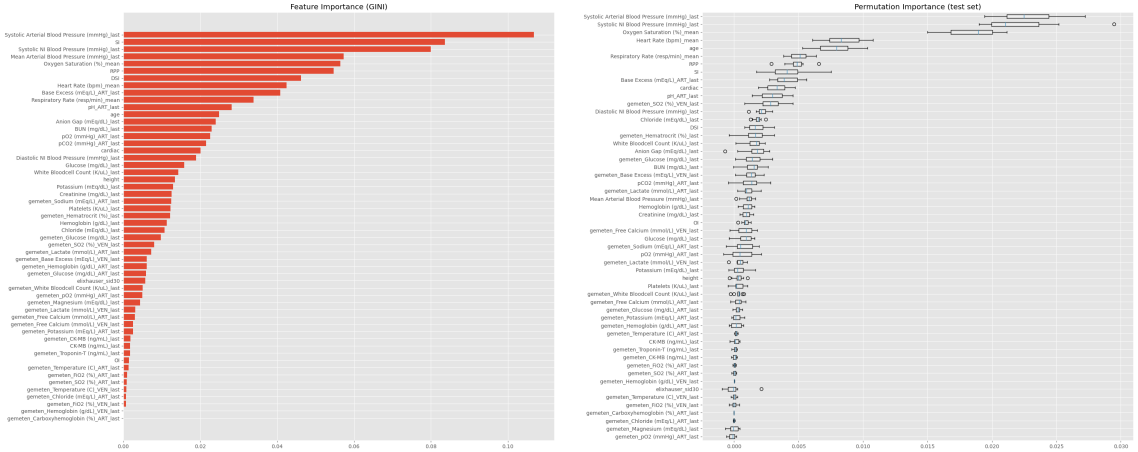
G.2.2 Adaptive Boosting



(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 21: Belangrijkheid van parameters: Adaboost (*subset: na VGA*)

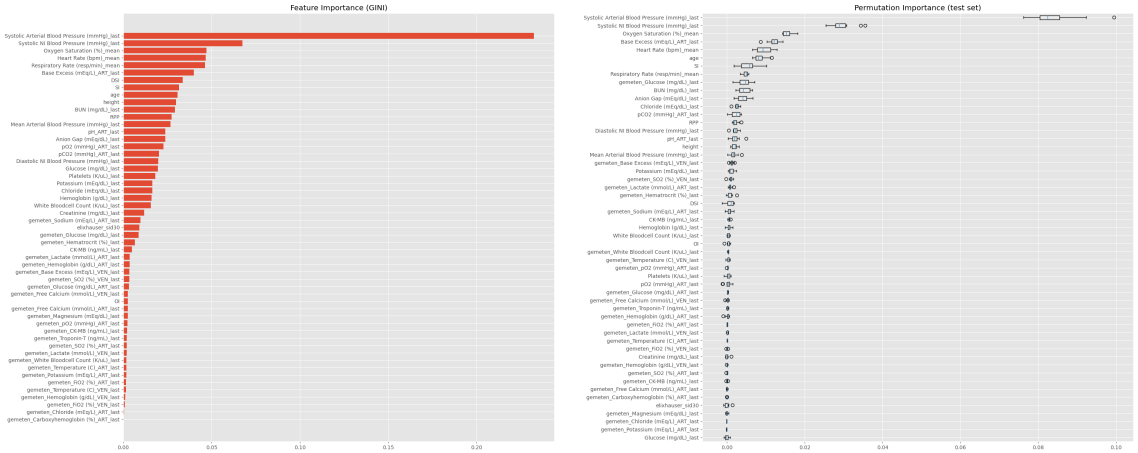
G.2.3 Gradient Boosting



(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 23: Belangrijkheid van parameters: GradientBoost (*subset: na VGA*)

G.2.4 Random Forest

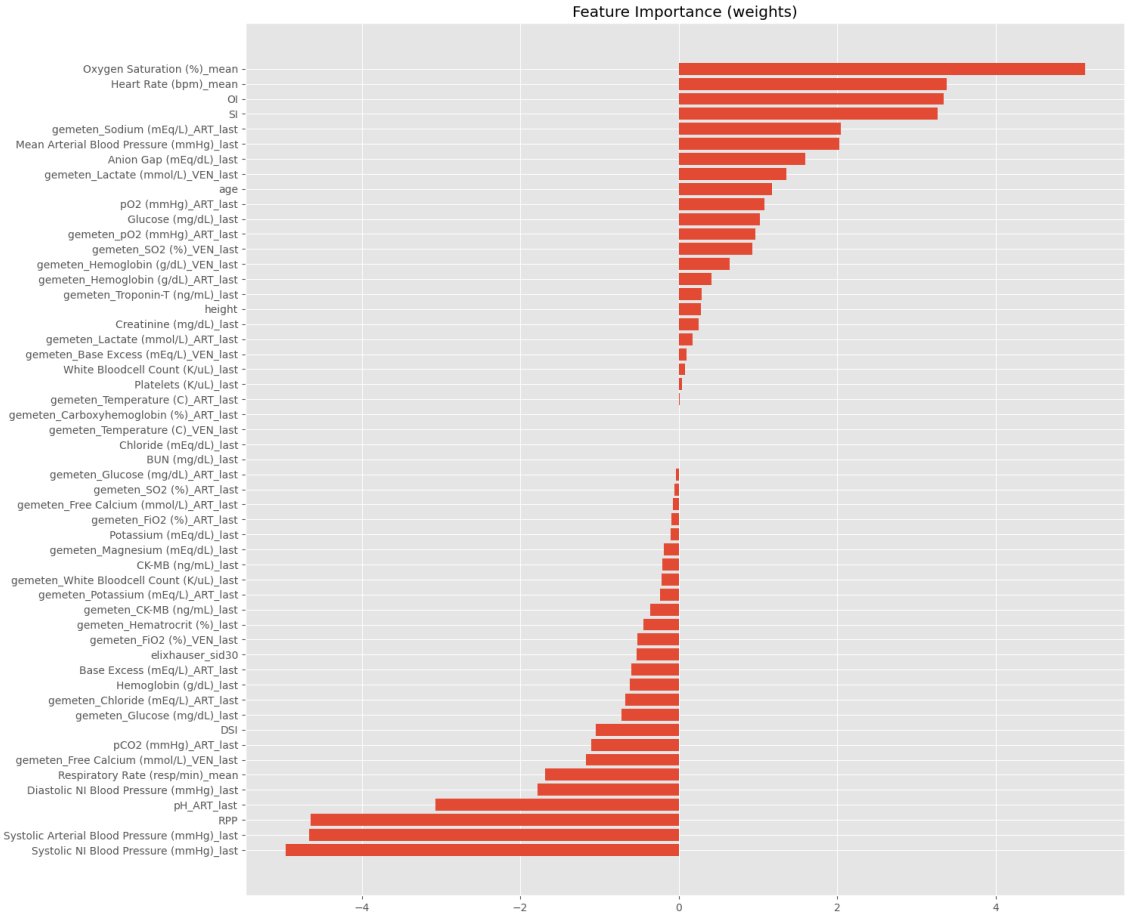


(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 25: Belangrijkheid van parameters: RandomForest (*subset: na VGA*)

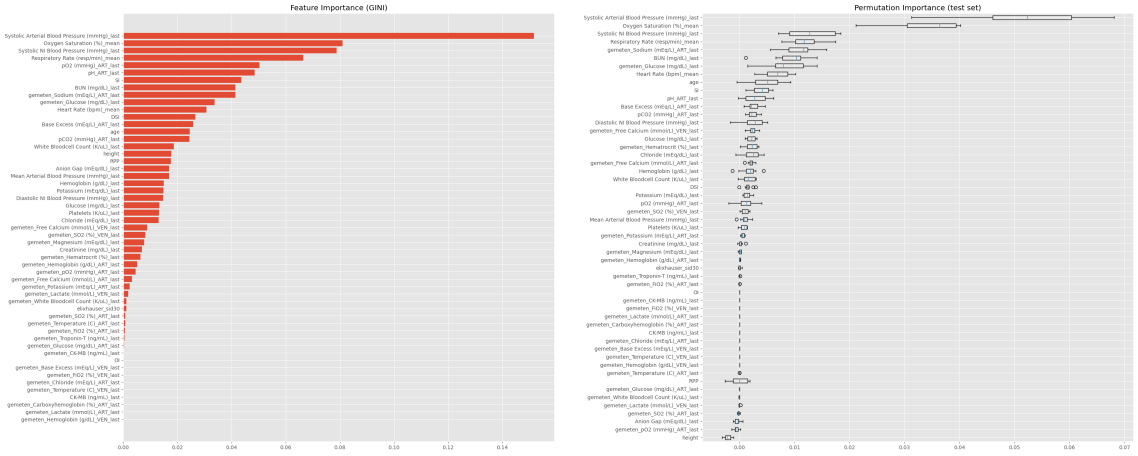
G.3 Parameter belangrijkheid cardiale patiënten

G.3.1 Logistische regressie



Figuur 26: Belangrijkheid van parameters (relatieve gewichten): Logistische regressie (*subset: cardiale patiënten*)

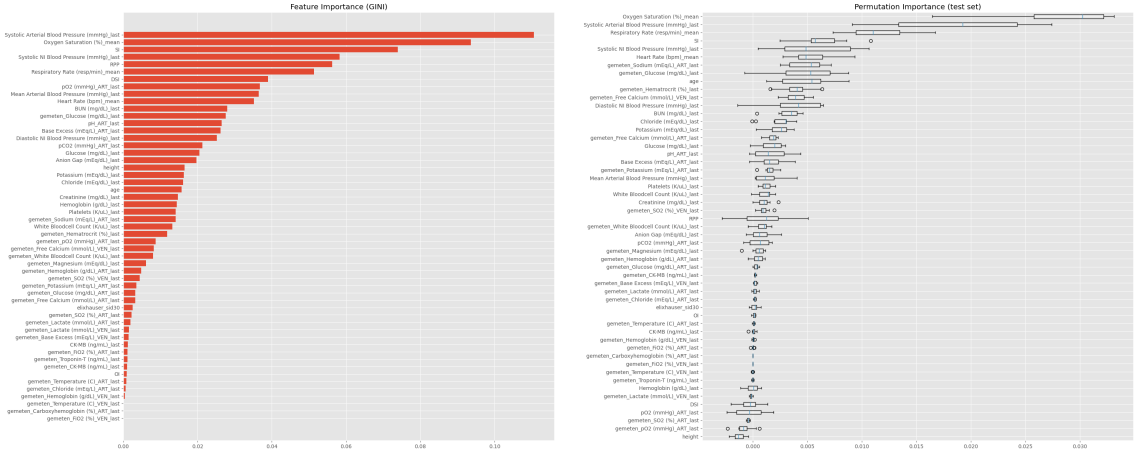
G.3.2 Adaptive Boosting



(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 28: Belangrijkheid van parameters: Adaboost (*subset: cardiale patiënten*)

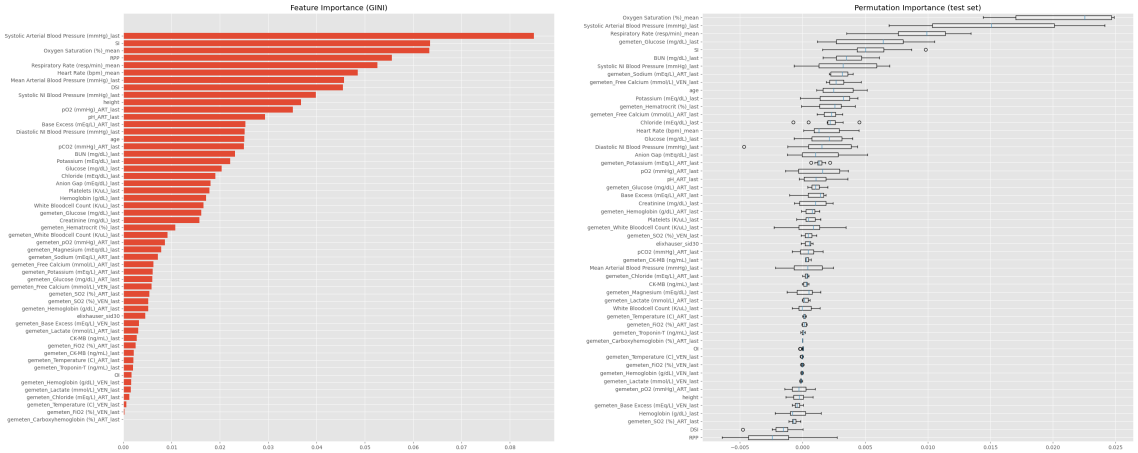
G.3.3 Gradient Boosting



(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 30: Belangrijkheid van parameters: GradientBoost (*subset: cardiale patiënten*)

G.3.4 Random Forest



(a) Gini impurity (relatieve belangrijkheid) (b) Permutatie (rel. afname nauwkeurigheid)

Figuur 32: Belangrijkheid van parameters: RandomForest (*subset: cardiale patiënten*)

H TRIPOD

De *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis* (TRIPOD)-richtlijn is opgesteld om de rapportage van klinische predictiemodellen transparanter en betrouwbaarder te maken. In deze appendix is een overzicht gegeven van alle onderdelen van de TRIPOD-checklist en hoe die zijn verwerkt in dit wetenschappelijk verslag.

Tabel 24: Toepassing van TRIPOD-checklist

Section/subject	Checklist	Toepassing
Title and abstract		
Title	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	Zie voorblad.
Abstract	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	Zie abstract.
Introduction		
Source of data	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	Zie sectie 4.1.
	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	Verwezen naar [23].
Participants	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	Verwezen naar [23].
	Describe eligibility criteria for participants.	Zie sectie 4.1 en 4.2.1.
	Give details of treatments received, if relevant.	Zie sectie 3.2.
Outcome	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	Zie sectie 4.2.4
	Report any actions to blind assessment of the outcome to be predicted.	N.v.t.
Predictors	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	Zie appendix A.

Gaat door op volgende pagina

	Report any actions to blind assessment of predictors for the outcome and other predictors.	N.v.t.
Sample size	Explain how the study size was arrived at.	Zie figuur 2.
Missing data	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	Zie sectie 4.3.1 en sectie 4.4.3
Statistical analysis methods	Describe how predictors were handled in the analyses.	Zie sectie 4.3.
	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	Zie heel sectie 4.4.
	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	Zie sectie 4.4.5.
Risk groups	Provide details on how risk groups were created, if done.	N.v.t.
Results		
Participants	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Zie figuur 2 en tabel 1.
	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	Zie tabel 1.
Model development	Specify the number of participants and outcome events in each analysis.	N.v.t.
	If done, report the unadjusted association between each candidate predictor and outcome.	Zie appendix E.
Model specification	Present the full prediction model to allow predictions for individuals (i.e., all 15a regression coefficients, and model intercept or baseline survival at a given time point).	Zie appendix G.
	Explain how to the use the prediction model.	Zie aanbevelingen, sectie 6.5.

Gaat door op volgende pagina

Model performance	Report performance measures (with CIs) for the prediction model.	Zie sectie 6.
Discussion		
Limitations	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	Zie sectie 6.3.
Interpretation	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	Zie sectie 6.2.
Implications	Discuss the potential clinical use of the model and implications for future research.	Zie sectie 6.5.
Other information		
Supplementary information	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	N.v.t.
Funding	Give the source of funding and the role of the funders for the present study.	N.v.t.