iStock

AUTEURS

Patrick S. Bäuerlein en Xin Tian
(KWR)

Frederic Béen
(KWR)

Yiqun Sun
(School of Earth Sciences and Engineering, Hohai University)

Peter van Thienen
(KWR)

NAUWKEURIGE IDENTIFICATIE POLYMEREN EN MICROPLASTICS MET MACHINE LEARNING

Microplastics vormen een groeiend probleem voor het milieu. Er zijn verschillende soorten, waardoor identificatie een uitdaging kan zijn. Infraroodspectroscopie in combinatie met machine learning kan nauwkeurige identificatie mogelijk maken.

Microplastics inclusief rubberdeeltjes vormen een groeiend probleem voor het milieu omdat ze zich overal verspreiden en langzaam afbreken. Het is daarom belangrijk deze kleine plastic deeltjes te identificeren om de omvang van het probleem vast te stellen en effectieve oplossingen te vinden. Dit wordt ook door de Europese Unie onderschreven en is vastgelegd in de drinkwaterrichtlijn (EU) 2020/2184).

Er zijn veel verschillende soorten microplastics, variërend in grootte, vorm, staat en materiaal, waardoor identificatie een uitdaging kan zijn. Infraroodspectroscopie in combinatie met machine learning kan hier helpen door patronen in de spectra te ontdekken en nauwkeurige identificatie mogelijk te maken, ook als de deeltjes al door het milieu zijn aangetast en hierdoor met andere methoden moeilijk te identificeren zijn.

Inleiding

Onderzoek naar microplastics is al jaren in ontwikkeling, maar de data en inzichten zijn nog niet op het niveau van die van opgeloste stoffen. De oorzaak ligt in de aard van deeltjes, die in tegenstelling tot moleculen, als uniek kunnen worden beschouwd. Net als bij sneeuwvlokken is het ene deeltje het andere niet. Chemisch identieke deeltjes kunnen verschillen qua grootte en vorm. Ook kunnen ze deels zijn aangetast (bijvoorbeeld geoxideerd aan het oppervlak). Hierdoor ontstaat feitelijk een mengsel van het originele materiaal en het aangetaste materiaal, in verschillende verhoudingen, met als gevolg dat identificatie aanzienlijk wordt bemoeilijkt.

Bij een opgeloste stof zijn we gewend om deze stof te kunnen vergelijken met dezelfde stof in een database/bibliotheek. Komen de data overeen, dan kunnen we zeker zijn dat we de stof hebben geïdentificeerd. Bij deeltjes is dit niet zo eenvoudig, ook al is het mogelijk deze te vergelijken met data in de database. Zodra een deeltje te veel afwijkt van de data in de bibliotheek, kan het soms niet of verkeerd worden geclassificeerd. Een database vullen met alle mogelijke staten van een deeltje is niet mogelijk omdat dat aantal oneindig is. Toch zijn deze data nodig voor een goede identificatie. Er moeten dus manieren worden bedacht om dit probleem op te lossen.

Voor de analyse van microplastics wordt vaak gebruikgemaakt van infraroodspectroscopie. Elk deeltje wordt bijvoorbeeld bestraald met een infraroodlaser. Dit levert een infraroodspectrum dat karakteristiek is voor dit deeltje. Vervolgens wordt dit spectrum met spectra in een database vergeleken. Als de afwijkingen van de spectra te groot zijn, is identificatie niet meer mogelijk. Om dit probleem aan te pakken, is een tweestappenaanpak gekozen. In eerste instantie zijn nieuw gegenereerde spectra aan een database toegevoegd. Vervolgens werden de spectra van echte deeltjes met de spectra in de nieuwe database met behulp van machine learning vergeleken.

Nieuwe spectra

Het creëren van nieuwe spectra vond op de volgende wijze plaats. Uit de bestaande database, in dit geval 210 spectra van verschillende deeltjes werden twee willekeurige spectra gekozen afkomstig van deeltjes met uiteenlopende karakteristieken maar wel van hetzelfde type polymeer (bijvoorbeeld polyethyleen). Hiervan werd door een lineaire combinatie een nieuw spectrum gemaakt. De individuele spectra krijgen hierbij een willekeurige weegfactor tussen 0,1 en 0,9. De som van beide weegfactoren mag echter niet meer dan 1 zijn. Op deze manier zijn de twee oorspronkelijke spectra altijd de uitersten. Deze stap kan met andere spectra worden herhaald tot men voldoende spectra heeft gecreëerd.

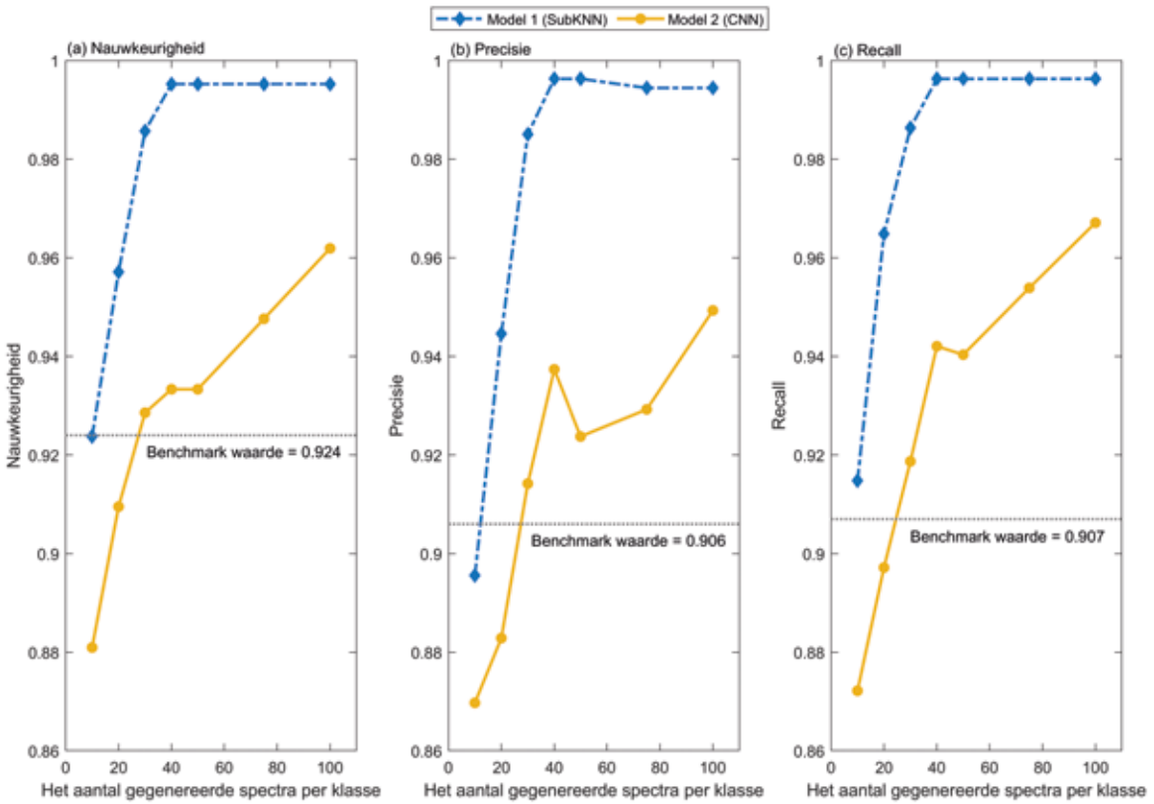
Een toevalsgenerator bepaalt welke spectra en welke wegfactoren er worden gebruikt. Bij het kiezen van het aantal spectra wordt rekening gehouden met de 'Pareto optimale waarde'. Dit wil zeggen dat de rekentijd, die nodig is voor het creëren van de spectra, redelijk moet zijn. Deze nieuwe spectra werden in een nieuwe database gekopieerd en vervolgens werd deze database gebruikt om de 210 originele spectra van echte deeltjes te classificeren. Hiervoor zijn twee verschillende machine learning modellen getest. Het eerste model is een ensamble sub-knn (subspace k-nearest neighbours) en het tweede CNN (convolutional neural network). Het sub knn model gebruikt de getallen zoals ze in de data staan. In dit model worden de getallen uit de spectra gebruikt om een punt in een multidimensionale coördinatensysteem te projecteren. Vervolgens wordt gekeken of de data van andere deeltjes hierbij in de buurt komen. Als dit zo is, dan zijn de deeltjes vergelijkbaar. Komt dus een onbekend deeltje in de buurt van bekende deeltjes, dan kan het worden geïdentificeerd.

Bij het tweede model worden de numerieke data van het spectrum omgezet in visuele data. De getallen worden in dit geval op een polair (circulair) coördinatensysteem weergegeven en als figuur opgeslagen. De figuren die ontstaan worden met elkaar vergeleken. CNN's zijn goed in het herkennen van patronen in een afbeelding, zoals lijnen, gradiënten, cirkels of kleuren. Wanneer er grote overeenkomsten zijn tussen twee afbeeldingen, dan zijn de bijbehorende deeltjes ook vergelijkbaar.

Complexiteit

Naast de twee verschillende manier van werken, onderscheiden CNN en sub-knn zich ook qua complexiteit. CNN, bekend van beeldherkenning, is een zogeheten deep learning algoritme, dat veel rekenkracht vraagt en een aanzienlijke hoeveelheid data nodig heeft om goed te kunnen werken. Sub-knn daarentegen heeft veel minder rekenkracht nodig en kan ook met kleine datasets al goede resultaten leveren.

Voor het beoordelen van de methoden zijn nauwkeurig-



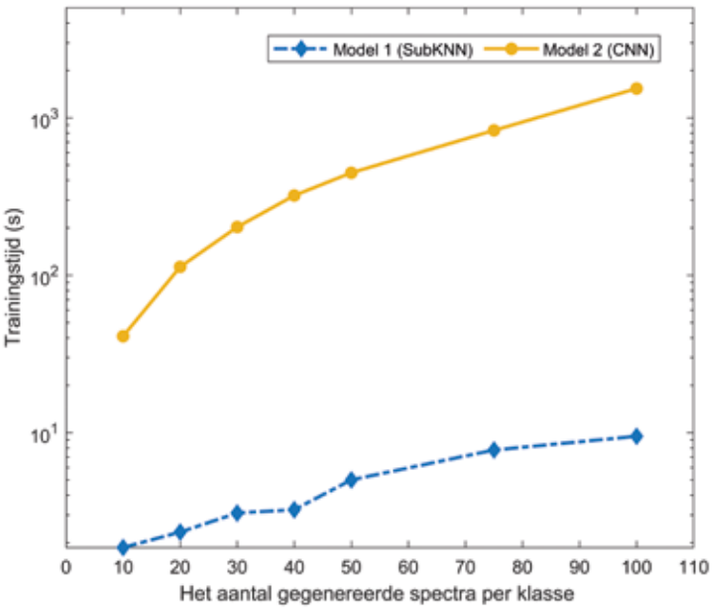
Afbeelding 1. Prestaties van de twee modellen. Alle modellen werden getraind op N gegenereerde spectra per klasse (N = 10, 20, 30, 40, 50, 75 of 100) en getest op 210 originele spectra. Een referentiewaarde werd berekend op basis van het SubKNN-model dat werd getraind en getest op de originele spectra. De prestaties van het model worden weergegeven als (a) nauwkeurigheid, (b) precisie en (c) recall.

heid, precisie en 'recall' van belang, evenals de hoeveelheid inspanning die vereist is, zoals de grootte van de dataset en de benodigde rekentijd en -kracht. De verhouding van correct geïdentificeerde polymeren ten opzichte van het totale aantal pogingen tot identificatie is de nauwkeurigheid. Precisie is de verhouding tussen het aantal keren dat het model een polymeer correct heeft geïdentificeerd en het totale aantal keren dat het model een poging tot voorspelling heeft gedaan voor een bepaalde uitkomst. Dus, als het model bijvoorbeeld 100 keer b.v. polyethylene heeft geïdentificeerd en daarvan zijn er 80 correct, dan is de precisie 80%. Recall is het aantal keren dat het model een polymeer daadwerkelijk correct identificeert, gedeeld door het totale aantal van dit polymeer. Dit geeft een idee van hoe goed het model presteert bij het correct identificeren van de polymeren. Het vinden van een goede balans tussen nauwkeurigheid, precisie en recall is essentieel.

Resultaten

Uit afbeelding 1 blijken de maximale waarden voor nauwkeurigheid, precisie en recall voor de twee modellen allemaal >0,94, wanneer de modellen 100 gegeneerde spectra per type polymeer als referentiespectra worden aangeboden. Dit betekent dat elk van deze modellen nauwkeurig polymeren kan identificeren met ten hoogste 6% foute voorspellingen. Het beste model met een nauwkeurigheidsscore van 0,995 (d.w.z. sub-knn, getraind op 40 gegenereerde spectra) heeft slechts één verkeerd geclassificeerd spectrum; dit model is het meest geschikt om polymeren te classificeren op basis van de infraroodspectra. Hoewel de modelarchitectuur van sub-knn eenvoudiger is dan die van de cnn, presteert de eerste beter dan de laatste, in termen van alle prestatie-indicatoren. Bovendien blijkt uit afbeelding 1 dat het aantal trainingsspectra (de x-as) een aanzienlijke invloed heeft op de modelprestaties. De prestaties van alle modellen nemen toe naarmate het aantal spectra in de bibliotheek toeneemt. Alleen een lichte afname van precisie bij sub knn kan worden gezien. Ofwel optimale modelprestaties kunnen in gevaar komen als modellen worden getraind met weinig spectra. Sub-knn bereikt de beste prestaties al bij 40 spectra, terwijl cnn zelfs bij 100 spectra nog aanzienlijk slechter presteert.

Er is ook een benchmarksimulatie (prestatietest) uitgevoerd met sub-knn, getraind en kruisgevalideerd (steekproefvalidatie) op basis van de oorspronkelijke dataset met alleen 210 spectra in totaal. Om het anders te zeggen: onderzocht wordt vanaf welk punt het gebruik van kunstmatig gegenereerde spectra tot betere resultaten leidt dan wanneer alleen de oorspronkelijke dataset wordt gebruikt. Afbeelding 1 toont de benchmark als referentielijn. In het algemeen kan het gebruik van 50 gegenereerde spectra of meer per polymeerklasse al leiden tot een prestatie



Afbeelding 2:
Trainingstijden van twee gebruikte modellen

die voor alle modellen beter is dan de benchmark. Met name sub-knn (het meest geschikte model) heeft slechts 20 spectra of meer nodig om beter te presteren dan de benchmark. Dit laat zien dat het genereren van spectra een oplossing biedt voor het geval er te weinig echte spectra beschikbaar zijn.
Afbeelding 2 toont de trainingstijd als functie van de grootte van de dataset. Sub-knn heeft 9,5 s nodig om modellen te trainen, terwijl het tweede cnn 5 min nodig heeft. Hoewel de langste periode (25 min) nog relatief kort is, is het belangrijk rekening te houden met toekomstige toepassingen waarbij het model wordt gebruikt om duizenden polymeren te classificeren in een online leermodus, waarbij het model herhaaldelijk moet worden getraind met toegevoegde spectra van nieuwe polymeertypes. Wat de trainingstijd betreft, is sub-knn ook de meest geschikte aanpak.
De hier voorgestelde methodes om deeltjes met machine learning te identificeren worden op dit moment in een programma geïntegreerd dat door iedereen zonder grote voorkennis van machine learning kan worden gebruikt. Hiervoor wordt een user interface gemaakt.

Conclusies

Het identificeren van microplastics in milieumonsters is een uitdaging vanwege de verschillen die optreden na afbraak in het milieu. Hierdoor vertonen microplastics sterk afwijkende eigenschappen ten opzichte van nieuwe plastics. Machine learning modellen kunnen helpen bij het nauwkeurig identificeren en tellen van microplastics in monsters. Het onderzoek toont aan dat het niet per se nodig is om de meest geavanceerde machine learning modellen te gebruiken. Uit de resultaten blijkt dat het eenvoudigere model zowel qua prestatie als rekentijd beter presteren. Dit benadrukt het belang van het testen van meerdere modellen voordat een definitieve keuze wordt gemaakt voor een bepaalde toepassing.

Patrick S. Bäuerlein, Xin Tian en Frederic Beén (*KWR*), Yiqun Sun (*School of Earth Sciences and Engineering, Hohai University*), Peter van Thienen (*KWR*)

SAMENVATTING

Gegevens en inzichten die voortkomen uit onderzoek naar microplastics zijn nog niet zo nauwkeurig als die van opgeloste stoffen. Dit komt door de aard van de deeltjes, die, in tegenstelling tot moleculen, als uniek kunnen worden beschouwd. Deeltjes kunnen verschillen in grootte, vorm en kunnen gedeeltelijk zijn aangetast. Dit maakt een juiste identificatie met bijvoorbeeld infraroodspectroscopie moeilijk.
Om dit probleem aan te pakken zijn machine learning modellen getest en gebruikt om de microplastics te identificeren. Het werd duidelijk dat al een relatief simpel model in staat is om microplastics nauwkeurig te identificeren. Dit maakt het gebruik van deze techniek geschikt voor een breder publiek en niet alleen maar voor specialisten. Er is wel kennis nodig voor het maken van een model en het trainen ervan, maar zodra dit is gebeurd, kunnen de modellen door iedereen met computerkennis worden gebruikt. De hier gepresenteerde aanpak kan ook worden toegepast op andere soorten spectrale gegevens, bijvoorbeeld ultraviolet-, Raman-, FTIR- en massaspectra.