مقاله مروری

فناوری پلیمر ایران

فصلنامه علمی
سال پنجم، شماره ۴، شماره پیاپی ۲۰، زمستان ۱۳۹۹
Vol. 5, No. 4, Issue No. 20
Winter 2021, Quarterly
صفحه ۵۶ - ۴۵

Iran Polymer Technology: Research and Development

# مروری بر روش‌های پیش‌بینی و تخمین ویژگی نمونه‌ها با استفاده از روش‌های تجزیه‌ای و الگوریتم‌های یادگیری ماشین

سید محمدرضا میلانی حسینی[1*]، سید مجید هاشمیان زاده[2]، بیتا یاراحمدی[3]

۱ تهران، دانشگاه علم وصنعت ایران، دانشکده شیمی، دانشیار شیمی تجزیه

۲ تهران، دانشگاه علم وصنعت ایران، دانشکده شیمی، دانشیار شیمی فیزیک

۳ تهران، دانشگاه علم وصنعت ایران، دانشکده شیمی، دانشجوی دکتری شیمی تجزیه

## چکیده ...

امروزه استفاده از یادگیری ماشین (Machine Learning) به‌علت مزایای بسیار از جمله سادگی، سرعت بالا، دقت زیاد در پیش‌بینی فرایندهای گوناگون، عدم نیاز به تجهیزات و وسایل پیچیده و در دسترس بودن کاربردهای زیادی در علوم و زمینه‌های مختلف از جمله آمار، ریاضیات، فیزیک، شیمی، بیوشیمی، مهندسی مواد، مهندسی پزشکی، داروسازی و... پیدا کرده است. بنابراین در عصر حاضر مطالعه و بررسی روش‌ها و الگوریتم‌های گوناگون یادگیری ماشین از اهمیت بسیاری برخوردار است. به‌عنوان زیرمجموعه‌ای از هوش مصنوعی، الگوریتم‌های یادگیری ماشین، مدل ریاضی بر اساس داده‌های نمونه یا داده‌های آموزشی به‌منظور پیش‌بینی یا تصمیم‌گیری بدون برنامه‌ریزی، ایجاد می‌کنند. یکی از جذاب‌ترین موضوعاتی که می‌توان با هوش مصنوعی روی آن متمرکز شد، پیش‌بینی و تخمین رخدادها در آینده است. یادگیری ماشین، توانایی یادگیری مستقل را برای ماشین‌ها ایجاد می‌کند. به‌عبارتی ماشین می‌تواند از تجربیات، مشاهدات و الگوهایی که بر اساس مجموعه‌ای از داده‌ها تجزیه و تحلیل می‌کند، آموزش ببیند. امروزه یادگیری ماشین کاربرد زیادی در شیمی تجزیه پیدا کرده است و از داده‌های حاصل از روش‌های مختلف تجزیه‌ای مانند طیف سنجی، فلورسانس، ولتامتری، طیف‌سنجی نشری، میکرواستخراج فاز جامد، سوانگاری مایع، سوانگاری گازی، طیف‌سنجی فروسرخ و ... برای مدل‌سازی، پیش‌بینی و طبقه‌بندی داده‌ها استفاده می‌شود. یادگیری ماشین همچنین به‌طور گسترده در سنتز، بهینه‌سازی پارامترها و کنترل خواص پلیمرها استفاده می‌شود. مدل‌های ساخته شده از دقت بسیار زیادی برخوردار هستند.

**واژه‌های کلیدی:**

**یادگیری ماشین**
**روش‌های تجزیه‌ای**
**ماشین‌بردار پشتیبان**
**پلیمرها**

*پست الکترونیکی مسئول مکاتبات:
drmilani@iust.ac.ir

## ۱ مقدمه

سه روش اصلی یادگیری ماشین شامل یادگیری با نظارت، یادگیری بدون نظارت و یادگیری عمیق است. هر یک ازاین روش‌های یادگیری شامل چندین نوع الگوریتم مختلف هستند که با توجه به نوع موضوع مورد بررسی از یکی از این الگوریتم‌ها باید استفاده شود. یادگیری ماشین، کمک فراوانی به صرفه‌جویی در هزینه‌های عملیاتی و بهبود سرعت عمل تجزیه‌وتحلیل داده‌ها می‌کند. به‌عنوان مثال در صنعت نفت و پتروشیمی با استفاده از یادگیری ماشین، داده‌های عملیاتی تمام حفاری‌ها اندازه‌گیری شده و با تجزیه‌وتحلیل داده‌ها، الگوریتم‌هایی تنظیم می‌شود که در حفاری‌های بعدی، استخراج پربازده و بهینه‌تری داشته باشند. یادگیری ماشین، مطالعه‌ی علمی الگوریتم‌ها و مدل‌های آماری مورد استفاده سامانه‌های رایانه‌ای است که به‌جای استفاده از دستورالعمل‌های واضح از الگوریتم‌ها و استنباط برای انجام وظایف سود می‌برند. به عبارت دیگر یادگیری ماشین یکی از زیر مجموعه‌های هوش مصنوعی است که به سامانه‌ها این امکان را می‌دهد تا بدون نیاز به برنامه‌نویسی به‌صورت خودکار یادگیری و پیشرفت داشته باشند. تمرکز اصلی یادگیری ماشینی بر توسعه برنامه‌های رایانه‌ای است که بتوانند به داده‌ها دسترسی پیدا کنند و از آن برای یادگیری خود استفاده کنند. فرایند یادگیری با مشاهدات یا داده‌ها آغاز می‌شود، مانند مثال‌ها، تجارب مستقیم یا دستورالعمل‌ها، تا به الگویی در داده‌ها برسند. هدف اصلی آن است که به رایانه این اجازه داده شود که بدون دخالت و کمک انسان به‌طور خودکار یادگیری داشته باشند و بتوانند اقدامات خود را بر مطابق با آن تنظیم کنند [۱].

## ۲ مراحل یادگیری ماشین

هدف یادگیری ماشین این است که رایانه‌ها و سامانه‌ها بتوانند به‌تدریج و با افزایش داده‌ها کارایی بهتری در انجام وظیفه مورد نظر پیدا کند. طیف پژوهش‌های یادگیری ماشین بسیار گسترده است. از لحاظ نظری پژوهشگران در تلاش هستند که روش‌های یادگیری تازه‌ای به وجود آورند و امکان‌پذیری و کیفیت یادگیری را برای روش‌هایشان مطالعه کنند و در سوی دیگر عده‌ای از پژوهشگران سعی می‌کنند روش‌های یادگیری ماشینی را بر مسائل تازه‌ای اعمال کنند. البته این طیف، گسسته نیست و پژوهش‌های انجام‌شده دارای مؤلفه‌هایی از هر دو رویکرد هستند. در ادامه به تشریح فرایندهای یادگیری ماشین پرداخته شده است. مراحل تجزیه‌وتحلیل فرایندهای یادگیری ماشین شامل جمع‌آوری و تهیه داده، انتخاب و آموزش مدل و تنظیم و پیش‌بینی ابرپارامترها است [۲].

### ۲–۱ جمع‌آوری و تهیه داده‌ها

اولین قدم در فرایند یادگیری ماشین این است که دانش و داده مورد نیاز برای ماشین تهیه شود. این داده‌ها به دو گروه تقسیم می‌شوند. گروه اول برای آموزش سامانه و گروه دیگر برای آزمایش سامانه استفاده می‌شود. باید به این مسئله توجه داشت که داده‌های انتخابی نماینده کل جمعیت باشند. داده‌ها معمولاً به میزان ۲۰/۸۰ یا ۳۰/۷۰ تقسیم می‌شوند، تا اطمینان حاصل شود که مدل پس از آموزش کافی می‌تواند بعداً آزمایش شود.

### ۲–۲ انتخاب و آموزش مدل

دومین مرحله و قدم بعدی در اصول یادگیری ماشین، انتخاب مدل و آموزش آن است. انواع مختلفی از الگوریتم‌ها و مدل‌های یادگیری ماشین وجود دارد که قبلاً ایجاد و اصلاح شده‌اند تا بتوانند نوع خاصی از مسئله یا مشکل را حل کنند. بنابراین، بسته به نیاز و مناسب بودن مدل برای حل مسئله مورد نظر، مدلی انتخاب و آموزش داده می‌شود.

### ۲–۳ ارزیابی مدل

ماشین الگوها و خصوصیات مختلفی را از داده‌هایی که به آن آموزش داده شده است یاد می‌گیرد و خود را برای تصمیم‌گیری‌هایی در زمینه‌های مختلف مانند شناسایی، طبقه‌بندی یا پیش‌بینی داده‌های جدید آموزش می‌دهد. برای بررسی دقیق اینکه ماشین چگونه قادر به گرفتن این تصمیمات است، پیش‌بینی‌ها را بر روی داده‌های آموزش داده شده، آزمایش می‌کنند. برای این کار ابتدا بر روی داده‌های آموزش داده شده کار می‌کنند و پس از آموزش مدل، از آن برای آزمایش بر اساس داده‌ها استفاده می‌شود تا مشخص شود چه میزان دقت دارد.

### ۲–۴ تنظیم و پیش‌بینی ابرپارامترها

در اصطلاحات مربوط به یادگیری ماشین، ابرپارامترها، پارامترهایی هستند که توسط خود مدل نمی‌توانند تخمین زده شوند، اما نیاز است که مورد بررسی قرار گیرند، زیرا نقش بسیار مهمی در افزایش عملکرد مدل دارند. اگر بخواهیم تعریفی ارائه دهیم، ابرپارامترها در یادگیری ماشین، پارامترهایی هستند که باید توسط کاربر برای اجرای الگوریتم، مشخص شوند. پارامترهای کلاسیک به‌وسیله داده‌ها آموزش داده می‌شوند، در حالی که ابرپارامترها ممکن است از داده‌ها یاد بگیرند.

یک مدل می‌تواند تعداد زیادی ابر پارامتر داشته باشد و فرایند انتخاب بهترین ترکیب ممکن از بین ابرپارامترها تنظیم ابرپارامتر نام دارد. برخی از روش‌های اساسی برای تنظیم ابرپارامترها

شامل جستجوی شبکه، جستجو تصادفی یا بهینه‌سازی مبتنی بر گرادیان است. پس از اتمام فرایند بهینه‌سازی ابر پارامترها، می‌توان گفت که مدل یادگیری ماشین ساخته شده است و بسته به میزان موفقیت آن یا به‌طور دقیق، توانایی پیش‌بینی آن، می‌توان آن را در دنیای واقعی اجرا و پیاده‌سازی کرد. بنابراین با کمک روش‌هایی که گفته شد، می‌توان الگوریتم یادگیری ماشین ساخت.

## ۳ تقسیم‌بندی مسائل

یکی از تقسیم‌بندی‌های متداول در یادگیری ماشینی، تقسیم‌بندی بر اساس نوع داده‌های در اختیار کارگزار هوشمند است. بر این اساس یادگیری‌ها به سه دسته متفاوت شامل یادگیری بدون نظارت، یادگیری با نظارت و یاد گیری عمیق تقسیم می‌شوند.

### ۳–۱ یادگیری با نظارت

یادگیری تحت نظارت، روشی عمومی در یادگیری ماشین است که در آن به سامانه، مجموعه‌ای از جفت‌های ورودی–خروجی ارائه شده و سامانه تلاش می‌کند تا تابعی از ورودی به خروجی را فرا گیرد. یادگیری تحت نظارت نیازمند تعدادی داده ورودی به‌منظور آموزش سامانه است. یادگیری تحت نظارت، خود به دو دسته رگرسیون و طبقه‌بندی تقسیم می‌شود. رگرسیون آن دسته از مسائل هستند که خروجی آن، عددی پیوسته یا مجموعه‌ای از اعداد پیوسته هستند مانند پیش‌بینی قیمت خانه بر اساس اطلاعاتی مانند مساحت، تعداد اتاق خواب‌ها، و غیره و دسته طبقه‌بندی به آن دسته از مسائل گفته می‌شود که خروجی آن، عضوی از یک مجموعه باشد، مانند پیش‌بینی این‌که ایمیلی هرزنامه هست یا خیر یا پیش‌بینی نوع بیماری فرد از میان ده بیماری.

### ۳–۲ یادگیری بدون نظارت

یادگیری بدون نظارت در مقابل یادگیری بانظارت، یکی از انواع روش‌های یادگیری ماشین است. اگر یادگیری بر روی داده‌های بدون برچسب و برای یافتن الگوهای پنهان در این داده‌ها انجام شود، یادگیری، بدون نظارت خواهد بود. در واقع برخلاف یادگیری نظارت‌شده، یادگیری بدون نظارت برای استنتاج و یافتن الگو از داده‌های ورودی بدون نیاز به نتایج برچسب‌گذاری شده، استفاده می‌شود. دو روش اصلی مورد استفاده در یادگیری بدون نظارت شامل خوشه بندی و کاهش ابعاد است [۳].

### ۳–۳ یادگیری تقویتی

هدف یادگیری تقویتی که بخشی از یادگیری ماشین است، این است که چگونه عامل‌های نرم‌افزاری، باید عملی را مناسب محیط انتخاب کنند تا پاداش بهینه بیشینه شود. یادگیری تقویتی مسئله‌ای است که عاملی می‌بایست رفتار خود را از طریق تعاملات آزمون و خطا با محیطی پویا فرا گیرد، با آن مواجه است. در یادگیری تقویتی هیچ نوع زوج ورودی–خروجی ارائه نمی‌شود. به جای آن، پس از انجام عمل، حالت بعدی و پاداش بلافصل به عامل ارائه می‌شود. هدف اولیه برنامه‌ریزی عامل‌ها با استفاده از تنبیه و تشویق است [٤]. امروزه یادگیری ماشین کاربرد زیادی در شیمی تجزیه پیدا کرده است و از داده‌های حاصل از روش‌های مختلف تجزیه‌ای مانند طیف–سنجی [۵]، فلورسانس [٦]، ولتامتری [۷]، طیف‌سنجی نشری [۸]، میکرواستخراج فاز جامد [۹]، سوانگاری مایع [۱۰]، سوانگاری گازی [۱۱]، طیف‌سنجی فروسرخ [۱۲] و ... برای مدل‌سازی، پیش‌بینی و طبقه‌بندی داده‌ها استفاده می‌شود. مدل‌های ساخته شده از دقت بسیار زیادی برخوردار هستند. اغلب مدل‌سازی داده‌ها به‌منظور طبقه‌بندی نمونه‌ها در دسته‌های مجزا انجام می‌شود و از مدل ساخته شده برای پیش‌بینی ویژگی‌های نمونه‌های ناشناخته استفاده می‌شود. در بین انواع روش‌های مختلف تجزیه‌ای، طیف‌سنجی فروسرخ به‌علت این‌که طیف حاصل از آن برای هر ترکیب، منحصربه‌فرد است و مانند اثرانگشت برای هر ترکیب عمل می‌کند، توجه ویژه‌ای را به خود جلب کرده، پژوهش‌های بسیاری در رابطه با مدل‌سازی طیف‌های فروسرخ تعداد زیادی از نمونه‌ها انجام شده است. استفاده از یادگیری ماشین، مزایای بسیاری دارد و از جمله مزایای آن می‌توان به سهولت شناسایی روندها و الگوریتم‌ها، بهبود مستمر، کاربردهای گسترده، عدم نیاز به دخالت انسان، صرفه‌جویی در وقت، صرفه‌جویی در هزینه‌ها و مدیریت داده‌های چندبعدی و چند متغیره اشاره کرد. یکی از ویژگی‌های یادگیری ماشین، حساس بودن به خطای موجود در داده‌ها است. بنابراین باید در مرحله جمع‌آوری داده‌ها حداکثر دقت را انجام داد، زیرا اگر داده‌ها با دقت جمع‌آوری نشوند مدل ساخته شده با خطا مواجه می‌شود و هنگام پیش‌بینی داده‌های ناشناخته دچار اشتباه می‌شود و دقت مدل ساخته‌شده کاهش می‌یابد. طیف پژوهش‌های یادگیری ماشین بسیار گسترده است.

## ۴ کاربردهای الگوریتم ماشین بردار پشتیبان

الگوریتم ماشین بردار پشتیبان جزء الگوریتم‌های یادگیری ماشین با نظارت است و دسته‌بندی را بر اساس تابع هسته با هدف طبقه‌بندی دو یا چند گروهی و رگرسیون انجام می‌دهد. این روش برای اولین بار توسط Vaping در سال ۱۹٦۳ برای تفکیک و دسته‌بندی داده‌هایی که جداپذیر خطی

بودند ساخته شد و در سال ۱۹۹۵ توسط Vaping و Cortes برای حالت غیرخطی تعمیم داده شد. امروزه ماشین‌های بردار پشتیبان به متداول‌ترین الگوریتم‌های پیش‌بینی و تشخیصی در یادگیری ماشین تبدیل شده‌اند. در سال ۲۰۱۹ F.R. Bertani و همکارانش مقدار سم افلاتوکسین موجود در نمونه بادام را با استفاده از سوانگاری مایع، طیف‌سنجی فلورسانس و الگوریتم ماشین بردار پشتیبان با دقت زیاد اندازه‌گیری و پیش‌بینی کردند.

در این روش ابتدا نمونه بادام آلوده با افلاتوکسین در آزمایشگاه تهیه شد و سپس طیف (HPLC) و فلورسانس نمونه در طول موج ۳۷۵ نانومتر اندازه‌گیری شد. مدل‌سازی‌ها در نرم‌افزار Matlab2018 انجام شد. با استفاده از مراحل پیش‌پردازش، تحلیل طیفی با استفاده از مدل طبقه‌بندی دودویی بر اساس الگوریتم ماشین بردار پشتیبان انجام شد. دقت طبقه‌بندی انجام شده با این روش ۹۴٪ به‌دست آمد [۱۳].

همچنین در سال ۲۰۱۶ Antonios Morellos و همکارانش با استفاده از طیف‌سنجی مرئی-فروسرخ نزدیک و الگوریتم‌های یادگیری ماشین موفق شدند مقدار نیتروژن، کربن و رطوبت موجود در خاک را اندازه‌گیری کنند. روش حداقل مربعات ماشین‌های برداری پشتیبان که اخیراً توسط Suykens و همکاران توسعه داده شده است، به‌عنوان روشی آسان اما قوی برای حل مسائل طبقه‌بندی، رگرسیون و تجزیه‌وتحلیل مسائل چندمتغیره خطی و غیرخطی با استفاده از معادلات خطی تنظیم‌شده عمل می‌کند و در این روش برخلاف روش ماشین بردار پشتیبان کلاسیک از معادلات درجه دوم استفاده نمی‌شود.

در این بررسی مقایسه بین الگوریتم‌های مختلف شامل الگوریتم رگرسیون مولفه اصلی، الگوریتم رگرسیون حداقل مربعات جزئی، الگوریتم حداقل مربعات ماشین بردار پشتیبانی و الگوریتم (Cubist) پیش‌بینی نیتروژن کل (TN)، میزان رطوبت (MC) وکربن آلی(OC) با استفاده از طیف‌سنجی فروسرخ نزدیک-مرئی انجام شد. نتایج نشان داد که برای مجموعه داده حاصل، الگوریتم (LS-SVM) عملکرد بهتری برای پیش‌بینی (MC) و (OC) نسبت به سایر الگوریتم‌ها داشت. اما مقدار (TN) با استفاده از الگوریتم (Cubist) بهتر و دقیق‌تر پیش‌بینی شد. روش‌های یادگیری ماشین مانند (Cubist) و (LS-SVM) قدرت توضیحی بهتری نسبت به الگوریتم‌های رگرسیون کلاسیک چند متغیره مانند (PCR) نشان دادند، به‌همین دلیل استفاده از آن‌ها برای طیف‌سنجی خاک توصیه شده است. در یادگیری ماشین می‌توان بعد مسائل را برای سهولت مدل‌سازی کاهش داد. برای مثال Xinhui Yu و همکارانش از هسته مرکزی خطی(RBF) برای نقشه‌برداری داده‌های موجود در فضای فعلی در فضایی با ابعاد بزرگ‌تر استفاده کردند و مدل مطلوبی ایجاد کردند که بیش‌ترین فاصله را نسبت به نزدیکترین نقاط در هر دسته دارد. محققان با استفاده از طیف فروسرخ نزدیک نمونه‌ها همراه با استفاده از الگوریتم ماشین بردار پشتیبان، روشی جایگزین،کارآمد و موثر برای روش‌های کلاسیک، برای شناسایی منشأ جغرافیایی زغال سنگ ارائه کردند.

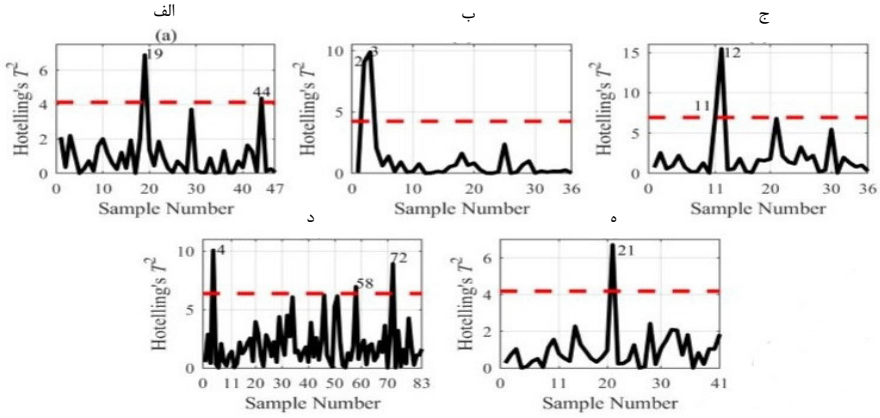
از اعمال تصحیح پراکنده میان ضربی (MSC) و متغیر استاندارد نرمال (SNV) بر روی داده‌ها جهت از بین بردن اثر پراکندگی استفاده شد. در تصحیح پراکنده میان ضربی طیف مرجع، میانگین همه طیف‌ها است و هر طیف نسبت به طیف مرجع تنظیم می‌شود. تصحیح پراکنده میان ضربی برای خذف تأثیر پراکندگی ناشی از اندازه ذرات مختلف زغال‌سنگ استفاده شده است. هنگام اعمال متغیر استاندارد نرمال، مقدار میانگین انحرافات و استاندارد همه نقاط داده برای طیف محاسبه می‌شود و هر طیف با استفاده از مقدار انحراف محاسبه شده تنظیم می‌شود. انحراف معیار با خطای استاندارد درواقع معیاری از پراکندگی اعداد است.

به دلیل خطای ابزار، دستگاه‌ها و نیروی انسانی، مجموعه داده‌ها اغلب حاوی داده‌های پرت و ناهنجار است که سبب انحراف می‌شوند. قابل توجه است که پایایی مدل‌سازی داده‌های حاصل از طیف‌سنجی فروسرخ نزدیک نمونه‌های زغال سنگ مختلف کاملاً وابسته به نوع و مقدار داده‌ها است و بنابراین تشخیص داده‌های پرت، گام مهمی قبل از نهایی کردن مدل است. در این مطالعه، آن‌ها از روش هتلینگ مربع (T-squared)، T برای تشخیص داده‌های پرت استفاده کردند. درواقع از (T-squared) هتلینگ برای مقایسه داده‌های چند متغیره استفاده می‌شود. در آمار، به‌ویژه در آزمون فرضیه، توزیع T مربعی هتلینگ، پیشنهاد شده توسط هارولد هتلینگ، توزیع احتمالی چند متغیره است که به‌شدت به توزیع F مربوط می‌شود. پارامتر $T^2$ به این صورت تعریف می‌شود [۱۴]:

$$T_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{P}_p \boldsymbol{\Lambda}_p^{-1} \boldsymbol{P}_p{}^{\mathrm{T}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \qquad (۱)$$

## ۴-۱ کاربرد یادگیری ماشین در صنایع غذایی

در سال‌های اخیر، تنوع و حجم داده‌های به‌دست آمده توسط ابزارهای تحلیلی مدرن برای اطمینان از کیفیت بهتر غذا به‌طور چشمگیری افزایش یافته است. چندین الگوریتم مختلف به‌عنوان ابزارهای شناسایی برای مقابله با حجم زیاد و پیچیدگی داده‌ها تولید شده‌اند. متداول‌ترین الگوریتم‌ها شامل الگوریتم تجزیه‌وتحلیل مولفه‌های اصلی، الگوریتم تجزیه‌وتحلیل حداقل




شکل ۱ نمودار $T^2$ هتلینگ برای (الف) نمونه زغال سنگ استرالیا، (ب) نمونه زغال سنگ روسیه، (ج) نمونه‌های زغال سنگ کانادایی، (د)نمونه‌های ذغال سنگ اندونزی و(ه) نمونه‌های زغال سنگ چینی را نشان می‌دهد.

مربعات جزئی، مدل‌سازی مستقل نرم بر اساس طبقه‌بندی مقایسه‌ای (SIMCA)، الگوریتم k نزدیک‌ترین همسایه (kNN)، الگوریتم تجزیه‌و‌تحلیل عامل موازی و الگوریتم ترکیبی قدرت تفکیک منحنی چندمتغیره–حداقل مربعات متناوب (MCR-ALS) است. با این حال الگوریتم‌های جایگزین برای بهبود داده‌ها، مانند الگوریتم ماشین بردار پشتیبان، طبقه‌بندی و درخت رگرسیون و جنگل تصادفی مزایای بیش‌تری نشان دادند.

در این زمینه مقاله منتشر شده توسط چنگ و همکاران حائز اهمیت است. در آن نویسندگان قبل از استفاده از الگوریتم ماشین بردار پشتیبان برای طبقه‌بندی انواع نوشیدنی‌های تولید کشور چین دو روش کاهش داده با استفاده از PCA و PLS را به کار بردند. به‌طورکلی، محققان نتیجه گرفتند که کاهش داده‌ها توسط الگوریتم PLS بهترین روش کاهش داده قبل از استفاده از الگوریتم ماشین بردار پشتیبان برای بررسی نوشیدنی‌ها است [۱۵].

یکی از مهمترین کارها در اندازه‌گیری دانه‌های کاکائو توسط Teye و همکارانش انجام شد. آن‌ها برای مدل‌سازی و طبقه‌بندی داده‌ها از الگوریتم‌های ماشین بردار پشتیبان و K همسایه نزدیک استفاده کردند. نتایج نشان داد که استفاده از الگوریتم ماشین بردار پشتیبان بهتر از الگوریتم K همسایه نزدیک است و آن‌ها با استفاده از این الگوریتم موفق شدند ۱۰۰٪ نمونه‌ها را به درستی طبقه‌بندی کنند.

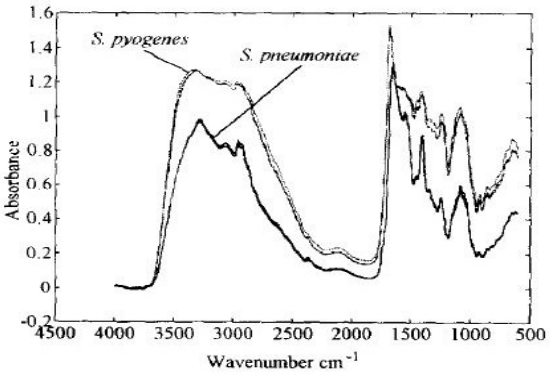
کیفیت چای به منطقه جغرافیایی برداشت چای و نوع بذر اولیه بستگی زیادی دارد. در میان مطالعات تجزیه‌ای انجام شده بر روی چای، مقاله منتشر شده توسط لیو و همکاران با ارزش است. آن‌ها از روش تجزیه‌ای الکتروشیمیایی ولتامتری برای تعیین کیفیت نمونه‌های چای استفاده کردند و برای مدل‌سازی و طبقه‌بندی داده‌های حاصل از ولتامتری از الگوریتم ماشین بردار پشتیبان استفاده کردند و موفق شدند داده‌های حاصل از نمونه‌های چای را با دقت بالایی طبقه‌بندی کنند.

شناسایی و طبقه‌بندی میکروارگانسیم‌ها مانند باکتری‌ها در مطالعات بالینی بسیار اهمیت دارد و یافتن روشی که بتواند این کار را با سرعت و دقت زیاد انجام دهد، بسیار ارزشمند است. Royston Goodacre و همکارانش با استفاده طیف فروسرخ باکتری و بهره‌گیری از شبکه عصبی مصنوعی، موفق شدند گونه‌های استرپتوکوک و انتروکوک را شناسایی و طبقه‌بندی کنند. آن‌ها از نرم افزار Matlab نسخه ۴,۲. ۱ برای مدل‌سازی استفاده کردند. در شکل زیر طیف جذب–بازتاب فروسرخ نمونه‌ها نشان داده شده است [۱۷].

## ۴–۲ اندازه‌گیری فلزات سنگین

امروزه به‌دلیل نگرانی‌های زیست‌محیطی اندازه‌گیری کمی نمونه‌ها سریع و در محل مورد نیاز است. طیف‌سنجی فروشکست القایی لیزری، روشی کارآمد برای رفع این مسئله است و به‌ویژه برای اندازه‌گیری عنصرهای فلزات سنگین کارآمد است. باید درنظر داشت که اثرات زمینه و ماتریس نمونه‌ها سبب پیچیدگی اندازه‌گیری می‌شوند. اندازه‌گیری کمی نمونه‌ها توسط طیف‌سنجی شکست لیزر از طریق منحنی کالیبره‌کردن انجام می‌شود.

از شبکه‌های عصبی مصنوعی برای کالیبراسیون غیرخطی و چند متغیره استفاده می‌شود. در سال ۲۰۱۳ J.El Haddad و همکارانش با استفاده از طیف‌سنجی فروشکست القایی لیزری و شبکه عصبی مصنوعی موفق شدند عناصر مس، آلومینیوم، کلسیم و آهن نمونه خاک را به سرعت و با دقت نیاز به‌طور کمی




شکل ۲ طیف جذب-بازتاب فروسرخ نمونه‌ها را نشان می‌دهد.

پیش‌بینی کنند. نمونه‌های خاک از یک منطقه در جنوب فرانسه و با استفاده از نمونه‌گیر هوشمند، دستگاه (XRF) قابل‌حمل مدل NitonXL3t800 نمونه‌برداری شدند. سپس نمونه‌ها در ماکروویو خشک شدند و با استفاده از الک صاف شدند و بعد آن‌ها را پرس کردند. هر نمونه به دو قسمت تقسیم شد، یک قسمت به مشخصه‌یابی مستقیم طیف‌سنجی شکست لیزر اختصاص داده شد و قسمت دیگر برای آزمون نشر پلاسمای جفت شده القایی (ICP-AES) برای اندازه‌گیری واقعی غلظت مورد استفاده قرار گرفت. در این مطالعه مدل‌سازی داده‌ها از طریق نرم‌افزار Igor Pro ٦,١١ انجام شد. مرحله اول شامل آموزش شبکه عصبی مصنوعی ۳ لایه با مقادیر تصادفی وزن‌ها بود. ثابت شده است که بین طیف شکست لیزر و غلظت نمونه، رابطه غیرخطی وجود دارد، بنابراین برای مدل‌سازی برای مثال از الگوریتم رگرسیون خطی نمی‌توان استفاده کرد و استفاده از الگوریتم‌های غیرخطی مناسب است. شبکه عصبی مصنوعی شبکه‌ای از نورون‌های بهم پیوسته است که توسط توابع فعال‌سازی غیرخطی مشخص می‌شود. از شبکه عصبی مصنوعی برای پیش‌بینی غلظت عناصر موجود در خاک استفاده شده است [۱۸].

## ۴-۳ کاربرد یادگیری ماشین در روش‌های الکتروشیمیایی

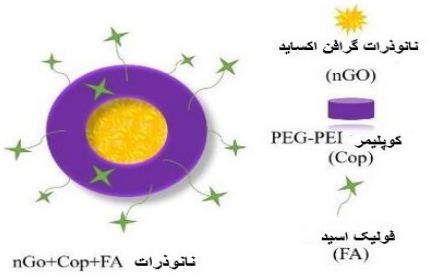
با استفاده از داده‌های حاصل از روش‌های الکتروشیمیایی مانند ولتامتری، پلاروگرافی و شبکه عصبی مصنوعی می‌توان مدل‌های بسیار کارآمد طراحی کرد. E.Cukrowska و همکارانش با استفاده از روش‌های الکتروشیمی مانند ولتامتری خطی و پلاروگرافی پالسی تفاضلی و شبکه عصبی مصنوعی موفق شدند مدلی دقیق برای پیش‌بینی مقدار آدنین و سیستئین در مخلوطی از آن‌ها طراحی کنند. میانگین خطای مطلق برای آدنین ۳/۷ درصد در روش ولتامتری، ٤/٦ درصد در روش پلاروگرافی و برای سیستئین ۵/۲ درصد در روش ولتامتری و ۵/۹ درصد در روش پلاروگرافی بدست آمد. در شکل زیر منحنی ولتاموگرام آدنین و سیستئین نشان داده شده است [۱۹].

## ۴-۴ تصویربرداری نوری از نانوذرات فلورسنت با استفاده از یادگیری ماشین

در سال ۲۰۱۸ Olga E. Sarmanova و همکارانش روش جدیدی برای اجرای تصویربرداری نوری از نانوذرات فلورسنت در یک محیط زیستی با استفاده از شبکه‌های عصبی مصنوعی پیشنهاد کردند. در این مطالعه آن‌ها از نانوکامپوزیت‌های سنتزشده جدید، نانوذرات اکسیدهای گرافن پوشش داده شده با لایه پلی (اتیلن ایمین) -پلی-(اتیلن گلیکول) کوپلیمر و اسیدفولیک استفاده کردند. درواقع آن‌ها، روش جدید برای تصویربرداری نوری نانوذرات در محیط زیستی با استفاده از نظارت بر حذف نانوکامپوزیت‌های فلورسنت و شبکه عصبی مصنوعی ارائه کردند. فعل و انفعالات مختلف بین زیست مولکول‌ها و نانوذرات، منجر به تغییر علائم فلورسانس و لومینسانس نانوذرات می‌شود و به دلیل فلورسانس زیاد بافت زیستی، استخراج علائم فلورسانس منفرد نانوذرات و بافت زیستی از طیف مشترک آن‌ها با استفاده از روش‌های کلاسیک، تقریباً غیرممکن به نظر می‌رسد.

در این مطالعه طیف‌های فلورسانس شامل بافت‌های زیستی بدون نانوذرات، نانوذرات خارج از بافت‌های زیستی و طیف‌های فلورسانس بافت‌های زیستی با نانوذرات با غلظت‌های مختلف اندازه‌گیری شدند و به‌عنوان داده‌های آموزشی و آزمایشی برای شبکه عصبی مصنوعی مورد استفاده قرار گرفتند. با توجه به چنین آموزشی شبکه عصبی مصنوعی، قادر به یادگیری چگونگی تقریبی تغییرات در طیف‌ها است که ناشی از تغییر طیف فلوئورسانس و فعل و انفعالات بین سطح نانوذرات و بافت زیستی است.

آن‌ها از اسیدفولیک به‌عنوان لیگاند برای سنتز نانوکامپوزیت‌ها استفاده کردند. اسیدفولیک برای رشد سلول‌های جدید از جمله سلول‌های آنکولوژی ضروری است. بنابراین تومورهای انواع خاصی از سرطان، گیرنده‌های غشایی آلفا برای اسیدفولیک دارند. در نتیجه تومور اسیدفولیک را از بدن می‌گیرد و برای ترمیم از آن استفاده می‌کند. در حضور اسیدفولیک گیرنده‌های سلول‌های سالم دردسترس نیستند و از اسیدفولیک معمولاً برای درمان سرطان استفاده می‌شود. شکل ۳ طیف‌های فلورسانس محلول‌های آبی nGO + Cop، nGO و nGO + Cop + FA در غلظت ۰/۰۱ گرم در لیتر را نمایش می‌دهد [۲۰].



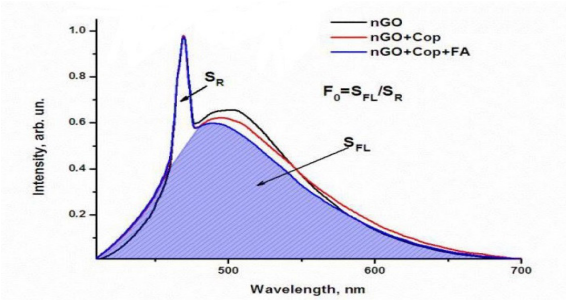

شکل ۳ نانوکامپوزیت‌های سنتز شده و اجزای آن‌ها را نشان می‌دهد.

## ۴-۵ سنتز مولکول‌های آلی جدید با استفاده از یادگیری ماشین

کاربرد مسائل بهینه‌سازی و جستجو در علوم و مهندسی از جمله طیف‌سنجی فراوان است. الگوریتم‌های ژنتیک با موفقیت در حل بسیاری از مسائل دشوار و بهینه‌سازی در انواع مختلف حوزه‌های تحقیقاتی شیمی از جمله طیف‌سنجی، فلورسانس، سنتز کمپلکس‌های معدنی، کمومتریکس و طیف‌سنجی فروسرخ نزدیک استفاده شده است. شکل زیر کاربرد الگوریتم ژنتیک را در سنتز ترکیبات آلی جدید نمایش می‌دهد. در ابتدا دو والد وجود دارد که ساختار ترکیبات آن‌ها شکسته می‌شود و در مرحله بعد انواع حالت‌های مختلفی که امکان دارد تا اجزای حاصل از قطعه‌قطعه شدن والدها باهم ترکیب شوند، مورد بررسی قرار می‌گیرد. سرانجام از بین فرزندان، آن‌هایی که پایدارتر و سازگارتر هستند باقی می‌مانند و خروجی الگوریتم را تشکیل می‌دهند [۲۱].

مولکول‌های آلی مزدوج بخش مهمی از ترکیبات شیمیایی برای اکتشافات علمی اساسی و کاربردهای فنی را شامل می‌شوند. غربالگری مولکول‌های محاسباتی در بسیاری از زمینه‌های تحقیقاتی از کشف دارو گرفته تا کشف مواد جدید در بسیاری از صنایع اهمیت ویژه‌ای دارد. الگوریتم ژنتیک، روش بهینه‌سازی مبتنی بر مفاهیم حاصل از تکامل است که در آن جمعیت کروموزوم‌ها در نسل‌های متوالی با استفاده از تکامل، جهش و انتخاب بهینه شده است. Ilana Y. Kanal و همکارانش با استفاده از الگوریتم ژنتیک موفق شدند ساختار شیمیایی تعداد زیادی مولکول آلی را پیش‌بینی کنند.

در مطالعه آن‌ها کروموزوم‌ها شامل ٦٤ الیگومر بودند و نسل‌های متوالی دارای کم‌ترین مقدار انرژی انتقال بودند. شبیه‌سازی با استفاده از الگوریتم ژنتیک و با استفاده از نرم‌افزار پایتون با Numpy و Pandas31modules انجام شد. در شکل ٤ توالی‌های ترامر در الگوریتم ژنتیک برای ترکیب دو مونومر

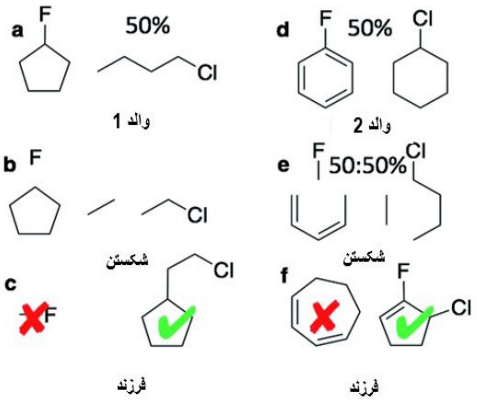



شکل ٤ طیف‌های فلورسانس محلول‌های آبی nGO، nGO+Cop و nGO+Cop+FA در غلظت ۰/۰۱ گرم در لیتر [۲۰].

نشان داده شده است. همچنین شکل ۵ جفت‌های مونومر که از ترامرها حاصل شدند را نمایش می‌دهد [۲۲].

## ۴-۶ شناسایی انواع سوخت‌های زیستی با استفاده از یادگیری ماشین

در سال ۲۰۱۹ Hong Ge و همکارانش روشی برای شناسایی انواع سوخت‌های زیستی با استفاده از طیف‌سنجی نشر شعله و الگوریتم‌های درخت ارائه کردند. آن‌ها ویژگی‌های مختلف از جمله طیف‌های نشری شعله از جمله رادیکال‌های شعله‌ای با شدت طیف‌های گروه‌های عاملی CH، CN، OH و C2، تشعشع شعاع شعله و تشعشع انرژی شعله را اندازه‌گیری کردند. مدل‌سازی با استفاده از چهار الگوریتم درختی به‌عنوان مثال درخت تصمیم، جنگل تصادفی، درختان بسیار تصادفی و درخت تصمیم Gradientboost ساخته شدند. چهار سوخت زیستی مختلف، از جمله چوب ذرت، بید، پوسته بادام زمینی و کاه گندم مورد استفاده قرار گرفتند. نتایج نشان داد که مدل‌های



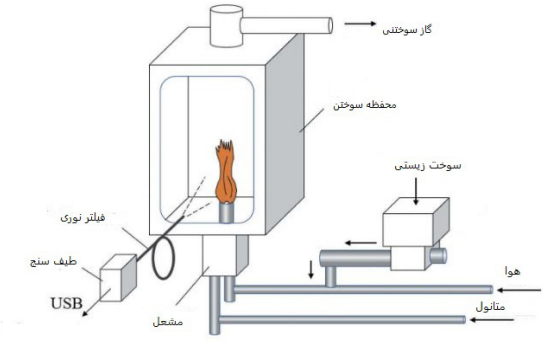

شکل ۵ کاربرد الگوریتم ژنتیک در سنتز ترکیبات آلی جدید [۲۱].

شناسایی ارائه شده قادر به شناسایی صحیح انواع سوخت‌های زیستی با متوسط میزان موفقیت شناسایی ۹۸٪ در ۱۰ آزمایش است [۲۳].

## ۴-۷ کاهش الکتروشیمیایی گاز $CO_2$

در سال ۲۰۱۸ M. ErdemGünay و همکارانش موفق شدند با استفاده از الگوریتم درخت تصمیم مدلی برای کاهش الکتروشیمیایی گاز $CO_2$ خارج شده از سامانه‌ها را ایجاد کنند. آن‌ها پایگاه داده‌ای از ٤٧١ نقطه داده تجربی استخراج شده از ٣٤ نشریه مختلف در مورد کاهش الکتروکاتالیستی $CO_2$ را جمع‌آوری کردند. در مرحله اول پایگاه داده با استفاده از نمودار جعبه و نمودارهای سبیل بررسی شد. سپس برای تعیین اهمیت متغیرها و تجزیه‌وتحلیل داده‌ها از الگوریتم درخت تصمیم برای تعیین شرایط منجر به کارایی بهتر واکنش الکتروشیمی، میزان تولید و انتخاب محصول استفاده شد. مشخص شد که محتوای مس کم‌تر از ۷۱٪، نوع کاتولیت، پتانسیل اعمال شده و pH الکترولیت باعث افزایش بازده واکنش می‌شود. در این مورد مقدار پتانسیل اعمال شده و محتوای مس بیش‌ترین اهمیت را در بین تمام متغیرهای ورودی داشتند. برای مدل‌سازی با استفاده از الگوریتم درخت تصمیم، از نرم افزار MATLAB (نسخه شماره ٨،٦) استفاده کردند.

آن‌ها به این نتیجه رسیدند که از میان محصولات تولید شده حاصل از واکنش الکتروشیمیایی دی‌اکسیدکربن CO بیش‌ترین مقدار را نسبت به ترکیبات $C_2H_6$، $C_2H_4$، HCOOH، CH4، و $CH_3OH$ دارد و توسط نشانگرهای X نشان داده شده است. آن‌ها با تغییر غشاهای مورد استفاده به این نتیجه رسیدند که استفاده از غشای Nafion کم‌ترین کارایی واکنش را ایجاد می‌کند ولی استفاده از غشای Membrane در مقایسه با غشاهای دیگر کارآیی بیش‌تری دارد. وقتی KOH برای کاتولیت و NaOH برای آنولیت در الکترولیزور استفاده شد ، روند مشابهی مشاهده شد. در شکل ٦ نمودارهای جعبه واکنش الکتروشیمیایی دی‌اکسیدکربن مربوط به تغییر نوع محصولات، نوع سل الکتروشیمی، نوع غشا، نوع آنولیت و کاتولیت نشان داده شده است [٢٤].




شکل ٦ طرح‌واره سامانه اندازه‌گیری طیف نشر سوخت‌های زیستی [٢٣].

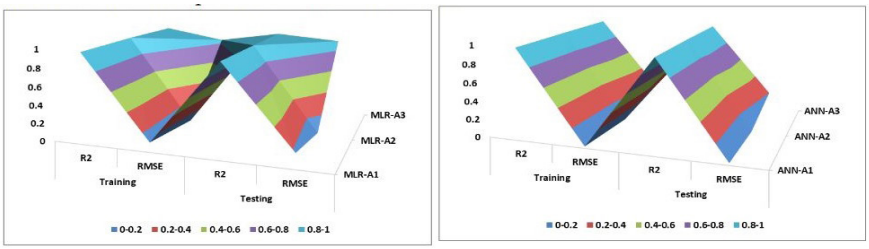
## ۴-۸ اندازه‌گیری غلظت‌های مختلف ایزوکوئرسیتین

اخیراً Hadiza Usman Abdullahi و همکارانش عملکرد دو مدل مختلف با استفاده از الگوریتم شبکه عصبی مصنوعی غیرخطی و الگوریتم رگرسیون چندخطی کلاسیک را برای شبیه‌سازی داده‌های جذب، مورد مقایسه قرار دادند. داده‌های جذب اندازه‌گیری شده از تجزیه‌وتحلیل با استفاده از دستگاه سوانگاری مایع با عملکرد بالا (HPLC) حاصل شدند و از غلظت‌های مختلف ایزوکوئرسیتین به‌عنوان متغیرهای ورودی استفاده شد. نتیجه شبیه‌سازی براساس دو معیار مختلف کارایی عملکرد یعنی ضریب تعیین($R^2$) و خطای مربع ریشه (RMSE) مورد ارزیابی قرار گرفتند. نتایج به‌دست آمده نشان داد که در توانایی مدل‌ها در شبیه‌سازی داده‌های حاصل از سوانگاری ترکیب ایزوکوئرسیتین دارای حداقل مقدار خطا ۰/۹۵۳۸ برای الگوریتم رگرسیون چند خطی کلاسیک و حداکثر مقدار خطا ۰/۹۹۹۳ برای الگوریتم شبکه عصبی مصنوعی در مرحله آزمایش است [۲۵].

## ۴-۹ یادگیری ماشین و پلیمرها

از یادگیری ماشین برای پیش‌بینی پارامترهای مختف پلیمرها استفاده می‌شود. توانایی طراحی کارآمد پلیمرهای دی‌الکتریک جدید و پیشرفته به‌دلیل کمبود اطلاعات کافی و قابل اعتماد در مورد فضای گسترده شیمیایی پلیمرها و مشکل تولید چنین داده‌های وسیعی با توجه به محدودیت‌های زمانی و محاسباتی/تجربی، مختل می‌شود. در سال ۲۰۱٦ Arun Mannodi-Kanakkithodi و همکارانش با استفاده از مدل‌های یادگیری ماشین و با استفاده از داده‌های تولید شده، به مسئله تسریع در طراحی دی‌الکتریک‌های پلیمری پرداختند. آن‌ها از الگوریتم ژنتیک برای بهینه‌سازی بلوک‌های تشکیل‌دهنده پلیمر به روش تکاملی استفاده کردند وموفق به طراحی پلیمرهایی با خواص دی‌الکتریک شدند [٢٦].

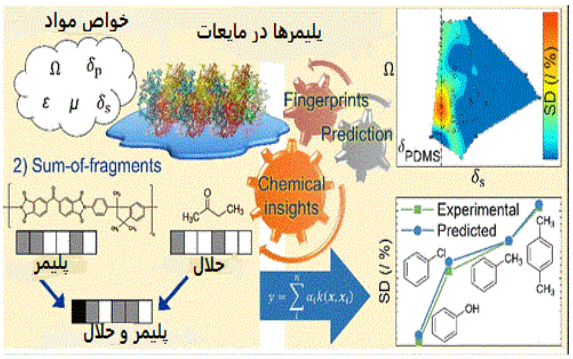
تورم در مایعات برای پلیمرهای مورد استفاده در بسیاری از کاربردهای فاز مایع از اهمیت فوق‌العاده‌ای برخوردار است. این ویژگی حیاتی انگیزه بسیاری از نظریه‌های تحلیلی، آزمایش‌های تجربی و شبیه‌سازی‌های اتمی اخیر را ایجاد کرده است. با این




شکل ۷ نمودار سطح نشان دهنده سطح $R^2$ و RMSE برای مدل‌های الف) شبکه عصبی مصنوعی و ب) رگرسیون چند متغیره کلاسیک

حال، در حال حاضر، روش داده محوری برای تورم پلیمر در دسترس نیست. اخیراً Qisong Xu و همکارانش از روش یادگیری ماشین برای بررسی تورم پلیمر در مایعات استفاده کردند. این روش برای تورم غشاهای نانو فیلتراسیون حلال آلی(OSN) و پلی‌دی‌متیل‌سیلوکسان (PDMS) در حلال‌های مختلف استفاده شد. ابتدا توصیفگرهای شیمیایی مانند پارامترهای حلالیت و خصوصیات حلال برای ساخت مدل‌های یادگیری ماشین پیشنهاد شد. آن‌ها با استفاده از الگوریتم رگرسیون، مدلی بر اساس پارامترهای حلالیت حلال و پلیمر طراحی کردند که بهترین پیش‌بینی کمی و همچنین رفتار تورم را برای غشاهای OSN نشان می‌دهد [۲۷].

طراحی معکوس چالش بزرگ در سامانه درشت‌مولکول‌ها مانند پلیمرها، به ویژه هنگام طراحی پلیمرهایی با رفتار فاز دلخواه است. در سال ۲۰۱۹ Gutin N. Kuma R و همکارانش با استفاده از یادگیری ماشین موفق به تنظیم دقیق نقطه ابری پلیمر پلی(۲-اگزازولین) شدند. آن‌ها با طراحی فضای چهار واحد تکرارشونده و طیف وسیعی از توده‌های مولکولی با دقت ٤ درجه سانتی‌گراد، خطای مربع میانگین را در یک محدوده دمایی ۲٤-۹۰ درجه سانتی‌گراد با استفاده از الگوریتم تقویت گرادیان با درختان تصمیم به‌دست آوردند. مدلی که آن‌ها ارائه دادند، توانایی کشف سریع و سامانه‌ای پلیمرهای جدید را دارد [۲۸].

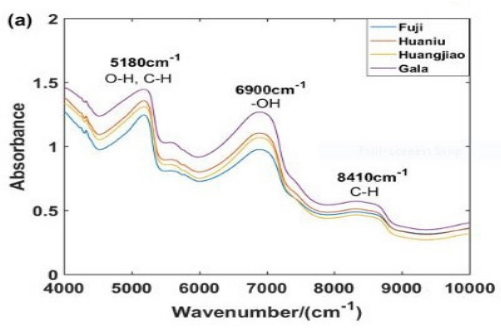



شکل ۸ استفاده از روش یادگیری ماشین برای بررسی تورم پلیمر در مایعات [۲۸].

تعیین درجه حرارت انتقال شیشه پلیمرها، $T_g$، به‌عنوان ویژگی مهم ترموفیزیکی، گاهی اوقات ممکن است به‌صورت تجربی دشوار باشد. روش‌های مدل‌سازی به ویژه رویکردهای داده محور، گزینه‌های امیدوارکننده برای پیش‌بینی $T_g$ به روشی سریع و قوی هستند. گشتاور چهار قطبی بدون ردیابی مولکولی و میانگین گشتاور هگزادکاپول مولکولی، با $T_g$ پلیمرها ارتباط نزدیک دارند. اخیراً Yun Zhang و همکارانش از این دو پارامتر به‌عنوان توصیف‌کننده مدل رگرسیون برای پیش‌بینی $T_g$ پلیمرها استفاده کردند و موفق شدند مدلی برای تخمین سریع و کم هزینه $T_g$ پلیمرها با دقت و ثبات بالا ارائه دادند [۲۹].

## ۴-۱۰ طبقه‌بندی نمونه‌های سیب با استفاده از الگوریتم خوشه‌بندی

سیب به‌عنوان محصول مهم کشاورزی دارای ارزش غذایی فوق‌العاده بالایی است. اخیراً Xiaohong Wu و همکارانش به‌منظور تشخیص سریع، دقیق و غیرمخرب انواع سیب، از الگوریتم خوشه‌بندی بهبودیافته (IPGG) برای طبقه‌بندی طیف‌های بازتابش فروسرخ نمونه‌های سیب ارائه کردند. در این پژوهش از طیف‌سنج فروسرخ آنتاریس شرکت ترمو الکترون آمریکا برای جمع‌آوری طیف فروسرخ چهار نوع سیب (فوجی، هوانیو، گالا و هوانگجیائو) استفاده شد. سپس برای از بین بردن




شکل ۹ طیف‌های مادون قرمز چهار نوع سیب مختلف [۳۰].

اطلاعات زائد و کاهش ابعاد طیفی به‌ترتیب اصلاح پراکندگی چندگانه (MSC) و تجزیه‌وتحلیل مولفه اصلی (PCA) استفاده شد. سرانجام الگوریتم خوشه‌بندی IPGG بر روی داده‌های طیفی پیش‌پردازش شده اجرا شدند. نتایج نشان داد که دقت خوشه بندی IPGG بالا بود و به ۹۶/۵ درصد رسید. نتایج تجربی نشان داد که طیف‌سنجی فروسرخ همراه با خوشه‌بندی MSC،PCA و IPGG روشی موثر برای شناسایی انواع سیب بود [۳۰].

## ۵ نتیجه‌گیری

امروزه از داده‌های تجزیه‌ای مختلف و الگوریتم‌های گوناگون برای مدل‌سازی مسائل مختلف استفاده می‌شود. داده‌های حاصل از روش‌های تجزیه‌ای مانند طیف‌سنجی فروسرخ، طیف سنجی، سوانگاری، فلورسانس، ولتامتری، پلاروگرافی، طیف‌سنجی نشر پلاسما، میکرو استخراج فاز جامد و ... پیش‌پردازش می‌شوند و پس از آماده‌سازی توسط الگوریتم‌های مختلف یادگیری ماشین مدل‌سازی می‌شوند. مدل‌های به‌دست آمده به این طریق از دقت بالایی برخوردار هستند. مدل‌سازی، مزایای بسیاری دارد و باعث صرفه‌جویی در وقت و هزینه، دقت بالا، سرعت بالا، تکرارپذیری خوب می‌شود. انجام برخی از مسائل پیچیده به تجهیزات عظیم و بسیار پرهزینه نیاز دارد که با استفاده از مدل‌سازی و شبیه‌سازی می‌توان بر این مشکل غلبه کرد.

# مراجع


1. Goodfellow I., Bengio Y., Courville A., Machine Learning Basics, *Deep learning* 1, **2016**.

2. Zhang Xian-Da., Machine learning., *A Matrix Alg Ebra Approach to Artificial Intelligence*. Springer, Singapore, 223-440, **2020**.

3. Bishop N., Christopher M., *Pattern Recognition and Machine Learning*. Springer, **2006**.

4. Szepesvári C., Algorithms for Reinforcement Learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning* 4, 1-103, **2010**.

5. Verbeeck, N., Richard M., Caprioli, and Raf Van de Plas., Unsu Pervised Machine Learning for Exploratory data Analysis in Imaging Mass Spectrometry, *Mass Spectrometry Reviews*, 3, 245-291, **2020**.

6. Leclerc P., et al., Machine Learning-based Prediction of Glioma Margin from 5-ALA Induced PpIX Fluorescence Spectroscopy, *Scientific Reports* 10, 1-9, **2020**.

7. Ye J.J., Chu-Hong L., and Xing-Jiu H., Analyzing the Anodic Stripping Square Wave Voltammetry of Heavy Metal Ions Via Machine Learning: Information Beyond a Single Voltammetric Peak, *Journal of Electroanalytical Chemistry*, 113934, **2020**.

8. Davari S A., and Anthony S., Quantification of Toxic Metals Using Machine Learning Techniques and Spark Emission Spectroscopy, *Atmospheric Measurement Techniques* 13, 5369-5377, **2020**.

9. Chen C.h., Joeska H., and Swen R., Predicting Fishiness off-flavour and Identifying Compounds of Lipid Oxidation in Dairy Powders by SPME-GC/MS and Machine Learning, *International Dairy journal* 77, 19-28, **2018**.

10. Haojie T., Zhao Y., Kang L., Electromagnetic Interference Diagnosis Based on HPLC Timing Sequence Topology and Machine Learning, *IOP Conference Series: Earth and Environmental Science*., 371, 5. IOP Publishing, **2019**.

11. Lebanov L., Random Forests Machine Learning Applied to Gas Chromatography–Mass Spectrometry Derived Average Mass Spectrum data Sets for Classification and Characterisation of Essential Oils, *Talanta* 208, 120471, **2020**.

12. Gastegger M., Jörg B., and Philipp M., Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra, *Chemical Science*, 8, 6924-6935, **2017**.

13. Bertani F. R.,Businaro L., Gambacorta L., Optical Detection of Aflatoxins B in Grained Almonds Using Fluorescence Spectroscopy and Machine Learning Algorithms, *Food Control,* 112, 107073, **2020**.

14. Yu X., Guo W., Zou L., Rapid Discrimination of Coal Geographical Origin via Near-infrared Spectroscopy Combined with Machine Learning Algorithms, *Infrared Physics & Technology*, 105, 103180, **2020**.

15. Cheng-Ming F., Genome-wide Expression Analysis of Soybean MADS Genes Showing Potential Function in the Seed Development, *PloS one,* 8.4, e62288, **2013**.

16. Ziqiang Yu., et al., Scalable Distributed Processing of K Nearest Neighbor Queries Over Moving Objects, *IEEE Transactions on Knowledge and Data Engineering,* 27.5, 1383-1396, **2014**.

17. Royston G., Timmins E.M., Rooney P.J., Rapid Identification of Streptococcus and Enterococcus Species using Diffuse Reflectance-absorbance Fourier Transform Infrared Spectroscopy and Artificial Neural Networks, *FEMS Microbiology Letters*, 14, 233-239, **1996**.

18. Haddad El., Josette., Artificial Neural Network for On-site Quantitative Analysis of Soils Using Laser Induced Breakdown Spectroscopy, *Spectrochimica Acta Part B: Atomic Spectroscopy*, 79, 51-57, **2013**.

19. Tutu H., Cukrowska E. M., Dohnal V., Havel J., Application of Artificial Neural Networks for Classification of Uranium Distribution in the Central Rand goldfield, South Africa, *Environmental Modeling & Assessment*, 10, 143-152, **2005**.

20. Sarmanova Olga E., A Method for Optical Imaging and Monitoring of the Excretion of Fluorescent Nanocomposites from the Body Using Artificial Neural Networks, *Nanomedicine: Nanotechnology, Biology and Medicine*, 14, 1371-1380, **2018**.

21. Kanal I. Y., and Geoffrey R., Hutchison, Rapid Computational Optimization of Molecular Properties using Genetic Algorithms: Searching Across Millions of Compounds for Organic Photovoltaic Materials, *Physics. Chem-ph,*1707.02949, **2017**.

22. Kanal I.Y., Owens S.G., Efficient Computational Screening of Organic Polymer Photovoltaics, *The Journal of Physical Chemistry Letters*, 4, 1613-1623, **2013**.

23. Hong Ge., Biomass Fuel Identification Using Flame Spectroscopy and Tree Model Algorithms ,*Combustion Science and Technology*, 1-18, **2019**.

24. Günay M., Erdem Lemi Türker., and Alper Tapan N., Decision Tree Analysis for Efficient $CO_2$ Utilization in Electrochemical Systems, *Journal of $CO_2$ Utilization,* 28, 83-95, **2018**.

25. Abdullahi H.U., Usman A.G., and Abba S.I., Modelling the Absorbance of a Bioactive Compound in HPLC Method Using Artificial Neural Network and Multilinear Regression Methods, *Dutse J Pure Appl Sci (DUJOPAS)*, 6.2, 362-371, **2020**.
26. Mannodi-Kanakkithodi A., Pilania G., Huan T.D., Machine Learning Strategy for Accelerated Design of Polymer Dielectrics, *Scientific Reports,* 6, 1-10, **2016**.
27. Qisong X., and Jianwen J., Machine Learning for Polymer Swelling in Liquids, *ACS Applied Polymer Materials*, 2, 3576-3586, **2020**.
28. Kumar J.N., Qianxiao L., Karen Y.T., Machine Learning Enables Polymer Cloud-point Engineering Via Inverse Design, *npj Computational Materials* 5,1-6, **2019**.
29. Zhang Y., and Xiaojie X., Machine Learning Glass Transition Temperature of Polymers, *Heliyon,* 6.10, e05055, **2020**.
30. Xiaohong W., Haoxiang Z., Bin W., Haijun., Determination of Apple Varieties by Near Infrared Reflectance Spectroscopy Coupled with Improved Possibilistic Gath–Geva Clustering Algorithm. *Journal of Food Processing and Preservation*, 44, 14561-14571, **2020**.