

روش های آماری در تحقیقات پزشکی 2016،  
جلد 25 (5) 1804-1823



انویسنده (گان) ۲۰۱۳ تجدید چاپ  
smm.sagepub.com

DOI: 10.1177/0962280213502437

/journalsPermissions.nav

sagepub.co.uk



# مقایسه روش های یادگیری ماشینی برای طبقه بندی با استفاده از شبیه سازی با چندین داده واقعی نمونه هایی از ذهنی مطالعات سلامت

میزانور خوندوکر<sup>۱،۲</sup>، ریچارد دابسون<sup>۲،۳</sup>، کارولین اسکیرو<sup>۴</sup>  
اندرو سیمونز<sup>۲،۳</sup> و دنیل استال برای طرح ابتکاری تصویربرداری عصبی بیماری  
آلزایمر\*

## چکیده

پیشینه: ادبیات اخیر در مورد مقایسه روش های یادگیری ماشین، سوالاتی را در مورد بی طرفی، بی طرفی و سودمندی بسیاری از مطالعات تطبیقی مطرح کرده است. گزارش نتایج روی مجموعه داده های مطلوب و خطای نمونه گیری در معیارهای عملکرد تخمینی بر اساس نمونه های واحد، منابع اصلی سوگیری در چنین مقایسه هایی تلقی می شوند. عملکرد بهتر در یک یا چند مورد لزوماً به معنای عملکرد بهتر در سطح میانگین یا در سطح جمعیت نیست و مطالعات شبیه سازی ممکن است جایگزین بهتری برای مقایسه عینی عملکرد الگوریتم های یادگیری ماشین باشند.

روش ها: ما عملکرد طبقه بندی تعدادی از الگوریتم های یادگیری ماشین مهم و پرکاربرد، یعنی جنگل های تصادفی (RF)، ماشین های بردار پشتیبان (SVM)، تحلیل تفکیک خطی (LDA) و k-نزدیک ترین همسایه (kNN) را مقایسه می کنیم. با استفاده از پردازش موازی گسترده بر روی ابررایانه های با کارایی بالا، خطاهای تعمیم را در ترکیبات مختلف سطوح چندین عامل مقایسه می کنیم: تعداد ویژگی ها، اندازه نمونه آموزشی، تنوع بیولوژیکی، تنوع تجربی، اندازه اثر، تکرار و همبستگی بین ویژگی ها.

<sup>۱</sup>کالج کینگز لندن، موسسه روانپزشکی، گروه آمار زیستی، لندن، انگلستان  
<sup>۲</sup>کالج کینگز لندن، موسسه روانپزشکی، مرکز تحقیقات زیست پزشکی NIHR برای سلامت روان در جنوب لندن و بنیاد NHS مادرلی، لندن، انگلستان

<sup>۳</sup>کالج کینگز لندن، موسسه روانپزشکی، واحد تحقیقات زیست پزشکی NIHR برای زوال عقل در جنوب لندن و بنیاد خدمات درمانی ملی مادرلی، لندن، انگلستان

<sup>۴</sup>کالج کینگز لندن، موسسه روانپزشکی، مرکز روانپزشکی اجتماعی، ژنتیکی و رشدی MRC، انگلستان

\* یکی از مجموعه داده های نمونه مورد استفاده در تهیه این مقاله از پایگاه داده ابتکار تصویربرداری عصبی بیماری آلزایمر (ADNI) ([adni.loni.ucla.edu](http://adni.loni.ucla.edu)) به دست آمده است. به همین ترتیب، محققان ADNI در طراحی و پیاده سازی ADNI مشارکت داشته و/یا داده ها را ارائه داده اند. اما در تجزیه و تحلیل یا نوشتن این گزارش مشارکتی نداشته اند. فهرست کاملی از محققان ADNI را می توانید در آدرس زیر بیابید:  
[http://adni.loni.ucla.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf)

نویسنده مسئول:

میزانور خوندوکر، گروه آمار زیستی، موسسه روانپزشکی، کالج کینگ لندن، لندن، انگلستان. ایمیل:

[mizanur.khondoker@kcl.ac.uk](mailto:mizanur.khondoker@kcl.ac.uk)

نتایج: برای تعداد کمتری از ویژگی‌های هم بسته، تعداد ویژگی‌هایی که تقریباً از نصف اندازه نمونه تجاوز نمی‌کنند، LDA از نظر میانگین خطاهای تعمیم و همچنین پایداری (دقت) تخمین خطا، روش انتخابی بود. SVM (با هسته RBF) با افزایش مجموعه ویژگی‌ها، به شرطی که اندازه نمونه خیلی کوچک نباشد (حداقل 20)، با اختلاف زیادی از LDA و همچنین RF و KNN بهتر عمل می‌کند. عملکرد KNN همچنین با افزایش تعداد ویژگی‌ها بهبود می‌یابد و از LDA و RF پیشی می‌گیرد، مگر اینکه تغییرپذیری داده‌ها خیلی زیاد باشد و/یا اندازه‌های اثر خیلی کوچک باشند. مشخص شد که RF فقط در برخی موارد که داده‌ها متغیرتر و اندازه‌های اثر کوچکتری دارند، از KNN بهتر عمل می‌کند، که در این موارد نیز تخمین‌های خطای پایداری نسبت به KNN و LDA ارائه می‌دهد. کاربردها در تعدادی از مجموعه داده‌های واقعی، یافته‌های مطالعه شبیه‌سازی را پشتیبانی می‌کنند.

#### کلمات کلیدی

یادگیری ماشین، اعتبارسنجی متقابل، خطای تعمیم، توزیع ناقص، ریزآرایه‌ها، الکتروانسفالوگرام (EEG)، تصویربرداری تشدیدمغناطیسی (MRI)

#### اپیشینه

پیشرفت‌های اخیر در فناوری‌های ژنومی، پروتئومیک، تصویربرداری عصبی و سایر فناوری‌های با توان عملیاتی بالا منجر به انفجار داده‌های با ابعاد بالا شده است که نیازمند توسعه روش‌های جدید یا اصلاح تکنیک‌های آماری و یادگیری ماشینی موجود برای به حداکثر رساندن بهره اطلاعاتی از چنین داده‌هایی است. افزایش تعداد روش‌های موجود، مقایسه روش‌ها را به طور منطقی ضروری کرده است تا بهترین روش در یک موقعیت خاص پیدا شود که منجر به انتشار مقالات متعددی با تمرکز بر مطالعات تطبیقی در ادبیات بیوانفورماتیک و زیست‌شناسی محاسباتی اخیر شده است. بخش بزرگی از این مطالعات، روش‌های آماری و یادگیری ماشینی تحت نظارت را برای طبقه‌بندی موضوعات، عمدتاً بر اساس بیان ژن ریزآرایه یا داده‌های طیف سنجی جرمی با ابعاد بالا، مقایسه کرده‌اند.<sup>۱-۱۰</sup>

ادبیات اخیر<sup>۱۱-۱۵</sup> در مورد این موضوع، سوالاتی در مورد بی‌طرفی، بی‌طرفی، سودمندی و روش‌های انجام اکثر این مقایسه‌ها مطرح شده است، زیرا بین یافته‌های چنین مطالعاتی اجماع کمی وجود دارد. مروری توسط بولستیکس و همکاران<sup>۱۱</sup> نشان دهنده تمایلی در برخی مطالعات تطبیقی برای نشان دادن برتری یک روش خاص با استفاده از مجموعه داده‌هایی است که به نفع روش انتخاب شده عمل می‌کنند. نگرانی‌های مشابهی در یک مطالعه اخیر نیز تکرار شد. بیوانفورماتیک مقاله یوسفی و همکاران<sup>۱۵</sup> پیشنهاد (i) گزارش نتایج روی مجموعه داده‌های مطلوب و (ii) به اصطلاح سوگیری چند قاعده‌ای که در آن چندین قانون طبقه‌بندی روی مجموعه داده‌ها مقایسه می‌شوند تا مزیت یک روش خاص را نشان دهند، زیرا منابع اصلی سوگیری در چنین مقایسه‌هایی هستند.

مطالعات تطبیقی وجود دارند که هدف آنها نشان دادن یک روش خاص به عنوان بهتر از بقیه نیست. چندین مطالعه از این دست، به اصطلاح خنثی‌مقایسه‌ها، توسط بولستیکس و همکاران ذکر شده‌اند.<sup>۱۱</sup> یکی از محدودیت‌های این مطالعات این است که مقایسه‌ها عمدتاً مبتنی بر مجموعه داده‌های واقعی هستند و مشکلی که در مقایسه عملکرد طبقه‌بندی تخمین زده شده روی مجموعه داده‌های واقعی وجود دارد، خطای نمونه‌گیری یا نویز در معیارهای عملکرد تخمین زده شده است. با توجه به این واقعیت که تخمین‌های عملکرد تابع تغییرپذیری نمونه‌گیری هستند،<sup>۱۲، ۱۵</sup> بهترین عملکرد در یک یا چند مورد لزوماً به معنای آن نیست که به طور متوسط یا در سطح جمعیت نیز همینطور باشد.

سه مسیر جایگزین را می‌توان به طور بالقوه برای مقایسه‌ای قوی‌تر و عینی‌تر بررسی کرد:

- (۱) استفاده از آزمون آماری برای در نظر گرفتن تغییرپذیری نمونه‌برداری یا نویز در تخمین‌های عملکرد، (۲) مقایسه تحلیلی بر اساس نظریه توزیع تخمین‌های عملکرد و (۳) تخمین مکرر معیار عملکرد روی تعداد زیادی از داده‌های شبیه‌سازی شده/مصنوعی برای میانگین‌گیری تغییرپذیری نمونه‌برداری یا نویز از تخمین

معیار عملکرد. هانزار و دوگرتی<sup>۱۲</sup> امکان استفاده از آزمون های آماری برای مقایسه عملکرد را بررسی کرد و نتیجه گرفت که مقایسه مستقیم بر اساس آزمون آماری غیر قابل اعتماد است و اغلب می تواند منجر به نتیجه گیری های اشتباه شود. مقایسه تحلیلی یک رویکرد زیبا خواهد بود اما نیاز به یافتن توزیع نمونه گیری از تخمین های عملکرد بر اساس توزیع مشترک متغیرهای ویژگی و کلاس دارد. با توجه به این واقعیت که بسیاری از الگوریتم های یادگیری ماشین مدرن پیچیده هستند (به اصطلاح جعبه سیاه اغلب مبتنی بر هیچ مدل آماری زیربنایی نیستند، تعیین ویژگی های توزیعی تحلیلی تخمین های عملکرد به جز برخی روش های کلاسیک تبعیض آماری مانند تحلیل تبعیض خطی (LDA) امکان پذیر نیست.<sup>۱۶</sup> همانطور که توسط هوا و همکاران نشان داده شده است،<sup>۱۷</sup> این امر، مسیر شبیه سازی را به عنوان ابزاری عملی و پایدار برای مطالعه عینی ویژگی های معیارهای عملکرد مبتنی بر الگوریتم های یادگیری با طیف وسیعی از پیچیدگی ها و اشکال، باز می گذارد. این امر در مطالعات دیگر در ادبیات مانند هانزار و دوگرتی نیز محقق شده است.<sup>۱۴</sup> که نتیجه گرفت که «مقایسه ی قواعد طبقه بندی در داده های واقعی بدتر از آزمایش های داده های مصنوعی است... قویاً پیشنهاد می کند که محققان مطالعات تطبیقی خود را روی مجموعه داده های مصنوعی انجام دهند».

ما همچنین فکر می کنیم که مطالعات شبیه سازی روشی مناسب و کاربردی برای فهمیدن این است که «کدام روش در چه شرایطی عملکرد بهتری دارد». حقیقت همیشه در داده های شبیه سازی شده مشخص است و بنابراین بررسی سوگیری، یعنی نزدیکی یک تخمین به حقیقت، که با داده های واقعی آسان نیست، آسان است. داده های مصنوعی همچنین امکان مطالعه خواص یک تخمین گر را در سطوح مختلف ویژگی های مختلف داده ها مانند تغییرپذیری، اندازه نمونه، اندازه اثر، همبستگی و غیره فراهم می کنند. با این حال، نقش داده های واقعی نیز مهم است و باید برای تکمیل تحقیقات مبتنی بر شبیه سازی استفاده شود، زیرا الگوها و ساختارها در داده های واقعی عموماً بسیار پیچیده تر هستند و هیچ مدل شبیه سازی نمی تواند الگوها، ابعاد و منابع تغییرپذیری در داده های تولید شده از یک سیستم بیولوژیکی واقعی را به طور کامل ثبت کند.

در این مطالعه، ما یک آزمایش شبیه سازی گسترده را برای مقایسه عملکرد طبقه بندی تعدادی از الگوریتم های یادگیری ماشین مهم و پرکاربرد، از کلاسیک ترین LDA گرفته تا ...، انجام می دهیم.<sup>۱۶</sup> به روش های مدرن مانند ماشین های بردار پشتیبان (SVM)<sup>۱۸-۲۰</sup> اگرچه مقایسه ها در مقالات عمدتاً بر روی داده های واقعی انجام شده است، اما داده های مصنوعی نیز تا حدودی قبلاً مورد توجه قرار گرفته اند.<sup>۳،۶،۷،۱۴</sup> با این حال، مطالعات شبیه سازی که قبلاً مورد استفاده قرار گرفته بودند، از نظر تعداد ویژگی های داده ها و پوشش آنها (فضای پارامتر) محدود بودند. کاربردهای طرح های چند عاملی برای آزمایش های شبیه سازی توسط اسکروندال مورد بحث قرار گرفت.<sup>۲۱</sup> در زمینه آزمایش های مونته کارلو. بررسی همزمان چندین عامل در سطوح مختلف، به بهبود اعتبار خارجی، یعنی میزان تعمیم نتایج به موقعیت های دیگر و به جمعیت، از نتایج مطالعه شبیه سازی کمک می کند. با استفاده از پردازش موازی گسترده بر روی ابررایانه های با کارایی بالا (مرکز محاسبات و داده های ادینبورگ، ECDF و مرکز تحقیقات زیست پزشکی NIHR برای خوشه لینوکس سلامت روان)، خطاهای تعمیم (خطاهای اعتبارسنجی متقاطع (CV) را برای تعداد زیادی از ترکیبات هفت عامل زیر ارزیابی و مقایسه می کنیم: تعداد متغیرها (ص)، حجم نمونه آموزشی (ن)، تنوع بیولوژیکی (یا بین آزمودنی ها) (-ب)، تغییرات درون آزمودنی (-ای)، اندازه اثر (تغییر برابر، -)، تکرار (رو همبستگی ( ) بین متغیرها. ما معتقدیم که هیچ تک سائز مناسب برای همه روش نوع در یادگیری ماشین، که مدت ها پیش توسط ولپرت محقق شد و<sup>۲۲</sup> در اطلاعیه اخیر تکرار شد و آنفورماتیک مقاله سرمقاله راک و همکاران.<sup>۱۴</sup> این نشان می دهد که هیچ روش طبقه بندی وجود ندارد که در همه شرایط از همه روش های دیگر بهتر عمل کند. انگیزه در نظر گرفتن چنین طیف وسیعی از عوامل، ارائه راهنمایی در مورد اینکه «کدام روش در چه شرایطی بهتر عمل می کند» است. ما همچنین یافته های خود را در مورد داده های شبیه سازی شده با ارزیابی عملکرد روی تعدادی از مجموعه داده های تجربی واقعی تولید شده از طیف وسیعی از پلتفرم های با توان عملیاتی بالا مانند داده های بیان ژن از ... تکمیل می کنیم.

ریزآرایه های DNA، داده های تصویربرداری عصبی از سیستم تصویربرداری تشدید مغناطیسی (MRI) با وضوح بالا، و داده های پتانسیل مرتبط با رویداد (ERP) که فعالیت مغز را از سیستم الکتروانسفالوگرام (EEG) اندازه گیری می کنند. برنامه شبیه سازی (simData) و برنامه تخمین عملکرد (classificationError) به عنوان بخشی از بسته R به نام optBiomarker ارائه شده اند که از شبکه جامع بایگانی (<http://www.cran.r-project.org/web/packages/optBiomarker/>) قابل دسترسی است.

## ۲ روش

### ۲.۱ روش های طبقه بندی

ما عملکرد طبقه بندی تعدادی از روش های طبقه بندی پرکاربرد را که مبتنی بر طیف متنوعی از الگوریتم ها و معماری ها هستند، یعنی درخت تصمیم و روش مبتنی بر نمونه گیری مجدد، جنگل های تصادفی، RF، مقایسه می کنیم. الگوریتم یادگیری مبتنی بر هسته، SVM<sup>۱۸-۲۰</sup> روش کلاسیک با انگیزه آماری، LDA<sup>۲۳-۲۶</sup> و الگوریتم مبتنی بر نمونه (نزدیک ترین نمونه های آموزشی) k-نزدیک ترین همسایه، kNN<sup>۲۷،۲۸</sup>.

### ۲.۲ بهینه سازی پارامترهای تنظیم

اکثر الگوریتم های طبقه بندی پارامترهای تنظیم کننده ی مخصوص به خود را دارند که در حالت ایده آل نیاز به بهینه سازی روی هر مجموعه داده ای که روش ها روی آن اعمال می شوند، دارند. ما پارامترهای مهم قابل تنظیم توسط کاربر را برای هر روش روی هر مجموعه داده ی شبیه سازی شده با استفاده از جستجوی شبکه ای روی فضاهای پارامتر ارائه شده، بهینه می کنیم. فضاهای جستجوی ما برای پارامترهای تنظیم کننده همیشه شامل مقادیر پیش فرض نرم افزار برای پارامترهای مربوطه بودند تا اطمینان حاصل شود که تخمین های عملکرد در پارامترهای بهینه شده حداقل به خوبی مقادیر در گزینه های پیش فرض هستند.

پیشنهاد می شود که روش RF نسبت به تغییرات پارامترهای تنظیم خود کاملاً مقاوم باشد. لئو بریمن (در راهنمای برنامه اصلی FORTRAN برای RF) پیشنهاد کرد که متری (تعداد متغیرهایی که به عنوان کاندید در هر گره استفاده می شوند) تنها پارامتری است که نیاز دارد

مقداری قضاوت برای تعیین. ما بهینه سازی می کنیم متری با استفاده از جستجوی شبکه ای روی یک نمونه تصادفی با اندازه ۵، شامل از مقدار پیش فرض (متری  $\frac{1}{4}$  پ-۱، از دنباله ف، ...،  $\frac{1}{2}$  پ-۲ گرم که در آن نمادگذاری [ایکس] نشان دهنده تابع کف، که بزرگترین عدد صحیح غیر بزرگتر از را می دهد ایکس. برای ص ۱۰۵، دنباله ف، ...،  $\frac{1}{2}$  پ-۲ گرم کمتر از پنج عنصر خواهد داشت که در این صورت جستجوی شبکه ای را روی خود دنباله انجام می دهیم. این موضوع در مقالات پیشنهاد شده است (مثلاً دیاز-اوریار و آلوارز د آندرس)<sup>۲۹</sup> که برای عملکرد عالی، نیاز کمی به تنظیم دقیق سایر پارامترهای RF وجود دارد. با این حال، ما بهینه سازی را در نظر گرفتیم. اندازه گره (حداقل اندازه گره های ترمینال) و نتری (تعداد درختان برای رشد) علاوه بر متری پارامتر. پارامتر گره بندی طول درختان را به صورت no node با تعداد موارد کمتر از ... کنترل می کند. گره بندی تقسیم خواهد شد. اندازه های گره کوچکتر معمولاً دقت بهتری ارائه می دهند و اندازه گره بزرگتر برای مجموعه داده های بزرگتر، معمولاً با هزینه کاهش اندک دقت، مزیت محاسباتی ایجاد می کند. با این حال، ما جستجوی شبکه ای را روی فضای پارامتر اجرا می کنیم. ف  $\frac{1}{4}$  پ-۲، ...، ۵ گرم برای هر مجموعه داده شبیه سازی شده برای یافتن مقدار بهینه اندازه گره پارامتر آن تری تعداد درختانی که قرار است در جنگل کاشته شوند را کنترل می کند. به طور شهودی، جنگل بزرگتر (درختان بیشتر) باید برای پایداری بهتر باشد. در اینجا، ما تنظیم می کنیم آن تری برای دقت بهینه در فضای پارامتر ف ۵۰۰، ۱۰۰، ۵۰، ... اگر کم که شامل مقدار پیش فرض (500) randomForest است.

ما SVM را با هسته های خطی، چندجمله ای و تابع پایه شعاعی (RBF) ارزیابی کردیم و RBF را برای همه محاسبات انتخاب کردیم زیرا مشخص شد که در شبیه سازی ما نسبت به هسته چندجمله ای کمتر بایاس دارد و همچنین به این دلیل که از هسته خطی عمومی تر است. این با استفاده از آرگومان هسته (kernel) انتخاب می شود. ۴٪ شعاعی) تابع R svm. عملکرد SVM با هسته RBF ممکن است به ... بستگی داشته باشد. هزینه (ثابت C جمله جریمه در ضرایب لاگرانژ) و

گاما(پارامتر عرض معکوس تابع هسته RBF). هزینه پارامتر حاشیه بردارهای پشتیبان را کنترل می کند - مقدار کوچکتر، جریمه خطاهای حاشیه را کاهش می دهد (نقاط جریمه نزدیک به مرز را نادیده می گیرد) و از این رو حاشیه طبقه بندی را افزایش می دهد. مقدار گاما انحنای مرز تصمیم گیری را کنترل می کند - مقادیر بالاتر، مرز تصمیم گیری را انعطاف پذیرتر (غیرخطی) می کند. ما بهینه می کنیم هزینه و گاما پارامترها با استفاده از جستجوی شبکه ای روی فضاهایی متشکل از نمونه های یکنواخت با اندازه ۵ (شامل مقادیر پیش فرض) از محدوده های (۱/۱۰، ۱) و (۱۰، ۱۰۰)، به ترتیب. ممکن است اشاره شود که فضاهای پارامتری طوری انتخاب می شوند که بین ... قرار گیرند.  $p=1$  پیش فرض و 10 -پیش فرض که در آن مقادیر پیش فرض برای هزینه و گاما پارامترها ۱ و ۱/ هستند، به ترتیب. kNN نیاز دارد ص

فقط تافی فی نینگک (تعداد

نزدیکترین همسایه ها) که ما در طول محدوده بهینه می کنیم، ۲، ... ن گرم ما از اعتبارسنجی متقاطع fold-10 استفاده می کنیم که از نظر محاسباتی نسبت به اعتبارسنجی متقاطع leave-one-out برای بهینه سازی همه پارامترها، کم زحمت تر است.

### ۲.۳ برآوردهای عملکرد

ما عملکرد روش ها را از نظر خطای طبقه بندی، حساسیت و ویژگی مقایسه می کنیم. ما از اعتبارسنجی متقابل out-leave-one برای تخمین این معیارهای عملکرد استفاده می کنیم. این انتخاب بر اساس ارزیابی و مقایسه چندین تخمین گر، یعنی CV leave-one-out، CV 10-fold، bootstrap 0.632 plus و انجام شده است. ۳۰ تخمین گرها. تخمین گرها از نظر بایاس و تغییرپذیری با استفاده از شبیه سازی برای نمونه های آموزشی با اندازه های مختلف مقایسه شدند. ما اندازه های نمونه های آموزشی (ن) از ۱۰ تا ۲۵۰ متغیر است، اما هر چهار تخمین گر را نمی توان برای کل بازه مقایسه کرد زیرا تخمین گرهای بوت استرپ و ۰.۶۳۲ به علاوه بوت استرپ اغلب برای مقادیر کوچک تر منجر به مشکلات محاسباتی می شوند. ن به دلیل ماهیت بسیار نامتوازن داده های باز نمونه گیری شده. برای اینکه برآوردهای مبتنی بر بوت استرپ در مقایسه باقی بمانند، اندازه نمونه را حداقل به ۵۰ محدود کردیم و مقایسه را بر اساس میانگین عملکرد در ۱۰ اندازه نمونه آموزشی مختلف انجام دادیم. (ن ۵۰٪، ۶۰، ...، ۱۴۰) نشان می دهد که برآوردهای مبتنی بر بوت استرپ کمتر اریب هستند، اما حتی برای مقادیر متوسط (ن

۵۰) اندازه نمونه های آموزشی (داده ها نشان داده نشده اند). همانطور که قصد داریم مقایسه کنیم روش هایی برای طیف وسیعی (با ن اندازه نمونه ای به کوچکی ۱۰)، ما بین برآوردهای واریانس یک طرفه (one-out-leave) و واریانس ده طرفه (fold CV-10) یکی را انتخاب می کنیم. برآوردهای واریانس یک طرفه برای نمونه های کوچک تر، نسبت به برآوردهای واریانس ده طرفه متناظر، اریبی کمتری دارند اما متغیرتر هستند. مقایسه ما روی داده های شبیه سازی شده نشان می دهد که برآوردهای واریانس یک طرفه تغییرپذیری بسیار مشابهی با برآوردهای واریانس ده طرفه دارد (شکل تکمیلی S1 را ببینید). در نهایت، برآوردهای واریانس یک طرفه را برای همه مقایسه ها انتخاب می کنیم، زیرا برای داده هایی با هر اندازه ای قابل اجرا است و نسبت به برآوردهای واریانس ده طرفه، اریبی کمتری دارد.

### ۲.۴ مجموعه داده های نمونه

ما عملکرد روش ها را روی تعدادی از مجموعه داده های واقعی تولید شده از طیف وسیعی از پلتفرم های با توان عملیاتی بالا مانند داده های بیان ژن از ریزآرایه های DNA، داده های تصویربرداری عصبی از سیستم MRI با وضوح بالا و داده های ERP اندازه گیری فعالیت مغز مشتق شده از سیستم EEG ارزیابی کردیم.

#### ۲.۴.۱ داده های بیان ژن دوقطبی

این مجموعه داده بر اساس مطالعه بیان ژن ریزآرایه است. ۳۱ بافت مغز پس از مرگ بزرگسالان (قشر پیش پیشانی پشتی-جانبی) از افراد مبتلا به اختلال دوقطبی و افراد سالم. Affymetrix

از پلتفرم HG-U133A GeneChips برای تعیین بیان تقریباً 22000 رونوشت mRNA از 61 فرد (30 فرد مبتلا به اختلال دوقطبی و 31 فرد سالم) استفاده شد. داده های نرمال شده RMA پیش پردازش شده این آزمایش از <http://www.ncbi.nlm.nih.gov/geo/> (GEO) شماره دسترسی GSE5388، به دست آمده است.

#### ۲.۴.۲ داده های تصویربرداری مغز MRI

داده های تصویربرداری مغز از پایگاه داده طرح تصویربرداری عصبی بیماری آلزایمر (ADNI) (PI Michael M Weiner، [www.loni.ucla.edu/ADNI](http://www.loni.ucla.edu/ADNI)) دانلود شدند. ADNI در سال ۲۰۰۳ توسط موسسه ملی پیری (NIA)، موسسه ملی تصویربرداری زیست پزشکی و مهندسی زیستی (NIBIB)، سازمان غذا و دارو (FDA)، شرکت های داروسازی خصوصی و سازمان های غیرانتفاعی، به عنوان یک مشارکت دولتی-خصوصی پنج ساله ۶۰ میلیون دلاری راه اندازی شد. هدف اصلی ADNI آزمایش این موضوع بوده است که آیا MRI سریالی، PET و سایر نشانگرهای بیولوژیکی در آزمایشات بالینی MCI و مراحل اولیه بیماری آلزایمر (AD) مفید هستند یا خیر. تعیین نشانگرهای حساس و اختصاصی پیشرفت بسیار زود هنگام آلزایمر به منظور کمک به محققان و پزشکان برای توسعه درمان های جدید و نظارت بر اثربخشی آنها و همچنین کاهش زمان و هزینه آزمایشات بالینی در نظر گرفته شده است. افراد ADNI ۵۵ تا ۹۰ ساله از بیش از ۵۰ مکان در سراسر ایالات متحده و کانادا در این تحقیق شرکت کردند و اطلاعات دقیق تر در [www.adni-info.org](http://www.adni-info.org) موجود است. داده های MRI با شدت 1.5 تسلا از وب سایت [adni.loni.ucla.edu/index.shtml](http://adni.loni.ucla.edu/index.shtml) (ADNI) [www.loni.ucla.edu/ADNI/search/Cores](http://www.loni.ucla.edu/ADNI/search/Cores) بیابید. به طور خلاصه، داده های اسکنرهای 1.5 تسلا با داده های جمع آوری شده از سیستم های مختلف MR با پروتکل های بهینه سازی شده برای هر نوع اسکنر استفاده شدند. پوشش کامل مغز و مجموعه برای مجموعه داده های MRI مورد نیاز بود و کنترل کیفیت دقیق بر روی تمام تصاویر MR مطابق با معیارهای کنترل کیفیت منتشر شده قبلی انجام شد. ۳۳.۳۳٪ از Freesurfer pipeline (نسخه 5.1.0) برای تصاویر MRI استفاده کردیم تا ضخامت قشر مغز منطقه ای و اندازه گیری های حجمی را تولید کنیم. تمام اندازه گیری های حجمی از هر فرد با حجم داخل مجموعه ای فرد نرمال سازی شدند، در حالی که اندازه گیری های ضخامت قشر مغز نرمال سازی نشدند و به شکل خام خود استفاده شدند. ۳۳.

#### ۲.۴.۳ داده های الکتروانسفالوگرافی

سیگنال های EEG نوسانات ولتاژ ثبت شده از الکترودهای روی پوست سر را اندازه گیری می کنند و شاخصی از فعالیت مغز را ارائه می دهند. این داده ها از ۴۱ بزرگسال با تشخیص فعلی ADHD (اختلال بیش فعالی کمبود توجه) و ۴۷ فرد بدون مشکلات سلامت روان به دست آمده است. ۳۵٪ ثبت شد (و همکاران شرح داده شده است McLoughlin؛ که به تفصیل در CPT-OX) در طول یک وضعیت استراحت ۳ دقیقه ای (چشمان باز) و یک تکلیف عملکرد مداوم نشانه گذاری شده EEG<sup>۳۶</sup> مونتاژ و ثبت EEG، و همچنین ارجاع مجدد، نمونه برداری کاهشی و روش های حذف آرتیفکت چشمی معادل مواردی بود که در Tye و همکارانش ذکر شده است. ۳۷٪ آزمایش هایی با تعداد مصنوعات بیش از ۲۰۰ پنجم پیک تا پیک در هر کانال از داده های فیلتر شده دیجیتالی پایین گذر (0.1-30 هرتز، فیلتر ناچ 50 هرتز، 12 دسی بل بر اکتاو) حذف شدند. داده های EEG پیوسته به فواصل 2 ثانیه ای تقسیم شدند و سپس طیف توان با استفاده از تبدیل سریع فوریه با پنجره هنینگ محاسبه شد. معیارهای کمی توان طیفی EEG در باندهای فرکانسی اصلی EEG در فواصل فرکانسی مربوطه (0.5-3.5 هرتز، 4-7.5 هرتز، 7.5-12.5 و 12.5-30 هرتز) میانگین گیری شدند. چگالی توان مطلق (پنجم=هرتز)، چگالی توان نسبی (نسبت هر دامنه فرکانسی منفرد که به چگالی توان جمع شده کمک می کند) و نسبت های - در مکان های الکتروود پوست سر به طور متوسط فرونتال، مرکزی و جداری محاسبه شدند. ERP ها از CPT-OX همانطور که در McLoughlin و همکاران توضیح داده شده است، استخراج شدند.



## ۲.۵ شبیه سازی

## ۲.۵.۱ شبیه سازی اصلی

مسائل طبقه بندی مبتنی بر داده های با ابعاد بالا عمدتاً با استفاده از داده های بیان ژن ریزآرایه نشان داده می شوند. بنابراین، ما مدل شبیه سازی خود را برای تولید داده های بیان ژن واقع گرایانه طراحی می کنیم که در آن می توانیم ویژگی های مختلف داده ها (تغییرپذیری، اندازه اثر، همبستگی و غیره) را به طور سیستماتیک تغییر دهیم تا اثرات چنین ویژگی های داده ای را بر عملکرد الگوریتم های طبقه بندی بررسی کنیم. برای اینکه داده های شبیه سازی شده ما تا حد امکان واقعی باشند، مطالعه شبیه سازی خود را بر اساس یک مجموعه داده ریزآرایه واقعی انجام می دهیم. ما مجموعه ای از عبارات پایه ( ) را با میانگین گیری لگاریتم نرمال شده تعریف می کنیم. داده های بیان ژن روی ۲۸ نمونه خون انسانی گرفته شده از ۲۸ فرد سالم. ۳۸ مجموعه بیان پایه شامل بیان ژن برای 54359 نشانگر است که با استفاده از پلتفرم بیان ژن ریزآرایه کل ژنوم انسان CodeLink سنجیده شده اند. تغییرپذیری سیستماتیک و نویز تصادفی در مجموعه بیان پایه را می توان حداقل فرض کرد زیرا داده ها نرمال سازی و میانگین گیری شده روی افراد زیادی هستند. بنابراین منطقی است که فرض کنیم سطوح بیان پایه ( ) متناسب با سیگنال فراوانی mRNA واقعی نشانگرهای مربوطه هستند. سپس از یک مدل اثرات تصادفی برای معرفی مقدار از پیش تعیین شده نویز تصادفی در داده ها استفاده می کنیم. ما دو سطح از تغییرپذیری تصادفی را در نظر می گیریم: تغییرات بین آزمودنی ها و  $\delta$  درون آزمودنی ها که بیشتر به عنوان شناخته می شوند. بیولوژیکی و فنی تغییرات در مقالات مربوط به ریزآرایه. برای یک اندازه نمونه آموزشی مشخص (ن)، ما در ابتدا داده ها را برای هر نشانگر به طور مستقل طبق ... شبیه سازی می کنیم.

 $p_{i\delta}$ ایکس ای جی  $p_{i\delta}$  ۱/۴ ب من ای جی.

کجا ایکس ای جی نشان دهنده لاگ شبیه سازی شده است مقدار بیان برای یک نشانگر درجی تکرار هفتم (جی ۱/۴، ۲، ... ) از من موضوع (من ۱/۴، ۲، ...). ما در ابتدا هر تکرار از هر ویژگی را به عنوان متغیر مستقل (غیرهمبسته) شبیه سازی می کنیم که می تواند به عنوان یک متغیره در نظر گرفته شود. سپس داده های مربوط به هر نشانگر بر اساس تکرارها میانگین گیری می شوند و یک ساختار چند متغیره به داده ها از چندین ( ) معرفی می شود. (ص) نشانگرها با اعمال یک پ-ساختار کوواریانس ابعادی از طریق تبدیل ریشه چولسکی (به زیر مراجعه کنید). پارامتر، عبارت پایه را نشان می دهد (که به طور تصادفی از مجموعه عبارت پایه گرفته شده است)، ب من اثر تصادفی برای من موضوع بیولوژیکی هفتم و ای جی نویز تجربی تصادفی است. فرض می کنیم که

ب من ای جی متغیرهای تصادفی مستقلی هستند که بر اساس توزیع شده اند  $\delta$ ، ۰-۲-  $p_{i\delta}$  و  $\delta$ ، ۰-۲-  $p_{i\delta}$  به ترتیب. مدل (1) امکان تولید داده های بیان ژن مستقل یا غیرهمبسته با مقادیر مختلف نویز تصادفی کنترل شده توسط پارامترهای - را فراهم می کند. ب و ای.

استقلال به ندرت یک فرض واقع بینانه برای هر داده چندبعدی از جمله بیان ژن است که در آن گروه هایی از ژن ها به صورت هماهنگ با هم عمل می کنند و روابط شبکه ای را برای انجام عملکردهای بیولوژیکی خاص تشکیل می دهند. متغیرهای درون چنین گروه ها یا شبکه هایی عموماً همبستگی بالایی با یکدیگر دارند، اما احتمالاً همبستگی های ناچیزی با متغیرهای گروه دیگر نشان می دهند. ما یک ماتریس همبستگی بلوکی-قطری را برای مدل سازی چنین روابط شبکه ای با استفاده از توپلیتز<sup>۳۹</sup> ساختار همبستگی برای بلوک هفتم:

$$\begin{array}{cccccc}
 & & & & & 0 \\
 & & & & & 1 \\
 & & & & & 2 \\
 & & & & & 3 \\
 & & & & & 4 \\
 & & & & & 5 \\
 & & & & & 6 \\
 & & & & & 7 \\
 & & & & & 8 \\
 & & & & & 9 \\
 & & & & & 10 \\
 & & & & & 11 \\
 & & & & & 12 \\
 & & & & & 13 \\
 & & & & & 14 \\
 & & & & & 15 \\
 & & & & & 16 \\
 & & & & & 17 \\
 & & & & & 18 \\
 & & & & & 19 \\
 & & & & & 20 \\
 & & & & & 21 \\
 & & & & & 22 \\
 & & & & & 23 \\
 & & & & & 24 \\
 & & & & & 25 \\
 & & & & & 26 \\
 & & & & & 27 \\
 & & & & & 28 \\
 & & & & & 29 \\
 & & & & & 30 \\
 & & & & & 31 \\
 & & & & & 32 \\
 & & & & & 33 \\
 & & & & & 34 \\
 & & & & & 35 \\
 & & & & & 36 \\
 & & & & & 37 \\
 & & & & & 38 \\
 & & & & & 39 \\
 & & & & & 40 \\
 & & & & & 41 \\
 & & & & & 42 \\
 & & & & & 43 \\
 & & & & & 44 \\
 & & & & & 45 \\
 & & & & & 46 \\
 & & & & & 47 \\
 & & & & & 48 \\
 & & & & & 49 \\
 & & & & & 50 \\
 & & & & & 51 \\
 & & & & & 52 \\
 & & & & & 53 \\
 & & & & & 54 \\
 & & & & & 55 \\
 & & & & & 56 \\
 & & & & & 57 \\
 & & & & & 58 \\
 & & & & & 59 \\
 & & & & & 60 \\
 & & & & & 61 \\
 & & & & & 62 \\
 & & & & & 63 \\
 & & & & & 64 \\
 & & & & & 65 \\
 & & & & & 66 \\
 & & & & & 67 \\
 & & & & & 68 \\
 & & & & & 69 \\
 & & & & & 70 \\
 & & & & & 71 \\
 & & & & & 72 \\
 & & & & & 73 \\
 & & & & & 74 \\
 & & & & & 75 \\
 & & & & & 76 \\
 & & & & & 77 \\
 & & & & & 78 \\
 & & & & & 79 \\
 & & & & & 80 \\
 & & & & & 81 \\
 & & & & & 82 \\
 & & & & & 83 \\
 & & & & & 84 \\
 & & & & & 85 \\
 & & & & & 86 \\
 & & & & & 87 \\
 & & & & & 88 \\
 & & & & & 89 \\
 & & & & & 90 \\
 & & & & & 91 \\
 & & & & & 92 \\
 & & & & & 93 \\
 & & & & & 94 \\
 & & & & & 95 \\
 & & & & & 96 \\
 & & & & & 97 \\
 & & & & & 98 \\
 & & & & & 99 \\
 & & & & & 100
 \end{array}$$

کجا

$$p\delta^3 \quad \text{ح.ل.} \frac{1}{4} \text{ حد اکثر } p \quad \text{د.ح.} \frac{2}{3} \quad \text{حد اکثر } p \quad \text{منفی } p$$

که از کاهش می یابد حد اکثر به دقیقه برای ۲ ل دح این پارامتر نرخ کاهش را کنترل می کند که برابر است با خطی برای ۱/۴. ساختار همبستگی هاب-توپلیتز، یک همبستگی شناخته شده بین یک هاب شبکه (معمولاً متغیر اول) و هریک از متغیرهای دیگر درون بلوکی که همبستگی بین هاب و ... وجود دارد را فرض می کند. ل متغیر ام به صورت زیر کاهش می یابد افزایش می یابد. این فرض واقع بینانه تری نسبت به ماتریس همبستگی بلوک-قطری با ساختار تبدیلی درون بلوک هایی است که توسط هوا و همکارانش در نظر گرفته شده است. ۱۷.

تعداد بلوک ها را تعیین می کنیم (ح) به ۱ برای ص ۵۵. برای ص ۵، تعداد بلوک ها تصادفی است از مجموعه انتخاب شده ف، ۲، ... ۱/۲ پ=۳- گرم کجا [ایکس] نشان دهنده تابع کف (بزرگترین عدد صحیح که بزرگتر از نباشد) است. ایکس). اگر ص مضربی از است اچ، همه بلوک ها دارای ابعاد یکسان در نظر گرفته می شوند، دح p=H. 1/4 اگر ص مضربی از نیست اچ، یکی از بلوک ها (معمولاً اولین بلوک) دارای بعد در نظر گرفته می شود دح 1/4 p=H. 1/2 وزارت دفاع پ، ح و دح 1/4 p=H. 1/2 برای بقیه (ح 1/4، ۳، ... ح) کجا وزارت دفاع پ، ح پ یادآور تقسیم است پ/ح.

ما از تبدیل ریشه چولسکی برای اعمال ساختار هاب-توپلیتز بلوک-قطری استفاده می کنیم. ر 1/4 دیاگ ف، ر، ۲، ... ر، گرم به داده های غیر همبسته. فرض کنید ایکس-۵ 1/4 ایکس، ...، ایکس ص پ یک است پ-بردار متغیرهای تصادفی که میانگین نشانگرها را در تکرارهای مربوطه نشان می دهد. سپس ماتریس کوواریانس ایکس-توسط داده شده است

$$p\delta^4 \quad \text{پنجم-} \delta^4 \text{ ب-} p \quad \text{ای-} p \text{ من ص،}$$

کجا من ص یک است پ-ماتریس همانی ابعادی. سپس ریشه چولسکی را محاسبه می کنیم سی از پنجم ل و ل وی ۱، و داده های تبدیل شده با ساختار کوواریانس مورد نظر به صورت زیر بدست می آیند 1/4 ایکس سی. به منظور معرفی یک منبع سیستماتیک از تغییرات (یعنی تفاوت گروهی)، نمونه آموزشی را به دو گروه تقسیم می کنیم، جی ۱ و جی ۲، از نظر اندازه ن 1/4 ن 1/2=۲- و ن 1/4 ن سپس گروه ها با اندازه اثر متفاوت می شود و ترتیب آن تفاوت (لگاریتم) ۵۲ تغییر شکل. P) یعنی مقداری اضافه می کنیم نشان دهنده گروه غیر استاندارد برای مقادیر بیان شبیه سازی شده در جی ۲ برای ایجاد داده ها در جی ۲. بسته به مقدار، می توان آن را افزایش یا کاهش داد. به طور تصادفی انتخاب شده افزا، ۱ گرم ما از یک توزیع نرمال کوتاه شده با پشتیبانی تولید می کنیم الف، ۱۲، 1/2 1/4. یعنی، از توزیع نرمال کوتاه شده ی زیر پیروی می کند:

$$p\delta^5 \quad \text{ف.} \delta^4 \text{ الف } 1/4 \text{ پ} \quad \text{۱۵:} \quad \frac{1}{-1}$$

کجا P: ۵ و P: ۵ به ترتیب نشان دهنده تابع چگالی و توزیع یک متغیر نرمال استاندارد است و مقیاس متغیر نرمال صفر میانگین اصلی (کوتاه نشده) است. آستانه تضمین می کند که نشانگرهایی که در تحلیل لحاظ می کنیم، اندازه اثر غیر صفر داشته باشند. ما عبارت - را بیان می کنیم. دقیقه 2/4 به عنوان حداقل تغییر برابر نشانگرها که باید در طبقه بندی در نظر گرفته شود. مقدار مورد انتظار معادله (5) به صورت زیر داده می شود.

$$p\delta^6 \quad \text{ای } 1/2 \text{ زول الف } 1/4 \quad \frac{\delta p}{-\delta p} \quad ۱$$



جدول ۱. مقادیر اندازه نمونه آموزشی (n)، تنوع زیستی (-b)، تغییرات تجربی (-a)، حداقل تغییر ناشدگی (-دقیقه) و تکرار (r) در شبیه سازی در نظر گرفته شده است.

ویژگی های داده ها (فاکتورها)				
ن	-ب	-ای	-دقیقه-δ	ر
۱۰	۰.۱	۰.۱	(۱.۷۴)۱.۰	۱
۲۰	۰.۵	۰.۵	(۲.۳۰)۱.۵	۳
۳۰	۱.۰	۱.۰	(۲.۸۸)۲.۰	۵
۴۰	۱.۵	۱.۵	(۳.۴۵)۲.۵	۷
۵۰	۲.۰	۲.۰	(۴.۰۳)۳.۰	۹
۷۵	۲.۵	۲.۵	(۴.۶۱)۳.۵	۱۱
۱۰۰	۳.۰	۳.۰	(۵.۱۸)۴.۰	۱۳
۱۵۰	۳.۵	۳.۵	(۵.۷۵)۴.۵	۱۵
۲۰۰	۴.۰	۴.۴	(۶.۳۳)۵.۰	۱۷

میانگین تغییر برابر (-) نشانگرهای زیستی موجود در طبقه بندی را می توان محاسبه کرد.

از  $\frac{1}{4} \times 10^4$  به عنوان تابعی از  $\delta$ ،  $P$ ، ما قصد داریم اثرات اندازه نمونه آموزشی (ن)، اندازه مجموعه ویژگی (ص)، تنوع زیستی (-ب)، تغییرات تجربی (-ای)، اندازه اثر (-)، تکرار (ر) و همبستگی ( ) بین متغیرها. ما مجموعه ویژگی هایی با اندازه های ۵، ۲۵، ۵۰، ۷۵ و ۱۰۰ را در نظر می گیریم. این باید یک محدوده معقول برای درک الگوهای تأثیر اندازه مجموعه ویژگی ها بر عملکرد طبقه بندی باشد. اگرچه داده های با توان عملیاتی بالا می توانند متغیرهای بسیار بیشتری داشته باشند، هدف معمول در طبقه بندی با چنین داده هایی، دستیابی به عملکرد پیش بینی کلاس خوب با کمترین تعداد ممکن ویژگی ها، به نام تعداد بهینه، خواهد بود. عوامل بازدارنده برای در نظر گرفتن مجموعه ویژگی های بزرگتر، هزینه محاسباتی است که برای هر متغیر اضافی در مجموعه ویژگی ها به طور قابل توجهی افزایش می یابد. مقادیر سایر عوامل در نظر گرفته شده در شبیه سازی ما در جدول ۱ خلاصه شده است. ما به طور تصادفی انتخاب می کنیم

حداکثر و دقیقه از توزیع های یکنواخت  $10^6$ ،  $10^8$  و  $10^4$  برای هر مجموعه داده  $10^2$ ،  $10^4$  به ترتیب، و مجموعه شبیه سازی شده. ما استفاده می کنیم  $10^4$  در تمام شبیه سازی ها و نشانگرهای زیستی به ترتیب نزولی اندازه اثر اضافه می شوند. تمام شبیه سازی ها ۵۰۰ بار تکرار شدند و میانگین نرخ خطا روی این ۵۰۰ مجموعه داده برای هر ترکیب از سطوح عوامل محاسبه شد.

۲.۵.۲ داده های غیر نرمال  
مایک تحلیل حساسیت انجام می دهیم تا پیامدهای احتمالی انحراف از داده های توزیع نرمال را بررسی کنیم. ما برای این تحلیل حساسیت، داده هایی از خانواده پواسون را شبیه سازی کردیم. داده های پواسون از چندین جهت با همتهای گاوسی خود متفاوت هستند: (i) پواسون یک توزیع گسسته است که داده های عدد صحیح را به جای داده های مقیاس بندی شده فاصله ای (شبیه گاوسی) تولید می کند، (ii) رابطه میانگین-واریانس برای داده های پواسون محدودتر از داده های گاوسی است زیرا آنها با پارامتر نرخ شدت ( ) یکسان هستند و (iii) داده های پواسون عموماً دارای چولگی مثبت هستند، به خصوص برای مقادیر کوچکتر. تطبیق دقیق میانگین و واریانس داده های پواسون با همتهای گاوسی آنها امکان پذیر نیست، اما ما مشابهت ( ) را حفظ کردیم. هاب-توپلیتز (ساختار کوواریانس قطری بلوکی بین متغیرها را مانند داده های گاوسی در نظر گرفت و شبیه سازی را در ترکیب های مختلفی از اندازه های نمونه آموزشی و مجموعه ویژگی ها انجام داد. (ن  $10^4$ ، ۲۰، ۳۰، ۴۰، ۵۰، ۷۵، ۱۰۰، ۱۵۰، ۲۰۰؛ ص  $10^4$ ، ۲۵، ۵۰، ۷۵، ۱۰۰). ما تنظیم می کنیم  $10^4$  برای شبیه سازی یک تکرار واحد از هر متغیر و معرفی تفاوت گروهی سیستماتیک ( ) به همان روشی که برای

داده های گاوسی. انتخاب به عنوان مصالحه ای بین عدم تقارن و تغییرپذیری داده ها انجام شد: ۴٪ به اندازه کافی بزرگ است تا تغییرپذیری معقول داده ها را تضمین کند و در عین حال به اندازه کافی کوچک است تا چولگی توزیع را تضمین کند.

### ۳ نتایج و بحث ۳.۱ نتایج شبیه سازی اصلی

میانگین تخمین خطای طبقه بندی با استفاده از اعتبارسنجی متقاطع با حذف یک باره در 500 تکرار از مجموعه داده های شبیه سازی شده برای هر چهار روش (RF، SVM، LDA و kNN) در مقابل مقادیر ویژگی های مختلف داده ها در شکل های 1 و 2 رسم شده است. پل های بالا و پایین در شکل 1 بر اساس ( ) هستند. ۱۰۰٪، ۲۵٪ و ۱۰۰٪/۴، ۱۰۰٪/۴، ۱۰۰٪/۴ به ترتیب ۷۵. نرخ خطا برای ترکیبات مختلف این دو عامل (اندازه نمونه آموزشی و اندازه مجموعه ویژگی ها) در شکل ۲ نمایش داده شده است. نتایج در شکل ۱ و نمودارهای اضافی در فایل تکمیلی (شکل S2) نشان می دهد که شرایط عملکرد بهینه برای LDA، که در این صورت از تمام روش های دیگر مورد مطالعه در اینجا بهتر عمل می کند، اندازه مجموعه ویژگی کوچکتر (نسبت به اندازه نمونه آموزشی) و همبستگی بالاتر است. به نظر می رسد ناحیه قدرت این روش پ=۵۵:۰ (تعداد ویژگی ها کوچکتر از تقریباً نصف اندازه نمونه) و همبستگی بالاتر از حد متوسط (۰:۶۴) بین ویژگی ها با بزرگتر شدن مجموعه ویژگی ها (پ=۴ را برای leave-one-out دارد. پل پایین شکل 1 میانگین خطای kNN و RF پیشی می گیرد و همچنین عملکرد بهتری نسبت به LDA از SVM، 0:5، ۱۰۰٪ و ۷۵٪، یعنی پ=۱/۴:۷۵: نشان می دهد که SVM با اختلاف قابل توجهی از سایر روش ها بهتر عمل می کند. اختلاف عملکرد در تغییرپذیری داده های بالاتر، اندازه اثر کوچکتر و همبستگی کمتر، بیشتر است. عملکرد kNN همچنین با افزایش اندازه مجموعه ویژگی ها بهبود می یابد و از LDA و RF بهتر عمل می کند، مگر اینکه تغییرپذیری داده ها خیلی زیاد و/یا اندازه اثر خیلی کوچک باشد. مشخص شد که RF فقط در برخی موارد که داده ها متغیرتر و اندازه اثر کوچکتری دارند، از kNN بهتر عمل می کند، که در این موارد، تخمین های خطای پایدارتری نسبت به kNN و LDA نیز ارائه می دهد (به شکل تکمیلی S3 مراجعه کنید).

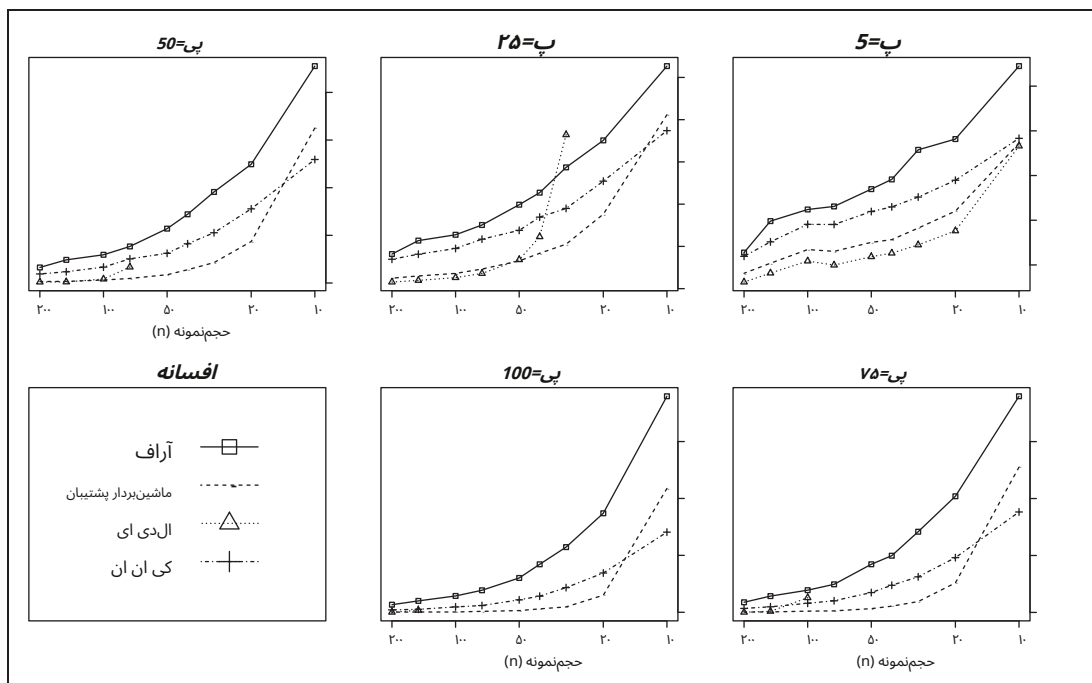
نمودارهای مربوطه برای حساسیت و ویژگی در شکل های تکمیلی S4 و S5 ارائه شده اند. عملکرد تمام روش های مورد مطالعه از نظر حساسیت و ویژگی متقارن بود، که انتظار می رود برای داده های توزیع شده متعادل و متقارن نیز صادق باشد. الگوها و ترتیب عملکرد از نظر حساسیت و ویژگی مشابه خطای کلی طبقه بندی بود.

شکل 2 نرخ خطا را در برابر نه مقدار مختلف نشان می دهد همانطور که در جدول 1 آمده است. پنج نمودار مربوط به پنج مقدار مختلف از پ (اندازه مجموعه ویژگی ها): به ترتیب ۵، ۲۵، ۵۰، ۷۵ و ۱۰۰. اگرچه مشخص شد که SVM برای نمونه های آموزشی بزرگتر و مجموعه ویژگی ها، عملکرد بهتری نسبت به سایر روش ها دارد (شکل ۱ را ببینید)، اما این روش برای نمونه های آموزشی کوچکتر (ن=20۵) نمونه. نمودار نشان می دهد که اندازه نمونه صرف نظر از تعداد متغیرها باید حداقل 20 باشد (ص) برای اینکه SVM عملکرد بهتری داشته باشد. نرخ خطا برای LDA برای اندازه های نمونه کوچکتر از تعداد متغیرها نشان داده نشده است (ص) از آنجایی که این روش برای ... منحنی است ص=۴. اگرچه LDA از نظر تئوری برای هر چیزی معتبر است ن=۴، وقتی اندازه مجموعه ویژگی ها بسیار نزدیک به اندازه نمونه آموزشی باشد، می تواند عملکرد ضعیفی داشته باشد. نمودار نشان می دهد که نباید حداقل به بزرگی ۲ باشد ص برای اینکه LDA عملکرد قابل مقایسه ای داشته باشد. این یافته با یافته های قبلی ما سازگار است. ۴۰.

برای درک بهتر الگوهای عملکرد طبقه بندی و سهولت مقایسه های بصری، نمودارهای سه بعدی میانگین خطای اعتبارسنجی متقاطع با حذف یک مورد خاص به عنوان تابعی مشترک از اندازه مجموعه ویژگی ها و تنوع زیستی در شکل 3 نمایش داده شده اند. همه نمودارها مربوط به ۱۰۰٪ و اندازه مجموعه ویژگی (ص) بین ۵ تا ۱۰۰ متغیر است. شکل، نرخ خطا را به عنوان تابعی مشترک از (-) نشان می دهد. ب، ص) و نشان می دهد که با افزایش اندازه مجموعه ویژگی ها، نرخ خطا کاهش می یابد (ص) برای همه روش ها به جز LDA رشد می کند. همانطور که قبلاً در روش دوبعدی دیده شد



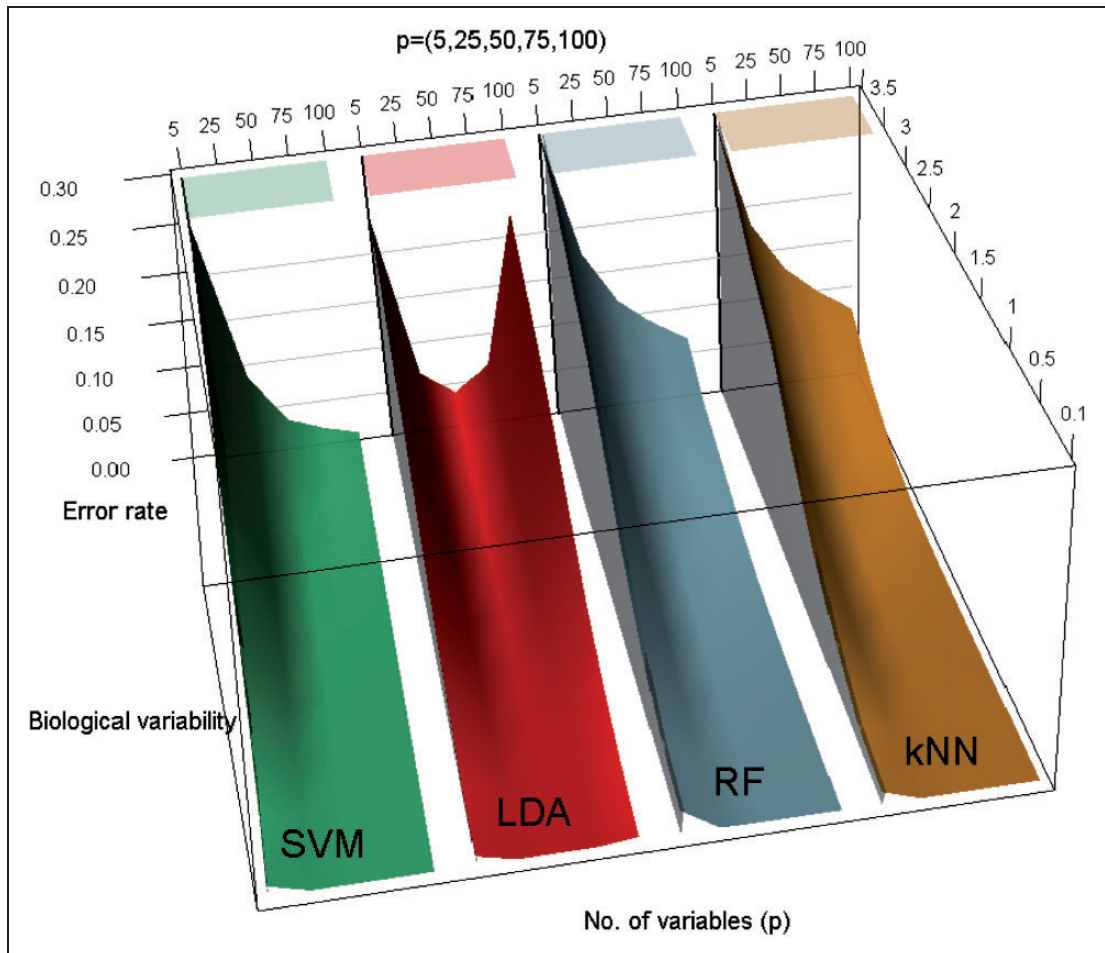
شکل ۱. میانگین خطای اعتبارسنجی متقابل با حذف یک باره در سطوح مختلف ویژگی‌های داده‌های مختلف: تغییرپذیری بیولوژیکی (ب)، نویز تجربی (ج)، حد پایین اندازه اثر (د)، تکرار (ر) و همبستگی بین متغیرها (ه). پل‌های بالا و پایین به ترتیب مربوط به اندازه‌های مجموعه ویژگی ۲۵ و ۷۵ متغیر و یک اندازه نمونه آموزشی رایج (۱۰۰/۴). نرخ نمودار، نرخ خطا را برای چهار روش در سطوح مختلف یک پارامتر خاص، همانطور که در شکل نشان داده شده است، مقایسه می‌کند. ایکس-محور برای مقادیر داده شده از پارامترهای دیگر. مقادیر داده شده از مجموعه انتخاب می‌شوند. ۱۰۰/۴، ب-۲۵/۴، ج-۱۵/۴، د-۱۵/۴، ه-۳/۴، ۳/۴. ساختار همبستگی نوع‌هاب-توپلیتز است (به جز نمودارهای مقابل، که بر اساس ماتریس همبستگی قابل تعویض تک بلوکی هستند تا نمودار مقابل را معنادار کنند). برای مجموعه ویژگی‌های کوچکتر (ص ۲۵/۴) و همبستگی بالاتر (۰/۶۳)، به نظر می‌رسد LDA در تمام سطوح ویژگی‌های داده‌ای مورد بررسی، به طور یکنواخت عملکرد بهتری داشته است. SVM برای مجموعه ویژگی‌های بزرگتر، از تمام روش‌های دیگر از جمله LDA بهتر عمل کرد (ص ۷۵/۴) و kNN در این سطح از اندازه مجموعه ویژگی، دومین الگوریتم برتر شناخته شد. مشخص شد که RF در بیشتر سناریوها عملکرد ضعیفی دارد، به جز داده‌های بسیار نویزی (دارای تغییرپذیری بالا و اندازه اثر کوچکتر) که در این موارد عملکرد بهتری نسبت به kNN دارد.



شکل ۲. میانگین خطای اعتبارسنجی متقابل با حذف یک باره در سطوح مختلف نمونه آموزشی و اندازه مجموعه ویژگی ها. نرخ خطا در برابر نه مقدار مختلف رسم شده است همانطور که در جدول ۱ آمده است. پنج نمودار مربوط به پنج مقدار مختلف از پ (اندازه مجموعه ویژگی): به ترتیب ۵، ۰.۲۵، ۰.۵۰، ۷۵ و ۱۰۰. سایر پارامترها روی مقادیر ثابت تنظیم شده اند: (-۰.۵/۴، -۰.۵/۴، دقیقه ۳/۴، ۱/۴). اگرچه عملکرد SVM برای نمونه های آموزشی بزرگتر و مجموعه ویژگی ها بهتر از سایر روش ها (شکل ۱ را ببینید) تشخیص داده شد، اما این روش برای نمونه های آموزشی کوچکتر (ن=۲۰۵) نمونه. نمودار نشان می دهد که اندازه نمونه باید حداقل ۲۰ باشد تا SVM عملکرد برتر خود را داشته باشد. نرخ خطا برای LDA برای اندازه های نمونه کوچکتر از تعداد متغیرها نشان داده نشده است (ص) از آنجایی که این روش برای ... منطبق است ص ۴۸. اگرچه LDA از نظر تئوری برای هر چیزی معتبر است ن ۴۸، وقتی اندازه مجموعه ویژگی ها بسیار نزدیک به اندازه نمونه آموزشی باشد، می تواند عملکرد بسیار ضعیفی داشته باشد. نمودار نشان می دهد که باید حداقل به بزرگی ۲ تصدیف برای اینکه LDA عملکرد قابل مقایسه ای داشته باشد.

باتوجه به نمودارها، عملکرد LDA با افزایش تعداد متغیرها شروع به کاهش می‌کند. ص. تقریباً از نصف حجم نمونه تجاوز می‌کند (2=). با این حال، LDA برای مقادیر کوچکتر ( ) از SVM بهتر عمل می‌کند. 0:55 و ویژگی بر اساس نسبت حجم نمونه تنظیم‌شده و برعکس.

خطاهای استاندارد (SE) تخمین‌های خطای اعتبارسنجی متقاطع با حذف یک باره در سطوح مختلف ویژگی‌های داده‌های مختلف در شکل تکمیل S3 رسم شده‌اند. به نظر می‌رسد SEهای تخمین‌های خطا از الگوهای بسیار مشابهی با تخمین‌های خطای میانگین پیروی می‌کنند. به عنوان مثال SVM پایدارترین (کمترین SE) تخمین‌های خطای حذف یک باره را در موقعیت‌هایی ارائه می‌دهد که مجموعه ویژگی‌ها بزرگتر از نصف اندازه نمونه باشد. آموزش‌باشند (پ=0.54)، در حالی که مشخص شد LDA دقیق‌ترین تخمین‌ها از نرخ خطا را برای پ=0.5 همبستگی‌های 0.5 و بالاتر بین ویژگی‌ها. با این حال، به نظر می‌رسد LDA برای نسبت اندازه مجموعه ویژگی‌های بزرگتر به اندازه نمونه (کمترین پایداری را دارد. پ=0.54). قدرت RF در موقعیت‌هایی که داده‌ها متغیرتر و اندازه اثر کوچکتری دارند، قابل مشاهده است، که در این موارد، تخمین‌های خطای پایداری نسبت به LDA و kNN ارائه می‌دهد.



شکل ۳. نمودار سه بعدی میانگین خطای اعتبارسنجی متقاطع با حذف یک متغیر به عنوان تابعی مشترک از اندازه مجموعه ویژگی ها (تعداد متغیرها) و تنوع زیستی. همه نمودارها مربوط به  $100\%$  و پنج اندازه مجموعه ویژگی مختلف (ص ۵۰، ۷۵، ۱۰۰، ۲۵، ۵۰) همانطور که در سربرگ نمودار نشان داده شده است. پارامترهای باقیمانده همانطور که در نمودارهای دو بعدی نشان داده شده است، روی مقادیر ثابت تنظیم شده اند. نمودارها نشان می دهند که نرخ خطا با افزایش اندازه مجموعه ویژگی ها (کاهش می یابد). ص) برای همه روش ها به جز LDA افزایش می یابد. همانطور که قبلاً در نمودارهای دوبعدی مشاهده شد، نرخ خطا برای LDA با افزایش تعداد متغیرها شروع به افزایش می کند. ص فراتر از  $2$ . اگرچه SVM برای مجموعه ویژگی های بزرگتر از LDA بهتر عمل می کند، اما LDA برای مجموعه ویژگی های کوچکتر بهتر از SVM عمل می کند (ص 205) ...

### ۳.۱.۱ نتایج روی داده های غیر نرمال

این بخش نتایج شبیه سازی بر اساس توزیع های غیر نرمال (پواسون) را همانطور که در بخش شبیه سازی توضیح داده شده است، ارائه می دهد. شکل تکمیلی S6 میانگین تخمین خطای طبقه بندی با روش اعتبارسنجی متقاطع یک مرحله ای را در 500 تکرار داده های شبیه سازی شده برای هر چهار روش (SVM، RF، kNN) در ترکیبات مختلف مجموعه ویژگی ها و اندازه نمونه آموزشی نشان می دهد. الگوها و ترتیب عملکرد برای داده های پواسون بسیار شبیه به آنچه برای داده های گاوسی مشاهده کردیم (شکل 2) است. به عنوان مثال، مانند داده های گاوسی، SVM برای نمونه های کوچکتر عملکرد خوبی ندارد (ن 205) اگرچه برای مجموعه ویژگی های بزرگتر، از سایر روش ها بهتر عمل می کند

### ۳.۲ نتایج روی داده های واقعی

### ۳.۲.۱ داده های بیان ژن دو قطبی

اندازه‌های اثر برای این مجموعه داده چندان بزرگ نیستند و به نظر نمی‌رسد عملکرد طبقه‌بندی با افزایش تعداد نشانگرها از حدود ۵ بهبود یابد. با این حال، از نظر عملکرد مقایسه‌ای، SVM در بیشتر موارد بهتر (و پس از آن LDA) به نظر می‌رسد که سازگار است.

جدول ۲. خلاصه ای از ویژگی های داده های تخمینی برای زیرمجموعه ای از 25 نشانگر برتر از داده های بیان ژن دوقطبی [31].

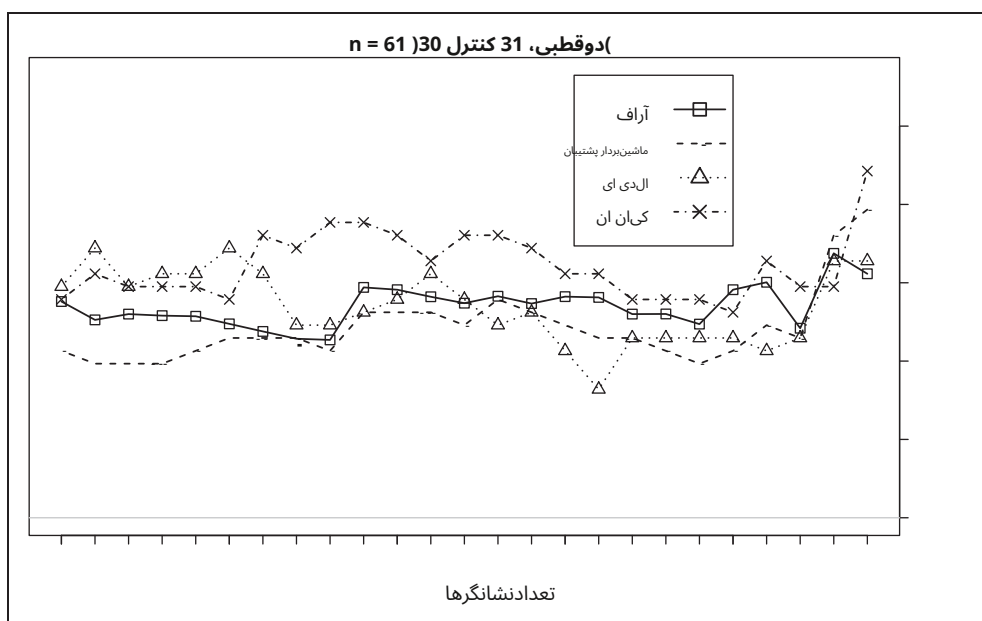
گروه	اندازه‌گیری	حداقل	میانگین	پزشکی	مکس
دوقطبی	تنوع* (-)	۰.۱۴	۰.۳۶	۰.۳۲	۰.۹۱
	همبستگی ( )	۰.۶۸	۰.۲۷	۰.۴۰	۰.۹۰
کنترل	واریاسیون (-)	۰.۱۴	۰.۲۷	۰.۲۱	۰.۴۹
	همبستگی ( )	۰.۶۲	۰.۱۸	۰.۲۴	۰.۸۴
ترکیبی	واریاسیون (-)	۰.۱۴	۰.۳۲	۰.۳۰	۰.۷۲
	همبستگی ( )	۰.۶۳	۰.۱۸	۰.۲۳	۰.۸۴
دوقطبی در مقابل کنترل	نسبت‌الگاریتمی (جی جی)	۰.۱۶	۰.۳۴	۰.۳۰	۰.۸۵

[illegible]



با یافته های حاصل از داده های شبیه سازی شده. لازم به ذکر است که عملکرد LDA زمانی که تعداد نشانگرها از ۱۸ فراتر رفت، که بسیار کمتر از آستانه است، شروع به بدتر شدن کرد. (تعداد=230) از مطالعات شبیه سازی پیش بینی شده است. این ممکن است به دلیل اندازه اثر کوچکتر نشانگرها باشد که به راحتی تحت تأثیر نویز قرار می گیرند، که همچنین با این واقعیت که RF همانطور که با داده های شبیه سازی شده از این نوع یافتیم، عملکرد بهتری نسبت به kNN داشت، پشتیبانی می شود.

۳.۲.۲ داده های تصویربرداری مغز بیماران آلزایمر و گروه کنترل  
پیشرفت های اخیر در فناوری های تصویربرداری عصبی مانند سیستم MRI با وضوح بالا، اندازه گیری مؤثر ضخامت قشر منطقه ای مغز و حجم منطقه ای را با استفاده از طرح تقسیم بندی خودکار تصویر عصبی مبتنی بر اتمس امکان پذیر ساخته است. چنین معیارهایی معمولاً برای ارتباط دادن بیماری آلزایمر به تغییرات فیزیکی در مناطق مختلف مغز، مانند هیپوکامپ و قشر انتورینال، استفاده می شوند. در این مطالعه، داده های تصویربرداری مغز از مطالعه ADNI مستقر در ایالات متحده ([www.loni.ucla.edu/ADNI](http://www.loni.ucla.edu/ADNI)) را برای طبقه بندی بیماران بر اساس ضخامت و حجم قشر منطقه ای آنها در نظر می گیریم. اطلاعات بیشتر در مورد داده ها و پروتکل های تولید چنین معیارهایی را می توان در بخش روش یافت. ما 25 معیار برتر را بر اساس نمونه ای از 418 بیمار (186 بیمار مبتلا به آلزایمر، 222 بیمار گروه کنترل) انتخاب می کنیم. پ-مقادیر آزمون بیز تجربی<sup>۳۱</sup> برای تفاوت میانگین ها بین گروه های AD و کنترل (CTR) برای تجزیه و تحلیل طبقه بندی. خلاصه ویژگی های داده های تصویربرداری بر اساس نشانگرهای انتخاب شده در جدول 3 آورده شده است. عملکرد طبقه بندی چهار روش (LDA، SVM، RF و kNN) برای طبقه بندی AD در مقابل CTR در شکل 5 نمایش داده شده است. اندازه های اثر ( ) برای این مجموعه داده بالاتر است و عملکرد روش ها بسیار نزدیک به یکدیگر به نظر می رسد. این مورد انتظار است و توسط یافته های شبیه سازی (به عنوان مثال، شکل 1) پشتیبانی می شود که نشان می دهد تفاوت های عملکرد با افزایش سن کمتر می شود.

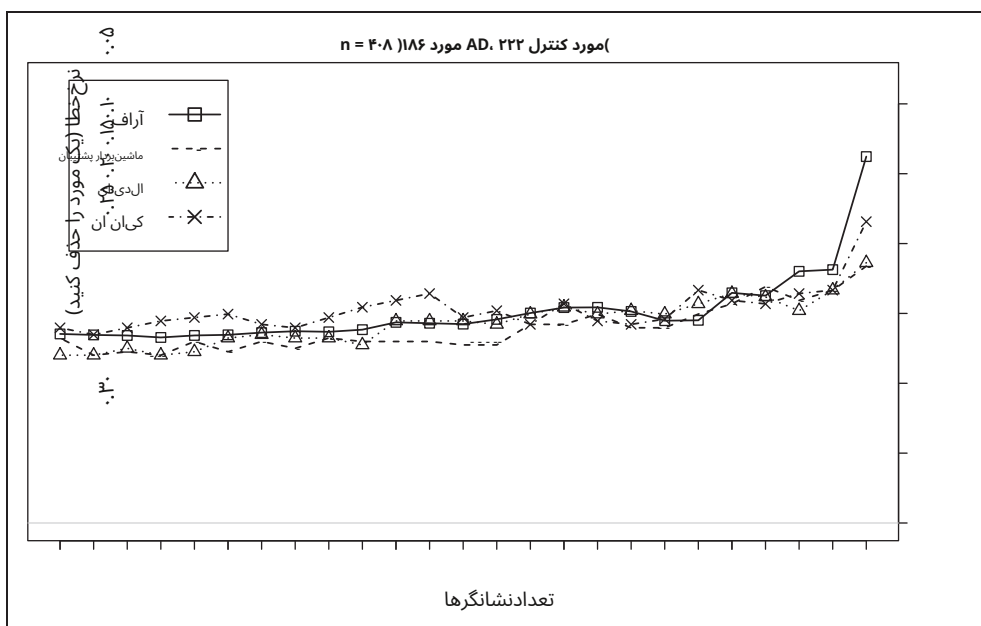


شکل ۴. خطاهای حذف یک مورد از CV برای طبقه بندی ۶۱ بیمار به گروه های دوقطبی و کنترل بر اساس ژن های با بیان متفاوت برتر از داده های بیان ژن دوقطبی<sup>۳۱</sup>.

جدول ۳. خلاصه ای از ویژگی های داده های تخمینی برای زیرمجموعه ای از ۲۵ نشانگر برتر تصویربرداری مغز از مطالعه ADNI.

خلاصه تخمین ها					گروه
مکس	پزشکی	میانگین	حداقل	اندازه گیری	
۱.۱۷	۱.۰۱	۱.۰۱	۰.۷۸	تنوع* (-)	آگهی
۰.۷۸	۰.۲۱	۰.۱۶	۰.۵۴	همبستگی ( )	
۱.۰۱	۰.۸۹	۰.۸۶	۰.۵۶	واریاسیون (-)	کنترل
۰.۷۲	۰.۱۳	۰.۱۴	۰.۵۲	همبستگی ( )	
۱.۰۰	۰.۹۵	۰.۹۳	۰.۷۸	واریاسیون (-)	ترکیبی
۰.۷۲	۰.۱۸	۰.۱۵	۰.۵۱	همبستگی ( )	
۱.۲۶	۰.۹۲	۰.۹۷	۰.۷۶	جیب	درمقابل کنترل AD

\* به پاورقی جدول ۲ مراجعه کنید.



شکل ۵. خطاهای حذف یک مرحله ای CV برای طبقه بندی ۴۰۸ بیمار به گروه های AD و کنترل بر اساس معیارهای تصویربرداری برتر. تفاوت های بسیار کمی بین عملکرد روش ها وجود دارد، که به دلیل اندازه های اثر بزرگتر مورد انتظار است و توسط یافته های حاصل از شبیه سازی پشتیبانی می شود که نشان می دهد تفاوت های عملکرد با افزایش اندازه اثر کمتر می شود.

اندازه اثر افزایش می یابد. در مجموع، به نظر می رسد SVM و LDA در بیشتر موارد عملکرد بهتری داشته اند، همانطور که الگوهای کلی مشاهده شده در شبیه سازی نیز آن را تأیید می کنند.

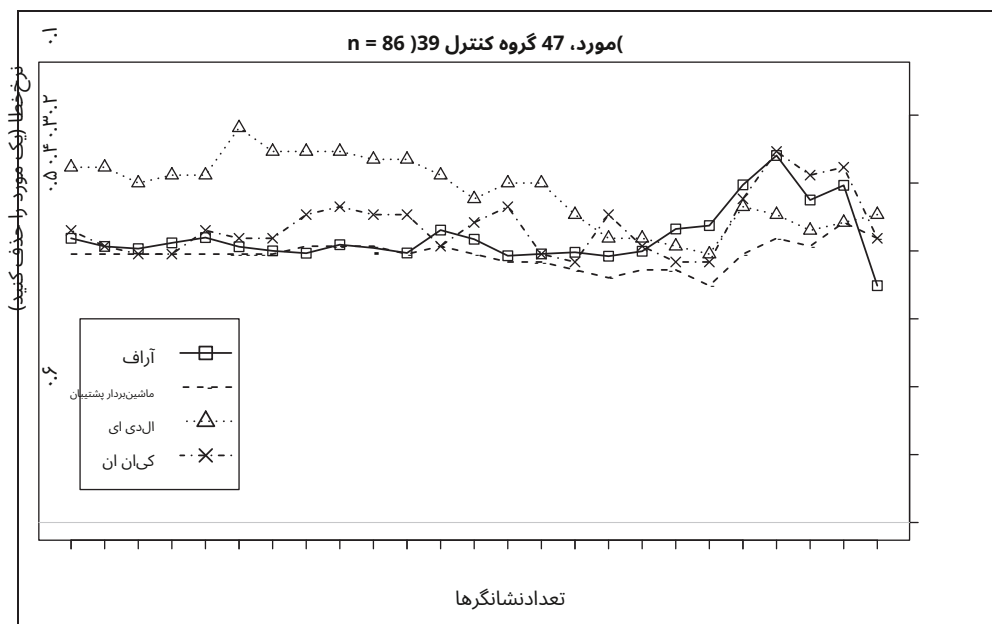
### ۳.۲.۳ داده های الکتروانسفالوگرافی

سیگنال های EEG نوسانات ولتاژ ثبت شده از الکترودهای روی پوست سر را اندازه گیری می کنند و شاخصی از فعالیت مغز را ارائه می دهند. داده های EEG برای این مثال از ۴۱ بزرگسال با تشخیص فعلی به دست آمده است.

جدول ۴. خلاصه ای از ویژگی های داده های تخمینی برای زیرمجموعه ای از ۲۵ ویژگی برتر از داده های EEG.

خلاصه تخمین ها					گروه
مکس	پزشکی	میانگین	حداقل	اندازه گیری	
۱.۱۲	۰.۹۳	۰.۹۳	۰.۷۴	تنوع* (-)	اختلال کم توجهی بیش فعالی
۰.۹۶	۰.۰۴	۰.۰۷	۰.۹۰	همبستگی (-)	
۱.۱۷	۱.۰۶	۱.۰۵	۰.۸۸	واریاسیون (-)	کنترل
۰.۹۷	۰.۰۴	۰.۱۱	۰.۹۱	همبستگی (-)	
۱.۰۰	۰.۹۹	۰.۹۹	۰.۹۸	واریاسیون (-)	ترکیبی
۰.۹۵	۰.۰۳	۰.۰۹	۰.۸۷	همبستگی (-)	
۰.۴۵	۰.۲۷	۰.۲۸	۰.۱۸	میانگین	درمقابل کنترل ADHD

\* به پاور ۲ جدول ۲ مراجعه کنید.



شکل ۶. خطاهای CV حذف یک طرفه برای طبقه بندی ۸۶ بیمار به گروه های ADHD و کنترل بر اساس معیارهای برتر EEG. این مثالی است که در آن اندازه اثر بسیار کوچک است و بنابراین نرخ خطا با افزایش تعداد متغیرها، الگوی کاهشی ثابتی را نشان نمی دهد. با این حال، این موضوع از مشاهده شبیه سازی در مورد نقاط قوت RF در برخورد با داده هایی با سیگنال ضعیف تر و نویز بالاتر پشتیبانی می کند، زیرا در این مثال، RF از LDA و KNN بهتر عمل می کند.

مبتلا به ADHD (اختلال بیش فعالی کمبود توجه) و ۴۷ نفر بدون مشکلات سلامت روان، ۳۵ توضیحات بیشتر داده ها در بخش روش آمده است. از بین ۶۳ معیار، تنها تعداد کمی از آنها بین گروه های مورد و کنترل تفاوت معنی داری داشتند. با این حال، ما ۲۵ معیار برتر را بر اساس ... انتخاب می کنیم. پ-مقادیر آزمون بیز تجربی ۴ برای تحلیل طبقه بندی

خلاصه‌ویژگی‌های داده‌های EEG بر اساس معیارهای انتخاب شده روی ۸۶ آزمودنی (دو مورد ADHD به دلیل مقادیر از دست رفته در برخی از متغیرهای انتخاب شده حذف شدند) در جدول ۴ آمده است. عملکرد طبقه بندی چهار روش (RF، SVM، LDA و kNN) برای اختلال کم توجهی بیش فعالی در مقابل کنترل طبقه بندی در شکل 6 نمایش داده شده است.

از بین سه مجموعه داده مثال، داده‌های EEG کوچکترین سیگنال (میانگین) را دارند (۰.۲۸٪) برای تمایز قائل شدن بین گروه‌های مورد و کنترل. به ندرت اندازه‌گیری‌های EEG وجود دارد که تفاوت معناداری بین گروه‌ها داشته باشد. این احتمالاً توضیح می‌دهد که چرا منحنی‌های خطا با افزایش تعداد ویژگی‌ها، کاهش ثابتی را نشان نمی‌دهند. با این حال، مشاهده می‌شود که RF از LDA و kNN بهتر عمل می‌کند و از مشاهدات حاصل از شبیه‌سازی در مورد نقاط قوت RF در برخورد با داده‌هایی با سیگنال ضعیف‌تر و نویز بالاتر پشتیبانی می‌کند.

#### ۴ نتیجه گیری

مایک مطالعه شبیه‌سازی گسترده انجام دادیم تا عملکرد طبقه بندی تعدادی از الگوریتم‌های یادگیری ماشین یا آماری پرکاربر را از نظر خطاهای تعمیم، حساسیت و ویژگی برای مسائلی طبقه بندی نظارت شده، به طور عینی مقایسه کنیم. تمرکز اصلی مطالعه ما بررسی این بود که «کدام روش در چه شرایطی عملکرد بهتری دارد» و این کار را با مقایسه عملکرد هادار ترکیبات مختلف سطوح عوامل متعدد (ویژگی‌های داده) انجام دادیم. نتایج مطالعه شبیه‌سازی ما و مثال‌های بعدی روی چندین مجموعه داده واقعی از پلتفرم‌های مختلف فناوری با توان عملیاتی بالا منجر به نتیجه‌گیری‌های زیر شد:

- برای تعداد کمتری از ویژگی‌های همبسته، تعداد ویژگی‌هایی که تقریباً از نصف اندازه نمونه تجاوز نمی‌کنند، LDA از نظر میانگین خطاهای تعمیم و همچنین پایداری (دقت) تخمین خطا، روش انتخابی شناخته شد. به نظر می‌رسد ناحیه قدرت LDA عبارت است از:  $p=0.55$  (تعداد ویژگی‌ها کوچکتر از تقریباً نصف اندازه نمونه) و همبستگی بالاتر (۰.۶۴) بین ویژگی‌ها.
  - همانطور که مجموعه ویژگی‌ها بزرگتر می‌شود ( $p=n$ ) پیشی می‌گیرد و همچنین LDA از RBF با هسته (SVM)  $\gamma=0.5$  با اختلاف فاحشی بهتر از RF و kNN عمل می‌کند. اختلاف عملکرد در تغییرپذیری داده‌های بالاتر، اندازه اثر کوچک‌تر و همبستگی کمتر، بیشتر است. با این حال، اندازه نمونه صرف نظر از تعداد ویژگی‌ها باید حداقل 20 باشد (ص) برای اینکه SVM به عملکرد برتر خود دست یابد.
  - عملکرد kNN همچنین با افزایش اندازه مجموعه ویژگی‌ها بهبود می‌یابد و از LDA و RF پیشی می‌گیرد، مگر اینکه تغییرپذیری داده‌ها خیلی زیاد و/یا اندازه‌های اثر خیلی کوچک باشند.
  - مشخص شد که RF فقط در برخی موارد که داده‌ها متغیرتر بودند و اندازه اثر کوچک‌تری داشتند، از kNN بهتر عمل می‌کند، که در این موارد، تخمین‌های خطای پایداری نسبت به kNN و LDA نیز آرائه می‌دهد.
  - همه روش‌ها تمایل به عملکرد بهتر در همبستگی بالاتر را نشان دادند، اما به نظر می‌رسد RF وقتی متغیرها همبستگی بسیار بالایی دارند، عملکرد نسبتاً بدتری دارد.
  - عملکرد تمام روش‌های مورد مطالعه از نظر حساسیت و ویژگی متقارن بود، که انتظار می‌رود برای داده‌های توزیع شده متعادل و متقارن نیز صادق باشد.
- هیچ‌یک از روش‌های مورد مطالعه (به جز LDA) نیازی به پیروی داده‌ها از توزیع احتمال خاصی ندارند و نتایج شبیه‌سازی باید در برابر انحراف از فرض نرمال بودن مقاوم باشند. ما این را با شبیه‌سازی از توزیع غیر نرمال (پواسون) نشان دادیم که نشان می‌دهد LDA در برابر برخی انحراف از نرمال بودن مقاوم است و سایر روش‌ها نیز به طور مشابه روی داده‌های نرمال و غیر نرمال عمل می‌کنند.

## تقدیرنامه ها

ازپروفسور اندرو پیکلز و داوران ناشناس برای نظرات مفیدشان در مورد پیش نویس اولیه که به بهبود مقاله کمک کرد، تشکر می کنیم. از دکتر گرین مک لافلین و پروفسور فیلیپ اشرسون برای ارائه داده های EEG و از بخش پزشکی مسیر (دانشگاه ادینبورگ) برای داده های بیان ژن پایه سپاسگزاریم.

## اعلام تضاد منافع

نویسنده(گان) هیچ گونه تضاد منافع بالقوه ای را در رابطه با تحقیق، تألیف و/یا انتشار این مقاله اعلام نکردند.

## بودجه

نویسنده(گان) دریافت حمایت مالی زیر را برای تحقیق، تألیف و/یا انتشار این مقاله اعلام کردند: MK، RD و AS از مرکز تحقیقات زیست پزشکی سلامت روان موسسه ملی تحقیقات سلامت (NIHR) در جنوب لندن و بنیاد NHS مادرلی و کالج کینگ لندن، حمایت مالی دریافت می کنند. دیدگاه های بیان شده متعلق به نویسنده(گان) است و لزوماً دیدگاه های NIHR، NHS یا وزارت بهداشت نیست.

این کار با استفاده از منابع ارائه شده توسط (<http://www.ecdf.ed.ac.uk/>) ECDF و مرکز تحقیقات زیست پزشکی سلامت روان (<http://compbio.brc.iop.kcl.ac.uk/cluster/>) NIHR انجام شده است. ECDF تا حدی توسط طرح (<http://www.edikt.org.uk/>) پشتیبانی می شود. جمع آوری و اشتراک گذاری داده های تصویربرداری مغز که در یکی از نمونه های این مطالعه استفاده شده است، توسط طرح ابتکاری تصویربرداری عصبی بیماری آلزایمر (ADNI) (مؤسسات ملی بهداشت، کمک هزینه U01 AG024904) تأمین مالی شده است. ADNI توسط مؤسسه ملی پیری، مؤسسه ملی تصویربرداری زیست پزشکی و مهندسی زیستی و از طریق کمک های سخاوتمندانه موارد زیر تأمین مالی می شود: Abbott؛ انجمن آلزایمر؛ بنیاد کشف داروی آلزایمر؛ Amorfis Life Sciences Ltd؛ AstraZeneca؛ Bayer HealthCare؛ BioClinica، Inc؛ Biogen Idec Inc؛ Bristol-Myers Squibb Company؛ Eisai Inc؛ Elan Pharmaceuticals Inc؛ Eli Lilly and Company؛ F. Hoffmann-La Roche Ltd؛ شرکت وابسته آن Research & Development، LLC؛ IXICO Ltd؛ Innogenetics، NV؛ GE Healthcare؛ Genentech، Inc؛ Johnson & Johnson Pharmaceutical Research & Development LLC؛ Janssen Alzheimer Immunotherapy؛ Medpace، Inc؛ شرکت مرک و شرکا؛ شرکت تشخیص بیماری آلزایمر مزو اسکیل؛ شرکت داروسازی نووارتیس؛ فایزر؛ سرویر؛ سینارک؛ و شرکت داروسازی تاکدا. مؤسسات تحقیقات سلامت کانادا بودجه ای را برای حمایت از مراکز بالینی ADNI در کانادا فراهم می کنند. کمک های مالی بخش خصوصی در ۷ نوامبر ۲۰۱۲ توسط بنیاد مؤسسات ملی بهداشت ([www.fnih.org](http://www.fnih.org)) تسهیل می شود. سازمان دریافت کننده کمک هزینه، موسسه تحقیقات و آموزش کالیفرنیا شمالی است و این مطالعه توسط مطالعه مشارکتی بیماری آلزایمر در دانشگاه کالیفرنیا، سن دیگو هماهنگ شده است. داده های ADNI توسط آزمایشگاه تصویربرداری عصبی در دانشگاه کالیفرنیا، لس آنجلس منتشر می شود. این تحقیق همچنین توسط کمک های مالی NIH P30 AG010129 و K01 AG030514 پشتیبانی شده است.

## منابع

۱. دودیت اس، فریدلاند جی و اسپید تی. مقایسه روش های تمایز برای طبقه بندی تومورها با استفاده از داده های بیان ژن. انجمن آمریکای ۲۰۰۲؛ ۹۷: ۷۷-۸۷.
۲. تان ای و گیلبرت دی. مقایسه تجربی تکنیک های یادگیری ماشین تحت نظارت در بیوانفورماتیک. در: مجموعه مقالات اولین کنفرانس بیوانفورماتیک آسیا و اقیانوسیه در مورد بیوانفورماتیک ۲۰۰۳ - جلد ۱۹، دارلینگهست، استرالیا. انجمن کامپیوتر استرالیا، ۲۰۰۳، صفحات ۲۱۹-۲۲۲.
۳. من ام، دایسون جی، جانسون کی و همکاران. ارزیابی روش های طبقه بندی داده های بیان ژن. مجله بیوفارم استات ۲۰۰۴؛ ۱۴: ۱۰۶۵-۱۰۸۴.
۴. استانتیکوف آ، آلفریس سی، تساماردینوس آی و همکاران. ارزیابی جامع روش های طبقه بندی چنددسته ای برای تشخیص سرطان با استفاده از ریزآرایه بیان ژن. بیوانفورماتیک ۲۰۰۵؛ ۲۱: ۶۳۱-۶۴۳.
۵. لی جی، لی جی، پارک ام و همکاران. مقایسه ای گسترده از ابزارهای طبقه بندی اخیر که بر روی داده های ریزآرایه اعمال می شوند. تحلیل داده های آماری محاسباتی ۲۰۰۵؛ ۴۸: ۸۶۹-۸۸۵.

۶. هوانگ ایکس، پن دلیو، گریندل اس و همکاران. مطالعه تطبیقی تشخیص علت نارسایی قلبی انسان با استفاده از پروفایل های بیان ژن. بیوانفورماتیک BMC ۲۰۰۵: ۲۰۵: ۲۰۵-۲۰۵.
۷. دیاز-اوربارت آر و دی آندرس آ. انتخاب ژن و طبقه بندی داده های ریزآرایه با استفاده از جنگل تصادفی. بیوانفورماتیک BMC ۲۰۰۶: ۲۰۷: ۳۰۷-۳۰۷.
۸. دوسات ن، مانگه آ، سولاسول جی و همکاران. مقایسه روش های طبقه بندی نظارت شده برای پروفایل پروتئین در تشخیص سرطان. سرطان اینفورم ۲۰۰۷: ۳: ۲۹۵-۳۰۵.
۹. پیروزنیا م، یانگ ج، یانگ م و همکاران. مطالعه تطبیقی روش های مختلف یادگیری ماشین بر روی داده های بیان ژن ریزآرایه. ژنومیک BMC ۲۰۰۸: ۹(۲): ۵۱۳.
۱۰. استانتیکوف آ، وانگ ال و آلفیرس سی. مقایسه جامع جنگل های تصادفی و ماشین های بردار پشتیبان برای طبقه بندی سرطان مبتنی بر ریزآرایه. بیوانفورماتیک BMC ۲۰۰۸: ۹: ۳۹۹.
۱۱. بولستیکس ای، استروبل سی، آگوستین تی و همکاران. ارزیابی طبقه بندی کننده های مبتنی بر ریزآرایه: یک مرور کلی. سرطان اینفورم ۲۰۰۸: ۶: ۷۷-۹۷.
۱۲. راک دی، ایدرک تی، تروپاناسکیا او و همکاران. مقالاتی در مورد نرمال سازی، انتخاب متغیر، طبقه بندی یا خوشه بندی داده های ریزآرایه. بیوانفورماتیک ۲۰۰۹: ۲۵: ۷۰۱-۷۰۲.
۱۳. یوسفی م، هوا ج، سیما س و همکاران. گزارش سوگیری هنگام استفاده از مجموعه داده های واقعی برای تحلیل عملکرد طبقه بندی. بیوانفورماتیک ۲۰۱۰: ۲۶: ۶۸-۷۶.
۱۴. هانزار بی و دوگرتی ای. در مورد مقایسه طبقه بندی کننده ها برای داده های ریزآرایه. بیوانفورماتیک ۲۰۱۰: ۲۰: ۳۹۵-۳۹۹.
۱۵. یوسفی م، هوا ج و داگرتی ای. سوگیری چند قاعده ای در مقایسه قوانین طبقه بندی. بیوانفورماتیک ۲۰۱۱: ۲۷: ۱۶۷۵-۱۶۸۳.
۱۶. فیشر آر. استفاده از اندازه گیری های چندگانه در مسائل طبقه بندی. آن هیومن ژنه ۱۹۳۶: ۷: ۱۷۹-۱۸۸.
۱۷. هوا جی، شیونگ زد، لویی جی و همکاران. تعداد بهینه ویژگی ها به عنوان تابعی از اندازه نمونه برای قوانین طبقه بندی مختلف. بیوانفورماتیک ۲۰۰۵: ۱۵: ۱۵۰۹-۱۵۱۵.
۱۸. واینیک وی. ان. ماهیت نظریه یادگیری آماری. نیویورک: اشپرینگر، ۱۹۹۵.
۱۹. کورتس سی و واینیک وی. شبکه های بردار پشتیبان. یادگیری ماشین ۱۹۹۵: ۲: ۲۷۳-۲۹۷.
۲۰. واینیک وی. ان. نظریه یادگیری آماری. نیویورک: جان وایلی و پسران، ۱۹۹۸.
۲۱. اسکروندال ای. طراحی و تحلیل آزمایش های مونت کارلو: حمله به خرد متعارف. پژوهش های رفتار چندمتغیره ۲۰۰۰: ۳۵: ۱۳۷-۱۶۷.
۲۲. ولپرت دی. فقدان تمایزات پیشینی بین الگوریتم های یادگیری. محاسبات عصبی ۱۹۹۶: ۱۳۴۱: ۸۰-۱۳۹.
۲۳. بریم ل. جنگل های تصادفی. یادگیری ماشین ۲۰۰۱: ۴۵: ۵-۳۲.
۲۴. فیشر آر. ای. استفاده از اندازه گیری های چندگانه در مسائل طبقه بندی. آن یوگن ۱۹۳۶: ۷: ۱۷۹-۱۸۸.
۲۵. مک لاکلان جی. جی. تحلیل تشخیصی و تشخیص الگوی آماری. نیویورک، ایالات متحده: انتشارات جان وایلی و پسران، ۱۹۹۲.
۲۶. دودا آر. او. هارت پی. ای. و استورک دی. جی. طبقه بندی الگو. نیویورک: وایلی، ۲۰۰۱.
۲۷. داسارانی بی وی. هنجارهای نزدیکترین همسایه (NN): تکنیک های طبقه بندی الگوی NN. لوس آلامیتوس، کالیفرنیا، ایالات متحده: انجمن کامپیوتر IEEE، ۱۹۹۱.
۲۸. شاخاروویچ جی، دارل تی و ایندیک پی. روش های نزدیکترین همسایه در یادگیری و بینایی: نظریه و عمل. کمبریج، ایالات متحده آمریکا: MIT Press، ۲۰۰۵.
۲۹. دیاز-اوربارت آر و آلوارد د آندرس اس. انتخاب ژن و طبقه بندی داده های ریزآرایه با استفاده از جنگل تصادفی. بیوانفورماتیک BMC ۲۰۰۶: ۲۰۷: ۳۰۷-۳۰۷.
۳۰. افرون ب. تخمین نرخ خطای یک قانون پیش بینی: بهبود اعتبارسنجی متقابل. انجمن آمار آمریکا ۱۹۸۳: ۷۸: ۳۱۶-۳۳۱.
۳۱. رایان ام، لاکستون اچ، هافاکر اس و همکاران. تحلیل بیان ژن اختلال دوقطبی، کاهش تنظیم چرخه یوبیکوئیتین و تغییرات در ژن های سیناپسی را نشان می دهد. روانپزشک مول ۲۰۰۶: ۱۱: ۹۶۵-۹۷۸.
۳۲. S, et al. AddNeuroMed بیماری آلزایمر و مطالعه MRI اندازه گیری. Westman E, Muehlboeck 32. Simmons A, ۱۱: ۴۷-۵۵.
۳۳. سیمونز ای، وستمن ای، موهلوبک اس، و همکاران. چارچوب AddNeuroMed برای ارزیابی چند مرکزی بیماری آلزایمر: تجربه از ۲۴ ماه اول. روانپزشک سالمندان بین المللی ۲۰۱۱: ۲۱: ۷۵-۸۲.
۳۴. وستمن ای، اگیلار سی، موهلوبک جی، و همکاران. معیارهای تصویربرداری تشدید مغناطیسی منطقه ای برای تحلیل متغیره در بیماری آلزایمر و اختلال شناختی خفیف. نئوپزگرافی مغز ۲۰۱۲: ۱-۱۵.
۳۵. اسکیرو سی و اشرسون پی. بی ثباتی عاطفی، همبودی و اختلال در بزرگسالان مبتلا به اختلال پیش فعالی-کمبود توجه. اختلال عاطفی ل ۲۰۱۳: ۱۴۷: ۸۰-۸۶.
۳۶. مک لافلین جی، آلبرشت بی، باناشفسکی تی و همکاران. شواهد الکتروفیزیولوژیکی برای حالت های آماده سازی غیرطبیعی و پردازش مهاردر ADHD بزرگسالان. عملکرد مغز بهاو ۲۰۱۰: ۶: ۶۶.
۳۷. تای سی، ریجسیدیک اف، گرون سی و همکاران. تأثیرات ژنتیکی مشترک بر علائم ADHD و فعالیت EEG یا فرکانس بسیار پایین: مطالعه ای بر روی دوقلوها. مجله روانشناسی کودک، روانپزشک ۲۰۱۲: ۵۳: ۶-۷: ۷۱۵-۷۱۵.
۳۸. اسمیت سی ال، دیکینسون پی، فورستر تی و همکاران. ارزیابی کمی RNA خون کامل انسان به عنوان یک نشانگر زیستی بالقوه برای بیماری های عصبی. تحلیلگر ۲۰۰۷: ۱۳۲: ۱۲۰۰-۱۲۰۹.
۳۹. هاردین جی، گارسیا اس آر و گولان دی. روشی برای تولید ماتریس های همبستگی واقع بینانه. preprint arXiv:1106.5834. ۲۰۱۱arXiv (۱۳۹۰).
۴۰. خوندوکرو ام، باخمن تی، میویسن ام و همکاران. تحلیل چند عاملی خطای پیش بینی کلاس: تخمین تعداد بهینه نشانگرهای زیستی برای قوانین طبقه بندی مختلف. مجله بیوانفورماتیک محاسباتی ۲۰۱۰: ۸: ۹۴۵-۹۶۵.
۴۱. اسمیت جی کی. مدل های خطی و روش های تجربی بیز برای ارزیابی بیان افتراقی در آزمایش های ریزآرایه. کاربرد آماری ژن مول بیولوژی ۲۰۰۴: ۳: ۳.