دانشکده صنایع و سیستم ها
گروه مهندسی فناوری اطلاعات

رساله جهت دریافت مدرک دکتری

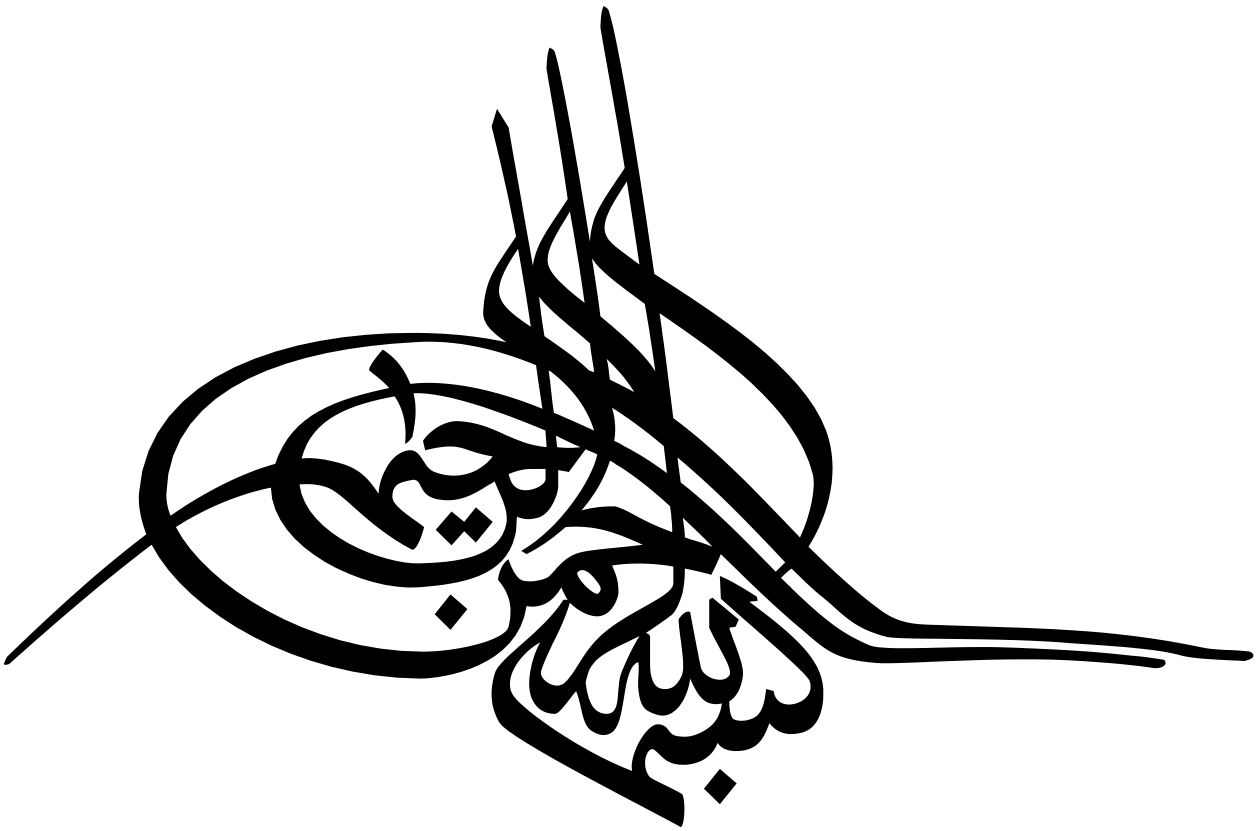
طراحی و ارائه مدلی هوشمند جهت پیش‌بینی خطر پرفشاری خون مبتنی بر SNP در شرایط عدم قطعیت

نگارش
سید علی لاجوردی

استاد راهنما
دکتر مهرداد کارگری

استاد راهنما دوم
دکتر مریم السادات دانشپور

خرداد ۱۴۰۲


بسم الله الرحمن الرحيم

دانشکده

مهندسی صنایع و سیستم ها

بسمه‌تعالی

تاییدیه اعضای هیات داوران حاضر در جلسه دفاع از رساله دکتری

بدینوسیله گواهی می شود **آقای سیدعلی لاجوردی** به شماره دانشجویی **۹۶۶۶۲۳۲۰۰۳** رشته **مهندسی فناوری اطلاعات - مدیریت سیستمهای اطلاعاتی** در تاریخ **۱۴۰۲/۰۳/۲۷** از رساله دکتری ۱۹ واحدی خود با عنوان: **طراحی و ارائه مدلی هوشمند جهت پیش‌بینی خطر پرفشاری خون مبتنی بر SNP در شرایط عدم قطعیت**

Title: **Design and presenting an intelligent model for predicting SNP-based hypertension in uncertainty**

دفاع کرده است. اعضای هیأت داوران نسخه نهایی این رساله را از نظر فرم و محتوا بررسی نموده و پذیرش آنرا برای دریافت درجه دکتری تخصصی (Ph.D) تایید می نماید.

اعضای هیأت داوران	نام و نام خانوادگی	رتبه علمی	امضاء
استاد راهنما	دکتر مهرداد کارگری	دانشیار	
استاد راهنمای دوم	دکتر مریم السادات دانشپور	دانشیار	
استاد مشاور	دکتر مهدی اکبرزاده	استادیار	
استاد داور داخلی	دکتر محمدمهدی سپهری	استاد	
استاد داور داخلی	دکتر توکتم خطیبی	دانشیار	
استاد داور خارجی	دکتر بهروز مینایی بیدگلی	دانشیار	
استاد داور خارجی	دکتر مهرداد روانشاد	دانشیار	
نماینده شورای تحصیلات تکمیلی	دکتر توکتم خطیبی	دانشیار	

آیین‌نامه حق مالکیت مادی و معنوی در مورد نتایج پژوهش‌های علمی دانشگاه تربیت مدرس

مقدمه: با عنایت به سیاست‌های پژوهشی و فناوری دانشگاه در راستای تحقق عدالت و کرامت انسان‌ها که لازمه شکوفایی علمی و فنی است و رعایت حقوق مادی و معنوی دانشگاه و پژوهشگران، لازم است اعضای هیئت‌علمی، دانشجویان، دانش‌آموختگان و دیگر همکاران طرح، در مورد نتایج پژوهش‌های علمی که تحت عناوین پایان‌نامه، رساله و طرح‌های تحقیقاتی با هماهنگی دانشگاه انجام‌شده است، موارد زیر را رعایت نمایند:

ماده ۱ -حق نشر و تکثیر پایان‌نامه/ رساله و درآمدهای حاصل از آن‌ها متعلق به دانشگاه هست ولی حقوق معنوی پدیدآورندگان محفوظ خواهد بود.

ماده ۲ -انتشار مقاله یا مقالات مستخرج از پایان‌نامه/ رساله به‌صورت چاپ در نشریات علمی و یا ارائه در مجامع علمی باید به نام دانشگاه بوده و با تائید استاد راهنمای اصلی، یکی از اساتید راهنما، مشاور و یا دانشجو مسئول مکاتبات مقاله باشد. ولی مسئولیت علمی مقاله مستخرج از پایان‌نامه و رساله به عهده اساتید راهنما و دانشجو هست.

تبصره: در مقالاتی که پس از دانش‌آموختگی به‌صورت ترکیبی از اطلاعات جدید و نتایج حاصل از پایان‌نامه/ رساله نیز منتشر می‌شود نیز باید نام دانشگاه درج شود.

ماده ۳ -انتشار کتاب، نرم‌افزار و یا آثار ویژه (اثری هنری مانند فیلم، عکس، نقاشی و نمایشنامه) حاصل از نتایج پایان‌نامه/ رساله و تمامی طرح‌های تحقیقاتی کلیه واحدهای دانشگاه اعم از دانشکده‌ها، مراکز تحقیقاتی، پژوهشکده‌ها، پارک علم و فناوری و دیگر واحدها باید با مجوز کتبی صادره از معاونت پژوهشی دانشگاه و بر اساس آئین‌نامه‌های مصوب انجام شود.

ماده ۴ -ثبت اختراع و تدوین دانش فنی و یا ارائه یافته‌ها در جشنواره‌های ملی، منطقه‌ای و بین‌المللی که حاصل نتایج مستخرج از پایان‌نامه/ رساله و تمامی طرح‌های تحقیقاتی دانشگاه باید با هماهنگی استاد راهنما یا مجری طرح از طریق معاونت پژوهشی دانشگاه انجام گیرد.

ماده ۵ -این آیین‌نامه در ۵ ماده و یک تبصره در تاریخ ۸۷/۴/۱ در شورای پژوهشی و در تاریخ ۸۷/۴/۲۳ در هیئت‌رئیسه دانشگاه به تائید رسید و در جلسه مورخ ۸۷/۷/۱۵ شورای دانشگاه به تصویب رسیده و از تاریخ تصویب در شورای دانشگاه لازم‌الاجرا است.

این‌جانب سید علی لاجوردی دانشجوی رشته مهندسی فناوری اطلاعات - مدیریت دستگاه‌های اطلاعاتی ورودی سال تحصیلی ۱۳۹۶ مقطع دکتری دانشکده مهندسی صنایع و دستگاه‌ها متعهد می‌شوم کلیه نکات مندرج در آئین‌نامه حق مالکیت مادی و معنوی در مورد نتایج پژوهش‌های علمی دانشگاه تربیت مدرس را در انتشاریافته‌های علمی مستخرج از پایان‌نامه / رساله تحصیلی خود رعایت نمایم. در صورت تخلف از مفاد آئین‌نامه فوق به دانشگاه وکالت و نمایندگی می‌دهم که از طرف این‌جانب نسبت به لغو امتیاز اختراع بنام بنده و یا هرگونه امتیاز دیگر و تغییر آن به نام دانشگاه اقدام نماید. ضمناً نسبت به جبران فوری ضرر و زیان حاصله بر اساس برآورد دانشگاه اقدام خواهم نمود و بدین‌وسیله حق هرگونه اعتراض را از خود سلب نمودم.

امضا:

تاریخ: ۱۴۰۲/۴/۴

آیین‌نامه چاپ پایان‌نامه (رساله)های دانشجویان دانشگاه تربیت مدرس

نظر به اینکه چاپ و انتشار پایان‌نامه (رساله)های تحصیلی دانشجویان دانشگاه تربیت مدرس، مبین بخشی از فعالیت‌های علمی - پژوهشی دانشگاه است بنابراین به‌منظور آگاهی و رعایت حقوق دانشگاه، دانش‌آموختگان این دانشگاه نسبت به رعایت موارد ذیل متعهد می‌شوند:

ماده ۱: در صورت اقدام به چاپ پایان‌نامه (رساله)ی خود، مراتب را قبلاً به‌طور کتبی به دفتر نشر آثار علمی دانشگاه اطلاع دهد.

ماده ۲: در صفحه سوم کتاب (پس از برگ شناسنامه) عبارت ذیل را چاپ کند:

کتاب حاضر، حاصل رساله دکتری نگارنده سید علی لاجوردی در رشته مهندسی فناوری اطلاعات – مدیریت دستگاه‌های اطلاعاتی است که در سال ۱۴۰۱ در دانشکده مهندسی صنایع و سیستم‌ها دانشگاه تربیت مدرس به راهنمایی جناب آقای دکتر مهرداد کارگری و سرکار خانم دکتر مریم السادات دانشپور و مشاوره جناب آقای دکتر مهدی اکبرزاده از آن دفاع شده است.

ماده ۳: به‌منظور جبران بخشی از هزینه‌های انتشارات دانشگاه، تعداد یک درصد شمارگان کتاب (در هر نوبت چاپ) را به دفتر نشر آثار علمی دانشگاه اهدا کند. دانشگاه می‌تواند مازاد نیاز خود را به نفع مرکز نشر در معرض فروش قرار دهد.

ماده ۴: در صورت عدم رعایت ماده ۳، ۵۰٪ بهای شمارگان چاپ‌شده را به‌عنوان خسارت به دانشگاه تربیت مدرس، تأدیه کند.

ماده ۵: دانشجو تعهد و قبول می‌کند در صورت خودداری از پرداخت بهای خسارت، دانشگاه می‌تواند خسارت مذکور را از طریق مراجع قضایی مطالبه و وصول کند؛ به‌علاوه به دانشگاه حق می‌دهد به‌منظور استیفای حقوق خود، از طریق دادگاه، معادل وجه مذکور در ماده ۴ را از محل توقیف کتاب‌های عرضه‌شده نگارنده برای فروش، تأمین نماید.

ماده ۶: این‌جانب سید علی لاجوردی دانشجوی رشته مهندسی فناوری اطلاعات – مدیریت دستگاه‌های اطلاعاتی مقطع دکتری تعهد فوق و ضمانت اجرایی آن را قبول کرده، به آن ملتزم می‌شوم.

نام و نام خانوادگی: سید علی لاجوردی

تاریخ و امضا: ۱۴۰۲/۴/۴

دانشکده صنایع و سیستم ها
گروه مهندسی فناوری اطلاعات

رساله دکتری

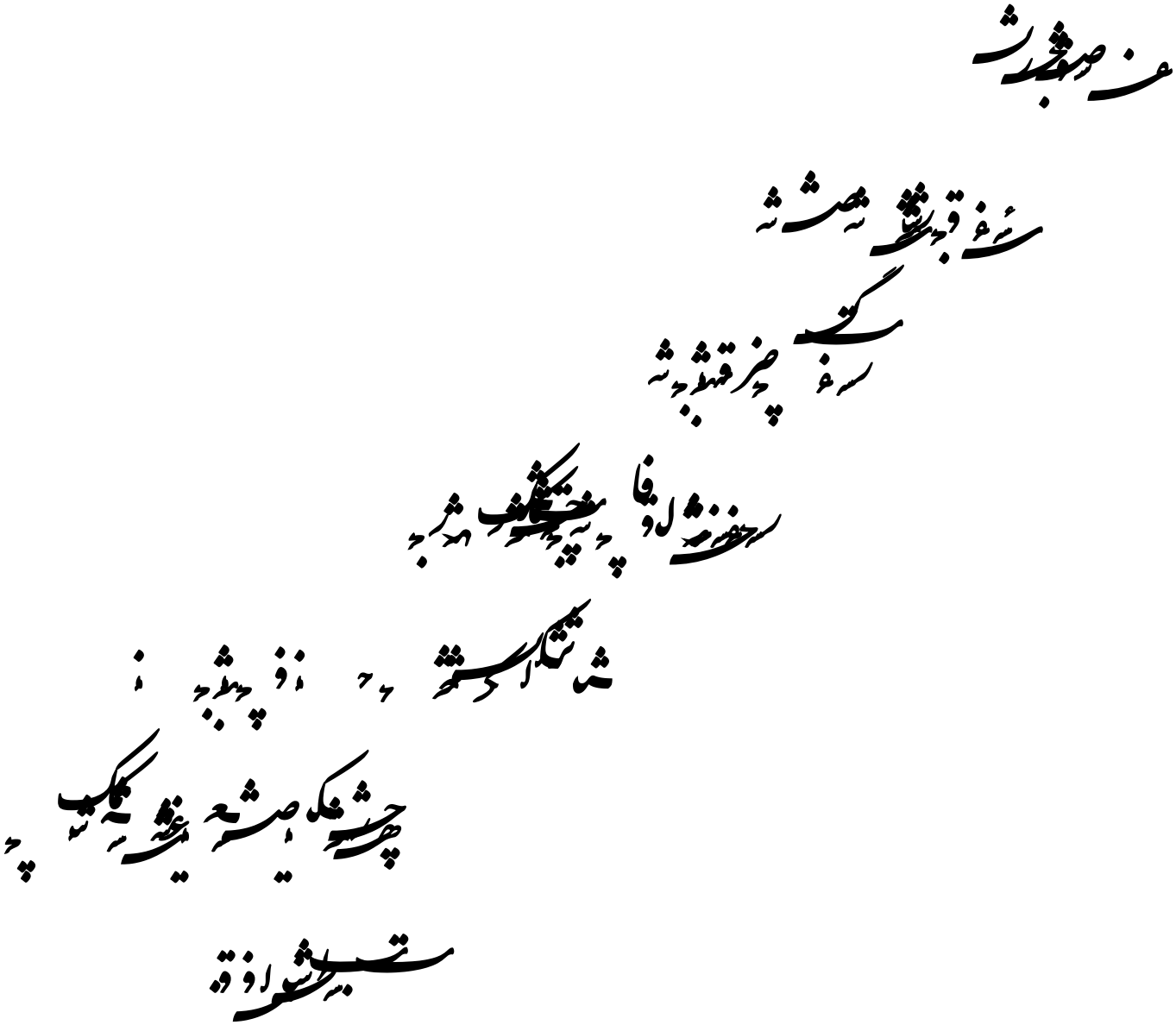
طراحی و ارائه مدلی هوشمند جهت پیش‌بینی خطر پرفشاری خون مبتنی بر SNP در شرایط عدم قطعیت

نگارش
سید علی لاجوردی

اساتید راهنما
دکتر مهرداد کارگری
دکتر مریم السادات دانشپور

استاد مشاور
دکتر مهدی اکبرزاده

خرداد ۱۴۰۲



چکیده

در دهه‌های اخیر، پزشکی شخصی بر مبنای ژنومیک برای ارائه خدمات مناسب و مؤثر برای بیماران بسته به ویژگی‌های ژنتیکی آن‌ها ارائه‌شده است. همچنین مطالعات گسترده ارتباط ژنومی انواع عوامل خطر ژنتیکی مبتنی بر جمعیت را برای بیماری‌های رایج و پیچیده شناسایی کرده است. جهت ارتقای سطح پزشکی شخصی، پژوهش‌های صورت گرفته در تلاش‌اند تا علاوه برافزایش درک عمیق ژنومیک، به بهبود مراقبت‌های پزشکی شخصی با استفاده از مدل‌های پیش‌آگهی دقیق‌تر بیماری‌ها کمک کنند. امتیاز خطر چندژنی و یادگیری ماشین، دو روش اولیه برای پیش‌آگهی خطر بیماری هستند. علیرغم پیشرفت‌های اخیر، نتایج به‌دست‌آمده از خطرهای چندژنی به دلیل روش‌هایی که در حال حاضر مورداستفاده قرار می‌گیرند، محدود است. این در حالی است که در طول زمان، الگوریتم‌های یادگیری ماشین توانایی‌شان برای پیش‌آگهی خطر بیماری‌های پیچیده بیشتر شده است. این افزایش توان قدرت پیش‌آگهی از توانایی الگوریتم‌های یادگیری ماشین برای بررسی داده‌های چندبعدی است.

در این تحقیق، روش‌های مختلف یادگیری ماشین و یادگیری عمیق جهت مدل‌سازی خطر بیماری پرفشاری خون بر اساس SNP بر روی‌داده‌های ۲۰ ساله طرح ژنتیک کاردیومتابولیک تهران، پژوهشکده علوم غدد درون‌ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی که شامل داده‌های ژنومی به‌صورت تراشه با ۶۵۲۹۱۹ SNP برای ۱۱۰۸۸ نفر در دوره‌های مختلف هست، پیاده‌سازی و در ادامه نتایج و دستاوردهای آن بیان می‌گردد.

روش پردازش اولیه و مدل‌سازی داده به خصوص در شبکه یادگیری عمیق از ادبیات موضوع متفاوت و نتایج تحقیق در دو بخش یادگیری ماشین و یادگیری عمیق دارای AUC به ترتیب ۰.۸۷ و ۰.۸۹ هست و ضمناً این مدل‌سازی امکان پیش‌آگهی در سنین جوانی را با دقت مناسبی فراهم کرده است.

کلمات کلیدی:

یادگیری عمیق، نشانگرهای ژنومی، پرفشاری خون، خطر بیماری

فهرست

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

جدول اختصارات

اخصار	معادل انگلیسی	معادل فارسی
AUC	Area Under Curve	ناحیه زیر منحنی
BMI	Body Mass Index	شاخص توده بدن
BP	Blood Pressure	فشارخون
CNN	Convolutional Neural Network	شبکه عصبی پیچشی
CNV	Copy number variation	تغییرات تعداد کپی
DBP	Diastolic Blood Pressure	فشارخون دیاستولیک
DL	Deep learning	یادگیری عمیق
DNN	Deep Neural Network	شبکه عصبی عمیق
DP	Dynamic Programming	برنامه نویسی پویا
EHR	Electronic Health Record	پرونده الکترونیک سلامت
FCN	Fully Connected Network	شبکه کاملا متصل
GDL	Genome Deep Learning	یادگیری عمیق ژنوم
GRM	Genetic Relation Matrix	ماتریس رابطه ژنتیکی
GWAS	Genome Wide Association Study	مطالعه ارتباط گسترده ژنوم
HGP	Human Genome Project	پروژه ژنوم انسان
HTN	Hypertension	پرفشاری خون
IRGP	Iranian Reference Genome Project	پروژه ژنوم مرجع ایران
ML	Machine Learning	یادگیری ماشین
MLM	Mixed Linear Model	مدل خطی مختلط
MSP	Multiple Sequence Alignment	همترازی توالی چندگانه
NPV	Negative Predictive Value	ارزش پیش‌بینی منفی
PCA	Principal Component Analysis	تحلیل مؤلفه های اصلی
PCR	Polymerase chain reactions	واکنش های زنجیره‌ای پلیمراز
PheWAS	Phenome-wide association study	مطالعه ارتباط گسترده پدیده
PPV	Positive Prediction Value	ارزش پیش‌بینی مثبت
PRS	Polygenic Risk Score	امتیاز خطر چند ژنی
PSA	Pairwise Sequence Alignment	همترازی توالی دوگانه
RFLPs	Restriction fragment length polymorphisms	چند ریختی طول قطعه محدود
RNN	Recurrent neural network	شبکه عصبی گردشی
ROC	Receiver Operating Characteristic	ویژگی عملکرد گیرنده
SBP	Systolic Blood Pressure	فشارخون سیستولیک
SN	Social networks	شبکه‌های اجتماعی
SNP	Single Nucleotide Polymorphism	چند ریختی تک نوکلئوتیدی
SNV	Single nucleotide variation	تغییرات تک نوکلئوتیدی
STR	short tandem repeats	تکرارهای پشت سر هم کوتاه
SVM	Support Vector Machine	ماشین بردار پشتیبان
TNR	True Negative Rate	نرخ منفی واقعی
TPR	True Positive Rate	نرخ مثبت واقعی
VNTRs	Variable number of tandem repeats	تعداد متغیر تکرارهای پشت سر هم
WGS	Whole genome sequencing	توالی یابی کل ژنوم

۱- مقدمه

با هر مرحله از پیشرفت علم پزشکی، دانش و ابزارهای مورداستفاده برای تشخیص بیماری‌ها تغییر نموده و از متافیزیک به فیزیک، از سلول به مولکول و اخیراً هم به سطحی رسیده است که می‌توان واکنش بین مولکول‌ها را مطالعه و بررسی نمود. در زمان سقراط علم پزشکی در ارتباط نزدیکی با فلسفه و علوم طبیعی بود اما بقراط این شیوه را تغییر داد و پزشکی را به‌صورت حرفه‌ای و به‌دوراز فلسفه طبیعی بررسی کرد. بقراط تشخیص بیماری را بر پایه تحقیق در مورد احوالات بیمار قرارداد و اولین فردی بود که بسیاری از بیماری‌ها را توضیح و تشخیص داد، همچنین استعداد ابتلا به بیماری‌های مختلف در جمعیت‌های مختلف آسیایی و اروپایی را که از طریق وراثت باقی می‌ماندند، را موردتوجه قرارداد (Adams, 1922). بنابراین احتمالاً بقراط اولین فردی بود که زمینه‌هایی را ایجاد کرد که درنهایت نظریه اصلی پزشکی منحصربه‌فرد بر پایه تشخیص بیماری برای هر فرد، شکل گیرد. اما علاوه بر تشخیص بیماری، تجویز داروی مناسب هم در درمان بیماری تأثیر به سزایی دارد. به‌طورکلی داروهایی که برای بیماران تجویز می‌شوند اصطلاحاً به‌صورت "یکی برای همه" است و همواره این سؤال مطرح می‌شد که چرا افراد مختلف در پاسخ به داروهایی که برایشان تجویزشده است، اثرات مطلوب و نامطلوب متفاوتی نشان می‌دهند؟ دانشمندان متوجه شدند که راز این امر احتمالاً در ژن‌ها نهفته است.

۱-۱- پزشکی شخصی

برای بالا بردن کیفیت درمانی باید خصوصیات ژنتیکی افراد را هم در نظر گرفت (Rafiei and Amjadi, 2013). با این دیدگاه واژه پزشکی شخصی[1] وارد علم شد. در سال ۱۹۹۸، زمانی که برای اولین بار رساله‌ای با عنوان شخصی کردن پزشکی انتشار یافت، توجه کمی به این موضوع شد (Jain, 2006).

پیشرفت‌های روزافزون علم پزشکی سبب شد که تشخیص و درمان بیماری‌ها کارآمدتر شود. یکی از این پیشرفت‌ها پروژه ژنوم انسان بود که با تکمیل شدن آن دریچه تازه‌ای از دیدگاه علم ژنتیک به سمت دنیای پزشکی باز شد. تکمیل این پروژه سرآغاز یک دوره جدیدی به نام پسا - ژنومی بود که سبب شد مفاهیم علمی جدید هم چون ژنومیک، پروتئومیک، فارماکوژنومیک و موارد مشابه وارد دنیای پزشکی شود. توالی یابی ژنوم انسان به همراه پیشرفت و توسعه فنّاوری‌های پربازده که عمدتاً به واژه Omics ختم می‌شود، دانش بشری را در مورد بیماری و سلامتی بهبود بخشیده و پایه‌ای را برای پزشکی انفرادی فراهم کردند.

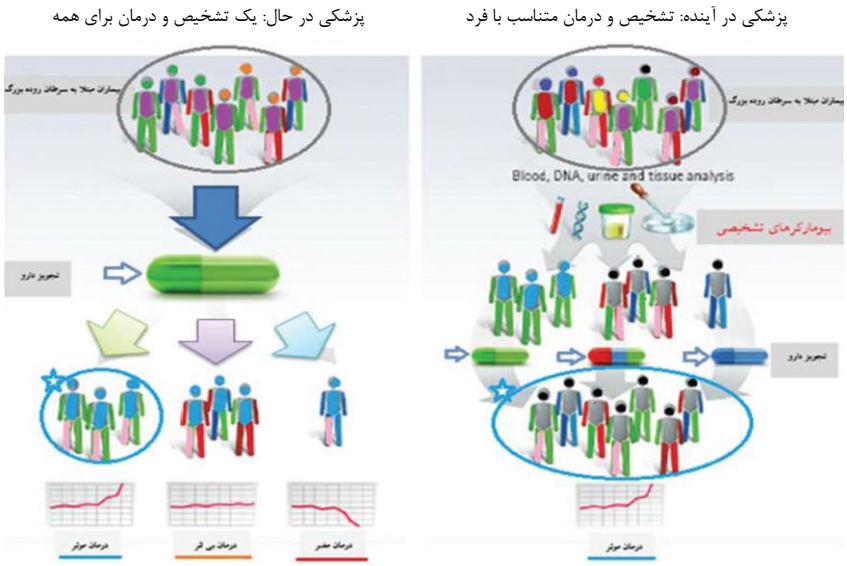
[1] Personalized Medicine

دامنه پزشکی فردی بسیار گسترده‌تر از آن است که با واژه طب ژنومی تعریف شود. علاوه بر ژنتیک، باید اپی ژنتیک و عوامل زیست‌محیطی نیز در نظر گرفته شود. بهترین راه برای تعبیر پزشکی فردی ادغام فن‌آوری‌های جدید برای عملکرد بالینی و تأثیر مهم آن در تشخیص مولکولی است. سامانه‌های زیست‌شناسی با رویکرد پزشکی برای توسعه اختصاصی نمودن درمان مهم است (Bauer *et al.*, 2015).

در حال حاضر، دارویی که برای درمان یک فرد بیمار تجویز می‌شود تقریباً برای تمام مبتلایان به آن بیماری یکسان است. این در حالی است که واکنش بیمار به دارو اثرات مطلوب و نامطلوب و یا اصطلاحاً ناسازگاری دارویی وجود دارد. از طرفی بروز و شدت بیماری‌ها در افراد مختلف با توجه به دخالت عوامل ژنتیکی، محیطی و اپی ژنتیکی متفاوت است.

بااین‌حال درمان‌های رایج برای همه این بیماران به‌صورت یکنواخت انجام می‌پذیرد. آمار نشان می‌دهد واکنش دارویی ناسازگار سالانه عامل مرگ بیش از صد هزار نفر در ایالات‌متحده هست و میلیون‌ها نفر به دلیل عوارض جانبی شدید بستری می‌شوند (Strachan and Read, 2018).

در پزشکی فردی الگوهای خاص سلولی و متابولیک در فاز تشخیص به‌عنوان زیست‌نشانگرهای[1] تشخیصی استفاده می‌شود که بیماران را بر اساس شباهت‌های موجود معرفی می‌نماید و بهترین اطلاعات درمان‌های فردی را ارائه می‌دهد (شکل ۱-۱) (Strachan and Read, 2018). پیشرفت‌های چشمگیر در تشخیص و درمان بیماری بر اساس تفاوت‌های ژنوم افراد در حوزه پزشکی فردی ایجادشده است (Ginsburg and Willard, 2009).



شکل ۱-۱ اثربخشی درمان تابع ژنوم و محیط در وضعیت حال و آینده (Strachan and Read, 2018)

ژنوم انسان‌ها علیرغم وجود شباهت‌های زیاد دارای تفاوت‌های جزئی است که آن را منحصربه‌فرد می‌نماید. درنتیجه باعث بروز درمان فردی کاملاً با تحقیقاتی که برای تعیین تغییرات و تفاوت‌های ژنوم انسانی انجام می‌شود گره‌خورده است.

[1] Biomarker

۱-۱-۱- پزشکی دقیق

هدف پزشکی دقیق، ارائه درمان‌های پزشکی مناسب برای بیماران با توجه به ویژگی‌های ژنتیکی آن‌ها هست. این امر در درجه اول شامل شخصی‌سازی مراقبت‌های پیشگیرانه و فعال به‌منظور به حداکثر رساندن کارایی درمان و کاهش هزینه‌ها است. شخصی‌سازی از طریق استفاده از انواع مختلف اطلاعات اومیکس[1] و ادغام آن‌ها باهدف درک خطرات بیماری حاصل می‌شود.

کاربرد پزشکی دقیق در فارماکوژنومیک باعث موفقیت‌های قابل‌توجهی درزمینهٔ ارائه داروهای سفارشی و مصرف آن‌ها در اندازه‌های مناسب گردیده است. به‌عنوان‌مثال، اطلاعات ژنتیکی به‌طور مرتب در استراتژی‌های درمانی برای درمان سرطان‌های پستان HER2 مثبت با داروی trastuzumab، درمان سرطان‌های ریه دارای افزایش بیان EGFR با erlotinib، یا درمان لوسمی میلوژنی مزمن دارای کروموزوم فیلادلفیا با imatinib مورداستفاده قرارگرفته است.

بااین‌حال، در مورد این موضوع که آیا ژنومیک دقیق در نقطه‌ای قرارگرفته است که بتواند منافع بیشتری را در رویکردهای بهداشت عمومی استاندارد ارائه دهد یا خیر، در حوزه سلامت عمومی هنوز هم بحث وجود دارد.

۱-۲- ژنوم انسان

ژنوم انسان مجموعه‌ای کامل از DNA ، شامل تمام اطلاعات موردنیاز برای حیات انسان است، که از ۳ میلیارد جفت باز تشکیل‌شده است. در هسته سلول‌های انسان ۲۳ جفت کروموزوم وجود دارد که از هر جفت کروموزوم یک نسخه از مادر (تخم) و دیگری از پدر (اسپرم) به فرزند می‌رسد. هر کروموزوم متشکل از یک مولکول درشت به نام داکسی ریبونوکلئیک‌اسید یا همان DNA است. مجموع ۲۳ جفت مولکول DNA فرد حاوی تمام اطلاعات لازم برای تعیین صفاتی است که فرد را می‌سازد. DNA خود را به بخش‌های خاصی به نام ژن سازمان‌دهی می‌کند که پروتئین‌های ضروری برای رشد و عملکرد بدن ما را رمزگذاری می‌کند. کدگذاری اختصاصی درون ژن‌ها از طریق ترکیب چهار باز خاص به نام‌های آدنین، تیمین، گوانین و سیتوزین انجام‌شده است.

۱-۲-۱-ژنومیک و ژنتیک

توالی جفت بازها در رشته DNA، محتوی پیام ژنتیکی را تعیین می‌کند. به عبارتی ژن، توالی خاصی از این بازها است که حاوی اطلاعات لازم جهت ساخت یک محصول مانند یک هورمون پروتئینی یا آنزیم را فراهم می‌کند. انسان‌ها هزاران ژن دارند که در سراسر طول DNA و در قالب ۲۳ جفت کروموزوم بسته‌بندی‌شده است. ژنومیک، مطالعه تمام ژن‌های ژنوم انسان است.

ژنتیک معمولاً به بررسی تک‌تک ژن‌ها و نقش آن‌ها در بیماری یا وراثت اشاره دارد. ژنومیک به تمامیت اطلاعات ژنتیکی یک فرد اشاره دارد و به‌توالی ژنتیک ژن‌ها، ساختار و عملکرد آن‌ها و همچنین تعاملات بین ژن‌ها می‌پردازد.

[1] omics

امروزه تعیین توالی کل ژنوم انسانی ممکن شده است، تا بتواند اساساً "نقشه" ژنوم شخصی انسان را، تحت عنوان توالی کل ژنوم (WGS)، ارائه کند. دانستن بیشتر در مورد DNA، به انسان این امکان را می‌دهد که بالقوه به‌سوی زندگی سالم‌تری حرکت کند. بررسی کامل ژنوم انسان ازلحاظ تاریخی مبتنی بر برنامه‌های تشخیصی بالینی بود و در آن توجه کمتری به پیشگیری از بیماری و مدیریت سلامت داده‌شده بود. بااین‌حال، این نوع از اطلاعات در حال حاضر به‌عنوان بخشی از یک استراتژی مراقبت‌های بهداشتی پیشگیرانه فردی به‌سرعت در حال رشد است و با توجه به تشخیص زودهنگام بیماری، فرصتی برای درمان و نتایج بهبود بهتر به وجود می‌آورد.

۱-۲-۲- آلل[1]

در هر رشته DNA مجموعه‌ای از اطلاعات به نام ژن‌ها قرار دارند. این ژن‌ها هستند که صفات مختلف جانوران یا انسان‌ها را تعیین می‌کنند. ژن‌ها برای همه انسان‌ها یکسان هستند بنابراین همه انسان‌ها ژن‌های یکسانی را به ارث می‌برند. اما این آلل‌ها هستند که تعیین می‌کنند ژن‌ها چگونه در یک فرد بروز پیدا کنند و آشکار شوند. به‌عنوان‌مثال وجود رنگ چشم به ژن‌ها مربوط است اما این‌که چه رنگ چشمی از والدین به ارث برده می‌شود به‌وسیله آلل‌ها تعیین می‌شود.

آلل‌ها به‌صورت جفت هستند و به عبارتی موجودات زنده برای هر صفت دو آلل دارد. (به‌عنوان‌مثال آلل قهوه‌ای و آبی برای چشم در یک نفر.) یک آلل روی یک کروموزوم و آلل دیگر روی کروموزوم دیگر قرار دارد. هر فرد یک آلل را از مادر و آلل دیگر را از پدر به ارث می‌برد.

تفاوت آلل‌ها و ژن‌ها در این است که ژن‌ها جفت نیستند چراکه چنان‌که گفته شد ژن‌ها در همه انسان‌ها یکسان هستند. به همین دلیل که گفته شد یک جفت آلل رخ‌نمودهای متفاوت به وجود می‌آورند اما این موضوع درباره ژن‌ها صادق نیست.

قانون وراثت مندل فرایندی که طی آن‌یک آلل انتقال می‌یابد را فرمول‌بندی کرده است.

چنان‌که گفته شد موجودات زنده دو آلل برای هر صفت دارند. به وضعیتی که دو آلل متفاوت برای یک صفت وجود داشته باشد هتروزیگوت[2] گفته می‌شود. در نوعی از هتروزیگوت یکی از دو آلل غالب است و دیگری مغلوب و درنتیجه آلل غالب آشکار می‌شود و آلل مغلوب پنهان. به‌عنوان‌مثال، در انسان آلل برای رنگ چشم قهوه‌ای غالب است و آلل برای رنگ چشم آبی مغلوب است.

غالب بودن رنگ قهوه‌ای بر رنگ آبی در چشم یک نمونه از تسلط کامل است. در روابط هتروزیگوتی که در آن هیچ‌کدام از آلل‌ها غالب نیستند اما هر دو به‌طور کامل آشکارند، وضعیت هم بارز گفته می‌شود. نمونه وضعیت هم بارز، به ارث رسیدن گروه خونی AB است. وقتی یک آلل به‌طور کامل بر دیگری مسلط نیست وضعیت تسلط ناقص است.

[1] Allele
[2] Heterozygote

۱-۲-۳-ژنوتیپ[1] یا ژن‌نمود

ژن‌نمود یکی از سه عاملی است که رخ‌نمود فرد را مشخص می‌نماید (دو عامل دیگر اپی ژنتیک ارثی و عوامل محیطی غیر ارثی می‌باشند). بنابراین شبیه بودن ژن‌نمود دو موجود زنده تضمین‌کننده شباهت رخ‌نمودی آن دو موجود نیست. ژن‌نمودها بیانگر نحوه چینش تفاوت‌ها در افراد و این تفاوت می‌تواند باعث افتراق موجودات از یکدیگر شود.

هر انسان بیش از ۲۰هزار ژن دارد. معمولاً، هرکدام از این ژن‌ها چندین نوع آلل در میان افراد مختلف دارند. همین موجب متفاوت بودن ژن‌نمود آن‌ها، و درنهایت گوناگونی می‌شود، چراکه احتمال این‌که دو فرد برای همه ژن‌های خود یک آلل را به ارث برده باشند تقریباً صفر است! مگر دوقلوهای همسان که حاصل تقسیم میوز سلول تخم می‌باشند، و بنابراین اطلاعات ژنتیکی آن‌ها کاملاً مشابه است.

۱-۲-٤-فنوتیپ[2] یا رخ‌نمود

رُخ‌نمود خصوصیات قابل‌مشاهده یا صفت یک ارگانیسم است. مانند خصوصیات بیوشیمیایی یا فیزیولوژیکی. رخ‌نمود از بیان ژن‌های یک ارگانیسم و همچنین تأثیر عوامل زیست‌محیطی و اپی ژنتیکی و تعامل بین این سه نتیجه می‌شود. هیچ‌کدام از موجودات زنده با ژن‌نمود همانند به یک شیوه عمل نمی‌کنند به این دلیل که ظاهر و رفتار با شرایط زیست‌محیطی و توسعه تغییر می‌کند. به همین ترتیب، تمام موجودات زنده که به نظر یکسان می‌رسند لزوماً ژن‌نمودهای یکسانی ندارند. تمایز ژن‌نمود، رخ‌نمود توسط ویلهلم جانسون در سال ۱۹۱۱ پیشنهاد شد تا تفاوت بین وراثت موجود زنده و آنچه وراثت تولید می‌کند روشن شود که این تفاوت مشابه چیزی است که توسط اوت ویزمن پیشنهادشده بود، که بین قسمت قابل توارث نطفه (وراثت) و سلول‌های سوماتیک بدن تمایز قائل شده بود.

باوجودآن تعریف به‌ظاهر ساده، مفهوم رخ‌نمود پیچیدگی‌های پنهانی دارد. بسیاری از مولکول‌ها و ساختارهایی که با مواد ژنتیکی کد شده‌اند قابل‌مشاهده در ظاهر یک ارگانیسم نیستد، اما می‌توان آن‌ها را قابل‌مشاهده کرد (برای مثال توسط وسترن بلات) و درنتیجه بخشی از رخ‌نمود هستند. گروه‌های خونی انسان جز این دسته هستند؛ بنابراین، رخ‌نمود باید شامل ویژگی‌هایی باشد که بتواند توسط برخی از روش‌های فنی به‌صورت قابل‌مشاهده درآید. توسعه بعدی اضافه کردن رفتار به رخ‌نمود است، رفتارها نیز از ویژگی‌های قابل‌مشاهده هستند. درواقع روی ارتباط بالینی رخ‌نمودهای رفتاری پژوهش‌هایی انجام‌شده است که به طیف وسیعی از سندرم‌ها مربوط می‌شوند.

بعضی از رخ‌نمودهای در طول سن ثابت هستند مانند رنگ چشم. اما بعضی از رخ‌نمودها در طول عمر انسان در حال تغییر هستند. مثلاً فشارخون نرمال در سنین مختلف (بدون در نظر گرفتن تأثیر بیماری‌ها) تغییر می‌نماید. به رخ‌نمودهایی که در طول سن انسان تغییر می‌نمایند رخ‌نمود طولی[3] گفته می‌شود.

---

[1] Genotype
[2] Phenotype
[3] Longitudinal Phenotype

۱-۲-۵- تنوع رخ‌نمودی

تنوع رخ‌نمودی (ناشی از تنوع ژنتیکی ارثی) یک پیش‌نیاز اساسی برای تکامل توسط انتخاب طبیعی هست. این موجود زنده است که به‌عنوان یک کل در تولید نسل بعدی کمک می‌کند (یا نمی‌کند)، بنابراین انتخاب طبیعی به‌طور غیرمستقیم از طریق رخ‌نمودها بر ساختار ژنتیکی یک جمعیت تأثیر می‌گذارد. بدون تنوع رخ‌نمودی، هیچ تکاملی توسط انتخاب طبیعی وجود ندارد. تعامل بین ژن‌نمود و رخ‌نمود اغلب توسط رابطه زیر معرفی می‌شود:

به‌طور دقیق‌تر می‌توان گفت:

به‌رغم اینکه ژن‌نمود انعطاف‌پذیری بیشتری در بیان و اصلاح نسبت به رخ‌نمود دارد، اما بسیاری از موجودات زنده به دلیل قرار گرفتن در شرایط مختلف زیست‌محیطی دارای رخ‌نمودهای متنوعی شدند. گیاه آمبرلاتام[1] می‌تواند در دو محیط مختلف در سوئد رشد کند. در زیستگاه‌های صخره‌ای، صخره‌های سمت دریا، گیاه با برگ انبوه و گل‌های بزرگ رشد پیدا می‌کند، ولی در میان تپه‌های شنی رشد گیاه همراه با برگ‌های باریک و گل‌های فشرده هست. این زیستگاه‌ها هستند که رخ‌نمود گیاه را مشخص می‌کنند. مفهوم رخ‌نمود را می‌توان به تغییراتی که در سطح ژن بر سازگاری موجود زنده تأثیر می‌گذارد گسترش داد.

۱-۲-۶- تفاوت‌های ژنومی

ژنوم هر فرد ویژه خود اوست و یک عامل مهم برای تعیین رخ‌نمودهای بیولوژیکی است. ۹۹.۹ درصد از توالی DNA ژنوم انسان‌ها یکسان است. ۰.۱ درصد باقی‌مانده باعث می‌شود یک شخص به‌طور مثال در ویژگی‌ها، رفتارها و بیماری‌ها منحصربه‌فرد شود (Ziegler *et al.*, 2012).

در هنگام مقایسه ژنوم دو فرد، تفاوت‌های بسیاری (به لحاظ تعداد) دیده می‌شود که این تفاوت‌ها حاصل تغییرات ژنومی‌مانند چند مورفیسم تک نوکلئوتیدی[2] و تغییرات بزرگ‌تر مانند حذف[3]، جایگزینی[4] و تغییر در تعداد کپی‌های ژنی[5] هستند. هرکدام از این تغییرات می‌توانند منجر به بروز یک رخ‌نمود متفاوت در فرد شوند که این رخ‌نمود می‌تواند مربوط به صفات ظاهری فرد مانند قد و یا رنگ چشم و یا در ارتباط با خطر بروز یک بیماری باشد.SNP یک تغییر ژنتیکی کوچک در DNA است و رایج‌ترین نوع تنوع ژنتیکی هست که در میان افراد رخ می‌دهد.

---

[1] Umbellatum Hieracium
[2] Single nucleotide variation
[3] Deletion
[4] Insertion
[5] Copy number variation

هر SNP نشان‌دهنده تفاوت در یک بلوک از ساختمان DNA تک‌رشته‌ای به نام نوکلئوتید است. DNA با چهار نوع نوکلئوتید که دارای باز آلی T، C،A و G مشخص می‌شود. SNP زمانی رخ می‌دهد که باز یک نوکلئوتید به‌عنوان‌مثال نوکلئوتید دارای باز A جایگزین یکی دیگر از سه باز نوکلئوتید (مثلاً G،C ، یا T) می‌شود (Ganesalingam and Bowser, 2010).

SNPها به‌طورمعمول در سراسر DNA افراد و به‌طور متوسط یک‌بار در هر ۳۰۰ نوکلئوتید رخ می‌دهد. بدین معنی که تقریباً ۱۰ میلیون SNP در ژنوم هر فرد وجود دارد (Diamandis, 2012) و یک تغییر باید حداقل در ۱ درصد از جمعیت اتفاق بیافتد تا جزء SNPهای رایج شود. حدود ۴۰ درصد SNP موجود در ژن‌ها باعث تغییر یک آمینواسید خواهد شد. SNP رایج‌ترین نوع یلی مورفیسم است که باثبات تکاملی و فراوانی بالا در ژنوم توزیع‌شده‌اند.

البته این تغییرات می‌توانند: بی‌ضرر (تغییر در رخ‌نمود)، مضر (دیابت، سرطان، بیماری‌های قلبی، بیماری هانتینگتون و هموفیلی) و نهفته (تغییرات در ژنوم ازجمله مناطق تنظیم‌کننده ژن به‌خودی‌خود مضر نیست، و تنها تحت شرایط خاصی مانند استعداد ابتلا به سرطان ریه و تغییر پاسخ فرد به یک درمان آشکار می‌شود) باشد.

بدین ترتیب کشف SNPها جهش عظیمی درزمینهٔ ی ارتباط ژنتیک با اثربخشی دارویی و استعداد ابتلا به بیماری ایجاد کرده است. تعیین ژن‌نمود SNPها در درمان فردی اهمیت بسزایی دارد (Mahrokh *et al.*, 2017).

۱-۲-۷- چندریختی تک نوکلئوتید[1] (SNP)

یک نگاه کوتاه در زمان آغاز پلی مورفیسم ژنتیکی انسان متعلق به ژن B-globin در سال ۱۹۷۸ بود که از آن برای تشخیص یک بیماری ارثی استفاده کرد. پس از ۲ سال، در سال ۱۹۸۰، تمایزهای کوتاهی در DNA کشف‌شده در کل ژنوم انسان گسترش یافت. این با استفاده از پلی مورفیسم طول محدود[2] استفاده شد. اطلاعات جالب پیچیده دیگری درباره پلی مورفیسم DNA در سال ۱۹۸۵ گزارش شد. آن‌ها به‌عنوان ماهواره‌های کوچک[3] شناخته شدند. به‌طورکلی، پلی مورفیسم ژنتیکی می‌تواند در طرح‌های بی‌شماری موجود باشد، شامل: چندریختی تک نوکلئوتیدی، چندریختی تکرار پشت سرهم که شامل تعداد متغیر تکرارهای پشت سرهم[4] و تکرارهای پشت سرهم کوتاه[5]. برای مطالعه انواع مختلف چندریختی‌های DNA، می‌توان از فن‌های مختلفی

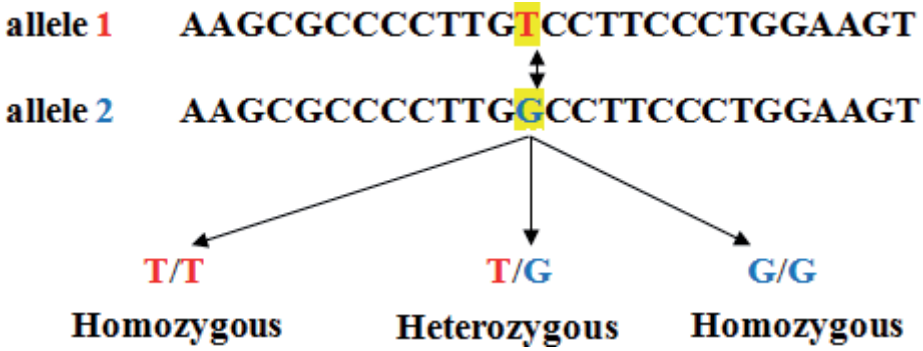
[1] Single Nucleotide Polymorphism
[2] Restriction fragment length polymorphisms (RFLPs)
[3] Mini satellites
[4] Variable number of tandem repeats (VNTRs)
[5] short tandem repeats (STR)

استفاده کرد، مانند پلی مورفیسم طول محدودکننده، واکنش‌های زنجیره‌ای پلیمراز[1]، روش‌های و ترتیب توالی ژنوم کل[2] (Çalışkan, Erol and Öz, 2020).

SNP یک تغییر در واحد بلوکی ساختار دستورالعمل DNA است. این ساده‌ترین فرمول تفاوت ژنتیکی در بین افراد است. SNP شایع‌ترین اتفاق در تفاوت ژنتیکی در بین DNA افراد است.

به‌طورمعمول ممکن است در هر ۳۰۰ نوکلئوتید یک‌بار اتفاق بیفتد، یعنی به‌طور متوسط حدود ۱۰ میلیون SNPدر ژنوم فرد وجود دارد. بیشترین تعداد، این SNPها بین ژن‌ها یا درون ژن‌ها تنظیم می‌شوند. آن‌ها ممکن است به‌عنوان نشانه‌های زنده و یا شاخص‌های ارثی عمل کنند، متخصصان توالی‌هایی را پیدا می‌کنند که با بیماری در ارتباط هستند. به‌محض اینکه SNPها در داخل یک ژن یا در یک ناحیه تنظیم‌کننده در نزدیکی یک ژن اتفاق می‌افتد، با تحریک کردن نقش ژن می‌توانند تأثیر بیشتری در بیماری نشان دهند. بااین‌حال، SNPها به‌طورکلی هیچ تأثیری در وضعیت عمومی سلامت ندارند. علاوه بر این، محققان اعلام کردند که SNP ها ممکن است به واکنش شخص در برابر داروهای قطعی کمک کند. هم‌چنین آن‌ها برای مسیری از عوامل ژنتیکی بیماری نادر در بستگان استفاده می‌شوند (Çalışkan, Erol and Öz, 2020).

شکل ۱-۲ نمونه‌ای از SNP برای دو آلل را نشان می‌دهد.



شکل ۲-۱ چندریختی تک نوکلئوتید برای دو آلل

۱-۲-۸-نقش عوامل ژنتیکی و غیر ژنتیکی در بروز صفات

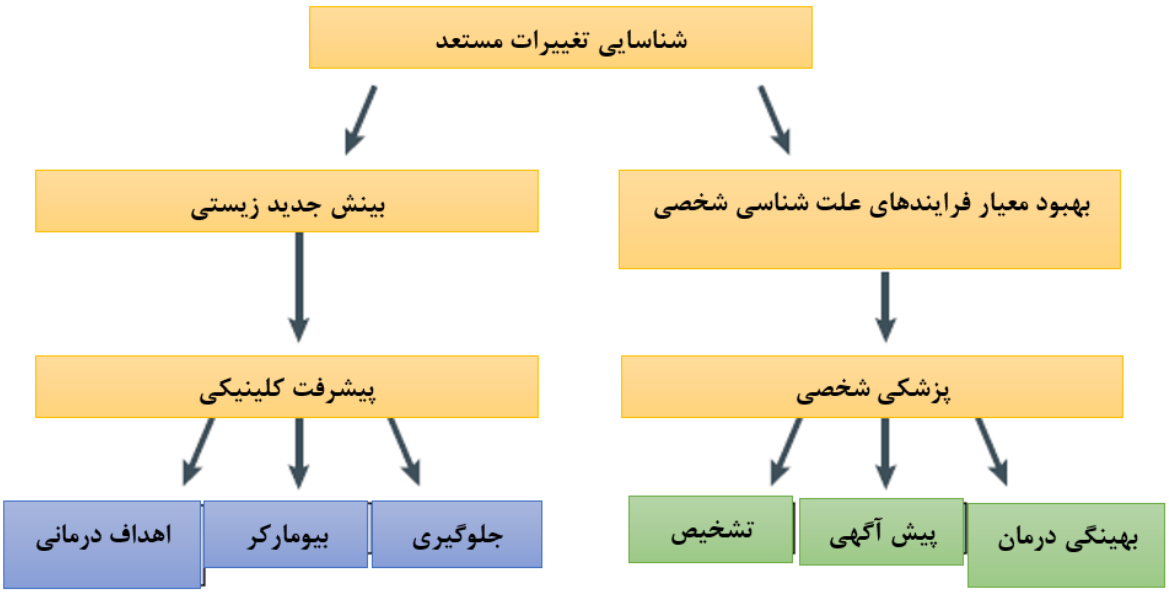
تا مدت‌ها فقط در مورد صفات و شرایطی که توسط ژن‌ها در یک جایگاه منفرد در ژنوم انسان ایجاد می‌گردد، بحث می‌شد. اما امروزه طبق تحقیقات (Filshtein *et al.*, 2019) و (Pattin and Moore, 2009) بسیاری از اختلالات شناخته‌شده با توارث مندلی مطابقت ندارند و هر صفت تنها توسط ژن‌هایی در یک لکوس تعیین نمی‌شود. درواقع، بسیاری از صفات توسط ترکیبی از ژن‌ها در بیش از یک محل در ژنوم انسان و همچنین عوامل غیر ژنتیکی تعیین می‌شوند. صفات چندعاملی صفاتی هستند که توسط عوامل متعدد ایجاد می‌شوند. در اغلب موارد، هر دو عوامل ژنتیکی و غیر ژنتیکی در ایجاد یک رخ‌نمود کمک می‌کند. به‌طور مثال قد انسان یک نمونه از یک صفت چندعاملی است. بااین‌حال، بسیاری از متخصصان ژنتیک انسانی در مورد قد انسان نگران نیستند، در عوض، در عرصه‌ی ژنتیک انسانی به‌طور عمده در مورد بیماری‌های انسان نگران هستند که

[1] Polymerase chain reactions (PCR)
[2] Whole genome sequencing (WGS)

در حال حاضر رایج‌ترین آن‌ها دیابت، بیماری قلبی، آلزایمر، آرتریت روماتوئید و سرطان هستند که تصور می‌شود در همه این بیماری‌ها چند جایگاه ژن و همچنین عوامل غیر ژنتیکی متعدد درگیر باشند و یک متخصص ژنتیک ممکن است آن‌ها را تحت رده‌بندی بیماری‌هایی که توسط عوامل متعدد ایجاد می‌شود به‌عنوان بیماری‌های چندعاملی مطرح نماید. یکی از ویژگی‌های مشترک بیماری‌های چندعاملی، طیف وسیعی از شدت و سن بیماری است. اضافه بر این پیچیدگی این واقعیت که برخی از این بیماری‌ها درواقع به وقوع یکدیگر کمک می‌کنند را نیز باید عنوان کرد، به‌عنوان‌مثال، دیابت خود یک عامل خطر برای بیماری قلبی است.

یک مشکل اساسی در ژنتیک کشف ارتباط بین ژن‌نمود و رخ‌نمود است. حصول اطمینان از وراثت یک صفت یک گام مهم به سمت امکان احتمال پیش‌بینی و ارزیابی خطر بیماری‌های انسان و درمان آن‌ها با استفاده از اطلاعات ژنتیکی هست. دانستن تمام عناصری که در وراثت دخالت دارند امکان وجود یک رابطه معلولی بین اطلاعاتی که نسل به نسل منتقل و منجر به رخ‌نمود موردنظر می‌شود را تسهیل می‌کند. مطالعات مرتبط با گستره ژنوم موفق به شناسایی بسیاری از چند مورفیسم‌های انسان وابسته به صفات قد، رنگ چشم، و یا استعداد ابتلا به بیماری‌های شایع را نشان داده‌اند اما این واریخت‌ها تنها بخش کوچکی از وراثت وابسته به یک صفت را شرح می‌دهند (Ligon *et al.*, 2000; Trerotola *et al.*, 2015).

کشف ارتباط بین عامل ژنتیکی و رخ‌نمود مزایایی مختلفی دارد که در تصویر ۱-۳ آمده است (McCarthy *et al.*, 2008).



شکل ۳-۱ فواید کشف روابط ژنوم-صفت (McCarthy *et al.*, 2008)

۱-۳- مطالعات گسترده ارتباط ژنومی[1]

در حدود سال ۲۰۰۰ میلادی، قبل از آنکه مطالعات گسترده ارتباط ژنومی معرفی شوند، جهت بررسی ارتباط بین ژن‌نمود و رخ‌نمود از بررسی توارث[2]در خانواده و یا پیوستگی ژنتیکی [3] استفاده می‌شد. این روش‌ها در بررسی بیماری‌هایی که یک ژن عامل بروز آن‌ها بود، بسیار موفق بودند. اما این یافته‌ها در بیماری‌های چندعاملی پیچیده تعمیم‌پذیر نبودند.

بررسی ارتباطی گسترده ژنومی اولین بار در سال ۱۹۹۶ توسط Merikangas جهت بررسی ژنتیکی بیماری‌های چندعاملی پیشنهاد شد (Risch and Merikangas, 1996) اما چون ایده آن اجرایی نبود، ۱۰ سال طول کشید تا اولین مطالعه از این نوع انجام شود. پروژه تعیین توالی ژنوم انسان[4] نیز در سال ۱۹۹۰ آغاز شد اما ۱۳ سال به طول انجامید تا نهایتاً در سال ۲۰۰۳، دقیقاً ۵۰ سال بعد از شناسایی اولیه DNA توسط واتسون و کریک (WATSON and CRICK, 1953) گزارش کامل توالی ژنوم انسان منتشر شد. این گزارش نشان داد که ژنوم انسان حاوی حدود سه بیلیون جفت باز هست که این توالی می‌تواند حاوی ۲۰ تا ۲۵ هزار ژن کد کننده‌ی پروتئین باشند.

پیشرفت‌های اخیر در پایگاه داده‌های اختصاصی جایگاه ژنی، در این زمینه بسیار اثرگذار بوده است. در مطالعات ارتباط ژنوم گسترده کشف نشانگرهای زیستی، شناسایی افراد دارای ژن بیماری‌های خاص و شناسایی نوع واکنش به درمان به‌طور وسیع‌تری صورت گرفته است (Ginsburg, Donahue and Newby, 2005).

نشانگرهای زیستی را می‌توان به نشانگرهای زیستی DNA، تومور و عمومی تقسیم کرد. البته این رده‌بندی دارای اشتراک نیز هست. از بین این سه گروه، نشانگرهای زیستی DNA مزایای زیادی نسبت به سایر نشانگرهای زیستی دارد.

به‌طوری‌که: در طول زندگی فرد ثابت است، نگهداری DNA آسان است، در هرزمانی قابل‌اندازه‌گیری هست، قابل تکثیر است، می‌تواند برای مطالعات گذشته و آینده‌نگر به کار رود و در بانک‌های زیستی هم معتبر است. در بین نشانگرهای زیستی، رایج‌ترین نوع تغییرات DNA پلی مورفیسم‌های تک نوکلئوتیدی هستند که عمدتاً دو آللی‌اند و سه نوع ژنوتیپ (هوموزیگوت مرجع، هوموزیگوت تغییریافته و هتروزیگوت) را ایجاد می‌کنند. البته شناسایی و استفاده از یک نشانگر زیستی بستگی به نوع مطالعه دارد (Ziegler *et al.*, 2012).

این مطالعات اغلب از نوع مورد شاهد[5]ی انجام می‌گردد؛ بدین معنا که گروهی از بیماران در مقابل افراد سالم انتخاب می‌شوند و تمامی این افراد برای قسمت عمده چند مورفیسم‌های تک نوکلئوتیدی شناخته‌شده (تعداد آن بستگی به فنّاوری مورداستفاده دارد اما به‌طورکلی یک‌میلیون و یا بیشتر است) تعیین ژن‌نمود می‌شوند، سپس میزان فراوانی اللی هرکدام از این نشانگرها در گروه‌های مورد و شاهد مقایسه می‌شود. اساس گزارش

[1] Genome Wide Association Study (GWAS)
[2] Inheritance
[3] Genetic linkage
[4] Human Genome Project (HGP)
[5] Case - Control

میزان اثر[1] بر مبنای محاسبه نسبت شانس[2] است بدین معنا که حضور یا عدم حضور یک آلل چقدر به حضور یا عدم حضور رخ‌نمود موردبررسی در یک جامعه‌ی مشخص و معلوم، مربوط است. هنگامی‌که فراوانی اللی در گروه مورد بیشتر از گروه شاهد است، نسبت شانس بیشتر از یک می‌شود. برای گزارش این نسبت آزمون کای‌دو [3] جهت محاسبه p-value استفاده می‌شود. این نوع مطالعه به روش‌های دیگری نیز قابل انجام است. یکی ازاین‌روش‌ها بررسی رخ‌نمودهای کمی مانند قد، غلظت، نشانگرهای زیستی و یا بیان ژن است. در مسائل طراحی مورد شاهدی موضوعات مهمی وجود دارند (McCarthy *et al.*, 2008):

۱. اندازه نمونه و در این موضوع دیدگاه روشن است: نمونه‌های بیشتر بهتر است که البته افزایش زیاد آن‌هم در مواردی راهگشا نیست.

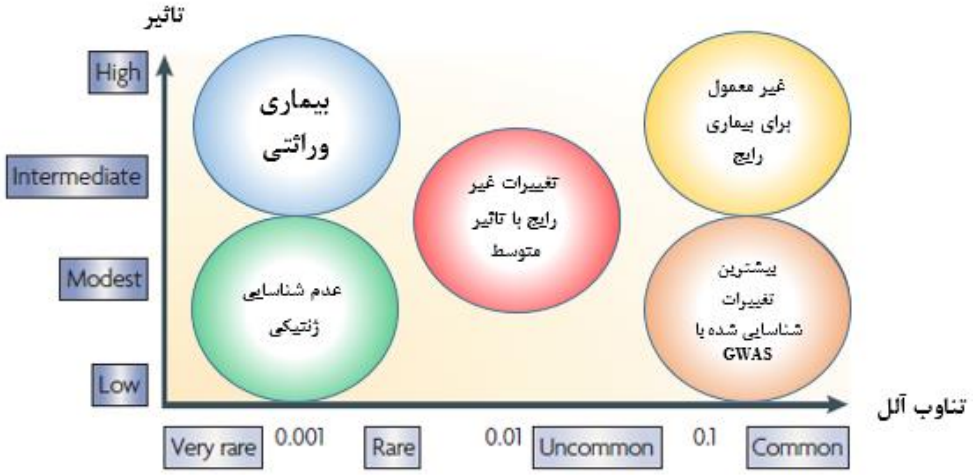
۲. گرایش[4] به زیر ساختار پنهان جمعی (رده‌بندی جمعیت و رابطه پنهان) که نرخ خطا نوع ۱ را افزایش می‌دهد و ادعاهای غلطی را درباره انواع مختلفی روابطی که در این زیر ساختار وجود دارد، بیان می‌کند. انتخاب جمعیت بهینه باید زیر ساختار جمعیت را محدود کند. در صورت امکان، چندین جمعیت کنترل باید مورداستفاده قرار گیرد و به‌صورت جداگانه انتخاب شوند تا بیشترین میزان شباهت به جمعیت بیماری را انتخاب کنند. به‌طورکلی، مطالعات آینده‌نگر، کنترل بیشتری را نسبت به مطالعات موردی، کنترل می‌کنند (Cardon and Bell, 2001).

۳. شایستگی نسبی روش‌های مرتبط با خانواده و ارتباطات مورد شاهدی. استراتژی‌هایی که از کنترل‌های خانوادگی استفاده می‌کنند و ارتباط را شناسایی می‌کنند می‌توانند ارزشمند باشند اما در استفاده از اطلاعات موجود ناکارآمد هستند. نتایج مثبت داده‌های معنی‌دار را فراهم می‌کنند، اما کمبود اهمیت آماری را ایجاد می‌کنند (Cardon and Bell, 2001).

۴. پتانسیل استفاده از ژن‌نمودهای کنترل تاریخ[5] برای جایگزینی یا تکمیل کنترل‌های جدید تایپ‌شده در مطالعات GWA آینده

روش‌های آماری مورداستفاده بسته به میزان نفوذ بیماری و الگوی غالب و یا مغلوبی آن متفاوت است که معمولاً با ابزارهای بیوانفورماتیک مانند PLINK و SNPTEST انجام می‌شود. بسیاری از برنامه‌های استاندارد محاسبات آماری (R و SPSS) برای محاسبات اولیه قابل بهره‌برداری است اما در مواجه با داده‌های GWAS با توجه به حجم بالای این نوع داده‌ها، بسیاری از برنامه‌های رایج غیر قابل‌استفاده خواهند بود. در این شرایط بسیاری از محققین از برنامه‌های متن‌باز و قابل‌تغییر استفاده می‌کنند که ازجمله می‌توان به برنامه‌های PLINK و GenABEL اشاره کرد (Aulchenko, De Koning and Haley, 2007).

[1] Effect size
[2] Odd ratio
[3] Chi squared
[4] propensity
[5] historical control genotypes



شکل ۴-۱ حوزه فعالیت تحقیقات GWAS (McCarthy *et al.*, 2008)

حوزه فعالیت تحقیقات GWAS در شکل ۱-۴ آمده است (McCarthy *et al.*, 2008). همان‌طور که در تصویر بیان‌شده است حوزه تحقیقات GWAS بر روی آلل‌هایی با احتمال رخداد بالا (آلل رایج) و تأثیر متوسط به پایین هست. می‌توان گفت GWAS معمولاً برای جفت‌پایه DNA با فراوانی آلل جزئی (MAF) بیشتر از ۵٪ برای جمعیت موردمطالعه تعریف می‌شود (Bush, 2019).

درنهایت می‌توان گفت صدها جهش در یک ژن منفرد منجر به بیماری‌های نادر می‌شود، اما اینکه چرا جهش‌ها حالت‌های شدید یا ضعیفی ایجاد می‌کنند، هنوز درک نشده است. ساخت مدل‌های بیماری ژنتیکی انسان، بخشی اساسی در درک بیماری، طراحی درمان‌ها و آزمایش روش‌های درمانی است. مطالعات ارتباط گسترده ژنوم در شناسایی صدها ژن زمینه‌ای و جهش‌های مرتبط قدرتمند بوده است؛ اما تجزیه‌وتحلیل عملکردی اکثریت‌قریب‌به‌اتفاق ژن‌ها و جهش‌های آن‌ها (حتی در ژن‌های کاملاً مشخص) وجود ندارد (Farris *et al.*, 2021).

۱-۳-۱- تحلیل GWAS و وراثت‌پذیری

وراثت‌پذیری به نسبت وراثتی بودن یک رخ‌نمود به یک ژن‌نمود خاص است که می‌توان از طریق تقسیم واریانس ژن‌نمود به واریانس رخ‌نمود آن را محاسبه کرد. این عددی که با نماد h^2 شناخته می‌شود عددی بین صفرتا یک است که هر چه عدد به یک نزدیک باشد قابلیت توارث پذیری بیشتری را برای رخ‌نمود موردنظر نشان می‌دهد.

اندازه‌گیری معیار h^2 یکی از مشکلاتی است که در روش‌های GWAS نیز بررسی‌شده‌اند. همان‌طور که اشاره شد، GWAS برای کشف ژن‌ها و مسیرهای مهم پزشکی به‌منظور شناسایی علل مولکولی و ژنتیکی بیماری‌زا طراحی‌شده است، اما به دلیل ناتوانی در توضیح وراثت‌پذیری بیشتر صفت‌های پیچیده موردانتقاد قرارگرفته است (Maher, 2008). تجزیه‌وتحلیل داده‌های GWAS نشان داد که نسبت زیادی از وراثت‌پذیری برای صفات کمی مانند قد، شاخص توده بدنی[1] و توانایی شناختی توسط SNPهای رایج توضیح داده می‌شوند (Yang *et*

[1] Body Mass Index

(*al.*, 2010, 2011; Davies *et al.*, 2011). این نتایج نشان می‌دهد که بیشتر وراثت‌پذیری به‌جای گم‌شدن، پنهان‌شده است (Gibson, 2010) و GWAS، SNPهایی را شناسایی نکرده که این نسبت وراثت پنهان را توضیح می‌دهند، زیرا اندازه اثرات SNPهای فردی برای رسیدن به سطح معنی‌داری دقیق ژنوم بسیار کم است (Yang *et al.*, 2010, 2017). به‌جز GWAS، روش دیگری برای پراکندگی واریانس ژنتیکی بر روی کروموزوم‌ها و بخش‌های ژنومی وجود دارد. در این روش، واریانس مربوط به یک کروموزوم یا بخش DNA متناسب با طول آن، مخصوصاً برای قد و اسکیزوفرنی محاسبه‌شده است (Yang *et al.*, 2013) و SNPهایی که در مناطق ژنی یافت می‌شوند، تغییرات بیشتری را نسبت به مناطقی که در ناحیه‌های بین ژنی مشاهده می‌شوند، نشان می‌دهند. تمام نتایج مطابق با الگوی ارثی چندژنی برای بیشتر ویژگی‌های پیچیده است.

علاوه بر این، طبق فرض بسیاری از ژن‌ها که دارای اثر کوچک هستند، انتظار می‌رود که نشانگر ژنتیکی بیشتری با افزایش اندازه نمونه مشخص شوند (Visscher *et al.*, 2012). طیف وسیعی از همکاری‌های بین‌المللی وجود دارد که نمونه‌های بسیاری از گروه‌های GWAS را در یک فرا تحلیل ترکیب می‌کند تا قدرت آماری کارآمد را برای تشخیص اثرات کوچک فراهم کند. بااین‌حال، برای یک فرا تحلیل، ژن‌نمود و اطلاعات رخ‌نمود سطح فردی به‌طورمعمول در دسترس نیستند. برای یک فرا تحلیل بزرگ در مقیاس بزرگ، سازمان‌دهی و انجام یک تجزیه‌وتحلیل شرطی و غیرعملی برای انجام تجزیه‌وتحلیل شرطی به‌طور تکراری[1] بسیار دشوار است. لازم به ذکر است در روش‌های آماری GWAS به‌جای محاسبه h^2 مقدار h^2_{SNP} را اندازه‌گیری می‌نمایند (Yang *et al.*, 2013).

۱-٤- هم‌ترازی توالی ژنوم

در چند دهه اخیر، پیشرفت در زیست‌شناسی مولکولی و تجهیزات موردنیاز آن باعث افزایش سریع تعیین توالی ژنوم بسیاری از گونه‌های موجودات شده است تا جایی که پروژه‌های تعیین توالی ژنوم‌ها از پروژه‌های بسیار رایج به‌حساب می‌آیند و هم‌ترازی توالی[2] (SA) معمولاً اولین گامی برای درک فیلوژنی مولکولی یک توالی ناشناخته در بیوانفورماتیک است (Chowdhury and Garai, 2017).

این کار باهم‌ترازی توالی ناشناخته با یک یا چند توالی شناخته‌شده در پایگاه داده برای پیش‌بینی بخش‌های رشته توالی برای انجام نقش‌های عملکردی و ساختاری بر مبنای تکامل انجام می‌شود. هم ترازی مطلوب، دو یا چند توالی را به‌گونه‌ای تنظیم می‌کند که حداکثر تعداد عناصر یکسان یا مشابه را باهم مقایسه کند. توالی‌ها می‌توانند توالی‌های نوکلئوتیدی (DNA یاRNA) یا توالی‌های اسیدآمینه (پروتئین) باشند (Reddy and Fields, 2022).

فرایند باز چینی ممکن است یک یا چند فضا یا شکاف در هم ترازی ایجاد کند. یک‌فاصله نشان‌دهنده از دست دادن یا افزایش احتمالی یک عنصر است بنابراین، درجه تکامل یا حذف[3]، انتقال و معکوس شدن در SA

[1] Iteratively
[2] Sequence Alignment
[3] indel

مشاهده می‌شود. دو نوع هم‌ترازی توالی ترتیب عبارت‌اند از: توازن جفتی[1] (PSA) و هم ترازی توالی چندگانه[2] (MSA). در PSA دو توالی را در نظر می‌گیرد درحالی‌که MSA چندین (بیش از دو) توالی مرتبط را تراز می‌کند. فواید MSA بیشتر از PSA است، زیرا چندین عضو یک خانواده توالی را در نظر می‌گیرد و درنتیجه اطلاعات بیولوژیکی بیشتری ارائه می‌دهد. همچنین MSA پیش‌نیاز تجزیه‌وتحلیل ژنوم مقایسه‌ای برای شناسایی و اندازه‌گیری مناطق محافظت‌شده و یا نقاط عملکردی در یک خانواده کامل توالی و برآورد کننده انحراف تکاملی بین توالی‌ها و حتی برای پروفایل توالی اجدادی هست. ترتیب توالی در سطح اسیدآمینه بیشتر مربوط به سطح نوکلئوتید است زیرا پروتئین کلید عملکردی مولکول‌های بیولوژیکی است و ازاین‌رو اطلاعات ساختاری و یا عملکردی را حمل می‌کند. به‌این‌ترتیب هم‌ترازی، دارای پیوند قوی با زیست‌شناسی ساختاری است. ازاین‌رو، SA، به‌ویژه MSA، نقطه شروع هر زمینه تحقیق بیولوژیکی است. که MSA به‌عنوان یک پنجره برای مشاهده دیدگاه تکاملی، کاربردی و ساختاری ماکرو مولکول‌های بیولوژیک در قالب مختصر هست (Reddy and Fields, 2022).

روش‌های مختلفی برای امتیازدهی در هم‌ترازی توالی به‌منظور شناخت یکسانی یا تشابه استفاده می‌شود. نمره نوکلئوتید یک طرح شناسایی ساده است که در آن موقعیت‌های یکسان در هر دو توالی، امتیازات مثبتی به دستمی‌دهند. در مقابل، برای پروتئین، یک نمره مشابهت نیز شمارش می‌شود (همراه با نمره یکسانی) که نشان‌دهنده اسیدآمینه با خواص فیزیکی و شیمیایی مشابه است. روش SA می‌تواند دو نوع باشد: هم‌ترازی سرتاسری و هم‌ترازی محلی. هم‌ترازی سرتاسری زمانی انجام می‌شود که شباهت در تمام طول دنباله شمارش شود. چندین فن MSA هم‌ترازی سرتاسری را انجام می‌دهند اما مشکلات زمانی به وجود می‌آیند که توالی‌ها تنها در مناطق محلی همخوانی داشته باشند، جایی که بلوک واضحی از ناپیوستگی‌های مشترک در همه‌ی توالی‌ها وجود دارد یا دامنه‌های تغییریافته در میان توالی‌های مرتبط وجود داشته باشد. در چنین مواردی، هم‌ترازی محلی برای شناخت مناطق مشابه در میان توالی‌ها انجام می‌شود. وقتی تفاوت زیادی در طول توالی‌هایی که باید مقایسه شود وجود دارد، هماهنگی محلی معمولاً انجام می‌شود.

برای PSA، برنامه‌ریزی پویا[3] (DP) همواره جهت‌گیری مطلوب برای یک تابع هدف مشخص را با یافتن راه بهینه تراز کردن با استفاده از فرایند ردیابی فراهم می‌کند. تابع هدف برای ارزیابی کیفیت هم ترازی مجموعه‌ای از توالی‌های ورودی استفاده می‌شود.

مشکلاتی محاسباتی در این روش‌ها بالاست و نیاز به منابع رایانه‌ای بالا دارد. علاوه بر این، DP از مشکلات بزرگ در MSA رنج می‌برد، زیرا تعداد توالی‌ها برابر با تعداد ابعاد است. اگر دو یا چند مسیر بهینه در دسترس باشند و نیاز به فرایند ردیابی باشد، پیچیدگی افزایش می‌یابد و محاسبه یک MSA دقیق NP-Complete است و بنابراین روش دقیق در MSA برای مجموعه‌های داده‌های کوچک غیرواقعی استفاده‌شده است.

[1] Pairwise Sequence Alignment
[2] Multiple Sequence Alignment
[3] Dynamic Programming

بنابراین روش‌های معمول MSA به‌صورت اکتشافی یا تقریبی است که راه‌حل ممکنی را در یک دوره زمانی کوتاه و محدود فراهم می‌کند. روش‌های اکتشافی موجود، بهترین راه‌حل را ارائه نمی‌دهند و به دلیل سرعت‌بالای رشد اندازه پایگاه‌های داده، توسعه فن جدید برای MSA هنوز تحت تحقیق است.

۱-۴-۱-روش‌های هم‌ترازی توالی چندگانه

به دلیل ارتباط پیچیده میان توالی‌های مرتبط و گاهی اوقات به دلیل فقدان تاریخ تکاملی، اغلب MSAها به‌سختی می‌توانند به‌توالی نهایی دست یابند. اغلب سه دسته از رویکردها در MSA استفاده می‌شود: رویکرد دقیق، پیشروندِ و تکرارپذیر. الگوریتم‌های دقیق معمولاً کیفیت بالایی را که بسیار نزدیک به مطلوب نهایی است ارائه می‌دهند. این روش‌ها تلاش می‌کنند به‌طور هم‌زمان چندین توالی را ترتیب دهند و بنابراین نیاز به DP دارند. از روش‌های دیگر می‌توان روش‌های درختی، ستاره‌ای و پیش‌رونده را نام برد که البته زمانی که طول توالی‌ها و نیز تعداد آن‌ها افزایش پیدا می‌کند، این روش‌ها چندان موفق نیستند.

به جهت اهمیت مسئله مدل‌سازی MSA در مقاله (Lajevardy and Kargari, 2022) مدلی مبتنی بر مدل‌سازی ریاضی مسئله MSA به همراه الگوریتم ژنتیک آن منتشرشده است[1].

۱-۵- فشارخون[2]

به فشار مایع درون رگ‌های خون، فشارخون می‌گویند. هم‌چنین به فشارخون شریانی توزیع‌شده در کل محفظه شریانی، فشارخون سیستمیک می‌گویند که توسط عوامل مختلفی تعیین می‌شود: ۱) نیروی خونی که با انقباض بطن چپ قلب وارد محفظه شریانی می‌شود. ۲) میزان گردش خون از محفظه شریانی به مویرگ‌ها. ۳) حجم کل خون. ۴) تنش ایجادشده توسط دیواره‌های بزرگ‌ترین رگ‌های خونی در مقاومت به پالس خون خارج‌شده از عروق توسط قلب.

۱-۵-۱-فشارخون سیستولیک[3]

به حداکثر فشارخون شریانی که در هنگام انقباض بطن چپ و خارج شدن خون به محفظه شریانی اندازه‌گیری می‌شود، فشارخون سیستولیک می‌گویند. این فشار در میلی‌متر جیوه سنجیده می‌شود.

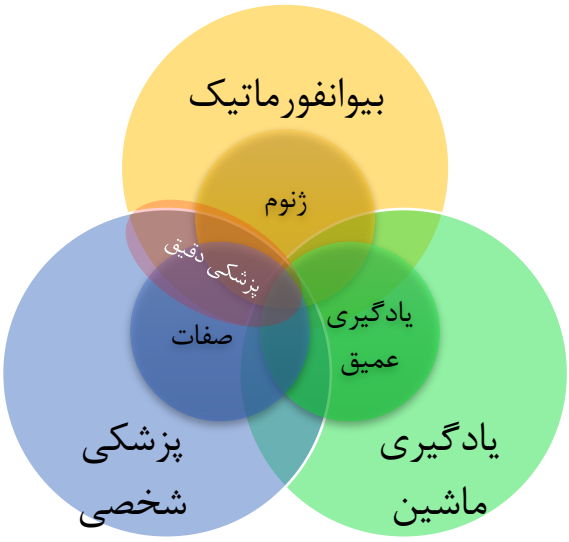
۱-۵-۲- فشارخون دیاستولیک[4]

به حداقل فشارخون شریانی که در مرحله استراحت قلب در شریان‌های بزرگ اندازه‌گیری می‌شود، فشارخون دیاستولیک می‌گویند. این فشار در میلی‌متر جیوه اندازه‌گیری می‌شود و تا حد زیادی توسط میزان جریان خون از محفظه عروقی شریانی به مویرگ‌ها تعیین می‌شود.

[1] https://github.com/l8026070/MSA
[2] Blood Pressure
[3] Systolic Blood Pressure (SBP)
[4] Diastolic Blood Pressure (DBP)

۱-۶- حوزه پژوهش

این تحقیق بر آن است که بر اساس ژنوم انسانی که از فن‌های مختلف بیوانفورماتیک تهیه می‌گردد و داده‌های طولی صفت پرفشاری خون، مدلی مبتنی بر فن‌های یادگیری ماشین خصوصاً یادگیری عمیق طراحی و ارائه نماید. این مدل می‌تواند بر اساس اطلاعات جمعیت شناختی بهینه گردد. نمودار حوزه پژوهش در شکل ۱-۵ آمده است.



شکل ۵-۱ نمودار حوزه پژوهش

۱-۷- ضرورت و اهداف پژوهش

بر اساس تحقیقات اشاره‌شده در بخش‌های قبل، مطالعه گسترده ارتباط ژنومی اهمیت بالایی در پزشکی شخصی دارد که آینده پزشکی نوین و نشان از اهمیت بسزای آن دارد.

بر همین اساس در این پژوهش اهداف کلی زیر پیگیری می‌شود:

- ✓ ارائه مدل مبتنی بر یادگیری ماشین در داده‌های حجیم جهت پیش‌بینی رخ‌نمودهای پرفشاری خون بر اساس SNP در رخ‌نمودهای:
  - فشارخون
    - فشار سیستولی SBP و فشار دیاستولی DBP
    - خطر پُرفشاری خون
- ✓ بهبود مدل مبتنی بر یادگیری ماشین جهت پیش‌بینی داده‌های طولی
- ✓ بهبود مدل بر اساس اطلاعات جمعیت شناختی
- ✓ انتخاب SNP بااهمیت بر اساس دقت و تعداد

با توجه به داده‌های ژنومی که دارای حجم بالایی هست، هدف نهایی این پژوهش، پیش‌بینی خطر پرفشاری خون بر اساس SNP و با روش‌های یادگیری ماشین در داده‌های حجیم مانند داده‌های توالی ChIP (مجموعه‌ای مشخص از SNPهای توالی یابی شده برای افراد نمونه) و یا توالی یابی کل ژنوم است.

۱-۷-۱-سؤالات تحقیق

بر اساس اهداف تعیین‌شده سؤال اصلی تحقیق به‌صورت زیر هست:

- ✓ آیا می‌توان از روش‌های یادگیری ماشین در پیش‌بینی خطر پرفشاری خون مبتنی بر SNP به شکل مؤثرتر از روش‌های موجود بهره برد؟

سؤال فرعی تحقیق نیز به‌صورت زیر هست:

- ✓ آیا می‌توان اطلاعات جمعیت شناختی و داده‌های طولی رخ‌نمود را در بهبود مدل پیش‌بینی کننده استفاده کرد؟
- ✓ با توجه به مدل ارائه‌شده، آیا می‌توان SNPهای مهم ژنومی را استخراج کرد؟

۱-۸- نوآوری تحقیق

همان‌طور که در ادبیات موضوع اشاره گردید، تحقیقات موجود در حوزه GWAS در حوزه آماری متمرکز است که این امر به‌واسطه رده‌بندی جمعیت، وابستگی ژنتیکی، تشخیص[1] و سایر منابع ناهمگونی منجر به سیگنال‌های جعلی و کاهش توان در مطالعات ارتباط ژنتیکی می‌شود (Runcie and Crawford, 2018).

همچنین مطالعاتی که تعداد اندکی SNP مرتبط با بیماری را در نظر می‌گیرند نسبت به مطالعات سنتی، رده‌بندی خطر ناکافی و غیر معینی را به دستمی‌آورند (Khan *et al.*, 2020). به همین دلیل در این تحقیق تعداد زیادی SNP با روش‌های یادگیری ماشین در نظر گرفته می‌شوند.

رویکرد پردازش داده و یادگیری ماشین خصوصاً یادگیری عمیق (با توجه به داده‌های حجیم ژنومیک) باعث بهبود خطاهای ذکرشده است (Esteva *et al.*, 2019). همچنین می‌توان با بهره‌گیری از لایه‌های مختلف در شبکه یادگیرنده داده‌های مختلف omic را به مدل افزود که باعث افزایش دقت مدل می‌گردد (Hebbring, 2019; Xu *et al.*, 2019). دستاورد این تحقیق با توجه به راهکار یادگیری ماشین، امکان برطرف کردن مشکلات ذکرشده برای روش‌های آماری و افزایش دقت را خواهد داشت (Sun *et al.*, 2019; Yin *et al.*, 2019). لازم به ذکر است انتخاب فشارخون به دلیل اهمیت آن و همچنین چالش موجود در ادبیات موضوع در خصوص پیش‌بینی خطر بیماری در جوانان هست (Wu *et al.*, 2020; Phulka *et al.*, 2023).

۱-۹- جمع‌بندی

در این فصل به بیان کلیات و مقدمه موضوع پرداخته شد و درنهایت بابیان اهداف و ضرورت تحقیق به پایان رسید. در فصل بعد با بررسی ادبیات موضوع روش‌های مختلف حل مسئله گفته‌شده بیان می‌گردد.

[1] ascertainment

۲- ادبیات نظری و پیشینه پژوهش

بسیاری از رویکردهای احتمالی برای ارزیابی ژن‌هایی که ریشه بیماری‌های شایع و صفات کمی هستند، به‌طور گسترده‌ای به دودسته تقسیم می‌شوند (Hirschhorn and Daly, 2005): مطالعات ژن‌های منتخب، که از روش‌های همبستگی یا باز توالی[1] استفاده می‌کنند و مطالعات گسترده‌ی ژنومی شامل هر دو نوع نقشه سازی پیوند[2] و مطالعات روابط گسترده ژنومی[3] هست. رویکردهای اصلی و مزایا و معایب آن‌ها در جدول ۲-۱ خلاصه‌شده است (Hirschhorn and Daly, 2005).

جدول ۲-۱ رویکردهای تشخیص تغییرات مبتنی بر صفات پیچیده[4] در بیماری‌های رایج (Hirschhorn and Daly, 2005)

مزایا	Association*	Resequencing*	Linkage‡	Admixture‡	Missense SNPs‡	Association‡	Resequencing‡
هیچ اطلاعات پیشین در مورد عملکرد ژنی مورد نیاز نیست	−	−	+	+	+	+	+
محلی سازی به منطقه ژنومی کوچک	+	+	−	−	+	+	+
ارزان	+	−	+	+	+/−	−	Prohibitive
خانواده ها مورد نیاز نیست	+	+	−	+	+	+	+
هیچ فرضیه ای در مورد نوع درگیر وجود ندارد	+	−	+	+	−	+	+
حساس به اثرات رده بندی نیست	−/+	−/+	+	+	−/+	−/+	−/+
نیازی به تغییر فرکانس آلل در میان جمعیت نیست	+	+	+	−	+	+	+
قدرت کافی برای تشخیص آلل های معمولی (5%<MAFs)	+	−	−/+	+	+	+	+
توانایی تشخیص آلل های نادر (1%>MAFs)	−	+	+	−	−	−	+
رکورد منطقی برای بیماری های شایع	+	−/+	+/−	N/A	N/A	N/A	N/A
ابزار برای تجزیه و تحلیل در دسترس است	+	+	+	+	+	+/−	−

باوجود روش‌های مختلف، اما بااین‌حال، آلل‌های نادر معمولاً بسیار سخت توسط روش‌های کشف روابط به دلایل مختلفی، شناسایی می‌شوند (Visscher *et al.*, 2014) که در ادامه به استراتژی‌های افزایش کارایی این روش‌ها پرداخته می‌شود.

۲-۱- استراتژی‌های افزایش بهره‌وری

در این بخش، در مورد استراتژی‌هایی که برای اجرای یک مطالعه گسترده ارتباط ژنومی پیشنهادشده‌اند بحث می‌شود (Hirschhorn and Daly, 2005) که با به حداقل رساندن تعداد ژن‌نمود موردنیاز، قدرت کافی درزمینه‌ی آزمایش آزمون‌های چندگانه رادارند. این استراتژی‌ها شامل طرح‌های چندمرحله‌ای برای به حداقل رساندن اندازه نمونه، استفاده از جمعیت خاص و جمع‌آوری نمونه‌ها هست. یکی دیگر از روش‌های بهبود، استفاده از اپی ژنتیک هست (Shah *et al.*, 2015) که در حوزه این تحقیق نیست.

[1] Resequencing
[2] Linkage Mapping
[3] GWAS
[4] Complex Trait

۲-۱-۱- انتخاب یک اندازه نمونه مناسب

واضح‌ترین محدودیت در مطالعات گسترده ژنومی هزینه‌های بالا و تلاش‌های قابل‌توجهی است که نیاز به ژن‌نمود صدها هزار SNPs در هر فرد است. به دلیل این هزینه بالا، فشار برای محدود کردن حجم نمونه وجود دارد که همراه با کاهش در قدرت نتایج است. برای درک پیامدهای این آستانه، یک آلل با بسامد ۱۵٪ و یک نسبت شانس از ۱.۲۵ را در نظر بگیرید. برای چنین نوعی، حتی با فرض اینکه SNP علّی (یا SNP دیگری که به‌عنوان یک پروکسی کامل عمل می‌کند) تایپ‌شده است، تقریباً ۶۰۰۰ مورد و ۶۰۰۰ کنترل برای ۸۰ درصد قدرت آماری تشخیص ارتباط با مقدار ارزش P برابر 5×10^{-8} برای ۵۰۰،۰۰۰ SNP مستقل موردنیاز است که این اندازه نمونه ۶ میلیارد ژن‌نمود نیاز دارد که هزینه‌های بالایی دارد.

۲-۱-۲- رویکرد چندمرحله‌ای

یک روش ساده برای بهبود، استفاده از یک فرآیند غربالگری دو یا سه مرحله‌ای است که در آن، آستانه معکوس بیشتری برای مثبت شدن نتیجه، در هنگام ارزیابی اسکن اولیه ژنوم ارتباط استفاده می‌شود .

۲-۱-۳- جمعیت اولیه[1] و نمونه‌های جمع‌آوری‌شده

روش‌های دیگر پیشنهادشده که باعث افزایش کارایی مطالعات ارتباطی می‌شود استفاده از جمعیت اولیه هست که باعث کاهش تعداد نشانگرهایی که نیاز به ژن‌نمود دارند، و نمونه‌های جمع‌آوری‌شده که تعداد نمونه‌های ژن‌نمود را کاهش می‌دهند، می‌شود. جمعیت اولیه به‌صورت جمعیت‌هایی که از یک مجموعه محدود از افراد در ۱۰۰ نسل گذشته یا کمتر ایجادشده‌اند، تعریف می‌شود. جمعیت اولیه مزایای خوبی برای ژن‌هایی که منجر به اختلالات مندلی می‌شوند دارند، اما ممکن است برای بیماری‌های رایج کمتر مؤثر باشند.

۲-۱-٤- اجتناب از گرفتن نتایج مثبت کاذب

در یک مطالعه که صدها هزار نشانگر را شامل می‌شود، به حداقل رساندن مثبت کاذب ضروری است. منابع ایجاد ارتباطات مثبت کاذب را می‌توان به سه دسته اصلی تقسیم کرد: نوسانات آماری که به‌وسیله شانس به وجود می‌آیند و درنتیجه باعث کم شدن p-value می‌شوند، انحرافات سامانمند پایه با توجه به طراحی مطالعه؛ و تجهیزات فنی.

روش‌های موجود در مطالعات گسترده ژنومی را می‌توان در سه رویکرد: آماری، فرا تحلیل و یادگیری ماشین رده‌بندی کرد که در ادامه به بررسی آن‌ها پرداخته می‌شود.

۲-۲- رویکرد آماری

رویکرد آماری، رویکرد غالب در تحلیل ارتباطات گسترده ژنومی هست. این تحلیل برآمده از رابطه زیر هست:

رخ‌نمود ⟶ تنوع تصادفی+ ژن‌نمود + محیط

در ادامه به بررسی معروف‌ترین مدل‌های ارائه‌شده در این حوزه پرداخته می‌شود.

[1] Founder populations

۲-۲-۱- مدل GRAMMAR[1]

این مدل در سال ۲۰۰۷ ارائه گردید (Aulchenko, De Koning and Haley, 2007). در مرحله اول، داده‌ها تحت مدل ترکیبی[2] رابطه ۲-۱ مورد تجزیه‌وتحلیل قرار می‌گیرند.

$$y_i = \mu + \sum_j \beta_j c_{ji} + G_i + e_i$$

رابطه ۱-۲ مدل GRAMMAR

که y_i رخ‌نمود فرد iام است، μ میانگین رخ‌نمود، c_{ij} مقدار هم وردایی jام یا اثر ثابت برای فرد iام است (مثلاً جنسیت یا سن)، β_j برآورد از اثر ثابت یا هم وردایی است، G_i اثرات تصادفی چندژنی و e_i خطای باقی‌مانده است.

این روش مبتنی بر اخذ مجدد از یک مدل چندژنی است و بر اساس تجزیه‌وتحلیل ارتباط این باقیمانده‌ها با چند مورفیسم های ژنتیکی با استفاده از روش‌های کلاسیک برای افراد مستقل است.

۲-۲-۲- مدل GCTA[3]

در تحلیل‌های آماری با رویکرد جزئی خطاهای متعددی ایجاد می‌گردد. یکی از این خطاها وراثت گم‌شده[4] هست که به دلایل مختلفی من‌جمله تعداد زیاد از انواع تغییرات رایج با اثر کم و یا انواع نادر تغییرات با اثر بالا و یا تغییرات ساختاری DNA است. درروش GCTA (Yang *et al.*, 2013) برخلاف تجزیه‌وتحلیل روابط تک SNP، به اثرات همه SNPsها به‌عنوان اثرات تصادفی با یک مدل خطی آمیخته (MLM[5]: بر اساس فیشر ۱۹۱۸) نگاه می‌شود و رابطه ۲-۲ آن را بیان می‌کند.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Wu} + \varepsilon \; with \; var(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{WW'}\sigma_u^2 + \boldsymbol{I}\sigma_\varepsilon^2$$

رابطه ۲-۲ مدل GCTA

در این رابطه y یک بردار از رخ‌نمودها با n نمونه است، β یک بردار از اثرات ثابت مانند جنسیت، سن و (یا) یک یا چند بردار اطلاعاتی خاص از تجزیه‌وتحلیل مؤلفه‌های اصلی (PCA[6]) ، u یک بردار از اثرات SNP، I یک ماتریس همانی است و e یک بردار از اثرات باقی‌مانده خطا است. W یک ماتریس ژن‌نمود استانداردشده و برآورد شده با حداکثر احتمال (Gilmour, Thompson and Cullis, 1995) است.

۲-۲-۳- مدل ACTA[7]

صفات پیچیده با تعداد زیادی از عوامل ژنتیکی و محیطی و همچنین تعاملات آن‌ها تعیین می‌شود. شناسایی ژن‌های مشارکتی و اندازه‌گیری اثرات آن‌ها درزمینه‌ی یک یا چند محیط، اهمیت اساسی در اهداف درمانی برای بیماری حیوانات و انسان و درک نحوه انتخاب طبیعی و مصنوعی ژنوم حیوانات و انسان را می‌دهد.

---

[1] Genome wide Rapid Association Using Mixed Model and Regression
[2] Mixed Model
[3] Genome-wide Complex Trait Analysis
[4] Missing Heritability
[5] Mixed Linear Model
[6] Principal Component Analysis
[7] Advanced Complex Trait Analysis

نرم‌افزار GCTA (Yang *et al.*, 2013) یک بسته نرم‌افزاری رایگان C++ برای اندازه‌گیری ارتباط تنوع ژنتیکی به تنوع رخ‌نمودی برای ویژگی‌های پیچیده است که شامل دو مرحله است: ارزیابی «ماتریس رابطه ژنتیکی[1]» (GRM)، که حاوی اندازه‌گیری همبستگی ژنتیکی بین افراد است و با استفاده از داده‌های تک ژن‌نمود چند مورفیسم (SNP) محاسبه و استفاده از حداکثر درستنمایی محدودشده[2] (REML) برای به حداکثر رساندن احتمال رخ‌نمودها به GRM داده می‌شود. پیاده‌سازی الگوریتم GCTA، زمانی که در پردازنده‌های چندهسته‌ای مدرن اجرا می‌شود، کارآمد نیست.

در مدل ACTA (Gray, Stewart and Tenesa, 2012) بهبود عملکرد چشمگیری ارائه می‌شود. در این مدل ابتدا مدل خطی آمیخته[3] (MLM) که توسط GCTA استفاده می‌شود، خلاصه می‌شود و از رابطه ۲-۳ استفاده می‌شود.

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{W}\boldsymbol{u} + \epsilon \; with \; \boldsymbol{V} = \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}\sigma_u^2 + \boldsymbol{I}\sigma_\varepsilon^2$$

رابطه ۳-۲ مدل ACTA

در این رابطه، i و j همیشه نشانگرهایی برای مجموع افراد P هست، و k نشان‌دهنده یک SNP درمجموع SNPها است. GCTA شامل تأثیرات مناسب SNP با استفاده از MLM است و y_i رخ‌نمود فرد i است و u یک بردار از اثرات SNP با توزیع نرمال است. U_k اثر رخ‌نمودی است که به k امین SNP مربوط می‌شود. W ماتریس رخداد است که اثرات SNP را به Y می‌دهد. این ماتریس شامل ژن‌نمود برای هر فرد است و توسط رابطه زیر محاسبه می‌شود:

$$W_{ik} = \frac{(r_{ik} - 2p_{k)}}{\sqrt{2p_k(1-p_{k)}}}$$

رابطه ۴-۲

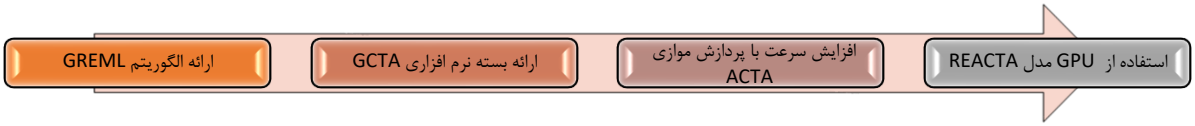
که R_{ik} تعداد کپی‌های آلل مرجع برای k امین SNP از فرد iام، و P_k فراوانی آلل مرجع است. به‌طور مشابه، X یک ماتریس بروز برای بردار اثرات ثابت است. β یک بردار از اثرات باقی‌مانده با توزیع نرمال است.

۲-۲-٤- مدل REACTA[4]

این مدل بر پایه مدل ACTA بناشده است و بر اساس استفاده از GPU سرعت پردازش اطلاعات را به‌طور متوسط از دو هفته در یک رخ‌نمود خاص، به ۱۰ دقیقه کاهش داده است (Cebamanos *et al.*, 2014). لازم به ذکر است حجم اطلاعات برای انجام یک GWAS در حدود ۴۰ گیگابایت هست که منجر به نیاز به حافظه بالا برای اجرای الگوریتم‌های آماری هست.

روند تحقیقات آماری در حوزه GWAS به‌صورت شکل ۲-۱ هست.

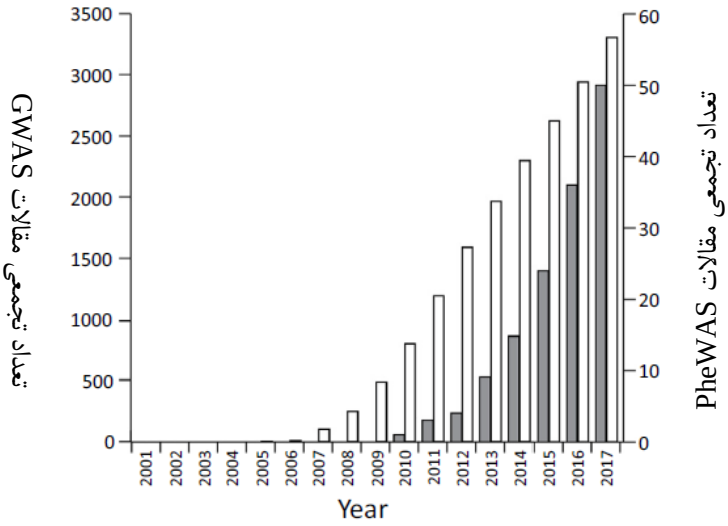
[1] Genetic Relation Matrix
[2] Restricted Maximum Likelihood
[3] Mixed Liner Model
[4] Regional heritability advanced complex trait analysis



شکل ۱-۲ روند زمانی و تکامل تحقیقات آماری بررسی‌شده در GWAS

۲-۲-۵- مدل PheWAS[1]

این مدل یک طرح کلی مطالعه یا روش آماری برای شناسایی رخ‌نمودهای مرتبط با یک ژن‌نمود هست. تحقیقات PheWAS اغلب به اطلاعات رخ‌نمودی عمیق و کامل، ازجمله داده‌های سامانه‌های سلامت الکترونیکی[2] یا داده‌های شیوع شناسی، برای تعیین وضعیت مورد شاهدی برای انواع رخ‌نمودها تکیه می‌کنند (Hebbring, 2019). در این تحقیقات سؤال اصلی این است که چگونه می‌توان داده‌های رخ‌نمودی را با ژن‌نمودی و دیگر لایه‌های داده‌های امیک ترکیب کرد تا بتوان به شناخت بهتر رسید و درمان بیماری‌های انسانی را فهمید و در این مسئله چالش‌های جدیدی در مدیریت داده‌ها، مسائل مربوط به حریم خصوصی و مراقبت‌های بالینی وجود دارد. روند این تحقیقات در شکل ۲-۲ آمده است.



شکل ۲-۲ روند تحقیقات PheWAS در مقایسه با تحقیقات GWAS (Hebbring, 2019)

۲-۲-۶- مدل OSCA[3]

افزایش سریع داده‌های اومیک در دهه‌های گذشته، بررسی ارتباطات بین پروفایل‌های اومیک مانند متیلاسیون در اپی ژنومیکس و ویژگی‌های پیچیده در گروه‌های بزرگ را تسهیل کرده است.

این پیشرفت‌ها عبارت‌اند از شناسایی تعداد زیادی از انواع ژنتیکی مرتبط بابیان ژن ، متیلاسیون DNA، اصلاح هیستون و فراوانی پروتئین؛ کشف اقدامات اومیک مربوط به صفات پیچیده ؛ بهبود دقت در پیش‌بینی صفت با استفاده از داده‌های اومیک؛ و اولویت‌بندی ژن‌های هدف برای صفات پیچیده با تلفیق داده‌های ژنتیکی و

[1] Phenome-wide association study
[2] Electronic Health Record
[3] omic-data-based complex trait analysis

امیک در نمونه‌های بزرگ هست (Zhang *et al.*, 2018). در این مدل اطلاعات مختلف اومیک در مدل آماری خطی MLM پیاده‌سازی و تأثیر عوامل مختلف بر روی رخ‌نمود بررسی‌شده است.

۲-۳- رویکرد فرا تحلیل

با گسترش تحقیقات GWAS و گزارش روابط متعدد بین عناصر ژنوم و صفات، تحقیقاتی مبتنی بر GWAS های انجام‌شده شکل گرفت که به فرا تحلیل[1] شناخته می‌شوند. مراحل فرا تحلیل در تصویر ۲-۳ آمده است. در این تصویر مراحل کار با کادر آبی و نکات مهم در هر مرحله با کادر قرمز نمایش داده‌شده است.

شکل ۳-۲ مراحل مدل‌های فرا تحلیل (Evangelou and Ioannidis, 2013)

در روش‌های فرا تحلیل از مقادیر مختلفی مانند p-vlaue و اندازه نمونه استفاده می‌شود. لازم به ذکر است p-vlaue تا دهه ۱۹۸۰ به‌طور گسترده‌ای درزمینه‌ی های مختلف علمی استفاده می‌شده است، اما پس‌ازآن در علوم زیست پزشکی به شاخصی غیر رایج و تقریباً ممنوع شده، تبدیل‌شده است (Evangelou and Ioannidis, 2013). در ادامه مهم‌ترین انواع فرا تحلیل بررسی می‌شوند.

[1] Meta-Analysis

۲-۳-۱- مدل METAL

فرا تحلیل در حال تبدیل‌شدن به یک ابزار فزاینده‌ی مهم در مطالعات ارتباط ژنومی (GWAS) بیماری‌ها و صفات ژنتیکی پیچیده است (de Bakker *et al.*, 2008) که می‌تواند یک راهبرد کارآمد و عملی برای تشخیص انواع تغییر با ابعاد متوسط باشد (Skol *et al.*, 2007). مدل متال (Willer, Li and Abecasis, 2010) به معرفی متغیرهای مختلف در بررسی روابط ژن- صفت می‌پردازد که کاربرد فراوانی در مدل‌های فرا تحلیل دارد (شکل ۴-۲).

	راهبرد تحلیل	
	مبتنی بر حجم نمونه	مبتنی بر معکوس واریانس
ورودی	N_i - sample size for study i $P_i - P$-value for study i Δ_i - direction of effect for study i	β_i- effect size estimate for study i se_i - standard error for study i
آمار میانی	$Z_i = \Phi^{-1}(P_i/2) * \text{sign}(\Delta_i)$ $w_i = \sqrt{N_i}$	$w_i = 1/\text{SE}_i^2$ $se = \sqrt{1/\sum_i w_i}$ $\beta = \sum_i \beta_i w_i / \sum_i w_i$
Overall Z-Score	$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$	$Z = \beta/SE$
Overall P-value	$P = 2\Phi(\lvert -Z \rvert)$	

شکل ۴-۲ پارامترهای معرفی‌شده در مدل METAL (Willer, Li and Abecasis, 2010)

۲-۳-۲- مقایسه مدل‌های فرا تحلیل

پس از انجام پروژه‌های GWAS در نقاط مختلف دنیا، مدل‌های متعددی به‌منظور فرا تحلیل منتشرشده است. در جدول ۲-۲ خلاصه‌ای از مهم‌ترین مدل‌هایی که تبدیل به بسته‌های نرم‌افزاری شده‌اند به همراه خواص آن‌ها بیان‌شده است (Evangelou and Ioannidis, 2013).

جدول ۲-۲ مقایسه بسته‌های نرم‌افزاری مدل‌های فرا تحلیل (Evangelou and Ioannidis, 2013)

	METAL	GWAMA	MetABEL	PLINK	R packages
قابلیت پردازش فایل های خروجی نرم افزارهای GWAS	No	Yes; SNPTEST, PLINK	Yes; ABEL	Yes; PLINK	No
پیاده سازی تاثیر ثابت	Yes	Yes	Yes	Yes	Yes
پیاده سازی تاثیر تصادفی	No	Yes	No	No	Yes
تولید معیار ناهمگونی	Q, I^2	Q, I^2	Q, I^2	Q, I^2	Q, I^2
مصورسازی نتایج فراتحلیل	No	Manhattan and QQ plots	Forest plots	No	Yes

۲-٤- رویکرد یادگیری ماشین

رویکردهای یادگیری ماشین برای تکمیل روش‌های تک SNP و چند SNP مؤثر در درک ساختار ژنتیک کلی بیماری‌های پیچیده انسان کاملاً امیدوارکننده است. بااین‌حال، به دلیل اینکه این روش‌ها برای داده‌های SNP

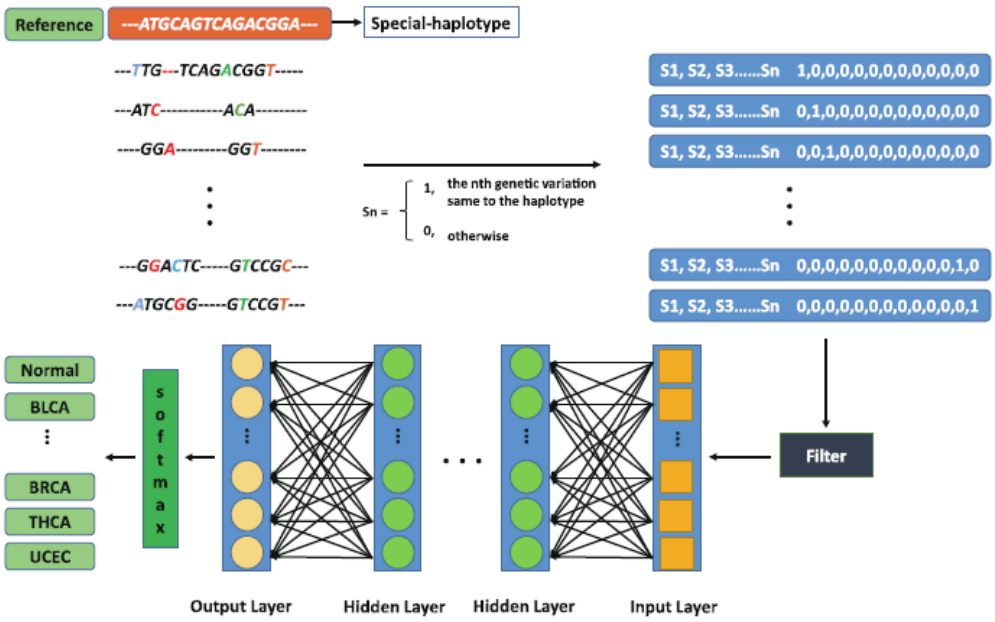
در ژنوم بهینه نشده‌اند، پیاده‌سازی‌های بهبودیافته و روش‌های انتخاب جدید متغیر موردنیاز است (Szymczak *et al.*, 2009).

در بررسی اثر مشترک چند متغیر ژنتیکی و محیطی، چالش‌های متعددی وجود دارد. اول، در GWAS، ژن‌نمودها تا یک‌میلیون SNP در چند هزار نفر تعیین می‌شود، که منجر به مشکل کوچک بودن نمونه و بزرگ بودن تعداد بعد می‌شود. دوم، هنگامی‌که تعداد زیادی از SNPهای ژن‌نمود در یک مقیاس ژنومی هستند، باید عدم تعادل ارتباط (LD) بین SNPها (که منجر به متغیرهای همبسته) شود، موردتوجه قرار گیرد. به همین علت، روش‌های آماری چند متغیره استاندارد مانند رگرسیون خطی چندگانه یا لجستیک برای داده‌های ژنومی مناسب نیستند (Szymczak *et al.*, 2009).

درنتیجه، به‌رغم محدودیت‌های فعلی، روش‌های یادگیری ماشین برای تکمیل پایه ژنتیکی بیماری‌های پیچیده، تکمیلی معقول برای آزمون استاندارد SNP است.

۲-۴-۱- مدل GDL[1]

صفات بیولوژیک نتیجه تعاملات بین توالی ژن و تعامل ژن با شرایط خاص محیطی است. مدل یادگیری عمیق برای مطالعه رابطه بین این عوامل و رخ‌نمود مناسب است. مدل GDL مبتنی بر یادگیری عمیق بر اساس اطلاعات ژنومی در پیش‌بینی سرطان طراحی‌شده است (Sun *et al.*, 2019). تصویر ۲-۵ دیاگرام عملکرد این مدل را نشان می‌دهد.



شکل ۵-۲ دیاگرام عملکرد مدل GDL (Sun *et al.*, 2019)

در این مدل رشته ورودی اطلاعات ۱۰ هزار SNP مرتبط با ۱۲ نوع سرطان است که برای هر نمونه با ژن مرجع مقایسه و تبدیل به رشته ۰و۱ می‌شود. جهت برچسب نیز از روش یک داغ[2] استفاده‌شده است. به این

[1] Genome Deep Learning
[2] One Hot

صورت که برچسب تولیدشده شامل ۱۲ صفر است که درصورتی‌که نمونه دارای سرطان خاصی باشد تنها یکی از صفرها به‌صورت یک ثبت می‌شود. دقت تحقیقاتی این مدل در تشخیص سرطان، ۷۰ درصد بوده است. خلاصه مدل‌های ارائه‌شده به تفکیک رویکرد و ارائه مدل در جدول ۲-۳ آمده است.

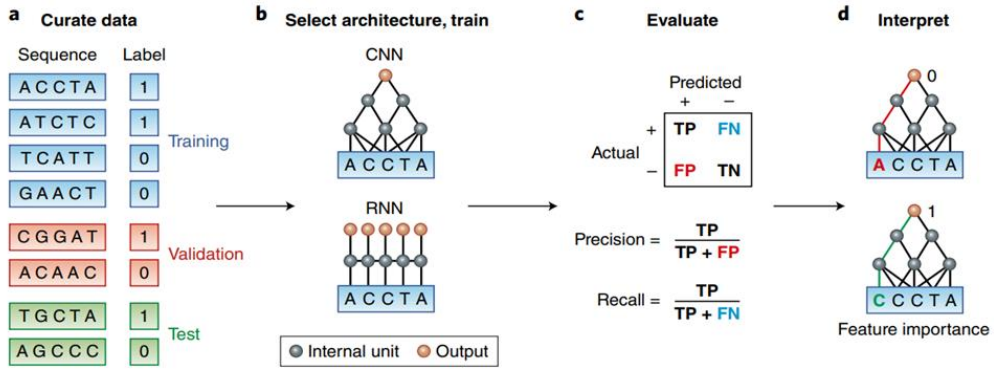
جدول ۳-۲ خلاصه مدل‌های بررسی‌شده

رویکرد	نام	سال	ارائه‌دهنده
آماری	GRAMMER	2007	et al،Aulchenko
	GCTA	2011	et al،Yang
	ACTA	2012	et al.،Gray
	REACTA	2014	et al.،Cebamanos
	PheWAS	2018	Hebbring
	OSCA	2018	et al.،Zhang
فرا تحلیل	METAL	2010	et al.،Willer
یادگیری ماشین	PENALIZED REGRESSION	2009	et al.،Szymczak
	GDL	2019	et al.،Sun

۲-۵- روش‌های یادگیری ماشین سنتی

یادگیری ماشین[1] یکی از زمینه‌های مهم هوش مصنوعی است که دستگاه‌های رایانه‌ای قدرتمندی را برای توصیف یادگیری و اجرای الگوریتم‌ها و مدل‌ها با دقت بی‌نظیر فراهم می‌کند. یک مزیت عمده یادگیری ماشین توانایی یادگیری و ساخت الگوریتم‌های بدون دخالت انسان است. علاوه بر این، دقت تجزیه‌وتحلیل یادگیری ماشین با افزودن داده‌های آموزش بهبود می‌یابد (Khorraminezhad *et al.*, 2020). فن‌های یادگیری ماشین به‌طور گسترده‌ای در تحقیقات ژنومیک استفاده‌شده است. وظایف یادگیری ماشین به دو دسته کلی با نظارت و بدون نظارت تقسیم می‌شود. در یادگیری با نظارت، هدف پیش‌بینی برچسب (رده‌بندی) یا پاسخ هر نقطه داده با استفاده از مجموعه‌ی ارائه‌شده از مثال‌های آموزشی دارای برچسب است. در یادگیری بدون نظارت هدف یادگیری الگوهای ذاتی درون داده‌ها است. هدف نهایی در بسیاری از کارهای یادگیری ماشین بهینه‌سازی عملکرد مدل نه بر روی‌داده‌های موجود (عملکرد آموزش) بلکه در عوض بر روی دادگان‌های مستقل (عملکرد تعمیم) است. با این هدف، داده‌ها به‌طور تصادفی به حداقل سه زیرمجموعه آموزش، اعتبار سنجی و مجموعه آزمون تقسیم می‌شوند. مجموعه آموزش برای یادگیری پارامترهای مدل، مجموعه اعتبار سنجی برای انتخاب بهترین مدل و مجموعه آزمون برای برآورد عملکرد تعمیم کنار گذاشته می‌شوند. یادگیری ماشینی باید به یک تعادل مناسب بین انعطاف‌پذیری مدل و میزان داده‌های آموزش برسد. چارچوب استفاده از یادگیری ماشین در داده‌های ژنومی به‌صورت شکل ۲-۶ است (Zou *et al.*, 2019).

[1] Machine Learning



شکل ۶-۲ چارچوب یادگیری ماشین در داده‌های ژنومی (Zou *et al.*, 2019)

الگوریتم‌های یادگیری نظارت‌شده را می‌توان به‌عنوان روش‌های مبتنی بر رگرسیون یا مبتنی بر درخت رده‌بندی کرد. رگرسیون لجستیک، رگرسیون خطی، شبکه‌های عصبی و SVM نمونه‌های رایج الگوریتم‌های یادگیری تحت نظارت رگرسیون هستند. در این بخش به بررسی روش‌های یادگیری با نظارت در مقالات متناسب با موضوع پرداخته‌شده است.

۲-۵-۱-الگوریتم درخت تصمیم[1]

درخت تصمیم‌گیری، روند نمایی[2] است با ساختار درخت مانند به‌صورت سلسله‌مراتبی که از سه عنصر اصلی تشکیل‌شده است (Quinlan, 1986). بالاترین گره[3] در درخت، گره‌ی ریشه است و گره داخلی (غیر برگ) آزمونی را بر روی یک ویژگی مشخص می‌کند و هر شاخه خارج‌شده از این گره، دستاورد این آزمون را نشان می‌دهد و گره‌ی پایانی را برگ می‌نامند که نشان‌دهنده کلاس‌ها هستند ؛ ازجمله اشیائی که به یک طبقه تعلق دارند یا اینکه خیلی مشابه هستند و می‌توان آن‌ها را در یک طبقه قرارداد. به‌طور عادی، پیچیدگی یک درخت تصمیم با افزایش تعداد ویژگی[4] افزایش می‌یابد. اگرچه در بعضی از شرایط دیده‌شده است که تنها تعداد کمی از ویژگی‌ها می‌توانند کلاسی را که هر شیء به آن تعلق دارد، تعیین کنند و بقیه ویژگی‌ها کم و یا بی‌تأثیرند. فرآیند دسته‌بندی هر نمونه با امتحان ویژگی بیان‌شده در گره ریشه شروع‌شده و سپس از شاخه‌های درخت با توجه به مقدار آن ویژگی پایین می‌آییم. سپس این فرآیند با امتحان گره بعدی که در انتهای شاخه انتخاب‌شده قرار دارد، ادامه می‌یابد تا نهایتاً به یک برگ برسیم. مهم‌ترین الگوریتم‌های درخت تصمیم شامل: ID3، C4.5،CART هست.

۲-۵-۲-الگوریتم جنگل تصادفی

روش جنگل تصادفی الگوریتمی است که توسط لئو بریمن در سال ۲۰۰۰ برای تولید مجموعه‌ای از پیش‌فرض‌ها و داده‌ها با استناد بر درختان جنگل که در فضای تصادفی رشد می‌یابند، مطرح شد (Breiman, 2000).

[1] Decision Tree
[2] Flowchart
[3] Node
[4] Feature

الگوریتم‌های جنگل تصادفی مدل‌های پیش‌بینی را با استفاده از یک روش تلفیقی و به کمک بسیاری از درختان تصمیم‌گیری می‌سازند. به‌ویژه، الگوریتم‌های جنگل تصادفی SNPهایی را انتخاب می‌کنند که در فرآیند تصمیم‌گیری ساخت درخت حاوی اطلاعات مفیدی هستند. قدرت مدل‌های جنگل تصادفی معادل با توانایی آن‌ها در به‌کارگیری مؤثر ساختار داده‌های به‌شدت چندبعدی و مفقودشده که حاوی تعاملات پیچیده هستند هست. به‌عنوان‌مثال، در یک مطالعه اخیر از یک الگوریتم جنگل تصادفی برای پیش‌بینی خطر T2D استفاده شد، که بهتر از هر دو مدل SVM و مدل‌های رگرسیون لجستیک عمل کرده است. در این مطالعه مجموعه‌ای از ۱۰۷۴ ژن‌نمود منفرد و ۱۰۱ SNP مربوط به T2D از پیش تعیین‌شده جمع‌آوری و تصحیح شدند. داده‌های تصحیح‌شده (۶۷۷ نمونه با ۹۶ SNP مرتبط) در یک الگوریتم یادگیری جنگل تصادفی قرار گرفتند و یک پیش‌بینی T2D بااعتبار AUC = 0.85 را با استفاده از متد اعتبارسنجی متقابل انجام دادند. با انجام این کار، همچنین مدل جنگل تصادفی SNP های از پیش تعیین‌شده را برای شناسایی زیرمجموعه‌هایی که به‌شدت با T2D همراهی دارند تصحیح کرده و می‌تواند به‌منظور تحقیق اتیولوژی بیماری مورداستفاده قرار گیرد. اجرای مدل جنگل تصادفی هنوز هم به‌عنوان یک روش یادگیری ماشینی برای مدل‌سازی خطر بیماری‌های پیچیده بسیار سودمند است.

۲-۵-۳-شبکه‌های بیزی[1]

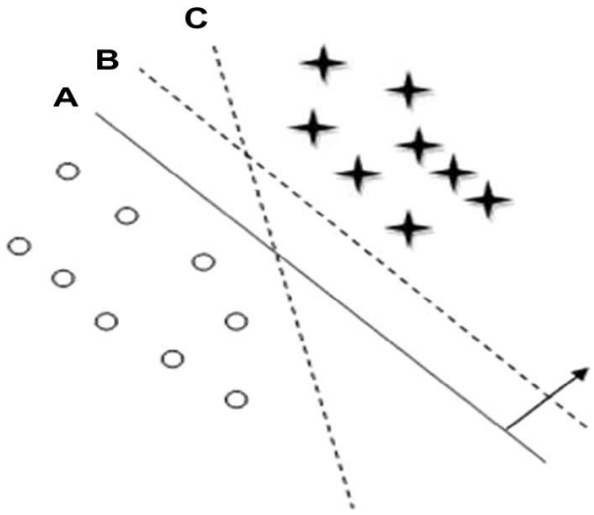
فرض اصلی رده‌بندی BN، استقلال ویژگی‌ها است. یکی دیگر از فرض‌های دور از دسترس این است که همه ویژگی‌ها به‌طور کامل وابسته هستند. این فرض منجر به مدل شبکه بیزی می‌شود که یک گراف خطی جهت‌دار است که گره‌های آن متغیرهای تصادفی هستند و لبه‌ها وابستگی‌های شرطی را نشان می‌دهند. BN یک مدل کامل برای متغیرها و روابط آن‌ها در نظر گرفته است. بنابراین، یک توزیع احتمال احتمالی مشترک (JPD) بر روی تمام متغیرها برای یک مدل مشخص‌شده است. در استخراج متن، پیچیدگی محاسبات BN بسیار سنگین است و به همین دلیل اغلب استفاده نمی‌شود. BN توسط Hernandez و Rodríguez مورداستفاده قرار گرفت (Hernández-Rodríguez *et al.*, 2016) با در نظر گرفتن یک مسئله در دنیای واقعی که در آن دیدگاه نویسنده با سه متغیر هدف اما مرتبط، موردنظر است. آن‌ها از رده‌بندی‌های چندبعدی بیزی استفاده کردند. به‌منظور استفاده از روابط بالقوه بین آن‌ها، متغیرهای هدف متفاوتی را در یک فعالیت رده‌بندی، به یکدیگر متصل نمودند. آن‌ها چارچوب رده‌بندی چندبعدی را به دامنه نیمه نظارتی گسترش دادند تا از مقدار زیادی از اطلاعات بدون برچسب در دسترس در این زمینه استفاده کنند.

۲-۵-۴-رده‌بند ماشین بردار پشتیبان[2]

ماشین بردار پشتیبان (SVM) که نخستین بار توسط واپنیک در سال ۱۹۹۸ ارائه شد، از یک مدل خط برای اجرای حدود کلاس غیرخطی، به‌وسیله‌ی مسیردهی به بردارهای ورودی غیرخطی در داخل یک فضای

[1] Bayesian Network
[2] Support Vector Machine

چندبعدی استفاده می‌کند. در این فضای جدید برای ایجاد یک سطح افزار بهینه همواره (OSH) انجام می‌شوند. نمونه‌های آموزشی که به بیشینه حد سطح خیلی نزدیک هستند، بردارهای پشتیبان نامیده می‌شوند. سایر نمونه‌های آموزشی در تعریف حدود کلاس دودویی بی‌تأثیرند. روش SVM به‌قدری ساده است که می‌توان آن را به‌وسیله‌ی ریاضیات تجزیه‌وتحلیل کرد. ازاین‌جهت، SVM می‌تواند به‌عنوان یک جایگزین مطلوب برای ترکیب نقاط قوت روش‌های آماری مرسوم که به‌طورمعمول دارای محیطی تئوری‌گرا هستند و به‌آسانی تجزیه‌وتحلیل می‌شوند و روش‌های فراگیری ماشین که محیطی بیشتر داده‌ای گرا هستند و توزیع آزاد و قوی دارند، مورداستفاده قرار گیرد. اصل مهم SVMها تعیین جدایی خطی در فضای جستجو است که می‌تواند بهترین کلاس‌های مختلف را جدا کند. در شکل ۲-۷، دو کلاس x، o و ۳ صفحه A، B و C وجود دارد. صفحه A بهترین جدایی بین کلاس‌ها را فراهم می‌کند، زیرا فاصله نرمال هر یک از نقاط داده بزرگ‌ترین است، بنابراین نشان‌دهنده حداکثر حاشیه جدایی است.



شکل ۷-۲ استفاده از ماشین بردار پشتیان در مسئله رده‌بندی

SVM دارای کاربردهای مختلفی در حوزه رده‌بندی هست. به‌عنوان‌مثال ، یک مدل یادگیری ماشین غیر پارامتری مبتنی بر رگرسیون SVM در مورد ژنتیک دیابت نوع ۱ ساخته‌شده و بر اساس ۳۴۴۳ نمونه ژن‌نمود فردی آموزش داده‌شده است. این مدل به AUC = 0.84 دست‌یافته است، که به‌طور قابل‌توجهی بالاتر از مدل نمره دهی خطر چندژنی AUC = 0.71 هست.

۲-۵-۵-رده‌بند درخت تصمیم

رده‌بندی درخت تصمیم، تقسیم سلسله مراتبی فضای داده آموزش را فراهم می‌کند که در آن، یک شرط بر مقدار ویژگی، برای تقسیم داده‌ها استفاده می‌شود. تقسیم فضای داده به‌صورت بازگشتی انجام می‌شود تا گره‌های برگ حاوی حداقل تعداد رکوردها باشند که برای رده‌بندی استفاده می‌شوند.

۲-۵-۶-شبکه‌های عصبی

شبکه عصبی شامل بسیاری از نورون‌ها است که نورون واحد اصلی آن است. ورودی‌های نورون توسط بردار X_i نشان داده‌شده است. مجموعه‌ای از وزن A وجود دارد که با هر نورون مورداستفاده برای محاسبه یک تابع از

ورودی‌های آن f، مورداستفاده قرار می‌گیرد. تابع خطی شبکه عصب به‌صورت مقابل است: $p_i = A.X_i$. در یک مسئله رده‌بندی باینری، فرض بر این است که برچسب کلاس X_i با y_i مشخص می‌شود و علامت تابع پیش‌بینی‌شده p_i برچسب کلاس را نشان می‌دهد.

شبکه عصبی چندلایه برای مرزهای غیرخطی استفاده می‌شود. این لایه‌های چندگانه برای ایجاد چندین خط مرزی تقسیم‌بندی شده استفاده می‌شود. این لایه‌ها برای تقریب محدوده‌های محصور متعلق به یک کلاس خاص استفاده می‌شود. خروجی نورون‌ها در لایه‌های قبلی به نورون‌ها در لایه‌های بعدی وارد می‌شود. فرآیند آموزش پیچیده‌تر است، زیرا خطاها باید بر روی لایه‌های مختلف بازپخش شوند. پیاده‌سازی NNها برای داده‌های متن وجود دارد که در این مقالات به آن اشاره‌شده است (Ng, Goh and Low, 1997; Ruiz and Srinivasan, 1999). مدل یادگیری عمیق ANN عملکرد پیش‌بینی کننده مهمی در مورد چاقی در مجموعه آزمایش‌ها با AUC = 0.9908 ارائه داد. مونتانز (Montañez, Fergus and Chalmers, 2018) توانایی الگوریتم یادگیری عمیق ANN را در ضبط اثرات ترکیبی SNPها و همچنین پیش‌بینی بیماری‌های پیچیده پلی ژنی به‌وضوح نشان دادند.

۲-۶- یادگیری عمیق[1] (DNN)

یادگیری عمیق که زیرمجموعه‌ای از یادگیری ماشین محسوب می‌گردد، رویکردی است که با استفاده از ساختار شبکه‌های عصبی عمیق، سعی بر یادگیری در چند سطح و یادگیری از آموخته‌های پیشین به‌صورت سلسله مراتبی دارد. این روش یک‌رشته نوظهور در حوزه یادگیری ماشین و نوعی رویکرد با سطوح مختلف یادگیری بازنمایی است که سعی دارد با تکیه‌بر روش‌های مبتنی بر یادگیری، همچون انسان اقدام به حل مسائل غیرخطی پیچیده در حوزه‌های مختلف مانند پردازش زبان‌های طبیعی، پردازش تصویر، پردازش صوت نماید (Lecun, Bengio and Hinton, 2015).

پیشرفت عمده اخیر در یادگیری ماشین، توسعه سریع الگوریتم‌های یادگیری عمیق[2] است که می‌تواند از طریق یک فرایند یادگیری انباشته و سلسله مراتبی، ویژگی‌های معنی‌دار را از دادگان‌های چندبعدی و پیچیده استخراج کند. در حال حاضر، تصور می‌شود که یادگیری عمیق پیشرفته‌ترین فن یادگیری ماشین برای کاربردهای مختلف است (Marcos-Zambrano *et al.*, 2021).

یادگیری عمیق برای پیش‌بینی دقیق داده‌های پیچیده مانند تصویر، متن یا فیلم به‌عنوان ابزاری قدرتمند ظهور کرده است. بااین‌حال، توانایی آن در پیش‌بینی مقادیر رخ‌نمودی از داده‌های مولکولی کمتر موردمطالعه قرارگرفته است. در بین معماری‌های یادگیری عمیق که در حال حاضر در پیش‌بینی ژنومی آزمایش‌شده‌اند، شبکه‌های عصبی پیچشی امیدوارکننده‌تر از گیرنده‌های چندلایه به نظر می‌رسند. محدودیت یادگیری عمیق

[1] Deep Neural Network
[2] Deep learning

در تفسیر نتایج است. این ممکن است برای پیش‌بینی ژنومی در اصلاح نژاد گیاهان و حیوانات مهم نباشد اما می‌تواند هنگام تعیین خطر ژنتیکی برای یک بیماری حیاتی باشد (Pérez-Enciso *et al.*, 2019).

انواع مختلفی از شبکه‌های عصبی عمیق مانند شبکه‌های عصبی بازگشتی، شبکه‌های عصبی پیچشی، ماشین بولتزمن محدودشده، رمزگذار خودکار و موارد دیگری وجود دارد.

۲-۶-۱- تحلیل داده‌های حجیم مبتنی بر یادگیری عمیق

داده‌های حجیم[1] به‌عنوان دادگان‌های بزرگ یا پیچیده شناخته‌شده‌اند که معمولاً سیستم پایگاه داده عادی برای انجام پردازش داده‌ها در یک‌زمان قابل‌قبول ناکافی است. این برنامه‌های پردازشی شامل تجزیه‌وتحلیل، ضبط، پردازش داده‌ها، جستجو، به اشتراک‌گذاری، ذخیره‌سازی، انتقال، مصورسازی، پرس‌وجو، به‌روزرسانی و مباحث حریم خصوصی اطلاعات می‌باشند.

داده‌هایی را می‌توان به‌عنوان داده‌های حجیم رده‌بندی کرد اگر داده‌ها شامل سه V است (Russom, 2018). اولین V حجم[2] است که داده‌ها باید تعداد زیادی داده داشته باشند که شامل اطلاعات، پرونده‌ها یا تراکنش‌ها هست. V دوم تنوع[3] داده است که داده‌ها در اشکال مختلف هستند و می‌توانند ساخت‌یافته یا غیر ساختاریافته باشند. آخرین V به‌سرعت[4] اشاره می‌کند که به‌سرعت انتشار داده اشاره می‌کند.

داده‌های ژنوم انسان، دارای حجم بالایی است (داده ژنوم هر انسان حدود ۳ گیگابایت هست) و از این منظر می‌توان گفت داده‌های ژنوم از جنس داده‌های حجیم هست.

فن‌های مختلف یادگیری ماشین مانند باهم آیی، داده‌کاوی حجیم، تجزیه خوشه‌ای و تجزیه‌وتحلیل متن در تحلیل این نوع از داده‌ها استفاده می‌گردد (Maltby, 2011).

بر اساس (Hordri *et al.*, 2017) تحقیقاتی که در حوزه پردازش داده حجیم و یادگیری عمیق انجام‌شده است حول ۴ محور به ترتیب : انتزاع سطح بالا، لایه‌های سلسله مراتبی، پردازش حجم بالای داده و مدل عمومی هست که نشان‌دهنده مراحل اجرای یادگیری عمیق هست.

الگوریتم‌های یادگیری عمیق، مانند آلفاگو[5] (Silver *et al.*, 2016) و تشخیص ابعاد (Russakovsky *et al.*, 2015)، از عملکرد انسان در وظایف بصری بهتر هستند و فن‌های تحلیل انعطاف‌پذیر و قدرتمند برای مقابله با مشکلات پیچیده دارند. یادگیری عمیق یک الگوریتم انتزاعی سطح بالا برای حل مسائل رده‌بندی و رگرسیون است. از طریق یادگیری عمیق و استخراج الگوهای داده‌ها، ساختارهای پیچیده را در مجموعه‌های داده‌های عظیم شناسایی می‌کند و پتانسیل زیادی برای برنامه‌های کاربردی در ژنتیک و ژنومیک دارد (Libbrecht and Noble, 2015).

۲-۶-۲- الگوریتم‌های یادگیری عمیق

---

[1] Big Data
[2] Volume
[3] Variety
[4] Velocity
[5] Alpha Go

مروری بر الگوریتم‌های یادگیری عمیق در جدول ۲-۴ ادامه آمده‌است.

جدول ۴-۲ جدول مقایسه شبکه DNN (Shrestha and Mahmood, 2019)

نوع شبکه	معماری	مدل شبکه	نوع آموزش	الگوریتم آموزشی	نمونه پیاده‌سازی شده	کاربرد رایج
شبکه عصبی پیشرفته[1]	CNN	متمایزکننده	با ناظر	شیب نزول بر اساس پردازش مجدد	شبکه سیامی، CNN عمیق	تشخیص عکس و رده‌بندی
	شبکه باقیمانده[2]	متمایزکننده	با ناظر	شیب نزول بر اساس پردازش مجدد	DeepResNet HighwayNet DenseNet	تشخیص عکس
	خود رمزنگار[3]	تولیدی	بدون ناظر	پردازش مجدد	خود رمزگذار حریص و خود رمزگذار متغیر	کاهش ابعاد، رمزگذاری
	شبکه‌های مخالف[4]	متمایزکننده و تولیدی	بدون ناظر	پردازش مجدد	شبکه تولیدی مخالف	تولید داده جعلی واقعی، بازسازی مدل‌های 3D، بهبود تصویر
	RBM	تولیدی با متمایزکننده قوی	بدون ناظر	شیب نزول بر اساس واگرایی متضاد	شبکه عقیده عمیق و ماشین بلتزمن عمیق	کاهش ابعاد، یادگیری ویژگی، مدل‌سازی موضوعی
شبکه عصبی گردشی[5]	LSTM	متمایزکننده	با ناظر	شیب نزول و پردازش مجدد در طول زمان	Deep RNN GRU NMT	پردازش زبان طبیعی، ترجمه زبان
شبکه عصبی عملکرد پایه شعاعی[6]	شبکه RBF	متمایزکننده	با ناظر و بدون ناظر	خوشه‌بندی k-means و تابع حداقل مربع	RNN	تقریب زدن تابع، پیش‌بینی سری زمانی
شبکه عصبی خودسازمان‌دهی کونونن[7]	گره‌های شکل یافته بر اساس شبکه‌های مستطیلی یا شش‌ضلعی	تولیدی	بدون ناظر	یادگیری رقابتی	شبکه عصبی خودسازمان‌دهی کونونن	کاهش ابعاد، مسائل بهینه‌سازی، تحلیل خوشه‌بندی

نقاط ضعف الگوریتم‌های یادگیری عمیق و فن‌های بهبود در جدول ۲-۵ آمده است.

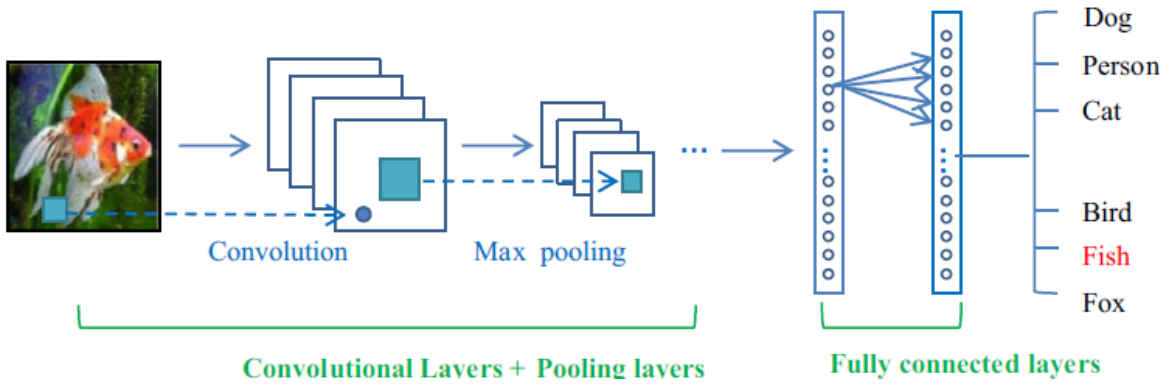
[1] Feed forward neural network
[2] Residual network
[3] Auto encoder
[4] Adversarial networks
[5] Recurrent neural network
[6] radial basis function neural network
[7] kohonen self organizing neural network

جدول ۵-۲ نقاط ضعف الگوریتم‌های یادگیری عمیق

نقاط ضعف	علت	اثر	بهینه‌سازی برای آدرس دادن نقاط ضعف
کاهش یا افزایش ناگهانی و شدید شیب	انتشار مشتق‌ها	زمان طولانی آموزش، پرت شدن کمینه عمومی، سکته آموزشی	تابع فعال‌سازی ReLu، وزن دهی ابتدایی، برقراری ارتباط بین فعال‌سازی درب فراموشی و محاسبه شیب در LSTM (RNN)، ارتباطات پرشی (ResNet)، سخت‌افزار سریع‌تر (GPUs)
کمینه محلی/ نواحی مسطح	شیب نزول فضای مسئله غیر محدب	همگرایی کمینه محلی	نرخ یادگیری انطباقی، مقداردهی مجدد شروع به پارامترها، فن‌های مقداردهی،
لبه‌های تیز	شیب نزول فضای مسئله غیر محدب و فضای تیز	پرت شدن به کمینه عمومی اشتباه	نرخ یادگیری انطباقی
پوشش بیش‌ازحد	گره‌های با اندازه بزرگ‌تر (شبکه) متناسب با دادگان، دادگان ضعیف	دقت ضعیف مدل	مقررات، از قلم افتادن، انتخاب اندازه مناسب برای شبکه، داده‌های آموزشی بهتر
زمان آموزشی زیاد	شبکه بزرگ (وزن‌ها و پارامترهای زیاد)، داده با ابعاد بالا و ...	استفاده ضعیف از منابع محاسبه، زمان هدر شده	نرخ یادگیری انطباقی، استفاده از ابر و GPUs
انتخاب بیش‌ازحد پارامتر	فضای جواب بسیار بزرگ (مسئله NP-hard) برای مسائل با ابعاد بالا	همگرایی به کمینه محلی	فرا ابتکاری (مانند الگوریتم ژنتیک) برای بهینه‌سازی با پارامترهای بالا، جستجو تصادفی، الگوریتم مبتنی بر مدل ترتیبی

۲-۶-۳-شبکه‌های عصبی پیچشی[1] (CNN)

شبکه‌های عصبی پیچشی یکی از مهم‌ترین روش‌های یادگیری عمیق هستند که در آن‌ها چندین لایه با روشی قدرتمند آموزش می‌بینند. این روش بسیار کارآمد بوده و یکی از رایج‌ترین روش‌ها در کاربردهای مختلف بینایی رایانه است. تصویر کلی یک معماری شبکه عصبی پیچشی در شکل ۲-۸ نمایش داده‌شده است.



شکل ۲-۸ معماری شبکه CNN

[1] Convolutional Neural Network

به‌طورکلی، یک شبکه CNN از سه لایه اصلی تشکیل می‌شود که عبارت‌اند از : لایه پیچشی[1]، لایه ادغام[2] و لایه تماماً متصل[3] (FCN). لایه‌های مختلف وظایف مختلفی را انجام می‌دهد (Krizhevsky, Sutskever and Hinton, 2017). در هر شبکه عصبی پیچشی دو مرحله برای آموزش وجود دارد. مرحله جلوبرنده[4] (FF) و مرحله پس انتشار[5] (BP). در مرحله اول تصویر ورودی به شبکه تغذیه می‌شود و این عمل چیزی جز ضرب نقطه‌ای بین ورودی و پارامترهای هر نورون و نهایتاً اعمال عملیات پیچشی در هر لایه نیست. سپس خروجی شبکه محاسبه می‌شود. در اینجا به‌منظور تنظیم پارامترهای شبکه و یا به‌عبارت‌دیگر همان آموزش شبکه، از نتیجه خروجی جهت محاسبه میزان خطای شبکه استفاده می‌شود. برای این کار خروجی شبکه را با استفاده از یک تابع خطا[6] (LF) با پاسخ صحیح مقایسه کرده و این‌طور میزان خطا محاسبه می‌شود. در مرحله بعدی بر اساس میزان خطای محاسبه‌شده مرحله BP آغاز می‌شود. در این مرحله گرادیانت هر متغیر با توجه به‌قاعده زنجیره‌ای محاسبه می‌شود و تمامی پارامترها با توجه به تأثیری که بر خطای ایجادشده در شبکه‌دارند تغییر پیدا می‌کنند. بعد از بروز آوری شدن پارامترها مرحله بعدی FF شروع می‌شود. بعد از تکرار تعداد مناسبی از این مراحل آموزش شبکه پایان میابد.

۲-۷- پیش‌بینی امتیاز خطر[7]

مدل‌های معمول خطر بیماری بر اساس همه‌گیرشناسی (باقدرت پیش‌بینی محدود) در درجه اول مبتنی بر خطر عامل‌های شیوه زندگی مانند وجود سابقه خانوادگی مثبت می‌باشند. اخیراً، در نظر گرفتن خطر عامل‌های ژنتیکی ازجمله SNPهای مرتبط با بیماری و یا مربوط به رخ‌نمود همراه با مدل‌سازی خطر، دقت پیش‌بینی فردی بیماری‌ها را بهبود بخشیده است.

شاید بهترین مدل‌های پیش‌بینی کننده خطر بر اساس پتانسیل آن‌ها به‌منظور راهنمایی در مورد پیشگیری از بیماری و درمان بدون نیاز به روش‌های غربالگری پزشکی پرهزینه و بالقوه نامطلوب (به‌عنوان‌مثال بیوپسی های تهاجمی) باشد.

در حال حاضر، توسعه مدل‌های خطر ژنتیکی بر روی دستیابی به قدرت پیش‌بینی دقیق برای شناخت افراد درخطر به شیوه‌ای قدرتمند تمرکز دارد. همان‌طور که قبلاً گفته شد، مطالعات GWA، SNPها را با توجه به ارتباط آن‌ها با یک بیماری / رخ‌نمود در سطح جمعیت تعریف می‌کند. بنابراین، ترکیب SNPها در یک مدل پیش‌بینی کننده‌ی خطر نیاز به ترکیب با مدل‌هایی دارد که ژن‌نمود یک فرد را به دستمی‌آورند تا بتوانند خطر را پیش‌بینی کنند. مدل‌های پیش‌بینی خطر ژنتیکی معمولاً توسط: (۱) امتیازدهی خطر چندژنی[8] (PRS) یا (۲) یادگیری ماشینی ساخته می‌شوند.

---

[1] Convolution
[2] Pooling
[3] Fully Connected Network
[4] feed forward
[5] back propagation
[6] loss function
[7] Risk Scroe
[8] Polygenic Risk Score

عملکرد پیش‌بینی کننده هر دو نوع مدل با استفاده از منحنی‌های مشخصه عملکرد سیستم[1] (ROC) ارزیابی می‌شود، که در آن حساسیت و اختصاصی پیش‌بینی‌ها در مقادیر مختلف رتبه‌بندی می‌شود.
مساحت زیر منحنی[2] ROC (AUC) برابر با این احتمال است که مدل موردبررسی، از میان یک جفت نمونه و شاهد (کنترل) تصادفی، یک مورد را به‌درستی شناسایی کرده است. نتایج AUC از ۰.۵ (به‌عنوان‌مثال تصادفی) تا ۱ (یعنی دقت ۱۰۰ درصد) متغیر است.

۲-۷-۱- امتیازدهی خطر چندژنی

مدل‌های معمول خطر بیماری بر اساس همه‌گیرشناسی (باقدرت پیش‌بینی محدود) در درجه اول مبتنی بر خطر عامل‌های شیوه زندگی مانند وجود سابقه خانوادگی مثبت می‌باشند. اخیراً، در نظر گرفتن خطر عامل‌های ژنتیکی ازجمله SNPهای مرتبط با بیماری و یا مربوط به رخ‌نمود همراه با مدل‌سازی خطر، دقت پیش‌بینی فردی بیماری‌ها را بهبود بخشیده است (Mosley *et al.*, 2020). امتیازدهی خطر چندژنی از یک رویکرد مدل ثابت برای جمع‌آوری مجموعه‌ای از آلل های خطر در یک بیماری پیچیده خاص استفاده می‌کند.در این روش تعداد زیادی SNP برای اندازه‌گیری خطر در بدو تولد استفاده می‌شوند (Phulka *et al.*, 2023). نمرات خطر چندژنی ممکن است وزنی یا غیر وزنی باشند. در نمرات خطر چندژنی وزنی، سهم آلل های خطر به‌طورمعمول با نسبت شانس یا اندازه اثر آن‌ها اندازه‌گیری می‌شود. در مقابل، نمرات خطر چندژنی غیر وزنی برابر با مجموع تعداد گونه‌های اللی پیوسته در یک ژنوم است. مدل غیر وزنی فرض می‌کند که همه گونه‌ها دارای اندازه اثر متعادل هستند. این فرض ساده‌انگاران، استفاده از نمرات خطر چندژنی غیر وزنی را برای صفات پیچیده با ساختارهای ژنتیکی زیربنایی که شامل اثرات نامتعادل گونه‌ها هست را محدود می‌کند.
دو مرحله برای تدوین امتیاز خطر چندژنی وجود دارد: (۱) مرحله کشف و (۲) مرحله اعتبار سنجی. مرحله کشف در مورد رتبه‌بندی خطر چندژنی وزنی آزمون ارتباط آماری (برای مثل رگرسیون خطی یا لجستیک) استفاده می‌کند تا اندازه اثر یک مجموعه بزرگ از داده‌ها و موارد کنترل پروفایل‌های ژن‌نمود فردی را برآورد کند. مرحله کشف یک نمره خطر چندژنی غیر وزنی نیازمند معیارهای انتخاب دقیق SNPها به‌منظور جلوگیری از اختلاط SNPها با اندازه اثرهای جزئی هست. در هر دو مورد نمره خطر چندژنی وزنی و غیر وزنی، پس از توسعه، مدل اکتشافی به مرحله اعتبارسنجی منتقل می‌شود. اعتبارسنجی نمره خطر چندژنی مستلزم استخراج SNPهایی حاوی اطلاعات مفید و اندازه‌های اثر از مجموعه کشف، با استفاده از آستانه p-value پیوستگی (به‌عنوان‌مثال ۵×۱۰^{-۸}) هست، که متعاقباً به مرحله اعتبار سنجی منتقل می‌شود. در طی این فرآیند، مدل نمره خطر چندژنی بر روی یک دادگان آزمایشی (یعنی مجموعه‌ای مستقل از داده‌های ژن‌نمودی مورد-شاهد) اعمال می‌شود. نمرات خطر چندژنی برای هر پروفایل ژن‌نمود فردی در داده‌های آزمایش محاسبه می‌شود.

[1] Receiver Operating Characteristic curves
[2] Area Under Curve

سپس قدرت پیش‌بینی نمرات خطر چندژنی فردی برای صفات پیچیده با استفاده از قدرت پیوستگی نمرات با نتایج کلینیکی (رخ‌نمودها) در دادگان‌های آزمایش محاسبه می‌شود.

تلاش‌های اولیه برای استفاده از نمرات خطر چندژنی وزنی مبتنی بر تعداد کمی از SNP های بسیار مهم بوده است که با استفاده از مطالعات GWA مشخص‌شده‌اند و در مورد بیماری‌های پیچیده فقط به ارزش پیش‌بینی کننده محدودی دست‌یافته‌اند. این امر یک محدودیت اصلی در مدل‌سازی نمره خطر چندژنی وزنی، به‌ویژه در مورد حد آستانه p-value برای انتخاب SNP در تأثیر دادگان‌های اکتشافی بر عملکرد مدل و قدرت پیش‌بینی را نشان می‌دهد. انتخاب تعداد محدودی از SNPها با اندازه اثر زیاد، با نادیده گرفتن بخش عمده‌ای از گونه‌ها که تأثیرات بسیار کمتری بر روی رخ‌نمود دارند، زیربناهای بیولوژیکی بیماری‌های پیچیده را بیش‌ازحد ساده می‌کند.

به‌عنوان‌مثال، نسبت شانس متوسط در هر آلل خطر T2D از ۱.۰۲ تا ۱.۳۵ متغیر است. مدل‌های امتیازات چندژنی اخیر، به‌منظور دستیابی به نتایج پیش‌بینی کننده بهتر در مورد صفات پیچیده ژنتیکی، SNP های گسترده‌تری را انتخاب می‌کنند.

به‌عنوان‌مثال، استفاده از حد آستانه p-value (به‌اندازه ۰.۰۱ ، ۰.۱ و ۰.۲ و غیره) منجر به توسعه مدل‌های بهبودیافته خطر چندژنی را برای بیماری‌های روانی با حداقل افزایش در خطاهای مثبت کاذب (یعنی مدل‌ها از نسبت قدرت به نویز قابل قبولی برخوردار هستند) شده است.

رویکرد نمره خطر چندژنی وزنی، پیش‌بینی خطر اسکیزوفرنی را با دستیابی به اثربخشی معقول AUC 0.65 امکان‌پذیر کرده است. به‌طور مشابهی، نتایج قابل‌توجهی از پیش‌بینی نمره خطر چندژنی وزنی برای سایر صفات پیچیده ازجمله دیابت نوع ۱ و بیماری سلیاک نیز به‌دست‌آمده است (Padmanabhan and Dominiczak, 2021). اخیراً، با توجه به اینکه SNPهایی که آستانه اهمیت بسیار سخت‌گیرانه برای ارتباط گسترده ژنومی را برآورده نمی‌کنند نیز می‌توانند پیش‌بینی کننده بیماری باشند، طیف وسیع‌تری از SNPها، از هزاران تا میلیون‌ها SNP برای ایجاد یک GRS بهبودیافته به‌عنوان PRS استفاده‌شده‌اند (Torkamani, Wineinger and Topol, 2018). باید تأکید کرد اندازه خطر ارائه‌شده توسط GRS یک محدوده احتمالی است و این با اطلاعات خطر از نشانگرهای ژنتیکی اختلالات تک ژنی متفاوت است.

۲-۷-۲- مدل‌های پیش‌بینی کننده بیماری بر پایه یادگیری ماشینی

روش‌های یادگیری ماشینی مجموعه‌ای از الگوریتم‌های آماری و محاسباتی پیچیده هست (به‌عنوان‌مثال ماشین بردار پشتیبان (SVM) یا جنگل تصادفی) که با استفاده از نقشه‌برداری ریاضی ارتباطات پیچیده بین یک مجموعه SNPهای خطرناک و رخ‌نمودهای بیماری پیچیده، پیش‌بینی‌هایی انجام می‌دهد.

این روش‌ها از رویکردهای نظارت‌شده یا نظارت‌نشده برای ترسیم ارتباط با بیماری‌های پیچیده استفاده می‌کنند. علیرغم سودمندی روش‌های یادگیری ماشین نظارت‌نشده و داده‌های غیر ژنتیکی در پیش‌بینی

بیماری‌ها، ما در این نوشتار بر روی مدل‌سازی نظارت‌شده که با داده‌های SNP اطلاع‌رسانی می‌شود، تمرکز خواهیم کرد.

مدل‌های یادگیری ماشینی نظارت‌شده به‌منظور پیش‌بینی بیماری‌ها با استفاده از آموزش الگوریتم‌های یادگیری از پیش تعیین‌شده برای ترسیم روابط بین داده‌های ژن‌نمودی فردی و بیماری مرتبط با آن‌ها ایجاد می‌شوند. قدرت پیش‌بینی بهینه برای بیماری موردنظر از طریق نقشه‌برداری از الگوی ویژگی‌های منتخب (متغیرها) در داده‌های ژن‌نمود آموزش حاصل می‌شود. برخی از مدل‌ها از روش‌های با شیب نزولی و دوره‌های تکراری برآورد پارامترها به‌منظور جستجو در میان داده‌های آموزشی برای قدرت پیش‌بینی کننده بهینه استفاده می‌کنند. این روند بازگشتی تا رسیدن به عملکرد بهینه پیش‌بینی کننده ادامه می‌یابد. در پایان مرحله آموزشی، مدل‌هایی با حداکثر قدرت پیش‌بینی کننده در دادگان‌های آموزش برای اعتبارسنجی انتخاب می‌شوند.

در مرحله اعتبار سنجی، عملکرد پیش‌بینی کننده‌ی مدل‌های یادگیری ماشین برای تعیین قدرت آن‌ها به جهت پیش‌بینی عمومی مورد ارزیابی قرار می‌گیرد. مشابه امتیازدهی خطر چندژنی، مرحله اعتبارسنجی با ارزیابی الگوریتم در یک دادگان‌های مستقل انجام می‌شود. مرحله اعتبار سنجی برای اطمینان از این‌که مدل‌های پیش‌بینی کننده داده‌های آموزش را بیش‌ازحد پوشش نمی‌دهند ضروری است. اعتبار سنجی متقابل روشی متداول برای اعتبارسنجی عملکرد مدل‌ها با استفاده از دادگان‌های اصلی است. بااین‌وجود، اعتبارسنجی (آزمایش) خارجی با استفاده از یک دادگان مستقل لازم است تا درنهایت قدرت پیش‌بینی کننده یک مدل یادگیری ماشین را تائید کند. سرانجام کارایی الگوریتم از طریق مقایسه‌های کنترل‌شده تصادفی با بهترین عملکرد بالینی فعلی مشخص می‌شود. تنها درصورتی‌که الگوریتم اطلاعاتی بیشتری را در جهت رده‌بندی دقیق‌تر جمعیت، پیش‌بینی خطر بیماری یا پاسخ‌های درمانی ارائه دهد، کارایی بالینی خود را درنهایت اثبات می‌کند (Ho *et al.*, 2019).

پیشرفت‌های روزافزون فنّاوری به‌طور مداوم در حال بهبود کیفیت و کمیت دادگان‌های پیوسته پیچیده‌ای هست که در مورد رخ‌نمودها و بیماری‌های انسانی جمع‌آوری‌شده است. ادغام این داده‌های با ابعاد ژنومی بسیار بالا در مدل‌های یادگیری ماشین می‌تواند منجر به بهبودهایی درزمینهٔ پیش‌بینی خطرات ژنتیکی فراتر ازآنچه در مورد نمرات خطر چندژنی حاصل می‌شود گردد. پیش‌بینی نمره خطر چند ژن بر اساس یک مدل رگرسیون پارامتری خطی است که شامل فرضیات سخت‌گیرانه هست، که شامل اثرات پیش‌بینی کننده افزایشی و مستقل، یک توزیع نرمال برای داده‌های زمینه‌ای و همچنین عدم همبستگی داده‌ها است. این فرضیات لزوماً در مورد ساختارهای ژنتیکی پایه‌ای بیماری‌های پیچیده چندژنی صحیح نیستند و بنابراین منجر به کاهش زیادی در کاراییِ پیش‌بینی می‌شوند. نکته قابل‌توجه این است که مدل‌سازی رگرسیون افزایشی خطی قادر به محاسبه اثرات متقابل پیچیده بین آلل های مرتبط، که بنا بر گزارش‌ها به مقدار زیادی با رخ‌نمودها مرتبط است نیست. ازاین‌رو، مدل‌سازی مبتنی بر رگرسیون افزایشی خطی، نمرات خطر چندژنی را به سمت پیش‌بینی‌های جانب‌دارانه و با کارایی کمتر سوق می‌دهد. در مقابل، الگوریتم‌های یادگیری ماشین از روش‌های

چند متغیره غیر پارامتری استفاده می‌کنند که به طرز قدرتمندی الگوها را از داده‌هایی که توزیع غیر نرمال دارند و به‌شدت به یکدیگر پیوسته‌اند تشخیص می‌دهند. ظرفیت الگوریتم‌های بر پایه یادگیری ماشین در جهت مدل‌سازی داده‌هایی با ساختارهای بسیار پیچیده و مرتبط باعث شده است که این رویکردها با اهداف پیش‌بینی بیماری‌های پیچیده از محبوبیت بسیار زیادی برخوردار باشند.

نقاط قوت و ضعف هر دو روش در جدول ۲-۶ خلاصه‌شده است (Ho *et al.*, 2019).

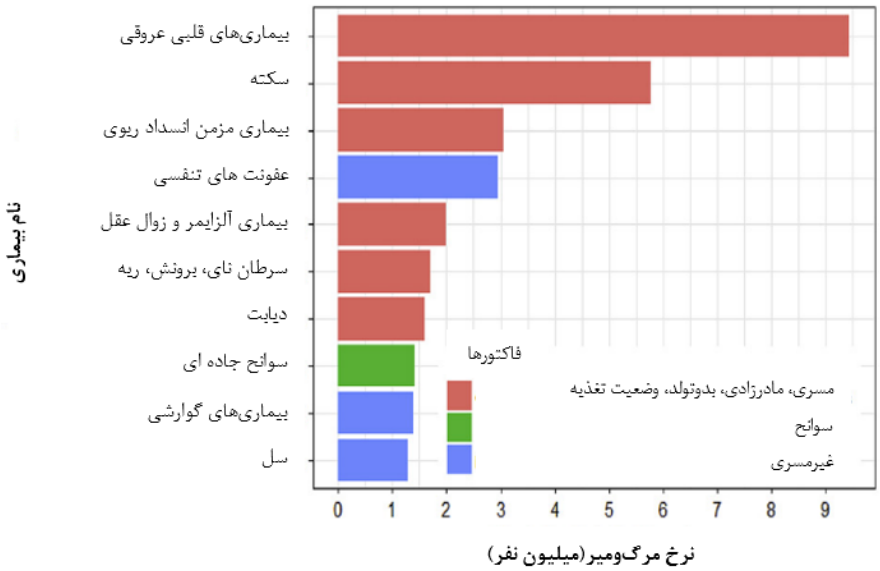
جدول ۶-۲ مقایسه روش GRS و یادگیری ماشین (Ho *et al.*, 2019)

	مدل یادگیری ماشین	امتیاز خطر چندژنی
نقاط قوت	✓ مؤثر برای داده‌های چندبعدی ✓ قابل‌استفاده در داده‌های پیچیده مرتبط ✓ عدم نیاز به فرض نرمال بودن	✓ سادگی اجرا ✓ سادگی تفسیر
نقاط ضعف	☒ سختی در اجرا ☒ سختی در تفسیر اجزای مؤثر ☒ نیاز به داده زیاد	☒ تأثیر افزایشی و مستقل پیش‌بینی کننده ☒ فرض نرمال بودن ☒ بی‌استفاده درداده پیچیده مرتبط

۲-۸- بیماری پرفشاری خون[1] (HTN)

بیماری پرفشاری خون یکی از مهم‌ترین مشکلات سلامت عمومی و در حال افزایش در سطح دنیا به شمار می‌آید (تیموری, ابراهیمی و علوی نیا, ۱۳۹۴). فشارخون بالا یکی از عوامل مهم در بیماری قلبی عروقی است که توسط عوامل ژنتیکی و غیر ژنتیکی ایجاد می‌شود و بیش از ۱.۲۸ میلیارد نفر را در جهان مبتلا کرده است. (Loganathan *et al.*, 2019)(Eshmurodova Dilrabo Khudayberdi kizi, Ergasheva Iroda Akbarali kizi and Siab, 2023). این بیماری با بیماری کلیوی، سکته مغزی نیز ارتباط دارد و شیوع بالایی در کشورهای درحال‌توسعه و کشورهای توسعه‌یافته دارد و سالانه جان بیش از ۱۰ میلیون نفر را می‌گیرد (Singh *et al.*, 2016) (Garimella *et al.*, 2023). به همین دلیل تشخیص بهنگام فشارخون برای درمان زودرس بسیار مهم است (Nour and Polat, 2020; Niu *et al.*, 2021). شکل ۲-۹ سهم هر بیماری در مرگ‌ومیر جهانی را نشان می‌دهد (Alizadehsani *et al.*, 2019) که بیان‌کننده اهمیت بیماری‌های قلبی و عروقی است.

[1] Hypertension



شکل ۹-۲ ده دلیل اصلی عامل مرگ در جهان (Alizadehsani *et al.*, 2019)

در مقاله " مدل پیش‌بینی فشارخون بالا مبتنی بر عوامل خطر ژنتیکی و محیطی در شمال هان چین" که در سال ۲۰۱۹ منتشر شد، پرفشاری خون به‌عنوان یک بیماری مزمن شناخته شد که شیوع بالایی در جهان دارد و از عوامل بیماری قلبی عروقی و مغزی عروقی است. در این مقاله علاوه بر عوامل و پارامترهای محیطی و کلینیکی، اطلاعات ژنتیکی مانند چندریختی نوکلئوتیدی[1] (SNP) نیز موردبررسی قرار گرفت. روش‌های آماری سنتی مانند آزمایشt و رگرسیون لجستیک چند متغیره برای تحلیل اطلاعات انجام شد. یک روش یادگیری ماشین مانند ماشین بردار پشتیبان نیز برای توسعه مدل پیش‌بینی بیماری فشارخون بالا استفاده شد. درنهایت دو مدل برای پیش‌بینی فشارخون سیستولیک و فشارخون دیاستولیک ساخته شد. مدل فشارخون سیستولیک شامل شش عامل محیطی مانند سن، شاخص توده بدن، دور مچ، فعالیت فیزیکی، سابقه خانوادگی فشارخون بالا و یک SNP و مدل فشارخون دیاستولیک شامل شش عامل محیطی مانند وزن، مصرف الکل، فعالیت فیزیکی، تری‌گلیسیرید، سابقه خانوادگی فشارخون بالا و سه SNP هست. ناحیه زیر نمودار برای مدل فشارخون سیستولیک ۰/۶۷۳ و مدل فشارخون دیاستولیک ۰/۸۱۷ به دست آمد (Li *et al.*, 2019).

هم‌چنین می‌توان با استفاده از روش‌های مختلف یادگیری ماشین، یک مدل پیش‌بینی بر مبنای هزینه طراحی کرد تا بهترین عملکرد در پیش‌بینی افراد دیابتی در معرض خطر پرفشاری خون را داشته باشد. در مقاله‌ای تحت عنوان " مقایسه روش‌های مختلف یادگیری ماشین در تشخیص پرفشاری خون در بیماران دیابتی با و بدون در نظر گرفتن هزینه‌ها از الگوریتم‌های مختلف داده‌کاوی مانند الگوریتم درخت تصمیم، ماشین بردار پشتیبان، شبکه عصبی و رگرسیون لجستیک استفاده شد. یکی از یافته‌های تحقیق این بود که افزایش فشارخون سیستول به میزان ۱۳۰میلی‌متر جیوه، فرد دیابتی را بیشتر در معرض پرفشاری خون قرار می‌دهد. با توجه به هدف حداقل سازی هزینه، درخت تصمیم و رگرسیون لجستیک بهترین عملکرد رادارند. درنتیجه

[1] Single Nucleotide Polymorphism

در مسائل پیش‌بینی بیماری‌ها حساس به هزینه کردن روش‌ها و در نظر گرفتن توزیع واقعی بیماری در جامعه اهمیت بیشتری دارد تا اینکه تنها هدف، کمینه کردن تعداد خطاهای دسته‌بندی روی‌دادگانی موجود باشد.
سابقه خانوادگی، سن، جنس، مصرف نمک، مصرف الکل، مصرف دخانیات، کم‌تحرکی و چاقی از عوامل خطر بیماری فشاری خون بالا به شمار می‌رود. این عوامل روی شاخص توده بدنی (BMI) ،دیابت، چربی خون بالا (شامل کلسترول، لیپوپروتئین با تراکم بالا (HDL)، لیپوپروتئین با تراکم پایین (LDL) و تری گلیسرید) تأثیر دارند (تیموری, ابراهیمی و علوی نیا, ۱۳۹۴).
آقای دانگ‌دونگ و همکارانش مقاله‌ای را تحت عنوان توسعه مدل‌های پیش‌بینی ریسک برای پرفشاری خون در سال ۲۰۱۷ منتشر کردند. ازآنجاکه بیماری پرفشاری خون از تهدیدهای پیشرو سلامت جهانی و از عوامل مؤثر در بیماری قلبی عروقی هست و هم‌چنین چون مداخلات کلینیکی در به تأخیر انداختن پیشرفت بیماری پرفشاری خون مؤثر است، مدل‌های پیش‌بینی تشخیصی برای شناسایی جمعیت بیمار پر ریسک فشارخون بالا مهم هست.
داده‌ها در ابتدا از پایگاه داده PubMed و Embase جمع‌آوری شدند، سپس عواملی ریسک، مدل‌های آماری، ویژگی‌های طراحی مطالعه و شرکت‌کنندگان و هم‌چنین معیار عملکرد موردبررسی قرار گرفتند. از عواملی ریسک سنتی مانند شاخص توده بدن، سن، مصرف سیگار، سطح فشارخون، سابقه خانوادگی فشارخون و عوامل زیست‌شیمیایی می‌توان نام برد، درحالی‌که در شش گزارش از نمره ریسک ژنتیکی به‌عنوان عامل پیش‌بینی استفاده‌شده است. دامنه محدوده زیر نمودار از ۰/۶۴ تا ۰/۹۷، C-Statistic از ۶۰٪ تا ۹۰٪ به‌دست‌آمده آمده است.
از یافته‌ها می‌توان به این نتیجه رسید که مدل‌های سنتی هنوز جزء مدل‌های پیش‌بینی ریسک غالب برای پرفشاری خون هستند، اما اخیراً مدل‌های بیشتری شروع به ترکیب و واردکردن عواملی ژنتیکی به‌عنوان قسمتی از پیش‌بینی کننده مدل استفاده می‌شود. هرچند که این پیش‌بینی کننده‌های ژنتیکی باید به‌خوبی انتخاب شوند. مدل‌های استفاده‌شده در این مقاله توانایی کالیبراسیون و تفکیک قابل‌قبول تا خوبی رادارند، اما برای اینکه در فعالیت کلینیکی و عملی مورداستفاده قرار بگیرند، به اعتبارسنجی و تنظیم بیشتری نیاز دارند (Li *et al.*, 2019).
در مقاله طبقه‌بندی خودکار انواع فشارخون بالا بر اساس ویژگی‌های شخصی توسط الگوریتم‌های یادگیری ماشین، به‌منظور شناسایی انواع فشارخون بالابر اساس اطلاعات و ویژگی‌های شخصی، چهار روش یادگیری ماشین شامل طبقه‌بندی درخت تصمیم C4.5 ، جنگل تصادفی، تجزیه‌وتحلیل تفکیک خطی و پشتیبانی خطی ماشین بردار استفاده‌شده و سپس با یکدیگر مقایسه شده است. در ادبیات، ابتدا رده‌بندی انواع فشارخون بالا با استفاده از الگوریتم‌های طبقه‌بندی بر اساس داده‌های شخصی انجام شد. برای توضیح بیشتر تنوع نوع طبقه‌بندی، چهار الگوریتم طبقه‌بندی کننده مختلف برای حل این مسئله انتخاب شد. در دادگان‌های فشارخون، هشت ویژگی ازجمله جنس، سن، قد (سانتی‌متر)، وزن (کیلوگرم)، فشارخون سیستولیک (میلی‌متر

جیوه)، فشارخون دیاستولیک (میلی‌متر جیوه)، ضربان قلب (دور در دقیقه) و BMI (کیلوگرم بر مترمربع) وجود دارد. برای توضیح وضعیت فشارخون بالا و سپس چهار کلاس وجود دارد که شامل فشارخون طبیعی (سالم)، پیش فشارخون، فشارخون مرحله ۱ و فشارخون مرحله ۲ است. نتایج به‌دست‌آمده نشان داده است که می‌توان با اطمینان از روش‌های یادگیری ماشین در تعیین خودکار انواع فشارخون بالا استفاده کرد (Nour and Polat, 2020).

هم‌چنین در مقاله رویکرد طبقه‌بندی بر اساس داده‌های ژنتیکی برای پیش‌بینی فشارخون بالا شناخته‌شده است که فشارخون بالا دلیل اصلی مرگ در سراسر جهان است. میزان مرگ‌ومیر ناشی از این بیماری از دهه گذشته تا ۹۴٪ افزایش‌یافته است. به دلیل این بیماری، شرایط خطرناک سلامتی مانند نارسایی قلبی، نارسایی کلیه و سکته مغزی ایجاد می‌شود. برای جلوگیری از بین رفتن دائمی ، نیازی به فن‌های خودکار برای کشف این بیماری وجود ندارد. اخیراً، اطلاعات ژنتیکی با روش‌های یادگیری ماشین برای تشخیص بیماری فشارخون بالا ترکیب‌شده است. تشخیص مبتنی بر ژنتیک نیز نقشی اساسی در تکامل این بیماری دارد. در این مقاله از داده‌های ژنتیکی و اطلاعات دموگرافیک برای تشخیص این بیماری استفاده‌شده است. برای شناسایی، دادگان-های ژنتیکی ۲۵۰ بیمار از پایگاه‌های داده مختلف (مخازن آنلاین) جمع‌آوری شد. تحقیقات از یک مجموعه ویژگی شامل ژن‌ها، آلل‌های SNP ، آلل‌های خطرناک، کروموزوم‌ها، منطقه و ملیت بیماران استفاده می‌کند. در اینجا، الگوریتمی برای ساختار داده ارائه‌شده است. نتایج به‌دست‌آمده نشان می‌دهد که بهترین دقت با استفاده از نایو بیزین یعنی ۹۸٪ حاصل می‌شود (Anis, Fahiem and Tauseef, 2016).

مقاله داده‌کاوی طبقه‌بندی-محور برای شناسایی الگوهای ریسک در ارتباط با پرفشاری خون در جمعیت خاورمیانه توسط آقای کبیر و همکارانش در سال ۲۰۱۶ منتشر شد. هدف، شناسایی الگوهای ریسک در ارتباط با بیماری پرفشاری خون با استفاده از داده‌کاوی بود و بررسی‌های موردنیاز با استفاده از مطالعات هم‌گروهی انجام‌شده است.

داده از ۶۲۰۵ شرکت‌کننده که ۴۴٪ مردان بزرگ‌تر از ۲۰سال و بقیه زنان بودند، هم‌چنین این افراد بدون فشارخون و سابقه بیماری قلبی عروقی بودند، به‌دست‌آمده آمد. برای توسعه یک مدل پیش‌بینی از ۳ نوع الگوریتم درخت تصمیم استفاده شد. اعتبارسنجی روی‌داده‌ها با سنجیدن و ارزیابی همه دسته‌بندها انجام شد. الگوریتم درخت آماری کارا سریع در میان مردان و زنان و هم‌چنین درخت رگرسیون و طبقه‌بندی در میان کل جمعیت بهترین عملکرد را داشته و مقدار C-Static و حساسیت برای مدل‌های پیش‌بینی ۷۰٪ و ۷۱٪ برای مردان، ۷۹٪ و ۷۱٪ زنان و ۷۸٪ و ۷۲٪ در جمعیت کل بود.

در مدل‌های درخت تصمیم، فشارخون سیستولیک، فشارخون دیاستولیک، سن، دور کمر شرکت‌کنندگان در هر دو جمعیت زنان و مردان، دور مچ، گلوکز پلاسما در میان زنان و گلوکز پلاسما در میان مردان از عوامل و عواملی استفاده‌شده در مدل ریسک است. در مردان، بیشترین ریسک فشارخون در آن‌هایی که فشارخون سیستولیک بیشتر از ۱۱۵ میلی‌متر جیوه و سنشان از ۳۰ سال به بالا بود، مشاهده شد. در زنان آن‌هایی که

فشارخون سیستولیک بیشتر از ۱۱۴ میلی‌متر جیوه و سن بالای ۲۳ سال داشتند جزء افراد با ریسک بالای پرفشاری خون بودند و در مجموع و کل جمعیت، ریسک بالا در فشارخون سیستولیک بالای ۱۱۴ میلی‌متر جیوه و سن بالای ۳۸ سال مشاهده شد (Ramezankhani *et al.*, 2016).

آقای هیان‌کیم و همکارانش مقاله وضعیت تشخیصی و سن در تشخیص فشارخون بالا در پیروی از توصیه‌های شیوه زندگی را در سال ۲۰۱۹ منتشر کردند. در این مقاله فعالیت فیزیکی منظم، ترک سیگار و مصرف کم الکل از رفتارهای مهم در شیوه زندگی است که می‌تواند باعث تغییر در مدیریت فشارخون شود. این مقاله ارتباط وضعیت تشخیصی و سن در تشخیص فشارخون با رفتارهای سبک زندگی در میان افراد با پرفشاری خون بررسی می‌کند. داده‌ها از بررسی آزمایش‌ها ملی بهداشتی و تغذیه‌ای از سال ۲۰۰۷ تا ۲۰۱۲ (n=5231) به‌دست‌آمده آمده است. مدل‌های رگرسیون لجستیک چندجمله‌ای برای تخمین ریسک مرتبط[1] با انطباق رفتارهای شیوه زندگی استفاده‌شده است.

تشخیص فشارخون بالا با سیگاری بودن فرد همراه بود. بین مدت‌زمان تشخیص از زمان تشخیص و سیگاری بودن ارتباط وجود دارد. نوشیدن بیش‌ازحد ارتباط معکوسی با مدت‌زمان تشخیص دارد. سن بالا در تشخیص با ریسک سیگاری بودن درگذشته ارتباط مستقیم و با نوشیدن بیش‌ازاندازه رابطه منفی دارد. افرادی که ورزش کرده‌اند، حتی اگر کمتر از زمان توصیه‌شده باشد، در مقایسه با افرادی که مشغول انجام فعالیت‌های بدنی نبودند، ازنظر تشخیص سن کمتری دارند و مدت‌زمان تشخیص نیز کوتاه‌تر است.

افراد بیمار باید با ویزیت‌های منظم توسط پزشک بر کنترل فشارخون و اصلاح رفتارهای بهداشتی متمرکز باشند (Kim and Andrade, 2019).

مقاله پیش‌بینی فشارخون بالابر اساس آنتروپومتری ، پارامترهای خون و اسپیرومتری، نیز فشارخون بالا را از عوامل خطر در بیماری‌های قلبی عروقی بیان می‌کند. ازآنجایی‌که ارتباط فشارخون بالا با آنتروپومتری، پارامترهای خون و اسپیرومتری بررسی نشده است، هدف از این مقاله مطالعه شناسایی عوامل ریسک فشارخون بالا در بزرگ‌سالان کره‌ای و بررسی مدل‌های پیش‌بینی فشارخون بالا همراه با آنتروپومتری، پارامترهای خون و اسپیرومتری است.

تجزیه‌وتحلیل رگرسیون لجستیک دودویی برای ارزیابی اهمیت آماری فشارخون بالا انجام شد و مدل‌های پیش‌بینی با استفاده از رگرسیون لجستیک، بیز و درخت تصمیم‌گیری تهیه شد. از میان تمام عوامل خطرساز فشارخون بالا، شاخص توده بدنی به‌عنوان بهترین شاخص در هر دو جنسیت شناخته شد. در مردان odd ratio = 1.429، 95% CI = 1.304-1.462 وزنان odd ratio =1.428، 95% CI = 1.204-1.453 است. در مدل پیش‌بینی فشارخون، مردان AUC برابر با ۰/۷۰۰ وزنان AUC برابر با ۰/۸۴۵ نشان دادند. این مطالعه عوامل خطر مختلفی را برای پیش فشارخون و فشارخون بالا پیشنهاد می‌کند و یافته‌ها می‌تواند به‌عنوان یک ابزار غربالگری در مقیاس بزرگ برای کنترل و مدیریت فشارخون بالا استفاده شود (Heo, 2018).

[1] Relative risk

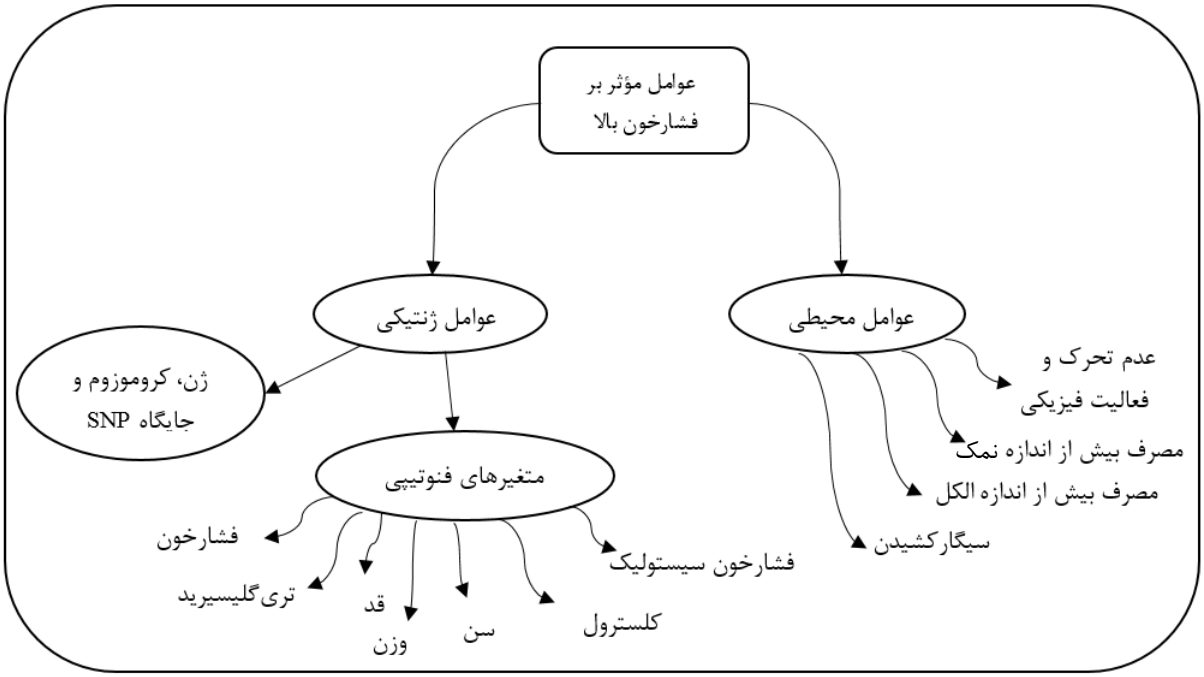
مقاله مدل‌های ریاضی برای اثرات فشارخون بالا و استرس روی کلیه و عدم قطعیت در سال ۲۰۱۷ منتشر شد. در این مقاله با استفاده از رویکردهای قطعی و تصادفی مجموعه‌ای از مدل‌های ریاضی برای بیماری مزمن کلیه ارائه‌شده است. به‌طور خاص، فشارخون بالا و استرس به‌عنوان دلایل اصلی آسیب رساندن و از بین بردن نفرون‌ها درنتیجه کاهش میزان پالایهاسیون گلومرولی در نظر گرفته می‌شود که منجر به صدمات جبران‌ناپذیری به عملکرد کلیه می‌شود و ممکن است منجر به نارسایی کلیه یا حتی از دست دادن کلیه شود. در این مقاله تعادل برای هر دو مدل قطعی و تصادفی تجزیه‌وتحلیل می‌شود. در این راستا، قضیه‌های ریاضی مربوط به ثبات و ثبات تصادفی مدل‌ها ارائه‌شده است. مسئله مهم دیگر، که در کار ارائه‌شده موردبحث قرارگرفته است، مشکل عدم اطمینان برای مدل‌های تصادفی است. برای محاسبه عدم قطعیت مدل‌های تصادفی از شبیه‌سازی‌ها در نرم‌افزار متلب استفاده شد. علاوه بر این، چارچوبی برای پیش‌بینی آینده مدل‌های پیشنهادی در موارد خاص ارائه گردید. سرانجام، مدل‌ها و نتایج نظری در برنامه با استفاده از آن‌ها در داده‌های بالینی واقعی مورد تائید واقع می‌شوند (Waezizadeh *et al.*, 2018). محدوده پرفشاری خون به‌صورت جدول ۲-۷ هست.

جدول ۲-۷ اعداد مربوط به پرفشاری خون (Hengel, Benitah and Wenzel, 2022)

	سیستولی (mm Hg)	**دیاستولی (mm Hg)**
عادی	کمتر از ۱۲۰	کمتر از ۸۰
مرتفع (پرفشاری خون)	۱۲۰-۱۲۹	کمتر از ۸۰
سطح ۱ پرفشاری خون	۱۳۰-۱۳۹	۸۰-۹۰
سطح ۲ پرفشاری خون	۱۴۰ و بالاتر	۹۰ یا بالاتر
نقطه بحرانی پرفشاری خون	بیشتر از ۱۸۰	بیشتر از ۱۲۰

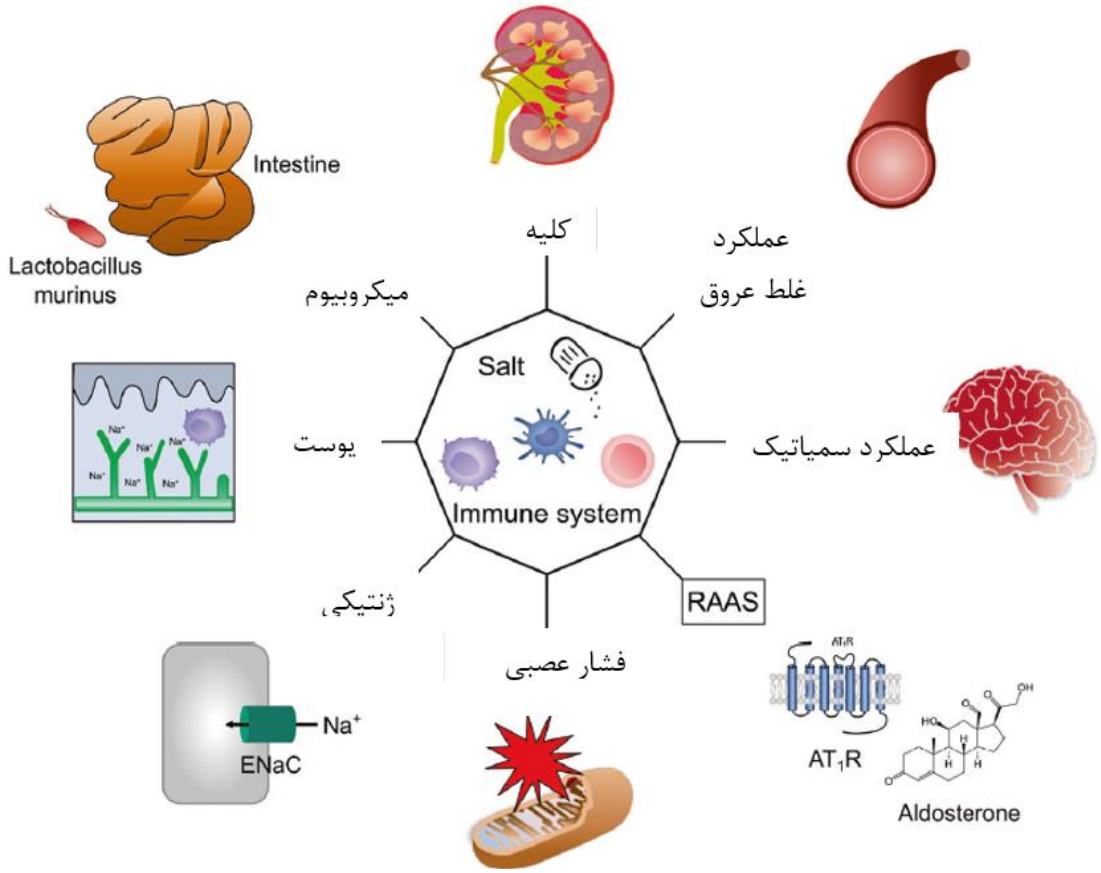
۲-۸-۱- عوامل محیطی پرفشاری خون

یکی دیگر از عوامل مؤثر بر بیماری فشارخون بالا، عوامل محیطی مانند مصرف نمک، مصرف الکل، سیگار کشیدن، عدم تحرک هست؛ که می‌توان با تغییر در این موارد از ابتلا و بروز بیماری فشارخون بالا جلوگیری کرد (Wang *et al.*, 2020). عوامل مؤثر بر بیماری فشارخون بالا را می‌توان به‌طورکلی به شکل ۲-۱۰ نمایش داد.



شکل ۱۰-۲ عوامل مؤثر بر پرفشاری خون (Wang et al., 2020)

یکی از تئوری‌هایی که در بروز بیماری پرفشاری خون مطرح است تئوری موزاییک هست (Hengel, Benitah and Wenzel, 2022). در این تئوری بیان می‌شود که رخ دادن شبکه‌ای متقابل از موارد ایجادکننده فشارخون باعث بروز این بیماری می‌شود که در شکل ۲-۱۱ آمده است.



شکل ۱۱-۲ نسخه اصلاح‌شده نظریه موزاییک فشارخون بالا. نمک و سلول‌های ایمنی در مرکز قرار می‌گیرند، زیرا بر سایر سلول‌ها تأثیر می‌گذارند (Hengel, Benitah and Wenzel, 2022)

در این بررسی در خصوص سن به‌عنوان پارامتر مؤثر بیان‌شده است: افزایش نمک با افزایش سن، اختلال یا از بین رفتن توانایی تعادل کل سدیم بدن در طول زمان است. سدیم فراوان‌ترین کاتیون در مایع خارج سلولی است که از مایعات بینابینی و داخل عروقی تشکیل‌شده است و از بین رفتن این تعادل باعث افزایش نمک و پرفشاری خون هست.

۲-۸-۲- مدل‌های پیش‌بینی کننده بیماری فشارخون

شناخت بیماری فشارخون تاریخچه‌ای بیش از ۵۰ سال دارد. فریزر رابرتز با پیکرینگ و همکارانش (McKusicK *et al.*, 1960) اشاره‌کرده بودند سطح فشارخون یک متغیر پیوسته است و هم‌چنین با رسم نقشه لگاریتمی آن خاطرنشان کردند که چولگی مثبت آن از بین می‌رود. وی هم‌چنین داده‌های بارنز و براون را در مورد فشارخون بستگان درجه‌یک بیماران مبتلابه مسمومیت خونی بارداری مجدداً تجزیه‌وتحلیل کرد و با استفاده از دستگاه امتیازدهی خود برای از بین بردن اثرات سن و جنس نتیجه گرفت که فشارخون در چنین اقوامی بیشتر از جمعیت در کل است.

مطالعات پیکرینگ و همکارانش (McKuSICK, 1960) آن‌ها را به این نتیجه رسانده است که توزیع فراوانی فشارخون پیوسته است و تعیین ژنتیکی فشارخون به‌صورت چندژنی است. در سال ۱۹۵۹ دو تحلیل گزارش داده‌شده است که یکی توسط پلات از منچستر و دیگری توسط موریسن و موریس از لندن هست که نشانگر دونمایی بودن[1] فشارخون هست که با این دیدگاه یک ژن غالب عامل مهمی در فشارخون بالا است.

کورتس و اسپنس (Kurtz and Spence, 1993) بیان می‌کنند که فشارخون یک ویژگی کمی پیچیده‌ای است که توسط عوامل محیطی چندگانه و عواملی ژنتیکی تعیین می‌شود.

هرچند که اشکال ساده‌ای برای فشارخون بالا در نظر گرفته‌شده است، اما فشارخون اساسی بر اساس روش‌های پیچیده وراثت‌پذیری مشخص می‌گردد.

بر اساس این تحقیق امکان جستجوی مناطق کروموزومی که شامل ژن بیماری‌زا فشارخون بالا باشد، فراهم‌شده است. برای مثال، مطالعات پیوند و ارتباط اخیر، احتمال اینکه یک مکان تنظیم‌کننده فشارخون ممکن است در ژن آنژیوتانسینوژن[2] روی کروموزوم یک یا در نزدیکی آن باشد را افزایش داده است.

واعظی زاده و دیگران (Waezizadeh *et al.*, 2018) به بررسی مدل‌های مختلف ریاضی برای بیماری مزمن کلیه که منجر به افزایش فشارخون می‌شود پرداخته‌اند و فشارخون بالا و استرس را به‌عنوان اصلی‌ترین علت صدمه و از بین بردن نفرون درنتیجه کاهش میزان پالایهاسیون گلومرولی بیان کرده‌اند که منجر به آسیب‌های جبران‌ناپذیری در عملکرد کلیه می‌شود و حتی ممکن است درنهایت باعث از دست دادن کلیه شود.

در سال ۲۰۲۰ وو و همکاران بر مبنای ملاک‌های سنتی پیش‌بینی فشارخون تحقیقی را مبتنی بر روش‌های یادگیری ماشین انجام دادند که نشان می‌دهد روش‌های یادگیری ماشین بهترین نتیجه را نسبت به روش‌های رگرسیون و FRS دارد (Wu *et al.*, 2020). همچنین در تحقیق دیگری با استفاده عوامل بالینی سن،

[1] bimodality
[2] angiotensinogen

مصرف سیگار، سابقه خانوادگی، فعالیت فیزیکی و شاخص جرمی بدن[1] دقت ۰.۸۵۴ و با اضافه کردن پارامترهای ژنتیکی دقت ۰.۸۷۱ به‌دست‌آمده است (Niu *et al.*, 2021; Chowdhury *et al.*, 2022). در تحقیقات بر مبنای PRS دقت به‌دست‌آمده بر اساس AUC برابر ۰.۷۱ (با در نظر گرفتن سن، جنسیت و BMI و ۸۵۸ SNP) و ۰.۷۶۴ (با در نظر گرفتن سن، نژاد، سابقه مصرف سیگار و دیابت و ۳،۰۴۳،۶۳۷ SNP) و ۰.۷۵ (با در نظر گرفتن جنسیت، BMI و ۲۱۶۶ SNP) هست (Giontella *et al.*, 2021; Kurniansyah *et al.*, 2022; Maj *et al.*, 2022).

۲-۹- جمع‌بندی

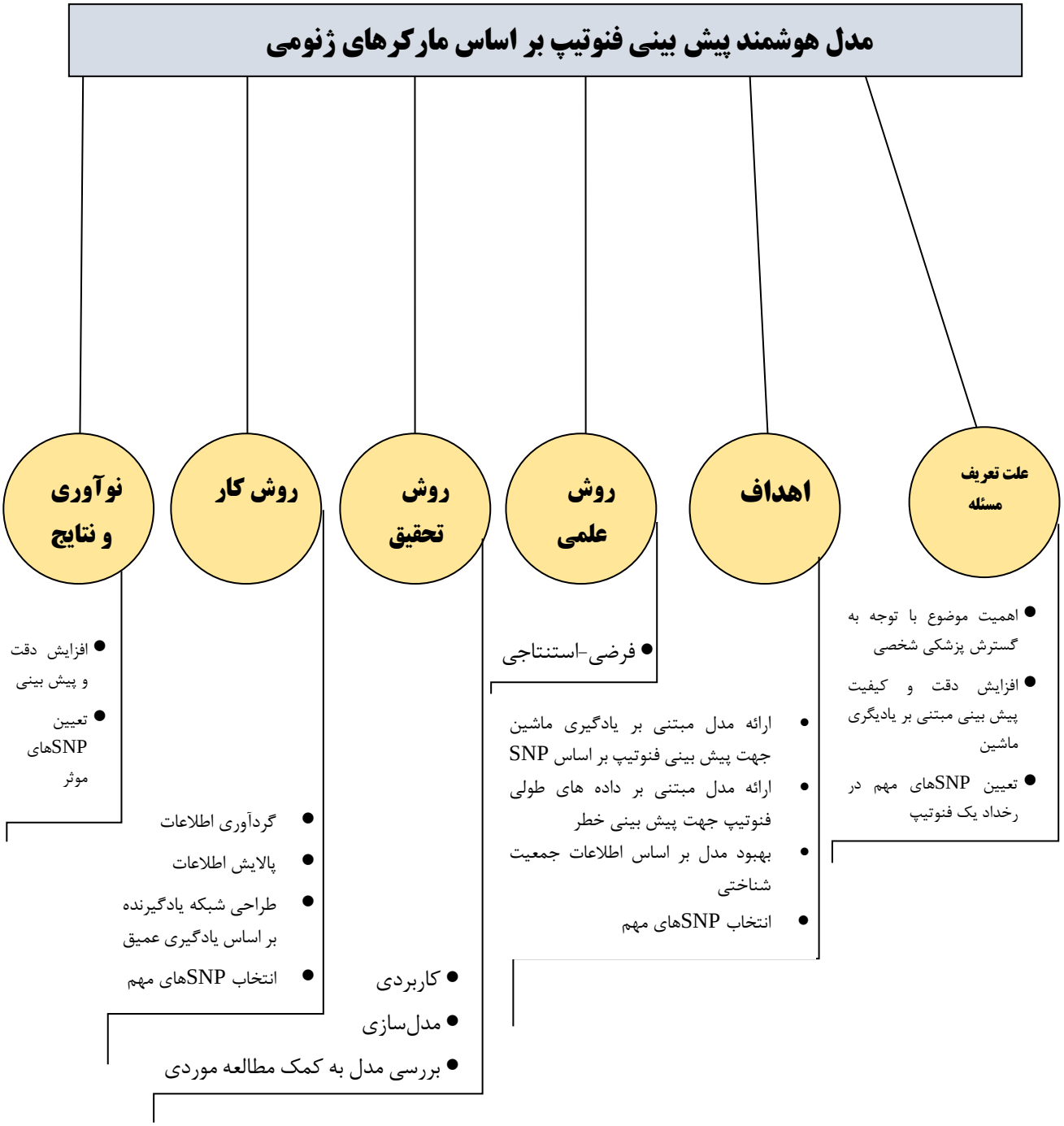
در این فصل بابیان تحقیقات مختلف حوزه ژنوم و فنوتیپ به بررسی مدل‌های متنوع پرداخته شد و بابیان نقاط قوت و ضعف روش‌ها و اهداف آتی که در تحقیقات اشاره‌شده بود، ساختار ادبیات موضوع شکل گرفت.

در ادامه بابیان دقت‌های گزارش‌شده و نقاط ضعف مدل‌های اشاره‌شده، ساختار مدل جدید پیشنهاد گردید که در فصل بعدی به‌طور کامل توضیح داده می‌شود.

[1] Body Mass Index (BMI)

۳- روش‌شناسی تحقیق

در این فصل به روش‌شناسی تحقیق پیش رو می‌پردازیم. بر اساس نوع تحقیق، این تحقیق یک پژوهش کاربردی است، چراکه در این پژوهش بر آنیم تا از طریق مطالعه ادبیات موضوع و حوزه تحقیق، مدلی مبتنی بر یادگیری ماشین بر اساس داده‌های طولی رخ‌نمود و نشانگرهای ژنومی، طراحی و ارائه نماییم. نمونه‌گیری انتخابی در این تحقیق به‌صورت کاملاً انتخابی بر روی‌داده‌های ژنومی و رخ‌نمودهای ذخیره‌شده انجام می‌شود. شمای کلی تحقیق در شکل ۳-۱ آمده است.



شکل ۱-۳ شمای کلی تحقیق (آونگ تحقیق)

بر اساس ادبیات بررسی‌شده در حوزه موردمطالعه، روش‌های آماری و تحلیل فراداده بیشترین کاربرد را نسبت به سایر روش‌ها دارند؛ اما همان‌طور که بررسی شد این روش‌ها دارای معایبی است که بر همین اساس روش‌های مبتنی بر یادگیری ماشین، پیشنهاد و دورنمای تحقیقات در این حوزه هست (Esteva *et al.*, 2019). بر همین اساس در این فصل ابتدا داده‌ها و محدوده‌های تحقیق بیان‌شده و درنهایت بابیان نوآوری‌ها فصل به اتمام می‌رسد.

۳-۱- داده‌های تحقیق

اگرچه بسیاری از ژن‌هایی که در ایجاد بیماری‌ها دخالت دارند، شناسایی‌شده‌اند، اما متخصصان بر این باور هستند که بسیاری از آن‌ها هنوز کشف نشده‌اند. داده‌های ژن‌نمود و رخ‌نمود موردنیاز برای طراحی مدل، می‌تواند از پایگاه داده‌های موجود و یا به‌وسیله شبیه‌سازی به دست آید.

با توجه به مسئله تحقیق، داده‌های تحقیق شامل سه بخش هست:

۱. داده‌های ژنومی
۲. داده‌های جمعیت شناختی و بالینی مانند سن، جنس، قومیت و موارد دیگر
۳. داده‌های طولی رخ‌نمود فشارخون در دوره‌های مختلف اندازه‌گیری شده

برای به دست آوردن داده‌های تحقیق از طرح ژنتیک کاردیومتابولیک تهران، پژوهشکده علوم غدد درون‌ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی استفاده می‌شود و این تحقیق با کد IR.SBMU.ENDOCRINE.REC.1400.008 توسط کمیته ملی اخلاق در تحقیقات زیست پزشکی در اردیبهشت ۱۴۰۰ تائید شده است.

مطالعه قند و لیپید تهران یک مطالعه گسترده در حوزه بهداشتی است که برای نخستین بار در کشور توسط مرکز تحقیقات غدد درون‌ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی از سال ۱۳۷۶ به بررسی عوامل خطرساز بیماری‌های غیر واگیر در شهروندان ساکن شرق کلان‌شهر تهران پرداخته است.

این مطالعه با رویکردی جامعه نگر، بر پایه اطلاعات اولیه مبنی بر شیوع پیشروندِ بیماری‌های غیر واگیر، روی بیش از ۱۵۰۰۰ نفر از جمعیت ساکن در شرق تهران انجام‌شده است و تاکنون در ۶ مقطع زمانی با بازه‌های سه‌ساله از این افراد اطلاعات جمع‌آوری‌شده است. این مطالعه، قدیمی‌ترین، بزرگ‌ترین و معتبرترین مطالعه‌ی آینده‌نگر در ایران است (Kheradmand *et al.*, 2015). جزئیات طراحی این مطالعه به‌تفصیل شرح داده‌شده است (Azizi *et al.*, 2009). جزئیات مطالعه گسترده‌ی ژنومی در مطالعه ژنتیک کاردیومتابولیک تهران که روی بیش از ۱۴۰۰۰ نفر از افراد شرکت‌کننده در طرح قند و لیپید تهران انجام‌شده است (Daneshpour *et al.*, 2017). در مطالعه مذکور، کلیه افرادی که تحت پوشش مطالعه قند و لیپید تهران قرارگرفته‌اند، به‌صورت خانوادگی بررسی شدند. درمجموع، ۳۹۵۹ خانواده شامل ۱۳۸۸۷ نفر در مطالعه گسترده ژنومی بررسی و هر نفر برای بیش از ۷۰۰ هزار نشانگر ژنتیکی تعیین ژن‌نمود شده است. از میان ۱۹۹۳ رخ‌نمود بررسی‌شده در کلیه GWASهایی که تاکنون انجام‌شده‌اند و اطلاعات آن در کاتالوگ GWAS موجود است، ۲۰۹ رخ‌نمود آن

در این جمعیت بررسی‌شده و اطلاعات رخ‌نمودی و ژن‌نمودی آن موجود است. امکان تعریف و یا اندازه‌گیری ۴۰۵ رخ‌نمود دیگر نیز در نمونه‌های سرمی این افراد وجود دارد.

در طی سال‌های گذشته با استفاده از این داده‌ها، تحقیقات گسترده‌ای درزمینهٔ ی تأثیر تغییرات رژیم غذایی، شیوع دیابت و تغییرات میزان لیپید، رخ‌نمودهای مرتبط با چاقی و عوامل ژنتیکی اثرگذار انجام‌شده است که باوجوداین زیرساخت و افزوده شدن داده‌های حاصل از فن‌آوری امیکس (دانشپور 1395 ,.*et al*) می‌توان گام مؤثری در جهت پزشکی شخصی برداشت. تمامی شرکت‌کنندگان یک نمونه خون پایه کافی برای آنالیز پلاسما و DNA ارائه کرده‌اند و برای آنالیزهای مبتنی بر خون و پیگیری طولانی‌مدت رضایت کتبی دادند. ضمناً برای اندازه‌گیری فشارخون، ابتدا شرکت‌کنندگان به مدت ۱۵ دقیقه روی صندلی می‌مانند و پس‌ازآن پزشک متخصص فشارخون را دو بار با استفاده از فشارسنج جیوه‌ای استاندارد اندازه‌گیری می‌کند. حداقل یک‌فاصله ۳۰ ثانیه‌ای بین این دو اندازه‌گیری جداگانه وجود دارد و میانگین دو اندازه‌گیری به‌عنوان فشارخون شرکت‌کننده ثبت می‌شود (Tohidi *et al*., 2012).

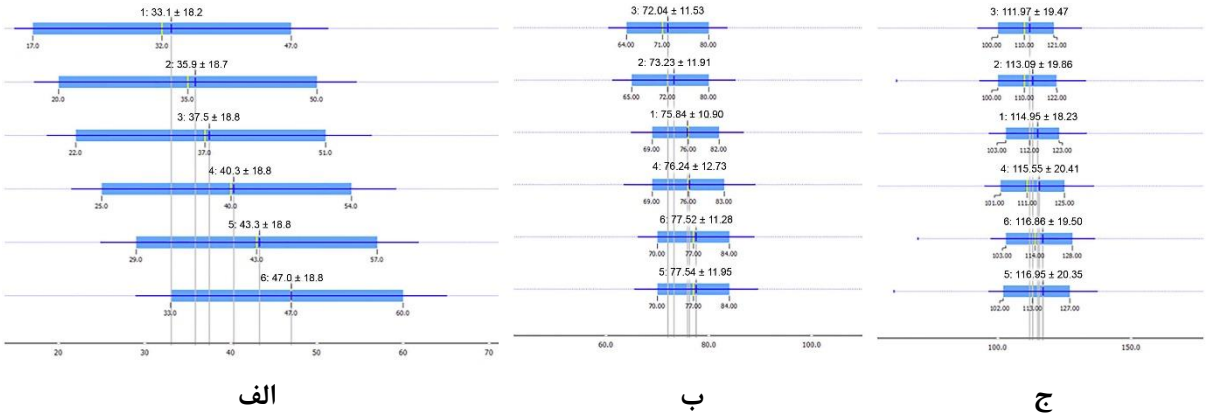
با توجه به محدودیت‌های سخت‌افزاری در پردازش حجم بالای داده‌های کامل ژنوم انسان، در این تحقیق داده‌های ژنومی به‌صورت چیپ[1] شامل ۶۵۲۹۱۹ SNP که برای ۱۱۰۸۸ نفر در دوره‌های مختلف هست، استفاده‌شده است.

لازم به ذکر است معیار ورود داده ثبت اطلاعات ژن‌نمودی و رخ‌نمودی در این طرح هست و معیارهای خروج شامل موارد زیر است:

۱- SNPهایی که دارای بیش از ۵ درصد داده گمشده می‌باشند از ژن‌نمود خارج‌شده‌اند.

۲- افرادی که در ژن‌نمود خود دارای بیش از ۵ درصد داده گمشده هستند از نمونه‌ها خارج می‌شوند.

۳- افرادی که با مصرف داروی خاص تغییرات معنی‌داری در فشارخون داشته باشند (این تغییرات ممکن است باعث کاهش یا افزایش باشد) از نمونه‌ها خارج‌شده‌اند. شکل ۳-۲ توزیع سن و مقادیر DBP و SBP را در فازهای مختلف نشان می‌دهد.

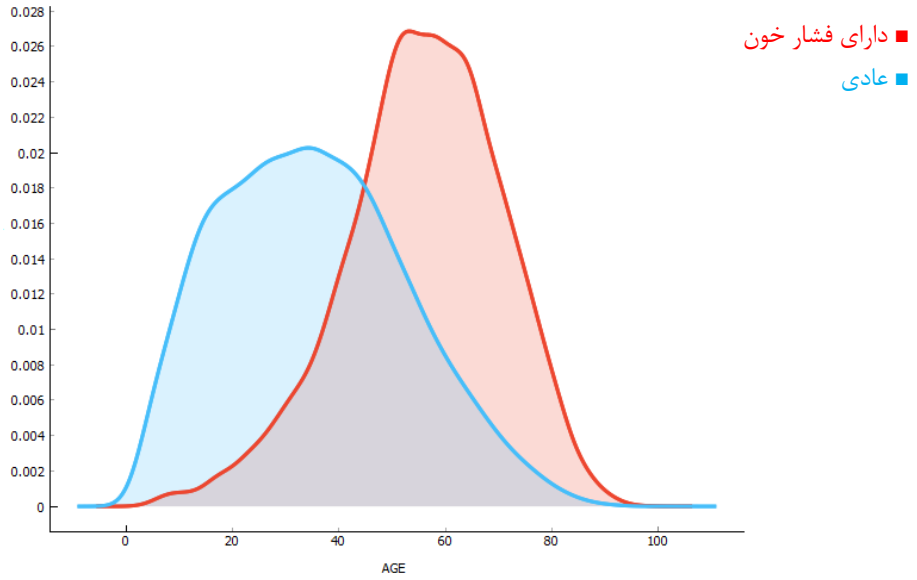


الف ب ج

شکل ۲-۳ الف) نمودار توزیع سن در فاز اول (بالا) تا فاز ششم (پایین) ب) نمودار توزیع DBP ج) نمودار توزیع SBP

[1] Chip

در شکل ۳-۳ توزیع سن با توجه به برچسب فشارخون را نشان می‌دهد که نقطه آستانه برابر ۴۸ سال را نمایش می‌دهد.



شکل ۳-۳ توزیع سن به تفکیک برچسب فشارخون

۳-۱-۱- پروژه ژنوم مرجع ایرانیان[1]

پروژه ژنوم مرجع ایرانیان برای ۸ قومیت ایرانی با کمک مطالعه ژنتیک کاردیومتابولیک تهران و مرکز تحقیقات ژنتیک انسانی کوثر بر روی بیش از ۱۴۰۰۰ نفر انجام‌شده است. این مطالعه با همکاری یکی از معتبرترین مراکز تحقیقات ژنتیک جهان یعنی، کمپانی ایسلندی deCODE Genetics ® در سال ۲۰۱۴ شروع و در سال ۲۰۱۷ با انتقال فناوری به ایران و در اختیار داشتن تمامی داده‌ها در پژوهشکده علوم غدد درون‌ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی به انتها و بهره‌برداری رسیده است (عزیزی, هدایتی و دانشپور, ۱۳۹۶).

نمونه ژنوم از گلبول‌های سفید با روش استاندارد پروتئیناز K استخراج‌شده با تراشه HumanOmniExpress-24-v1-0 توسط شرکت deCODE مطابق با دستگاه Illumina ژن‌نمود شدند (Kolifarhood *et al.*, 2021). داده‌های ژنومی برای هر فرد دارای ۶۰۰ هزار SNP می‌باشند که در پَروَنجا ped ذخیره‌سازی شده است.

۳-۲- ارزیابی مدل

دانشی که در مرحله یادگیری مدل تولید می‌شود، می‌بایست در مرحله ارزیابی مورد تحلیل قرار گیرد تا بتوان ارزش آن را تعیین نمود و درنتیجه، کارایی الگوریتم یادگیرنده مدل را نیز مشخص کرد. این معیارها را می‌توان هم برای دادگان‌های آموزشی در مرحله یادگیری و هم برای مجموعه رکوردهای آزمایشی در مرحله ارزیابی محاسبه نمود (Stehman, 1997).

[1] Iranian Reference Genome Project

در بحث دسته‌بندی[1] یک دادگان[2] با استفاده از روش‌های رده‌بندی، هدف دستیابی به بالاترین دقت ممکن در رده‌بندی و تشخیص دسته‌ها است. در برخی از مسائل، تشخیص صحیح نمونه‌های مربوط به یکی از دسته‌ها برای ما اهمیت بیشتری دارد. به‌عنوان‌مثال، تحقیقی را در نظر بگیرید که در آن، هدف شناسایی افراد مبتلابه یک نوع خاص از یک بیماری خطرناک است. فرض کنید برای افرادی که مبتلابه این بیماری هستند، خطر مرگ وجود دارد و جهت رفع این خطر، نیاز به دریافت نوعی داروی خاص دارند. در این شرایط، تشخیص درست بیماران دارای اهمیت بسیار زیادی است.

به این معنا که خطا در تشخیص افراد سالم قابل‌چشم‌پوشی است اما برای شناسایی افراد بیمار نمی‌توان این احتمال را به جان خرید. به‌عبارت‌دیگر، انتظار ما تشخیص تمام افراد بیمار است، بدون جا انداختن، حتی اگر فرد سالمی به‌اشتباه جز افراد بیمار رده‌بندی شود. در چنین مواقعی، که دقت تشخیص یک دسته در مقایسه با دقت تشخیص کلی، اهمیت بیشتری دارد، مفهوم ماتریس درهم‌ریختگی[3] استفاده می‌شود (Powers and Ailab, 2011).

۱-۲-۳-ماتریس درهم‌ریختگی

بر اساس مثالی که پیش‌تر بیان شد، فرض کنید تعلق به دسته افراد بیمار را مثبت بودن[4] و عدم تعلق به این دسته را منفی بودن[5] در نظر بگیریم. هر نمونه یا فردی در واقعیت، متعلق به یکی از برچسب‌های مثبت یا منفی است و از سوی دیگر، از هر الگوریتمی که برای رده‌بندی داده‌ها استفاده شود، درنهایت هر نمونه عضو یکی از این دودسته[6]، رده‌بندی خواهد شد. بنابراین برای هر نمونه داده، یکی از چهارحالتی که در ادامه بیان‌شده، ممکن است اتفاق بیفتد.

نمونه عضو دسته مثبت باشد و عضو همین کلاس تشخیص داده شود: مثبت صحیح یا True Positive

نمونه عضو کلاس مثبت باشد و عضو کلاس منفی تشخیص داده شود: منفی کاذب یا False Negative

نمونه عضو کلاس منفی باشد و عضو همین کلاس تشخیص داده شود: منفی صحیح یا True Negative

و درنهایت، نمونه عضو کلاس منفی باشد و عضو کلاس مثبت تشخیص داده شود: مثبت کاذب یا False Positive

پس از اجرای الگوریتم رده‌بندی، با توجه به توضیحات و تعاریف ذکرشده، می‌توان عملکرد یک رده‌بند را به کمک جدولی به شکل زیر بررسی کرد.

[1] Classification
[2] Data Set
[3] Confusion Matrix
[4] Positive
[5] Negative
[6] Class

جدول ۱-۳ ماتریس درهم‌ریختگی

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

این جدول را اصطلاحاً ماتریس درهم‌ریختگی می‌گویند. جدول یا ماتریس درهم‌ریختگی، نتایج حاصل از رده‌بندی را بر اساس اطلاعات واقعی موجود، نمایش می‌دهد. حال بر اساس این مقادیر می‌توان معیارهای مختلف ارزیابی رده‌بند و اندازه‌گیری دقت را تعریف کرد. پارامتر دقت[1]، متداول‌ترین، اساسی‌ترین و ساده‌ترین معیار اندازه‌گیری کیفیت یک رده‌بند است و عبارت است از میزان تشخیص صحیح رده‌بند درمجموع دودسته. این پارامتر درواقع نشان‌گر میزان الگوهایی است که درست تشخیص داده‌شده‌اند و بر اساس ماتریس ارائه‌شده در بالا، با رابطه زیر تعریف می‌شود.

$$Accuracy = (TP+TN) / (TP+FN+FP+TN)$$ رابطه ۱-۳

البته، پارامتر دقت معمولاً به‌صورت درصد بیان می‌شود. اما پارامترهای دیگری نیز علاوه بر معیار دقت وجود دارند که می‌توان به‌سادگی از این ماتریس استخراج کرد. یکی از متداول‌ترین آن‌ها، معیار حساسیت[2] است که آن را نرخ پاسخ‌های مثبت درست[3] نیز می‌گویند. حساسیت به معنی نسبتی از موارد مثبت است که آزمایش آن‌ها را به‌درستی به‌عنوان نمونه مثبت تشخیص داده است. این پارامتر به‌صورت زیر محاسبه می‌شود.

$$Sensitivity (TPR) =TP / (TP+FN)$$ رابطه ۲-۳

درواقع، «حساسیت» همان معیار بحث شده در مورد مثال بالا است. معیاری که مشخص می‌کند رده‌بند، به چه اندازه در تشخیص تمام افراد مبتلابه بیماری موفق بوده است. همان‌گونه که از رابطه فوق مشخص است، تعداد افراد سالمی که توسط رده‌بند به‌اشتباه به‌عنوان فرد بیمار تشخیص داده‌شده‌اند، هیچ تأثیری در محاسبه این پارامتر ندارد و درواقع زمانی که پژوهش‌گر از این پارامتر به‌عنوان پارامتر ارزیابی برای رده‌بند خود استفاده می‌کند، هدفش دستیابی به نهایت دقت در تشخیص نمونه‌های کلاس مثبت است.

در نقطه مقابل این پارامتر، ممکن است در مواقعی دقت تشخیص کلاس منفی حائز اهمیت باشد. از متداول‌ترین پارامترها که معمولاً در کنار حساسیت بررسی می‌شود، پارامتر خاصیت[4] است که به آن نرخ پاسخ‌های منفی درست[5] نیز می‌گویند. خاصیت به معنی نسبتی از موارد منفی است که آزمایش آن‌ها را به‌درستی به‌عنوان نمونه منفی تشخیص داده است. این پارامتر به‌صورت زیر محاسبه می‌شود.

$$Specificity (TNR) = TN / (TN+FP)$$ رابطه ۳-۳

[1] Accuracy
[2] Sensitivity
[3] True Positive Rate
[4] Specificity
[5] True Negative Rate

این دو پارامتر (حساسیت و خاصیت) نیز مشابه معیار دقت، معمولاً به‌صورت درصد بیان می‌شوند. واضح است که پیش‌بینی عالی، پیش‌بینی است که مقادیر حساسیت و ویژگی مربوط به آن، هر دو صد درصد باشند؛ اما احتمال وقوع این اتفاق در واقعیت بسیار کم است و همیشه یک حداقل خطایی وجود دارد. پارامترهای حساسیت و خاصیت، بنا بر ماهیتی که دارند همواره در رقابت با یکدیگر هستند. یعنی افزایش یکی با کاهش دیگری همراه است و برعکس. همین وضعیت منجر به تولید ابزاری دیگر برای ارزیابی کیفیت رده‌بندها شده است. منحنی مشخصه عملکرد سیستم[1] (ROC)، عبارت است از منحنی که ارتباط بین دو پارامتر حساسیت و خاصیت را بیان می‌کند (Fawcett, 2006).

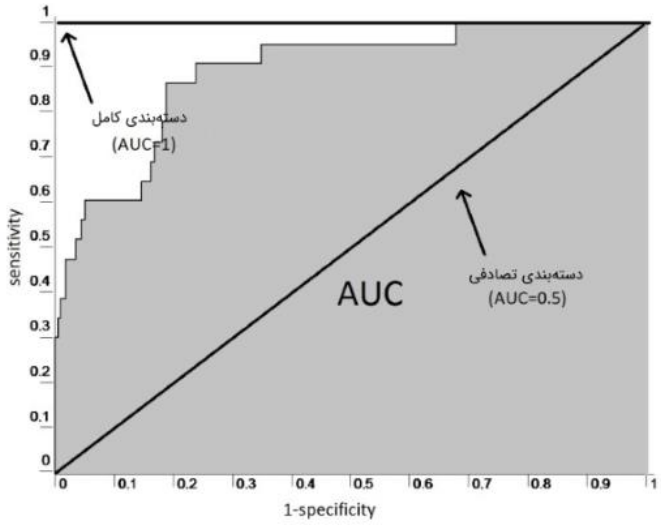
بسیاری از رده‌بندها همانند روش‌های مبتنی بر درخت تصمیم و یا روش‌های مبتنی بر قانون، به‌گونه‌ای طراحی‌شده‌اند که تنها یک خروجی دودویی (مبنی بر تعلق ورودی به یکی از دودسته ممکن) تولید می‌کنند. به این نوع رده‌بندها که تنها یک خروجی مشخص برای هر ورودی تولید می‌کنند، رده‌بندهای گسسته گفته می‌شود که این رده‌بندها تنها یک نقطه در فضای ROC تولید می‌کنند.

به‌طور مشابه رده‌بندهای دیگری نظیر رده‌بندهای مبتنی بر روش بیز و یا شبکه‌های عصبی نیز وجود دارند که یک احتمال و یا امتیاز برای هر ورودی تولید می‌کنند، که این عدد بیانگر درجه تعلق ورودی به یکی از دودسته موجود هست. این رده‌بندها پیوسته نامیده می‌شوند و به دلیل خروجی خاص این رده‌بندها یک آستانه جهت تعیین خروجی نهایی در نظر گرفته می‌شود.

چنانکه در شکل ۳-۴ مشاهده می‌کنید، محور عمودی این نمودار نشان‌دهنده نرخ مثبت صحیح، و محور افقی نشان‌دهنده مقدار نرخ مثبت غلط است. نتایج مختلف رده‌بندی نشانگر نقاط مختلف بر روی این نمودار هستند و درنهایت یک منحنی را تشکیل می‌دهند. با توجه به شکل زیر، در بهترین حالت و با فرض رده‌بندی صد درصد صحیح در هر دودسته، نقطه مربوطه عبارت است از نقطه گوشه بالای سمت چپ، یعنی نقطه (۰،۱) و نیز با فرض رده‌بندی به‌صورت تصادفی، نقطه متناظر در منحنی، یکی از نقاط موجود روی خط واصل نقطه (۰،۰) و نقطه (۱،۱) خواهد بود. در واقعیت، منحنی حاصل از یک رده‌بندی، منحنی بین این دو حالت است.

یک طبقه بند کامل دارای AUC برابر با ۱ خواهد و یک رده‌بندی ضعیف دارای AUC ۰.۵ است. یک رده‌بندی کننده خوب دارای AUC بیشتر از ۰.۵ خواهد بود و با رسیدن به ۱ عالی در نظر گرفته می‌شود (Martinez-Ríos *et al.*, 2021).

[1] Receiver Operating Characteristic: ROC



شکل ۴-۳ منحنی ROC و نمایش AUC

مساحت زیر نمودار[1] (AUC) که در شکل ۳-۸ آمده است، به‌عنوان یک معیار برای ارزیابی عملکرد رده‌بند مورداستفاده قرار می‌گیرد. با توجه به توضیحاتی که پیش‌تر ارائه شد، بدیهی است که در حالت ایدئال، مساحت زیر منحنی برابر با بیشترین مقدار خود، یعنی یک است. بنابراین، هر چه مساحت زیر نمودار به عدد یک نزدیک‌تر باشد، به معنای بهتر بودن عملکرد رده‌بند است. علاوه بر دو پارامتر حساسیت و خاصیت، پارامترهای دیگری هم از ماتریس درهم‌ریختگی استخراج می‌شوند که هر یک بیان‌کننده مفهومی هستند و کاربردهای متفاوتی دارند.

ازجمله آن‌ها، دو پارامتر مقدار پیش‌بینی‌شده مثبت[2] و مقدار پیش‌بینی‌شده منفی[3] هستند، که برای بیان «نسبت پاسخ‌های درست در هر دسته» استفاده می‌شوند. ارزش اخباری مثبت، بیان‌کننده این است که چند درصد از الگوهایی که رده‌بند آن‌ها را مثبت تشخیص داده، در واقعیت هم مثبت هستند و به همین ترتیب، ارزش اخباری منفی نشان می‌دهد که چند درصد از نمونه‌هایی که عضو دسته منفی تشخیص داده‌شده‌اند، در واقعیت هم عضو همین دسته هستند. این دو پارامتر نیز به‌سادگی از روی ماتریس درهم‌ریختگی به‌صورت رابطه زیر است.

$$NPV=TN / (TN+FN)، PPV=TP / (TP+FP)$$ رابطه ۳-۴

پارامتر مهم دیگری به نام معیار اف[4] وجود دارد که برای ارزیابی عملکرد رده‌بندها بسیار مورداستفاده قرار می‌گیرد و از ترکیب دو پارامتر حساسیت و ارزش اخباری مثبت حاصل می‌شود. با این توضیح که پارامتر ارزش اخباری مثبت را اصطلاحاً دقت[5] و حساسیت را اصطلاحاً بازخوانی[6] می‌نامند، «معیار اف» به‌صورت رابطه زیر است.

$$F\text{-measure}= 2 * (Recall * Precision) / (Recall + Precision)$$ رابطه ۳-۵

[1] Area Under Curve
[2] Positive Prediction Value
[3] Negative Predictive Values
[4] F-Measure
[5] Precision
[6] Recall

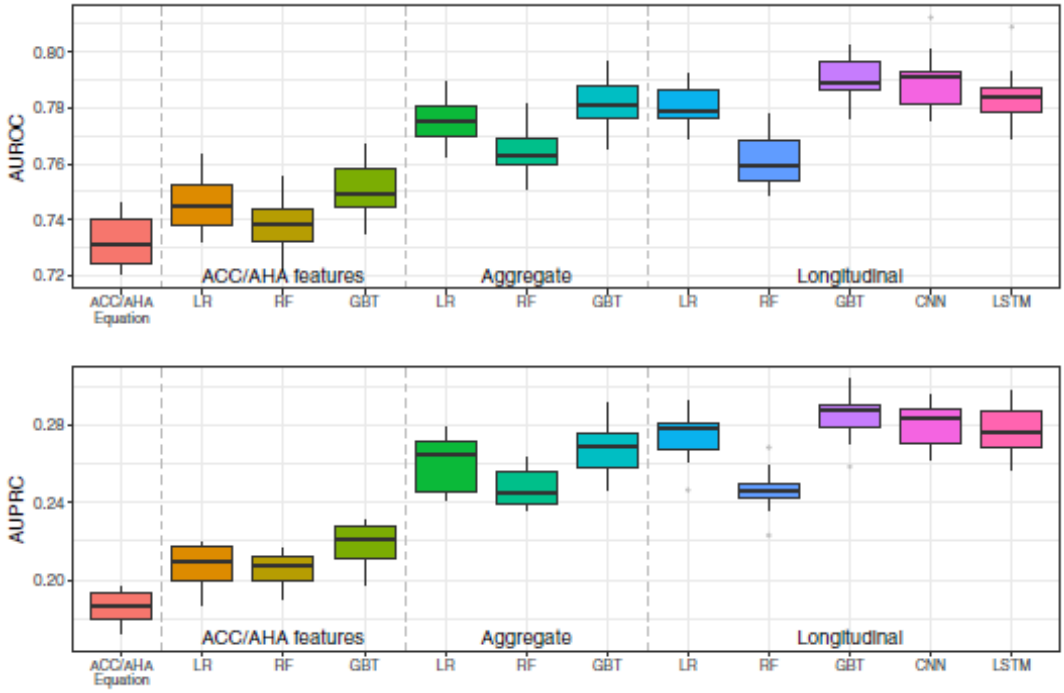
ماتریس درهم‌ریختگی، باوجود منطق و ساختار ساده‌ای که دارد، مفهومی قدرتمند است که در انواع تحقیقات، می‌تواند به‌تنهایی اطلاعاتی جامع از نحوه عملکرد رده‌بند ارائه کند.

یکی از دیگر از مشکلاتی که در ارزیابی مدل مؤثر است بیش برازش[1] داده‌ها بر مدل هست. برای حل این مشکل از رویه قاعده‌مند سازی[2] استفاده می‌کنند (Zhu *et al.*, 2018). هدف از رویه‌ی قاعده‌مند سازی جلوگیری از بیش برازش به داده‌ها توسط مدل است و درنتیجه با مشکل واریانس بالا طرف است. جدول ۳-۲ خلاصه‌ای از انواع روش‌های متداول قاعده‌مند سازی را ارائه می‌دهد.

جدول ۲-۳ روش‌های قاعده‌مند سازی

Elastic Net	Ridge	LASSO
بین انتخاب متغیر و ضرایب کوچک مصالحه می‌کند	ضرایب را کوچک‌تر می‌کند	ضرایب را تا ۰ کاهش می‌دهد برای انتخاب متغیر مناسب است

یکی دیگر از معیارهایی که در ادبیات موضوع مطرح‌شده است معیار AUPRC هست که سطح زیر منحنی نمودار دقت و بازخوانی است که در داده‌هایی که ازنظر برچسب نامتعادل هستند دارایی کارایی بهتری هست. در شکل ۳-۵ نمودار مقایسه پیش‌بینی بیماری قلبی مبتنی بر داده‌های ژنومی در دو معیار AUROC و AUPRC آمده است (Zhao *et al.*, 2018).



شکل ۵-۳ مقایسه AUROC و AURPC

۲-۲-۳-روش‌های بخش‌بندی داده جهت ارزیابی

از روش‌های ارزیابی الگوریتم‌های رده‌بندی (که در این الگوریتم روال کاری بدین‌صورت است که مدل رده‌بندی توسط دادگان آموزشی ساخته‌شده و به‌وسیله دادگان آزمایشی مورد ارزیابی قرار می‌گیرد) می‌توان به روش

[1] Overfit
[2] Regularization

بسط یافته[1] اشاره کرد که در این روش چگونگی نسبت تقسیم دادگان‌ها (به دو دادگان آموزشی و دادگان آزمایشی) بستگی به تشخیص تحلیلگر دارد که معمولاً دوسوم برای آموزش و یک‌سوم برای ارزیابی در نظر گرفته می‌شود. مهم‌ترین مزیت این روش سادگی و سرعت‌بالای عملیات ارزیابی است ولیکن روش بسط یافته معایب زیادی دارد ازجمله اینکه دادگان‌های آموزشی و آزمایشی به یکدیگر وابسته خواهند شد، درواقع بخشی از دادگان اولیه که برای آزمایش جدا می‌شود، شانسی برای حضور یافتن در مرحله آموزش ندارد و به‌طور مشابه در صورت انتخاب یک رکورد برای آموزش دیگر شانسی برای استفاده از این رکورد برای ارزیابی مدل ساخته‌شده وجود نخواهد داشت. همچنین مدل ساخته‌شده بستگی فراوانی به چگونگی تقسیم دادگان اولیه به دادگان‌های آموزشی و آزمایشی دارد. چنانچه روش بسط یافته را چندین بار اجرا کنیم و از نتایج حاصل میانگین‌گیری کنیم از روشی موسوم به زیر نمونه‌گیری تصادفی[2] استفاده نموده‌ایم. که مهم‌ترین عیب این روش نیز عدم کنترل بر روی تعداد دفعاتی که یک رکورد به‌عنوان نمونه آموزشی و یا نمونه آزمایشی مورداستفاده قرار می‌گیرد، است. به‌بیان‌دیگر در این روش ممکن است برخی رکوردها بیش از سایرین برای یادگیری و یا ارزیابی مورداستفاده قرار گیرند.

چنانچه درروش زیر نمونه‌گیری تصادفی به شکل هوشمندانه‌تری عمل کنیم به‌صورت که هرکدام از رکوردها به تعداد مساوی برای یادگیری و تنها یک‌بار برای ارزیابی استفاده شوند، روش مزبور در متون علمی بانام اعتبارسنجی متقابل[3] شناخته می‌شود.

همچنین درروش جامع اعتبارسنجی متقابل چند تا[4] کل دادگان‌ها به k قسمت مساوی تقسیم می‌شوند. از k-1قسمت به‌عنوان دادگان‌های آموزشی استفاده می‌شود و بر اساس آن مدل ساخته می‌شود و با یک قسمت باقی‌مانده عملیات ارزیابی انجام می‌شود. فرآیند مزبور به تعداد k مرتبه تکرار خواهد شد، به‌گونه‌ای که از هرکدام از k قسمت تنها یک‌بار برای ارزیابی استفاده‌شده و در هر مرتبه یک دقت برای مدل ساخته‌شده، محاسبه می‌شود. در این روش ارزیابی دقت نهایی رده‌بند برابر با میانگین k دقت محاسبه‌شده خواهد بود. معمول‌ترین مقداری که در متون علمی برای k در نظر گرفته می‌شود برابر با ۱۰ هست. بدیهی است هر چه مقدار k بزرگ‌تر شود، دقت محاسبه‌شده برای رده‌بند قابل‌اعتمادتر بوده و دانش حاصل‌شده جامع‌تر خواهد بود و البته افزایش زمان ارزیابی رده‌بند نیز مهم‌ترین مشکل آن هست. حداکثر مقدار k برابر با تعداد رکوردهای دادگان اولیه است که این روش ارزیابی بانام رها کردن یکی[5] شناخته می‌شود.

در روش‌هایی که تاکنون به آن اشاره‌شده، فرض بر آن است که عملیات انتخاب نمونه‌های آموزشی بدون جایگذاری صورت می‌گیرد. به‌بیان‌دیگر یک رکورد تنها یک‌بار در یک فرآیند آموزشی موردتوجه واقع می‌شود.

[1] Holdout
[2] Random Sub-sampling
[3] Cross Validation
[4] k-Fold Cross Validation
[5] Leaving One Out

چنانچه هر رکورد در صورت انتخاب شدن برای شرکت در عملیات یادگیری مدل بتواند مجدداً برای یادگیری مورداستفاده قرار گیرد روش مزبور بانام خود راه‌انداز[1] شناخته می‌شود.

۳-۳- کاهش خطا در یادگیری ماشین در بیوانفورماتیک

در این بخش موضوعات مختلف و چالش‌های مربوط به روش‌های یادگیری ماشین در بیوانفورماتیک موردبحث قرارگرفته است. حضور داده‌های تکراری ، مسئله مهمی در بیوانفورماتیک است.

با ظهور اینترنت ، داده‌ها به‌صورت آزاد در دسترس عموم است که تشخیص خطاها و اندازه‌گیری کیفیت داده‌ها را دشوار می‌کند. منبع داده‌ها، در زمان استفاده از آن‌ها، توسط محققان بیوانفورماتیک در نظر گرفته نشده است و اگر داده‌های موجود در منبع کثیف باشند، دقت رده‌بندی کاهش می‌یابد.

داده‌های بیولوژیکی کثیف می‌تواند به دلیل عوامل زیادی ازجمله موارد زیر باشد:

۱- خطاهای حین آزمایش

۲- تفسیر غلط داده‌ها توسط زیست شناسان

۳- تایپ اشتباه مربوط به خطای انسانی

۴- روش‌های غیراستاندارد در آزمایش‌ها استفاده شود.

الگوریتم‌های یادگیری، هنگام یادگیری در یک پایگاه داده‌های بیولوژیکی کثیف، باید با روندی پایدار داشته باشد. روش‌های ML (یادگیری ماشین) با تطابق‌پذیری خود باید بتوانند تصمیماتی بهینه بگیرند تا از بیش برازش داده‌ها جلوگیری کنند. ضمناً ازآنجاکه پایگاه داده‌های بیولوژیکی روزانه آپدیت می‌شوند، الگوریتم‌های ML باید مدت‌زمان یادگیری و آموزش کوتاهی داشته باشند.

در حین تجزیه‌وتحلیل بیولوژیکی، همکاری با متخصصان این حوزه (در این مورد زیست شناسان) ضروری است که برای حفظ کیفیت داده‌ها، داده‌های موجود در پایگاه‌های داده باید دائماً مورد بازبینی قرار گیرند.

انتخاب ویژگی داده‌ها مهم‌ترین عاملی است که بر عملکرد پیش‌بینی کننده مدل یادگیری ماشین تأثیر می‌گذارد. انتخاب ویژگی داده در مرحله آموزش یادگیری ماشین باهدف کاهش ابعاد داده‌ها، حذف نویز داده‌ها و داده‌های بی‌ربط و درنتیجه حفظ مفیدترین سیگنال‌ها از دادگان‌ها انجام می‌شود. فرآیند انتخاب ویژگی داده‌ها می‌تواند به‌طور گسترده با استفاده از روش‌های پالایش آن‌ها، ماژول‌های تعبیه‌شده یا روش‌های پوشانه[2] انجام شود. انتخاب این روش‌ها به ویژگی‌های اصلی داده و معیارهای مدل پیش‌بینی بستگی دارد. در مورد بیماری‌های پیچیده چندژنی، در حال حاضر در میان داده‌های ژن‌نمودی، SNPها دارای مفیدترین اطلاعات محسوب می‌شوند. فرض بر این است که SNPهایی که برای استفاده در مدل‌های پیش‌بینی کننده انتخاب می‌شوند با مکان‌هایی همراهی دارند که در اتیولوژی بیماری زمینه‌ای نقش مهمی ایفا می‌کنند.

[1] Bootstrap
[2] Wrapper

بااین‌وجود، ممکن است چگونگی ارتباط SNP با بیماری قابل درک نباشد. معمولاً در مرحله اول ساخت مدل، گونه‌های موجود در داده‌های ژن‌نمودی بر اساس آستانه‌های p-value مطالعه GWA پالایش‌شده و دسته‌بندی می‌گردند. در الگوریتم ساخت و عملکرد مدل، برای انتخاب SNPها و اثرات متقابل آن‌ها روش‌هایی دیده‌شده است که بدین ترتیب تنها به SNP های دارای اطلاعات مفید اجازه می‌دهد که در مدل پیش‌بینی کننده شرکت کنند. پوشش‌دهنده‌ها نیز همان هدفی رادارند که روش‌های تعبیه‌شده دنبال می‌کنند. بااین‌حال، پوشش‌دهنده‌ها ماژول‌های انتخاب SNP مستقل هستند که قبل از فرآیند ساخت مدل اجراشده‌اند.

بیش برازش[1] پدیده‌ای است که در آن مدل‌ها به حدی یک دادگان را پوشش می‌دهند که نمی‌توان آن‌ها را به داده‌های دیگر تعمیم داد. احتمال به وجود آمدن مدل‌های با بیش برازش را می‌توان با تنظیم آن‌ها کاهش داد، این فرآیندی است که قدرت تعمیم پیش‌بینی مدل‌های یادگیری ماشین را به حداکثر می‌رساند.

۳-٤- محدوده‌های تحقیق و روش اجرا

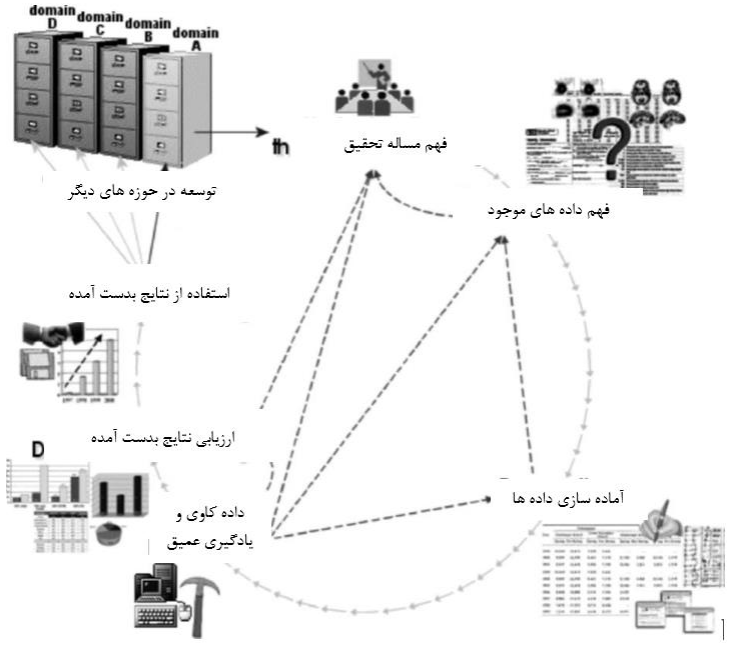
اطلاعات ژنوم انسان شامل بخش‌های مختلفی است. پردازش این بخش‌ها بر اساس نوع داده متفاوت می‌باشند. در این تحقیق اطلاعات تغییرات یک SNP از ژنوم هر انسان بررسی می‌شود و متغیر مستقل اندازه‌گیری شده تغییرات SNP ژنوم انسان نسبت به ژن مرجع هست. همچنین برای بررسی بهتر نقش عوامل محیطی اطلاعات وضعیتی انسان مانند جنس، سن، قومیت و موارد دیگر به‌عنوان متغیرهای مستقل وارد مدل می‌شوند (Zhao *et al.*, 2018).

متغیر وابسته در این تحقیق داده‌های طولی فشارخون (به‌صورت کمی) و خطر پرفشاری خون (به‌صورت برچسب ۰و۱) طی سال‌های اندازه‌گیری شده در تحقیق قند و لیپید تهران هست.

منظور از داده‌های طولی، داده‌های ثبت شده برای هر فرد در ۶ دوره هست که ماشین یادگیرنده بر مبنای دوره‌های گذشته به پیش‌بینی دوره‌های بعدی می‌پردازد.

با توجه به اینکه این تحقیق در حوزه داده‌کاوی در اطلاعات پزشکی هست، از روش گفته‌شده در (Cios and William Moore, 2002) استفاده‌شده که در تصویر ۳-۶ آمده است.

[1] Overfitting



شکل ۶-۳ نحوه اجرای تحقیق

در این روش ابتدا اطلاعات خام از منابع داده متناسب با متغیرهای مستقل و وابسته استخراج می‌شوند. برای این کار لازم است که اطلاعات و نوع مقادیر آن‌ها به‌خوبی شناخته شوند.

در مرحله بعدی آماده‌سازی داده‌ها انجام می‌گردد. در این مرحله داده‌هایی که حذف‌شده[1] و یا ناشناخته[2] می‌باشند باید معین شوند که روش‌های مختلفی مانند حذف کردن، جایگزینی با مقدار میانگین و یا پیش‌بینی مبتنی بر داده‌های هم‌گروه استفاده می‌گردد. ضمناً برای شروع مدل‌سازی بهتر است از روش‌های گسسته سازی[3] که باعث کاهش احتمال بیش برازش نیز هست استفاده کرد.

پس از آماده‌سازی داده‌ها، مدل‌سازی بر اساس ابزارهای مختلفی انجام می‌شود. با توجه به حجم داده و نوع الگوریتم استفاده‌شده از نرم‌افزار Orange جهت پیاده‌سازی مدل‌های اولیه استفاده می‌شود و برای پیاده‌سازی مدل نهایی از زبان برنامه‌نویسی php تحت Linux استفاده می‌شود. البته جهت نمایش اولیه داده‌ها و محاسباتی نظیر توزیع متغیرها، از نرم‌افزارهای Excel و SPSS استفاده می‌شود.

با توجه به مدل‌هایی که در حوزه ادبیات موضوع بررسی‌شده است، زیرساخت پیشنهادی جهت اجرای مدل نهایی، سروری شامل ۸ پردازشگر 3.2Ghz به همراه 64GB رم و 300GB هارددیسک هست و زمان تخمینی جهت اجرای نهایی یک هفته هست.

با اجرای مدل و استفاده از روش‌های ارزیابی معرفی‌شده در ادبیات، خروجی‌های مدل نهایی حاصل می‌شود که این خروجی‌ها با توجه به اهداف تحقیق شامل موارد زیر است:

۱- مدل پیش‌بینی کننده پرفشاری خون بر اساس SNPهای در نظر گرفته‌شده

۲- مدل پیش‌بینی کننده پرفشاری خون بر اساس SNP و داده‌های طولی

[1] Missing
[2] Unknown
[3] Binding

۳- تعیین SNPهای مهم و مؤثر بر پرفشاری خون و تعیین دقت پیش‌بینی مبتنی بر مجموعه پیشنهادی در خصوص مدل‌ها نتایج شامل جداول درهم‌ریختگی بر اساس برچسب واقعی نمونه هست.

۳-۵- فرایندهای مدل‌سازی

با توجه به اینکه در این تحقیق مدل‌های متنوعی ساخته و پیاده‌سازی شده است، فرایندهای مدل‌سازی در ادامه آمده است.

۳-۵-۱-فرایند مدل‌سازی بر مبنای SNP کاندید

مطالعات GWAS شامل تحقیقاتی است که حداقل ۱۰۰۰ نفر در آن مشارکت داشته باشند و با روش‌های علمی - آماری، SNP های مؤثر یک در رخ‌نمود شناسایی‌شده‌اند. به همین دلیل برای شناسایی بهتر SNPهای مؤثر تمامی SNPهای گزارش‌شده در تحقیقات فشارخون را که به‌عنوان SNP کاندیدا وارد مدل می‌کنیم.

پروفایل رخ‌نمود نیز مبتنی بر سن، جنسیت، فشارخون که متناسب با هر فرد تعداد یک الی شش دوره هست، وارد مدل می‌گردد (داده جنسیت ثابت می‌ماند و داده سن و برچسب‌های فشارخون مبتنی بر دوره تعیین می‌شود). برای برچسب‌زنی داده‌ها، از محدوده‌های زیر برای فشارخون استفاده می‌کنیم (Basile *et al.*, 2019).

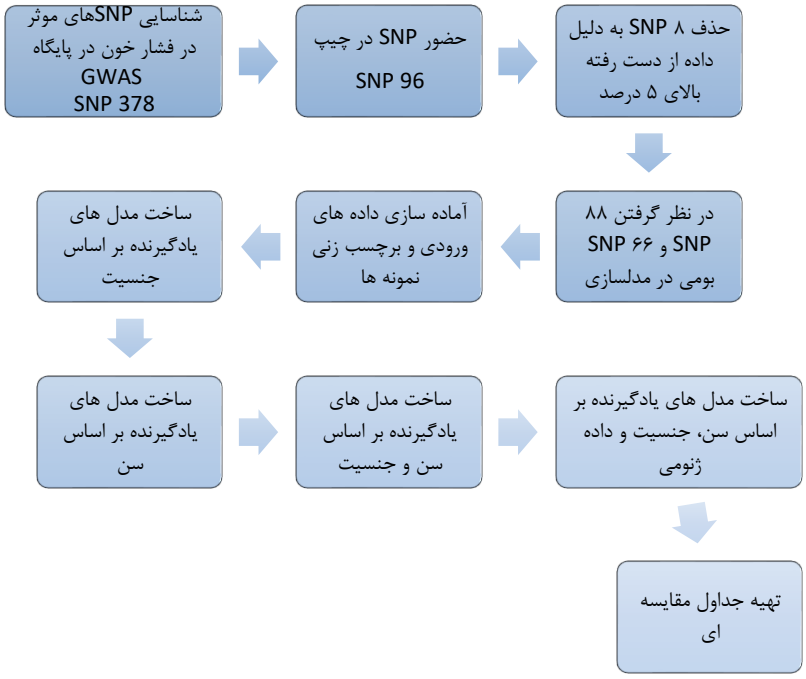
Normal: sbp<120+10 and dbp<80+5
High: sbp<140 or dbp<90
G1 Hyper: sbp<160 or dbp<100
G2 Hyper: sbp<180 or dbp<110
G3 Hyper: sbp>180 or dbp>110
Isolated: sbp>140 and dbp<90

برای آنکه خطایی در برچسب‌ها ایجاد نشود داده‌هایی با برچسب Isolated از نمونه‌ها حذف‌شده است (Mahajan *et al.*, 2019). ضمناً به جهت سهولت، در هر بار مدل‌سازی تنها یک برچسب وارد مدل می‌شود. تعداد کل نمونه‌ها برابر ۴۸۸۱۹ ردیف هست که نشان می‌دهد به‌طور متوسط افراد در ۳.۹۸ دوره شرکت کرده‌اند.

بر اساس فهرست GWAS نسخه GRCh38.p13 تمامی SNP های مرتبط با رخ‌نمود کلی فشارخون (صفت با کد EFO_1001132) برابر ۳۷۸، SNP هست که از این تعداد ۹۶ SNP در داده‌های ژنومی این تحقیق می‌باشند. ضمناً از این تعداد نیز ۸ SNP شرایط Miss<5% و MAF>5% را ندارند. ضمناً با توجه به تحقیق بومی که در انستیتو انجام‌شده است، علاوه بر ۸۸ SNP گفته‌شده، ۶۶ SNP را نیز از تحقیقات (Kolifarhood, M. Daneshpour, *et al.*, 2019; Kolifarhood, M. S. Daneshpour, *et al.*, 2019; Kolifarhood *et al.*, 2021) استخراج‌شده است و درنهایت تعداد ردیف‌های نهایی ۴۸۸۱۹ و تعداد ستون‌های نهایی ۱۵۸ ستون ورودی (۱۵۶ SNP به همراه جنسیت و سن) و ۴ برچسب هست.

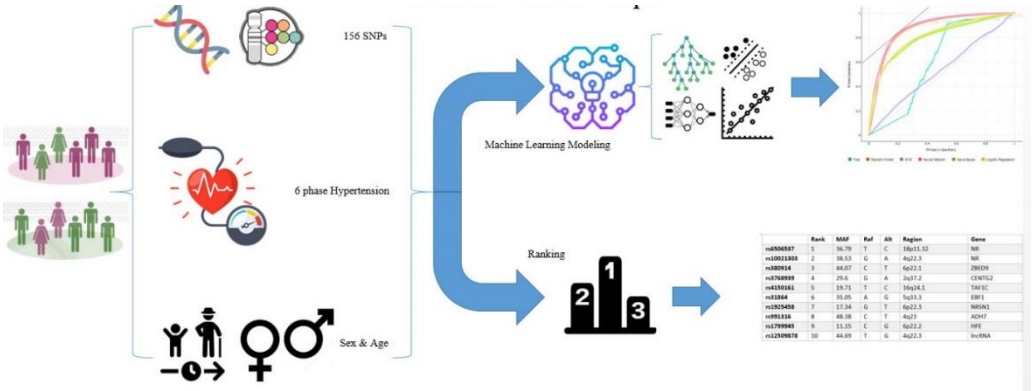
برای مقایسه روش‌های مختلف یادگیری ماشین در پیش‌بینی بیماری، از روش‌های شبکه عصبی، رگرسیون خطی، جنگل تصادفی، درخت تصمیم و ماشین بردار پشتیبان استفاده‌شده است و درنهایت ارزیابی نیز با روش

10-fold رده‌بندی‌شده انجام‌شده است. مراحل تعیین SNPهایی که در مدل‌سازی آمده‌اند و روند کار در شکل ۳-۷ آمده است.



شکل ۳-۷ مراحل اجرای مدل‌سازی یادگیری ماشین سنتی

مسیر اجرایی مدل‌سازی در شکل ۳-۸ آمده است.



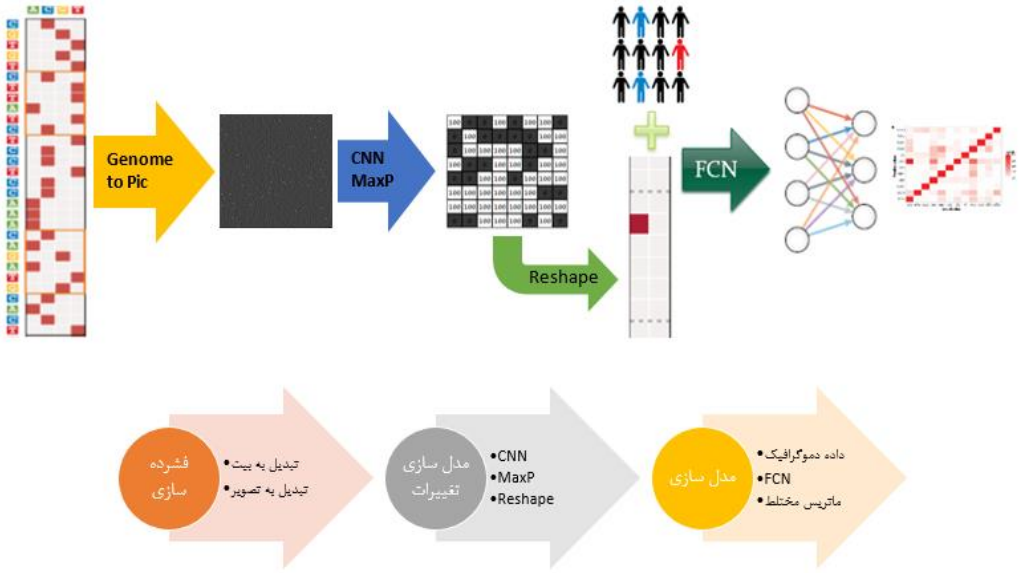
شکل ۳-۸ مسیر اجرای مدل یادگیری ماشین سنتی مبتنی بر SNP کاندید

پارامترهای اصلی برای مدل‌ها عبارت‌اند از: درخت تصمیم با حداقل اندازه زیرمجموعه ۵، حداقل نمونه در برگ دو، و آستانه توقف ۹۵٪ است. جنگل تصادفی دارای ۱۰۰ درخت با حداقل زیرمجموعه ۵ است. SVM با چندین تابع فعالیت استفاده‌شده و بهترین نتیجه تابع سیگموید[1] با مقدار خطای ازدست‌رفته ۰.۱، هزینه ۱ و اغماض[2] عددی ۰.۰۰۱ است. رگرسیون منطقی با تنظیم‌کننده مرز L2 و قدرت ۱ است. شبکه عصبی با سه لایه و تابع فعالیت ReLu توسط حل گر Adam، آلفا ۰.۰۰۰۱ و حداکثر تکرار ۲۰۰ است. لازم به ذکر است که مدل‌سازی با چهار و پنج لایه و سایر توابع فعال انجام‌شده و بهترین نتیجه با این پارامترها بوده است.

۳-۵-۲-فرایند مدل‌سازی CNN

[1] Sigmoid
[2] tolerance

مسیر اجرایی شبکه عمیق CNN به‌صورت شکل ۳-۹ هست.



شکل ۹-۳ مسیر اجرایی شبکه عمیق CNN

۳-۶- جمع‌بندی

این فصل به بیان روش کامل تحقیق پرداخت و با ارائه نمودارهای تصویری، بیان چگونگی تحقیق، انجام شد. بر اساس نمودارهای گفته‌شده در فصل بعدی نتایج گزارش می‌شود.

٤- نتایج و دستاوردها

این فصل به بیان نتایج و دستاوردهای به‌دست‌آمده مبتنی بر روش‌شناسی و داده‌های گفته‌شده در فصل قبل می‌پردازد.

٤-١- بخش‌بندی و کنترل کیفی داده‌های ژنومی

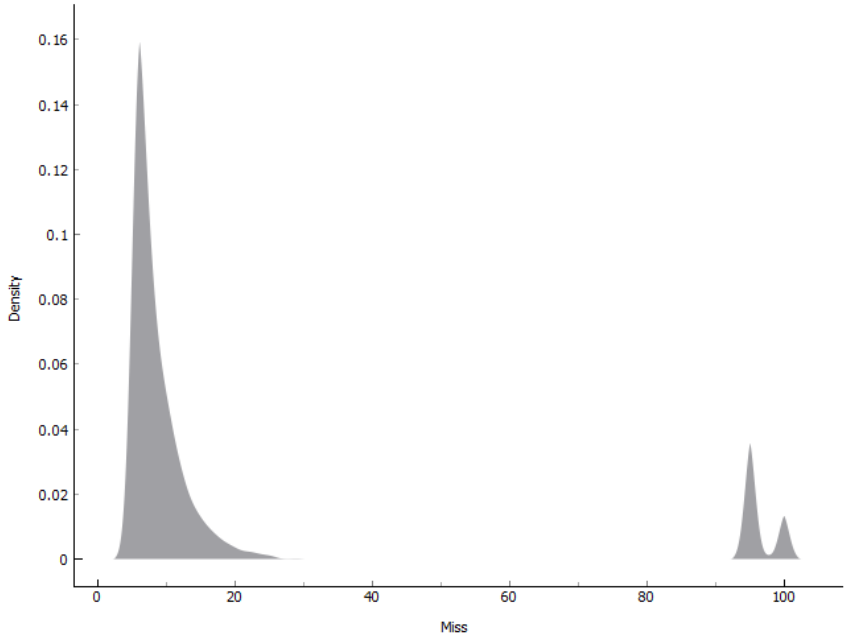
بر مبنای (Mattoo, 2019) جهت بهبود نتایج و کنترل کیفی، دسته‌بندی بر اساس سن (بالای ۱۸ سال و پایین‌تر از ۱۸ سال)، بخش‌بندی SNP از منظر نادر و رایج بودن انجام شد.

برای کنترل کیفی داده‌های ژنومی، SNPهایی که دارای مقادیر گمشده[1] کمتر از ۵ درصد هستند در نظر گرفته شد. در مرحله بعدی با توجه به اهمیت SNP های رایج، تنها SNPهایی با MAF بالاتر از ۵ درصد در نظر هست (Mattoo, 2019). جدول ۴-۱ تعداد SNP های نهایی را نمایش داده است.

جدول ۴-۱کنترل کیفی داده‌های SNP

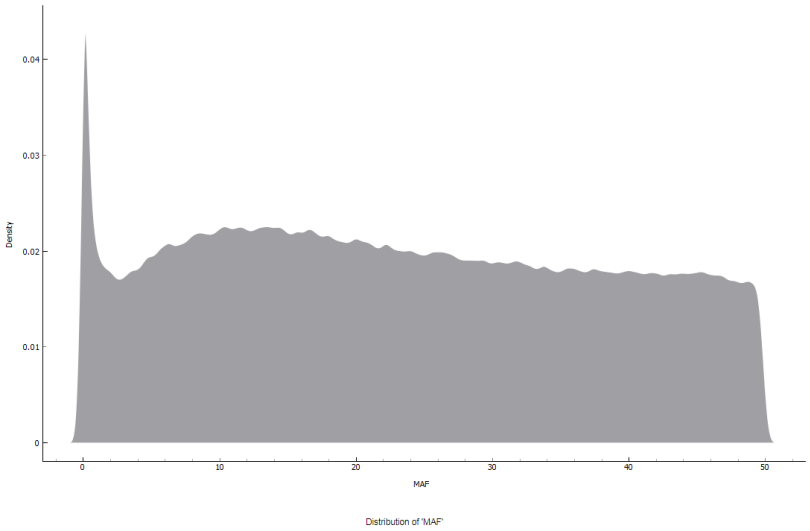
مجموع	652919	100.0%
Miss<5	647357	99.1%
MAF>5	571860	87.6%

تصویر ۴-۸ نمودار چگالی SNP های گمشده در شکل ۴-۱و تصویر ۴-۲ نمودار چگالی MAF را بیان می‌کنند.



شکل ۴-۱ نمودار تغییرات تعداد SNP با مقادیر گمشده

[1] Missed Value



شکل ۲-۴ نمودار چگالی MAF

از روش‌هایی که باعث افزایش دقت می‌گردد، انتخاب SNP های مهم بر اساس عوامل مختلف هست. یکی از این عوامل انتخاب SNPهای کاندیدا معرفی‌شده در GWAS های گذشته در خصوص فشارخون هست.

۲-٤- اجرای مدل بر مبنای SNP کاندیدا

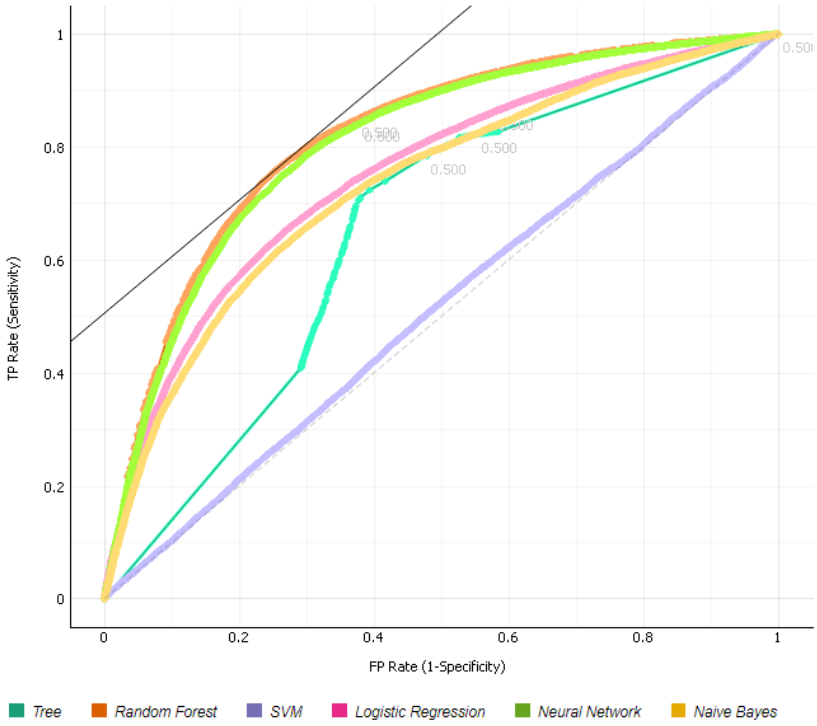
جدول ۴-۲ نتایج مدل‌سازی جنسیت، سن و جنسیت، سن و درنهایت سن، جنسیت و داده ژنومی را با روش‌های مختلف یادگیری ماشین، نشان می‌دهد. معیارهای اصلی بررسی مدل‌های پیش‌بینی کننده شامل مساحت زیر نمودار ROC، دقت، معیارF، صحت و بازخوانی هست که مهم‌ترین معیار AUC است.

جدول ۲-۴ نتایج مدل‌سازی روش‌های یادگیری ماشین سنتی

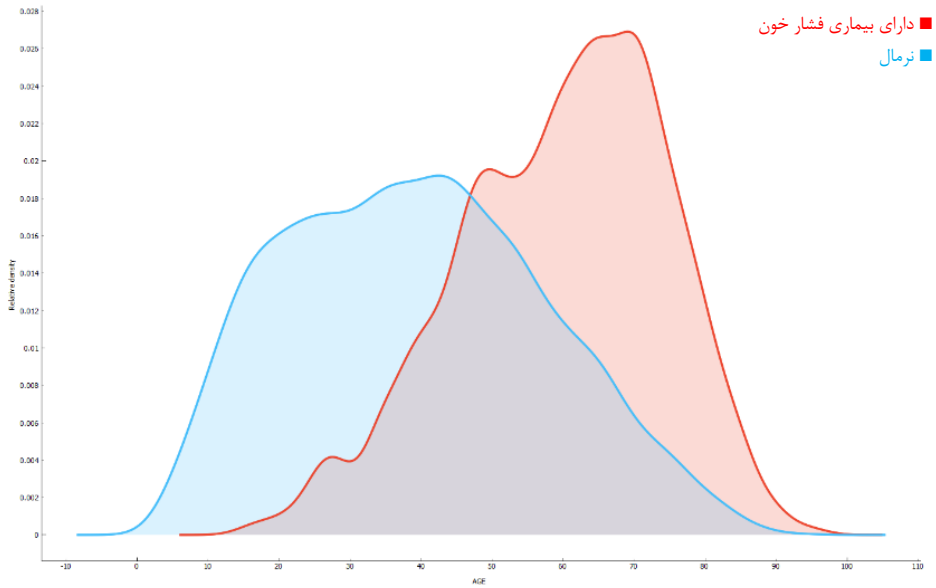
روش	مدل	AUC	صحت	مقدار F	دقت	بازخوانی
Random Forest	جنسیت	0.507	0.507	0.505	0.507	0.507
	سن	0.806	0.741	0.741	0.742	0.741
	سن و جنسیت	0.809	0.74	0.74	0.743	0.74
	جنسیت، سن و داده ژنومی	0.868	0.888	0.881	0.878	0.888
Neural Network	جنسیت	0.504	0.505	0.505	0.505	0.505
	سن	0.808	0.743	0.742	0.746	0.743
	سن و جنسیت	0.814	0.742	0.741	0.745	0.742
	جنسیت، سن و داده ژنومی	**0.885**	**0.891**	**0.889**	**0.887**	**0.891**
Logistic Regression	جنسیت	0.507	0.507	0.505	0.507	0.507
	سن	0.808	0.743	0.743	0.743	0.743
	سن و جنسیت	0.808	0.743	0.743	0.744	0.743
	جنسیت، سن و داده ژنومی	0.811	0.853	0.819	0.817	0.853
Tree	جنسیت	0.507	0.507	0.505	0.507	0.507
	سن	0.806	0.741	0.74	0.742	0.741
	سن و جنسیت	0.81	0.741	0.741	0.743	0.741
	جنسیت، سن و داده ژنومی	0.610	0.852	0.844	0.838	0.852

SVM	جنسیت	0.5	0.503	0.498	0.503	0.503
	سن	0.397	0.303	0.303	0.303	0.303
	سن و جنسیت	0.549	0.53	0.498	0.54	0.53
	جنسیت، سن و داده ژنومی	0.565	0.810	0.780	0.759	0.810

نمودار ROC حاصل از مدل‌سازی جنسیت، سن و داده ژنومی در شکل ۴-۳ و توزیع سنی بر اساس فشارخون در شکل ۴-۴ آمده است.



شکل ۴-۳ نمودار ROC مدل‌سازی جنسیت، سن و داده ژنومی



شکل ۴-۴ نمودار توزیع سنی بر اساس برچسب فشارخون

نتایج حاصل از بخش‌بندی سن در تصویر چندگانه ۴-۵ آمده است.

Scores

Method	AUC	CA	F1	Precision	Recall
Neural Network	0.848	0.868	0.861	0.858	0.868
Random Forest	0.844	0.869	0.860	0.856	0.869
Logistic Regression	0.755	0.842	0.791	0.789	0.842
Naive Bayes	0.672	0.756	0.760	0.765	0.756
Tree	0.606	0.833	0.825	0.819	0.833
SVM	0.505	0.784	0.762	0.745	0.784

`All field,Age<=18:3400,DBP+SBP`

Scores

Method	AUC	CA	F1	Precision	Recall
Neural Network	0.850	0.867	0.862	0.858	0.867
Random Forest	0.842	0.870	0.861	0.857	0.870
Logistic Regression	0.755	0.842	0.792	0.790	0.842
Naive Bayes	0.672	0.756	0.760	0.765	0.756
Tree	0.606	0.834	0.826	0.820	0.834
SVM	0.505	0.784	0.762	0.745	0.784

`All field,age>18:24000,DBP+SBP`

Scores

Method	AUC	CA	F1	Precision	Recall
Neural Network	0.672	0.990	0.986	0.981	0.990
Logistic Regression	0.665	0.990	0.986	0.981	0.990
Naive Bayes	0.652	0.990	0.986	0.981	0.990
SVM	0.632	0.990	0.986	0.981	0.990
Random Forest	0.607	0.990	0.986	0.981	0.990
Tree	0.500	0.990	0.986	0.981	0.990

`Age+Sex,Age<=18:3400,DBP+SBP`

Scores

Method	AUC	CA	F1	Precision	Recall
Neural Network	0.769	0.845	0.775	0.792	0.845
Logistic Regression	0.767	0.841	0.787	0.781	0.841
Random Forest	0.765	0.844	0.779	0.778	0.844
Tree	0.757	0.844	0.777	0.776	0.844
Naive Bayes	0.757	0.845	0.774	0.714	0.845
SVM	0.414	0.596	0.650	0.753	0.596

`Age+Sex,Age>18:24000,DBP+SBP`

Scores

Method	AUC	CA	F1	Precision	Recall
Neural Network	0.863	0.886	0.880	0.877	0.886
Random Forest	0.857	0.884	0.876	0.873	0.884
Logistic Regression	0.785	0.860	0.815	0.811	0.860
Naive Bayes	0.691	0.771	0.780	0.789	0.771
Tree	0.594	0.855	0.846	0.840	0.855
SVM	0.512	0.761	0.764	0.768	0.761

`All field,27400,DBP+SBP`

شکل ۵-۴ تصویر جداول مقایسه‌ای روش‌های با نظارت مبتنی بر بخش‌بندی سن

با توجه به نتایج به‌دست‌آمده، بخش‌بندی بر روی سن باعث کاهش دقت نهایی مدل می‌گردد.

نتایج این بخش در (S Ali Lajevardi *et al.*, 2022) منتشر گردید[1].

۴-۳- محاسبه امتیاز خطر ژنتیکی مبتنی بر GRS

برای مقایسه نتایج به‌دست‌آمده از روش‌های یادگیری ماشین با روش‌های آماری مرسوم، محاسبه امتیاز خطر ژنتیکی با ابزارهای GCTA و PLINK انجام‌شده است.

برای این کار ابتدا، داده‌ها به ۱۰ قسمت (بخش) تصادفی تقسیم شدند. در مرحله بعدی با ۹ قسمت محاسبه تأثیر فردی[2] و پس‌ازآن محاسبه تأثیر SNP توسط بسته GCTA انجام می‌شود. سپس با اجرای دستورات لازم در PLINK محاسبه امتیاز انجام می‌شود. درنهایت قسمت کنار گذاشته توسط یک دسته‌بند در هر بخش محاسبه و نتایج آن ادغام می‌گردد. برای این کار از نرم‌افزار Orange استفاده‌شده است. لیست دستورات در شکل ۴-۶ آمده است. پَروَنجا ژنوم ورودی به‌صورت genome.ped هست.

[1] https://github.com/l8026070/Hypertension-CNN

[2] Individual

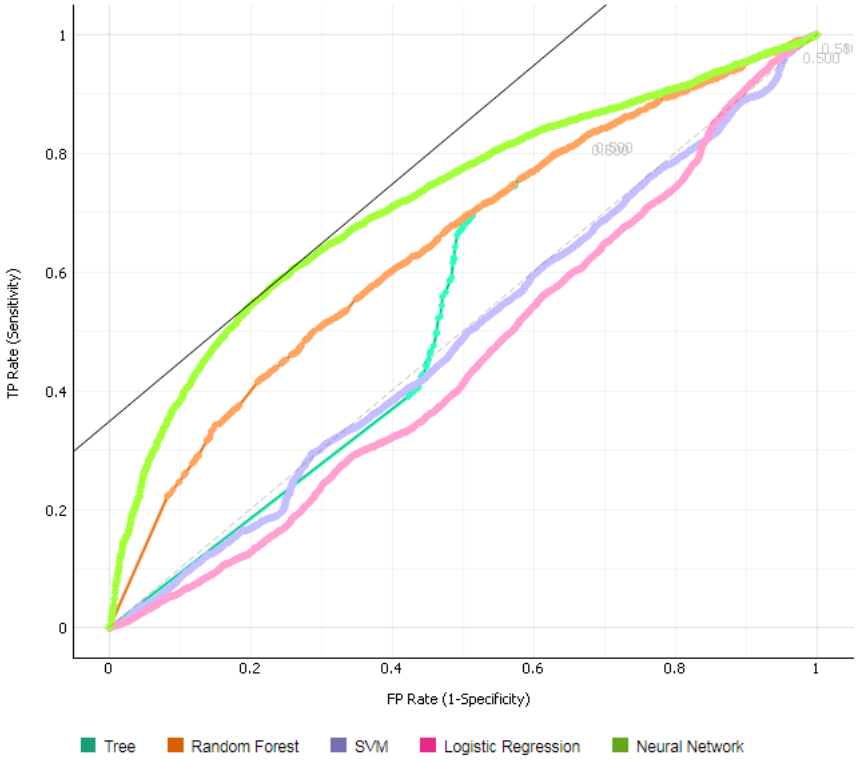
```
plink --noweb --ped genome.ped --map genome.map --make-bed --out genome
gcta64 --bfile genome --make-grm --reml --thread-num 10 --out genome
gcta64 --reml --grm genome --pheno genome.phen --cvblup --out genome
plink --noweb --score genome.tr_n.snp.blp --bfile genome --pheno
genome.ts_n.phen --covar genome.cov --out genome.t_n
```

شکل ۶-۴ نمونه کد محاسبه امتیاز خطر ژنتیکی

نتایج به‌دست‌آمده در جدول ۴-۳ آمده است و نمودار ROC نیز در شکل ۴-۷ آمده است. نتایج اولیه در مقاله (لاجوردی *et al.*, 1399) منتشر گردید.

جدول ۳-۴ نتایج دسته‌بندهای مختلف برای امتیاز خطر ژنتیکی

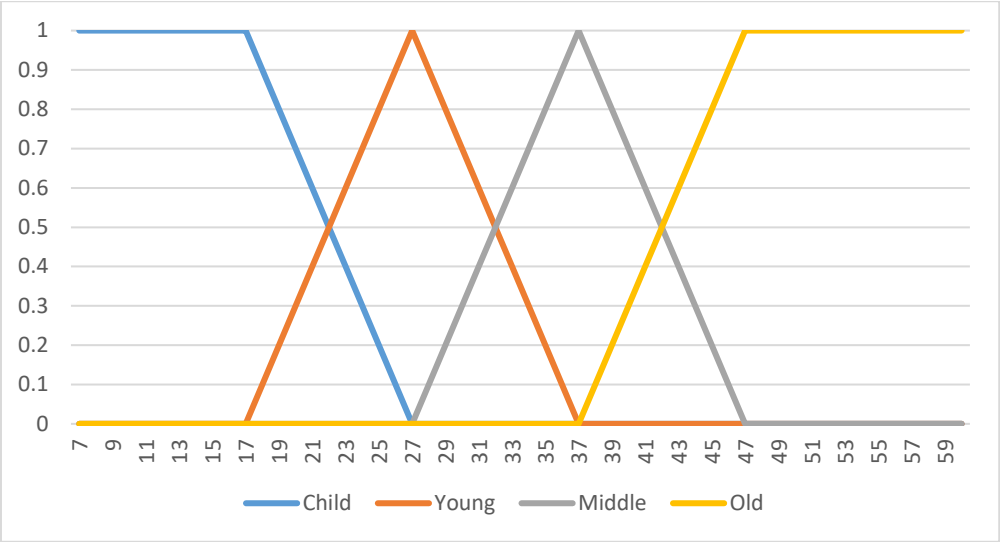
Method	AUC	CA	F1	Precision	Recall
Neural Network	0.714	0.742	0.648	0.652	0.742
Random Forest	0.641	0.702	0.692	0.685	0.702
Tree	0.550	0.704	0.693	0.685	0.704
SVM	0.464	0.675	0.647	0.630	0.675
Logistic Regression	0.425	0.747	0.639	0.558	0.747



شکل ۷-۴ نمودار ROC مدل‌سازی امتیاز خطر ژنومی

٤-٤- مدل‌سازی فازی

با توجه به متغیرهای در نظر گرفته‌شده در مدل‌سازی، تنها متغیری که امکان فازی سازی بر روی آن وجود داشت، سن است. به همین دلیل بر اساس ادبیات موضوع فازی سازی سن بر اساس شکل ۴-۸ انجام گردید (Satyanarayana *et al.*, 2020; Singh, Verma and Singla, 2020). درنهایت با در نظر گرفتن سن فازی و SNP های در نظر گرفته‌شده در مدل‌سازی یادگیری ماشین، مدل‌سازی دیگری نیز برای خطر پرفشاری خون انجام گردید.



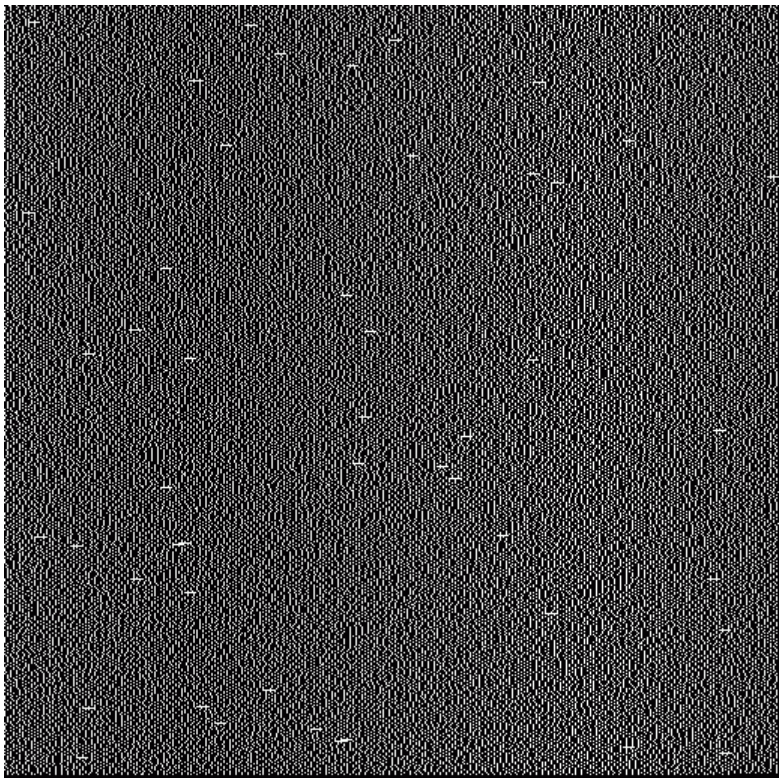
شکل ۸-۴ فازی سازی سن

٤-٥- اجرای شبکه عمیق CNN

همان‌طور که در فصل قبل ذکر شد برای شروع شبکه CNN ابتدا داده‌ها به‌صورت تصویر درآمده است.

٤-٥-١- تبدیل داده ژنومی به تصویر

برای اجرای الگوریتم CNN بر روی‌داده‌های ژنومی، اولین کار تبدیل داده ژنومی شامل ۶۵۲۹۱۹ SNP، به یک تصویر هست. برای این کار داده‌های ژنوم که شامل رشته‌ای از T،G،C،A هست تبدیل به بیت شده و به‌صورت تصویر سیاه‌وسفید در یک پَروَنجا تصویری با فرمت GIF ذخیره‌سازی انجام می‌شود. نتیجه حاصل‌شده به‌صورت شکل ۴-۹ هست.



شکل ۹-۴ تصویر ساخته‌شده از داده ژنومی

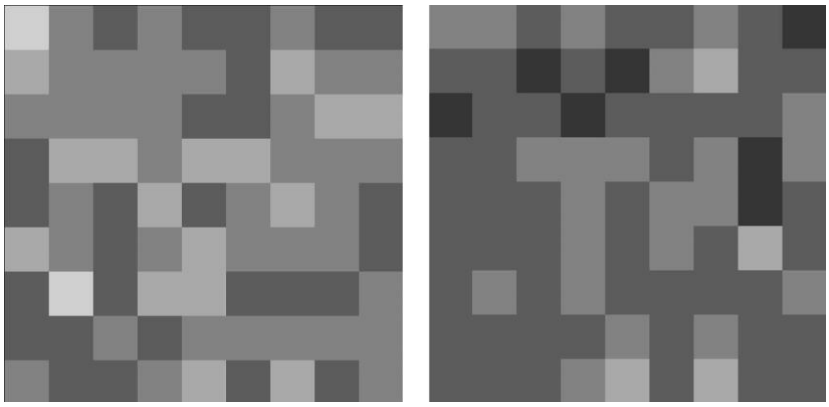
با این کار علاوه بر اینکه تصویر آماده طراحی شبکه CNN هست، داده ژنومی به شکل مطلوبی فشرده‌سازی شده است که نوع فشرده‌سازی به حالت بدون از دست رفتگی[1] و بدون نیاز به الگوریتم بازیابی جهت خارج کردن از حالت فشرده هست.

٤-٥-٢- اجرای شبکه CNN

طراحی شبکه عمیق مبتنی بر CNN بر اساس ۲۴ پالایه به‌صورت ۴*۴ که در هر پالایه در هر ردیف تنها یک ستون دارای مقدار یک و مابقی صفر هست و پس از اعمال پالایه، تمامی نتایج در همدیگر ادغام می‌شود (اجرای پیچشی CNN). تعداد لایه‌های CNN وابسته به پیچیدگی تغییرات هست که با توجه به داده‌های آزمون، یک‌لایه مناسب است.

در مرحله بعد MaxPool استفاده‌شده است که به‌صورت ۲*۲ هست که در هر ۴ خانه بیشینه را به لایه بعدی منتقل می‌کند و باعث کاهش بعد ۱/۲ در طول و در ارتفاع تصویر می‌شود که درنتیجه در هر مرحله تصویر ۱/۴ می‌شود که با انجام تعداد لایه معین می‌توان به مقدار دلخواه برای ورودی به لایه بعدی FCN رسید. با توجه به داده‌های آزمون، ۶ لایه استفاده‌شده است. تابع فعالیت FCN به‌صورت Relu هست.

درنهایت برای یکدست‌سازی اطلاعات تصاویر به‌دست‌آمده مبتنی بر max-min نرمال‌سازی شده و برای هر ورودی، مقیاس ۱۰۰ برای نمایش خروجی در نظر گرفته‌شده است. تصویر ۴-۱۰ دو نمونه از این پردازش هست.



شکل ۱۰-۴ خروجی CNN، MaxP، Normalize

ساخت هر تصویر نهایی در حدود ۳ دقیقه زمان می‌برد که این مرحله بالغبر ۲۳ روز به طول انجامید.

۳-۵-۴- آماده‌سازی داده‌های طولی رخ‌نمود

با توجه به اینکه تمامی داده‌های رخ‌نمودی در پژوهشکده بر روی یک پَروَنجا SPSS هست، در مرحله بعدی داده‌های SPSS رخ‌نمودهای ثبت‌شده بر مبنای رخ‌نمودهای موردنیاز و پردازش اولیه اطلاعات بر اساس توزیع رخ‌نمودها و داده‌های ثبت‌شده در ۶ دوره TLGS، دسته‌بندی گردید.

سپس داده‌های رخ‌نمود به ژن‌نمود مبتنی بر کدهای صادره از TLGS متصل شد که منجر به استخراج صحیح ۱۱۰۸۸ ردیف اطلاعاتی شامل داده‌های ژن‌نمود و رخ‌نمود گردید.

[1] Lossless

برای اجرای FCN بر مبنای (Basile *et al.*, 2019) برچسب‌های زیر مدنظر قرار گرفت:

۱- عادی: فشار دیاستولی کمتر از ۸۰ و فشار سیستولی کمتر از ۱۲۰

۲- فشار دیاستولی بالاتر از ۸۰

۳- فشار سیستولی بالاتر از ۱۲۰

۴- فشار دیاستولی بالاتر از ۸۰ یا فشار سیستولی بالاتر از ۱۲۰

لازم به ذکر است با توجه به اینکه داده سن و جنسیت تأثیر مهمی در فشارخون دارند (Basile *et al.*, 2019)، این دو داده نیز علاوه بر داده ژن‌نمود به‌عنوان داده مستقل وارد مدل می‌شوند که بر این اساس می‌توان داده‌های دوره‌های مختلف را نیز برای یک فرد پردازش نمود.

۴-۵-۴- اجرای شبکه تمام متصل[1]

برای اجرای شبکه تمام‌اتصال انتهای شبکه عمیق، از روش‌های شبکه عصبی ساده، جنگل تصادفی، درخت تصمیم، رگرسیون لجستیک و SVM استفاده گردید. نتایج حاصل در شکل‌های ۴-۱۱ الی ۴-۱۳ آمده است.

Method	AUC	CA	F1	Precision	Recall
Neural Network	0.809	0.939	0.932	0.929	0.939
Random Forest	0.785	0.936	0.929	0.925	0.936
Logistic Regression	0.729	0.932	0.901	0.878	0.932
Tree	0.661	0.928	0.918	0.913	0.928
SVM	0.633	0.920	0.909	0.900	0.920

شکل ۱۱-۴ جدول مقایسه‌ای روش‌های با نظارت در فشار دیاستولی بالا

Method	AUC	CA	F1	Precision	Recall
Neural Network	0.918	0.925	0.922	0.919	0.925
Random Forest	0.885	0.924	0.918	0.915	0.924
Logistic Regression	0.866	0.911	0.893	0.887	0.911
SVM	0.657	0.914	0.888	0.887	0.914
Tree	0.650	0.915	0.905	0.900	0.915

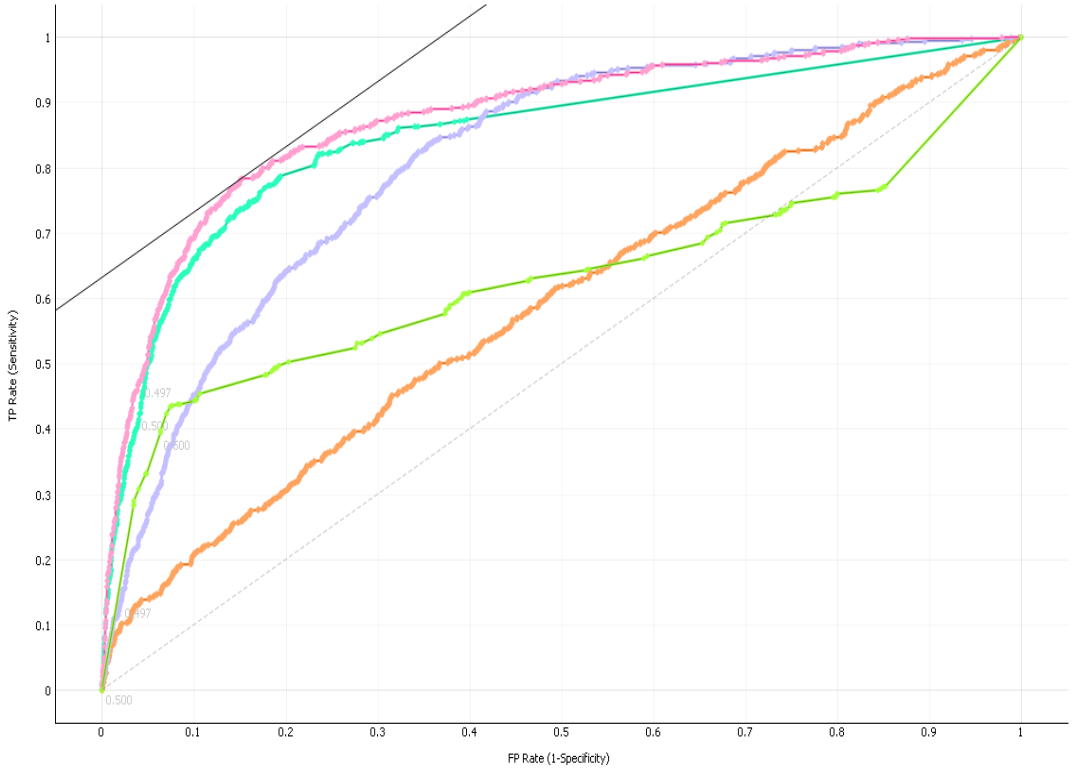
شکل ۱۲-۴ جدول مقایسه‌ای روش‌های با نظارت در فشار سیستولی بالا

Method	AUC	CA	F1	Precision	Recall
Neural Network	0.877	0.901	0.896	0.893	0.901
Random Forest	0.849	0.896	0.889	0.885	0.896
Logistic Regression	0.811	0.881	0.851	0.845	0.881
Tree	0.621	0.879	0.870	0.864	0.879
SVM	0.608	0.857	0.838	0.825	0.857

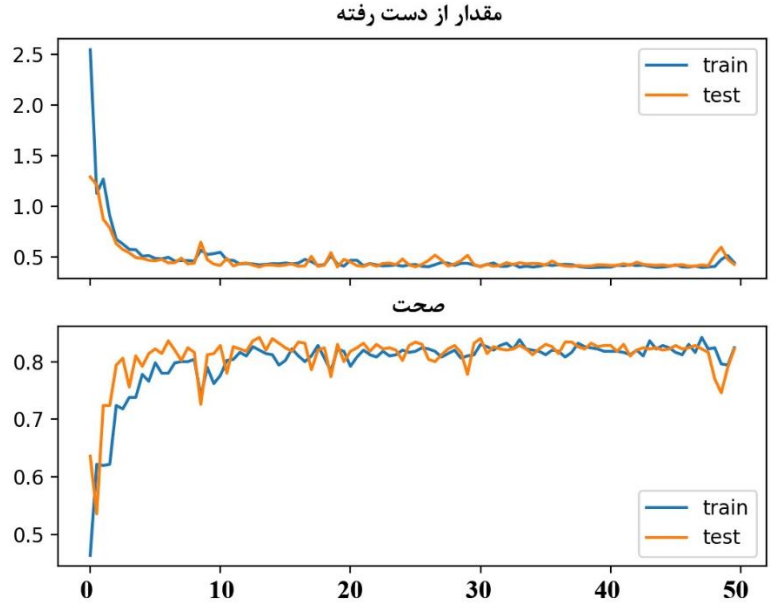
شکل ۱۳-۴ جدول مقایسه‌ای روش‌های با نظارت در فشار سیستولی یا دیاستولی بالا

نمودار ROC مدل‌سازی CNN با توابع مختلف در شکل ۴-۱۴ آمده است.

[1] Fully Connected Network (FCN)

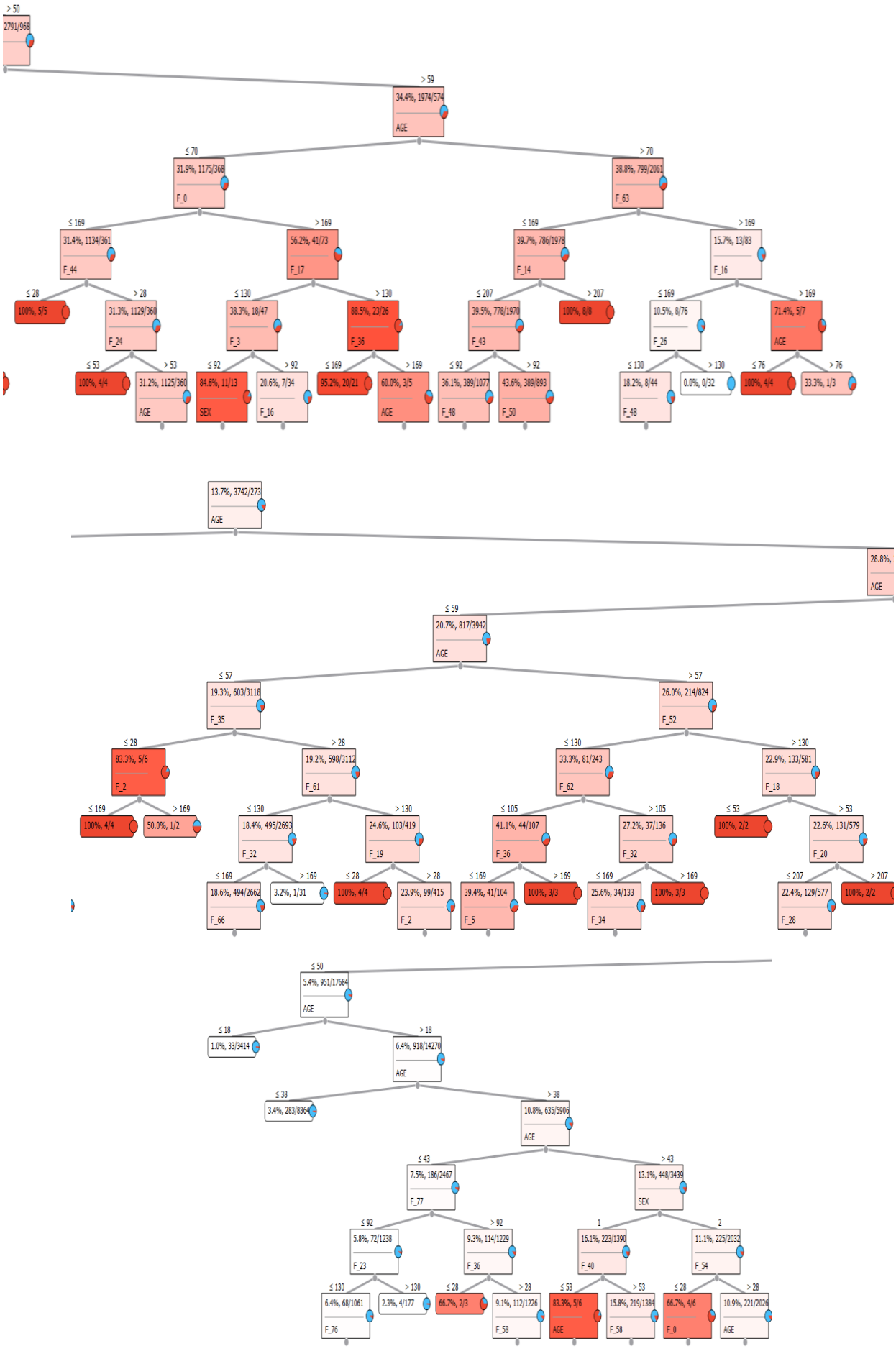


شکل ۱۴-۴ نمودار ROC مدل‌سازی CNN



شکل ۱۵-۴ نمودار صحت و مقدار ازدست‌رفته

با استفاده از الگوریتم درخت تصمیم بر روی‌داده‌ها، درختی حاصل می‌شود که در تصویر ۴-۱۵ آمده است. در این درخت تفکیک سنی به‌دست‌آمده کاملاً متناسب با تحقیقات زیستی در حوزه فشارخون هست و نشان می‌دهد که مهم‌ترین عامل در بیماری فشارخون سن هست و در سطوح اهمیت بعدی SNP های متفاوتی برای جنست زن و مرد نقش دارند. در این تصویر هر چه قدر کادری قرمزرنگ‌تر باشد احتمال وجود فشارخون در آن دسته افراد بالاتر است که به ترتیب برای بالای ۶۰ سال، بالای ۵۰ سال و بالای ۱۸ سال هست.



شکل ۱۶-۴ درخت تصمیم ناشی از مدل‌سازی فشارخون

٤-٦- جمع‌بندی نتایج

با توجه به اینکه روش‌های مختلفی در پیش‌بینی فشارخون بر مبنای داده‌های ژنومی در این تحقیق انجام‌شده است، خلاصه‌ی نتایج در جدول ۴-۴ آمده است.

جدول ۴-۴ خلاصه نتایج مدل‌سازی‌های انجام‌شده در پیش‌بینی بیماری پرفشاری خون

نوع	داده	تعریف	CNN	GRS	Fuzzy	Top10	ML (NN Best)					
متغیر	جنسیت	مرد/زن	*	*	*	*	*	*	*	*	*	
	سن	سال دوره‌ای	*			*		*	*	*		*
		سن فازی			*							
		سن تعیین‌شده		*							*	
	ژنوم	تصویر 600K	*									
		600K GRM		*								
		GWAS SNP			*	*			*			
		Farhood SNP										
برچسب	فشارخون	HTN	*	*	*	*	*	*	*	*	*	*
نتایج	AUC		**0.877**	**0.71**	**0.866**	**0.83**	**0.542**	**0.817**	**0.885**	**0.81**	**0.726**	**0.804**
	صحت		0.901	0.74	0.88	0.861	0.884	0.771	0.891	0.884	0.747	0.883
	مقدار F		0.896	0.65	0.878	0.832	0.829	0.768	0.889	0.834	0.639	0.838
	دقت		0.893	0.65	0.876	0.834	0.781	0.767	0.887	0.841	0.6	0.839
	بازخوانی		0.901	0.74	0.88	0.861	0.884	0.771	0.891	0.884	0.747	0.883

٤-٧- شناسایی SNP های مهم

برای شناسایی SNP های مهم روش‌های مختلف رتبه سنجی وجود دارد. این روش‌ها امکان شناسایی را با توجه به توزیع متغیر مربوطه در داده‌ها بر مبنای برچسب و دیگر متغیرهای ورودی انجام می‌دهد.

به جهت اهمیت استفاده از روش‌های مختلف، روش‌های بهره اطلاعات[1]، نسبت بهره[2]، شاخص جینی[3]، کای دو[4]، رلیف[5] وFCBF استفاده گردید. درنهایت با توجه به رتبه‌ای که هر SNP در هر روش به دست آورد، اهمیت میانگین SNP محاسبه گردید و بر مبنای آن لیست مرتب مهم‌ترین SNPها در جدول ۴-۵ آمده است. در این جدول ستون‌های TCGS بر اساس اطلاعات بومی در داده‌ها اندازه‌گیری شده است و اطلاعات GWAS نیز در ستون‌های مربوطه قرارگرفته است.

برای مقایسه با دیگر مدل‌ها، خطر پرفشاری خون بر اساس ۱۰ SNP برتر نیز محاسبه گردید. نتایج کامل در مقاله (S. Ali Lajevardi *et al.*, 2022) منتشر گردید[6].

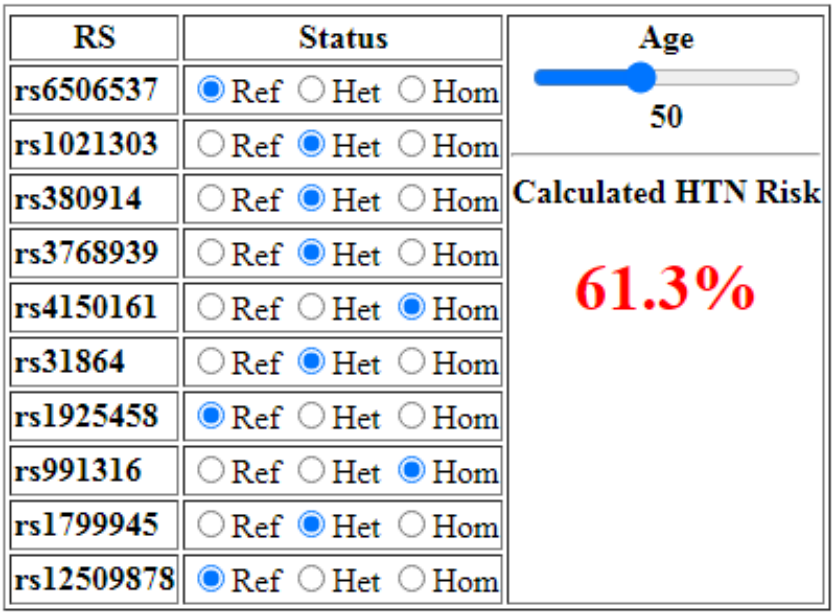
جدول ۵-۴ لیست مرتب شناسایی ۱۰ SNP مهم

RS		TCGS			GWAS				
		MAF	Ref	Alt	TRAIT	POS	Gene	RISK ALLELE	P-VALUE
rs6506537	1	36.79	T	C	Hypertension	18p11.32	NR	rs6506537-T	4.00E-09
rs10021303	2	38.53	G	A	Hypertension	4q22.3	NR	rs10021303-A	7.00E-06
rs380914	3	44.07	C	T					
rs3768939	4	29.6	G	A	Cardiovascular disease in hypertension (calcium channel blocker interaction)	2q37.2	CENTG2	rs3768939-A	9.00E-07
rs4150161	5	19.71	T	C	Systolic blood pressure response to hydrochlorothiazide in hypertension	16q24.1	TAF1C	rs4150161-T	1.00E-06
rs31864	6	35.05	A	G	Hypertension	5q33.3	EBF1	rs31864-A	3.00E-06
rs1925458	7	17.34	G	T	Alzheimer's disease in hypertension	6p22.3	NRSN1	rs1925458-?	1.00E-06
rs991316	8	48.38	C	T	Hypertension	4q23	ADH7	rs991316-T	5.00E-06
rs1799945	9	11.15	C	G	Hypertension	6p22.2	HFE	rs1799945-G	2.00E-10
rs12509878	10	44.69	T	G	Diastolic blood pressure night-to-day ratio in hypertension	4q22.3	lncRNA	rs12509878-C	6.00E-06

[1] Information gain
[2] gain ratio
[3] gini index
[4] χ^2
[5] relief
[6] https://github.com/l8026070/Hypertension-ML

٤-٨- ابزار سنجش خطر بیماری فشارخون بالا

بر مبنای SNP های مهم شناسایی‌شده درروش رتبه‌بندی و تأثیر عامل سن در محاسبه خطر، ابزاری به زبان PHP در آدرس https://work.feham.ir/CalH منتشرشده است. تصویر ابزار به‌صورت شکل ۴-۱۶ هست.



شکل ۴-۱۷ تصویر ابزار ساخته‌شده جهت پیش‌بینی خطر HTN

۴-۹- تحلیل ساختار علمی موضوع امتیاز خطر ژنتیکی

محققین در یک حوزۀ علمی به‌منظور دیدن فراسوهای دانش در حوزۀ تخصصی خود، معمولاً آثار پژوهشگران پیشین خود را مرور می‌کنند؛ به‌بیان‌دیگر، با اتکا به گذشته علم، آینده علمی حوزه تخصصی خود را ترسیم می‌کنند. داشتن درک و نمایی کلی از چارچوب علمی حوزه موردنظر، یکی از راه‌هایی است که پژوهشگران را برای رسیدن به اهداف پژوهشی در یک حوزۀ تخصصی کمک می‌کند. در این راستا ترسیم نقشۀ یا ساختار علمی آن حوزه، امری ضروری به نظر می‌رسد. در یک نقشۀ علمی که بر اساس بروندادهای علمی-پژوهشی محققین یک حوزه علمی ترسیم می‌شود، نویسندگان تأثیرگذار، خوشه‌ها و زیر حوزه‌های موضوعی شکل‌گرفته در طول زمان ، روند حرکتی آن حوزه و موضوعات پژوهشی آینده حول آن حوزه تخصصی تعیین و معرفی می‌شوند. در نقشه‌های علمی ظهور حیطه‌های جدید و توقف برخی حوزه‌های علمی اشباع‌شده به‌وضوح نمایان است. به بیانی ساده نقشه علمی مصورسازی نتایج برآمده از تجزیه‌وتحلیل انتشارات یک حوزۀ علمی از زوایای گوناگون و ارائه چارچوبی کلی از آن حوزه است.

در تعریف نقشه‌های علمی اعتقاد بر این است که نقشۀ علمی، نمایی از حوزه‌های علمی است که با تجزیه‌وتحلیل کمّی اطلاعات کتابشناختی تهیه می‌شود. عناصر تشکیل‌دهنده نقشه‌های علمی، بروندادهای حوزه‌های پژوهشی هستند. در این نقشه‌ها، حوزه‌های علمی که دارای ارتباط مفهومی قوی‌تری هستند، برکنار هم و حوزه‌هایی که ارتباط ضعیف‌تری دارند در فاصلۀ دورتری از هم قرار می‌گیرند (Noyons, 1999).

٤-٩-١- هم رخدادی واژگان

با توجه به رشد روزافزون حجم اطلاعات و سرعت فزاینده تولید علم، رصد روندهای علمی و شناسایی دانش پنهان‌شده در متون کاری دشوار است. لذا امروزه برای کشف لایه‌های پنهان دانش در علوم مختلف از روش‌های مختلفی ازجمله متن‌کاوی استفاده می‌شود. محققین توانسته‌اند با بهره‌گیری ازاین‌روی‌ها به کشف دانش پنهان موجود در متون علمی دست پیدا کنند. درحالی‌که دستیابی به چنین دانشی از طریق روش‌های سنتی امکان‌پذیر نیست. یکی از این فنون که طی سال‌های اخیر خیلی مورد استقبال و اقبال متخصصان رشته‌های مختلف قرارگرفته تحلیل هم رخدادی واژگان[1] است که به‌عنوان شاخصی نوین به جامعهٔ علمی عرضه شد (Rip and Courtial, 1984). هم رخدادی واژگان یک فن تحلیل محتوا هست که هم فراوانی موضوعات و هم ارتباط بین آن‌ها را بیان می‌کند . تحلیل هم رخدادی واژگان، ابزاری برای کشف الگوهای پنهان و رویدادهای نوظهور مفهومی است که از طریق آن می‌توان مفاهیم اصلی یک زمینه یا حوزهٔ علمی را شناخت و به‌واسطه این شناخت، الگوها و رویدادهای مفهومی، ساختار علمی، شبکهٔ مفهومی، روابط سلسله مراتبی مفاهیم، و مقولات مفهومی آن حوزه را کشف، ترسیم و مدیریت کرد. روش تحلیل هم رخدادی واژگان با تکیه‌بر واژه‌های پربسامد می‌تواند مهم‌ترین موضوعات پژوهشی در هر حوزه علمی را مشخص کند؛ یعنی میزان رخداد یک واژه، نشان‌دهندهٔ میزان اهمیت آن در حوزهٔ علمی موردنظر است. کالون[2] در سال ۱۹۸۳ اولین کسی بود که این روش را توصیف کرد و از آن در پژوهش‌های خود استفاده کرد (He, 1999). در این تحلیل، هم رخدادی کلیدواژه‌ها در عنوان، چکیده یا متن مقالات بررسی می‌شود. هم رخدادی کلیدواژه‌ها میزان ارتباط شناختی میان یک مجموعه مدارک را نشان می‌دهد. با مقایسهٔ نقشه‌های حاصل در دوره‌های زمانی مختلف، پویایی علم ردیابی می‌شود. با به‌کارگیری روش تجزیه‌وتحلیل هم رخدادی واژگان می‌توان موضوعات علمی را استخراج و ارتباط میان آن‌ها را به‌صورت مستقیم از محتوای موضوعی کشف کرد . تحقیقات نشان داده که مصورسازی دانش از طریق تحلیل هم رخدادی واژگان می‌تواند تصویر روشن‌تری از آن حوزه نمایان سازد چراکه تحلیل واژگان موضوعات اساسی آن حوزه را در یک شمای کلی در اختیار پژوهشگر قرار می‌دهد. ازآنجاکه نقشه‌های علمی دارای ساختاری مشابه ساختار شبکه‌های اجتماعی[3] هستند، برای مصورسازی و تحلیل و تفسیر آن‌ها از فنون تحلیل شبکه‌های اجتماعی استفاده می‌شود (Guns, Liu and Mahbuba, 2011).

در این تحقیق از روش تحلیل هم رخدادی واژگان که یکی از روش‌های علم‌سنجی است، استفاده‌شده تا از حوزهٔ امتیاز خطر ژنتیکی اطلاعات کمّی قابل‌تحلیل به دست آید. جامعه پژوهش ۱۱۷۵ مقاله در حوزه امتیاز خطر ژنتیکی است که در پایگاه داده ScienceDirect در بازه سال‌های۲۰۱۶تا۲۰۲۰ نمایه شده است. جهت انجام این تحقیق، عبارت امتیاز خطر ژنتیکی به‌عنوان هستهٔ اصلی در نظر گرفته‌شده است. تمام کلمات کلیدی مربوط به ۱۱۷۵ مقاله از پایگاه داده ScienceDirect که شامل این عنوان در مجموعه کلیدواژه‌های خود بودند،

[1]Co-word analysis
[2]Callon
[3]Social networks

جمع‌آوری شد. سپس هر یک از کلمات کلیدی به‌عنوان یک گره در شبکه در نظر گرفته‌شده است. یال‌های شبکه نیز به‌صورت ارتباط بین کلمات تعریف‌شده‌اند. به‌طوری‌که بین دو گره یال وجود خواهد داشت درصورتی‌که آن دو گره(کلمه) به‌صورت مشترک در لیست کلمات کلیدی یک مقاله آمده باشند. پس از جمع‌آوری تمام کلمات ، ارتباط دوبه‌دو بین تمام کلمات بررسی شد و درصورتی‌که دو کلمه دارای مقاله مشترک بودند، بین آن‌ها یک یال تعریف‌شده است. با توجه به این‌که محققین برای انتخاب کلمات کلیدی، استاندارد معینی را در نظر نمی‌گیرند و انتخاب این کلمات به‌صورت سلیقه‌ای انجام می‌شود. در دادگان جمع‌آوری‌شده، بسیاری از کلمات کلیدی علی‌رغم یکسان بودن ازنظر مفهوم و معنا، در مقالات متفاوت در نظر گرفته‌شده است که همه یک مفهوم را متبادر می‌سازد. این موضوع باعث می‌شود در شبکه حاصل از ارتباط کلمات کلیدی مقالات، به ازای هر کلمه کلیدی یک گره در نظر گرفته شود و درنتیجه شبکه دارای گره‌های تکراری ازلحاظ مفهوم باشد. این امر علاوه بر ایجاد گره‌های ایزوله در شبکه و تشکیل اجتماعات بسیار کوچک با حداکثر ۳ یا ۴ گره، باعث می‌شود نتایج تحلیل منعکس‌کننده واقعیت مسئله نباشد. برای حل این مشکل، تمام کلمات ازنظر ظاهر و مفهومی بررسی‌شده و کلمات هم‌معنا به‌عنوان یک گره واحد در نظر گرفته‌شده‌اند. این کار به‌صورت ماشینی و با استفاده از الگوریتم فاصله لون‌اشتاین انجام‌شده است.

٤-۹-۲- نتایج و ارزیابی آن

در جدول ۴-۶ ده مجله که بیشترین ارتباط و همسویی با موضوع موردبررسی را داشته‌اند مشاهده می‌شود. که می‌تواند جهت تصمیم‌گیری برای ارسال مقالات مرتبط با موضوع به مجلات جهت چاپ مورداستفاده قرار گیرد.

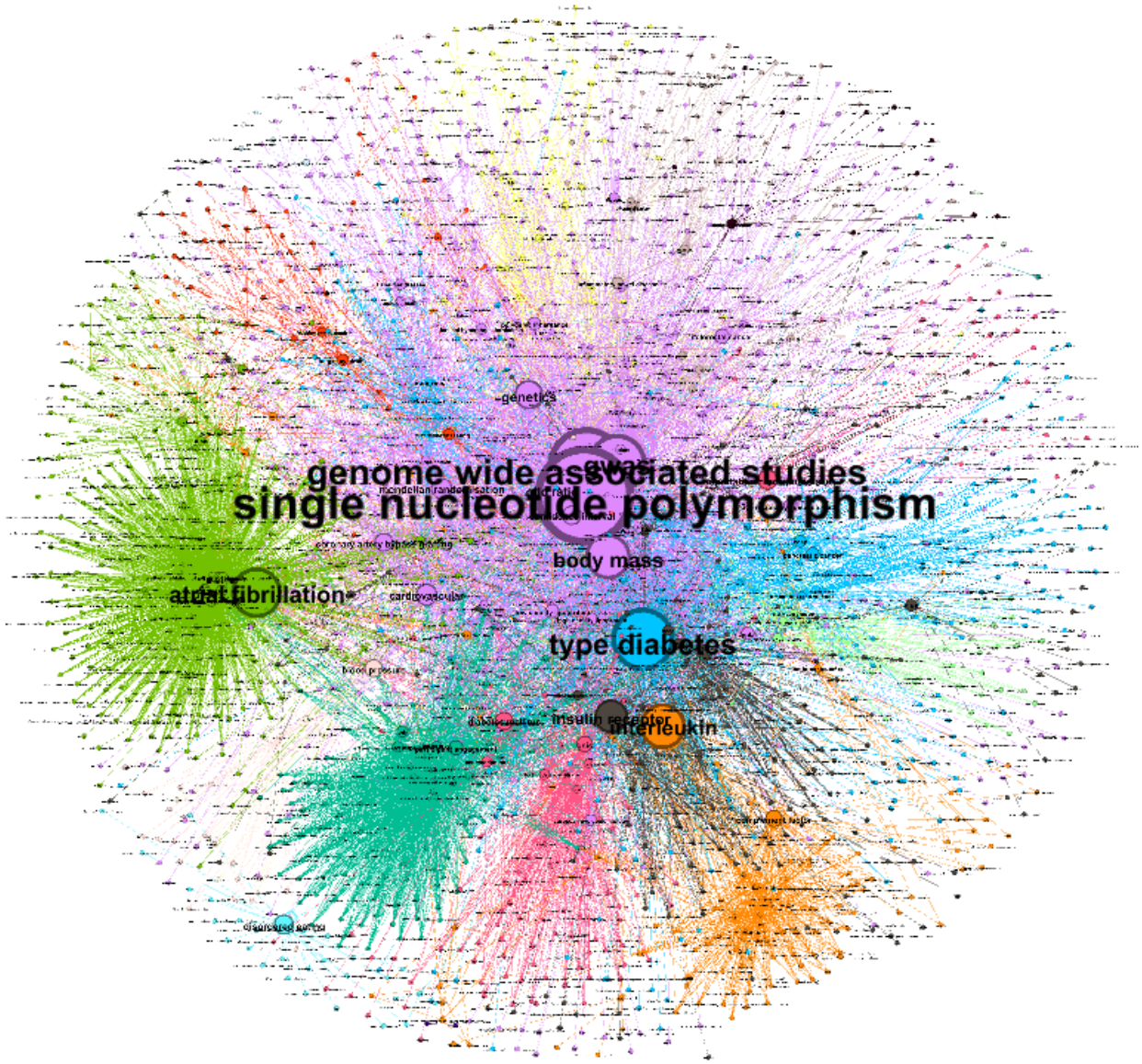
جدول ۶-۴ نام مجلات با بیشترین ارتباط با موضوع

ردیف	نام مجله	تعداد مقالات چاپ‌شده	IF[1]	۵ ساله IF
۱	Gastroenterology	۸۵	۱۹.۲۳۳	**۱۸.۸۴۴**
۲	European Neuropsychopharmacology	۵۱	۴.۴۶۸	**۴.۳۹۴**
۳	Journal of the American College of Cardiology	۴۳	۱۸.۶۳۹	**۱۹.۰۶۸**
۴	Atherosclerosis	۴۱	۳.۹۱۹	**۴.۶۳۹**
۵	Ophthalmology	۲۱	۷.۷۳۲	**۷.۸۴**
۶	Alzheimer's & Dementia	۲۱	۱۴.۴۲۳	**۱۳.۵۳۳**
۷	The American Journal of Human Genetics	۱۷	۱۰.۵۰۲	**۱۰.۳۴۴**
۸	The Lancet Diabetes & Endocrinology	۱۶	۲۴.۵۴۰	**۲۳.۰۱۳**
۹	Journal of Clinical Lipidology	۱۶	۳.۸۶۰	**۴.۰۸۰**
۱۰	Journal of Allergy and Clinical Immunology	۱۵	۱۰.۲۲۸	**۱۱.۵۹۲**

برای اجتماع یابی از الگوریتم اجتماع یابی[2] بکار گرفته‌شده در نرم‌افزار Gephi 0.9.2 موسوم به الگوریتم لووین[3] که توسط بلاندل[4] و همکاران در سال ۲۰۰۸ ارائه شد، استفاده‌شده است (Blondel *et al.*, 2008).

[1] Impact Factor
[2] Community detection
[3] Louvain
[4] Blondel

در شکل ۴-۱۷ خروجی اجتماع یابی شبکه بر اساس الگوریتم لووین مشاهده می‌شود. در شبکه به‌دست‌آمده مقدار ماژولاریتی ۰.۵۹هست. در این شکل هر اجتماع با یک رنگ مجزا مشخص‌شده است. بر اساس (Xie and Szymanski, 2011) ماژولاریتی بین ۰.۳ تا ۰.۷ نشان‌دهندهٔ ساختار قوی خوشه‌هاست. نتایج اجتماع یابی نشان می‌دهد در حوزهٔ پژوهش‌های این موضوع به ترتیب زیر حوزه‌های بیماری‌های مرتبط با دیابت (۲۶.۳۲٪) ، بیماری‌های قلبی (۱۳.۲۹٪) ، بیماری‌های خونی (۶.۳۱٪) و فشارخون (۵.۳۴٪) از منظر مسائل خطر ژنتیکی بیشترین پژوهش‌ها را به خود اختصاص داده است. نتایج به‌دست‌آمده در مقاله (لاجوردی, علوی و کارگری, ۱۳۹۹) منتشرشده است.



شکل ۱۸-۴ نمایی از شبکه ارتباط کلمات کلیدی در مقالات

٤-١٠-جمع‌بندی

نتایج کامل تحقیق در این فصل به ترتیب گفته‌شده در فصل قبل ارائه گردید و درنهایت نقشه علمی موضوع تحقیق ارائه شد. ضمناً در این فصل مقالاتی مستخرج از تز نیز ارائه شد و درنهایت در فصل بعدی به بیان نقاط قوت و ضعف تحقیق و همچنین کاربرد آن پرداخته می‌شود.

۵- بحث و نتیجه‌گیری

در این فصل به تحلیل نتایج به‌دست‌آمده در فصل قبل پرداخته می‌شود و درنهایت با جمع‌بندی مطالب مستند تز به اتمام می‌رسد.

۵-۱- بررسی مدل پیش‌بینی کننده خطر HTN مبتنی بر SNP

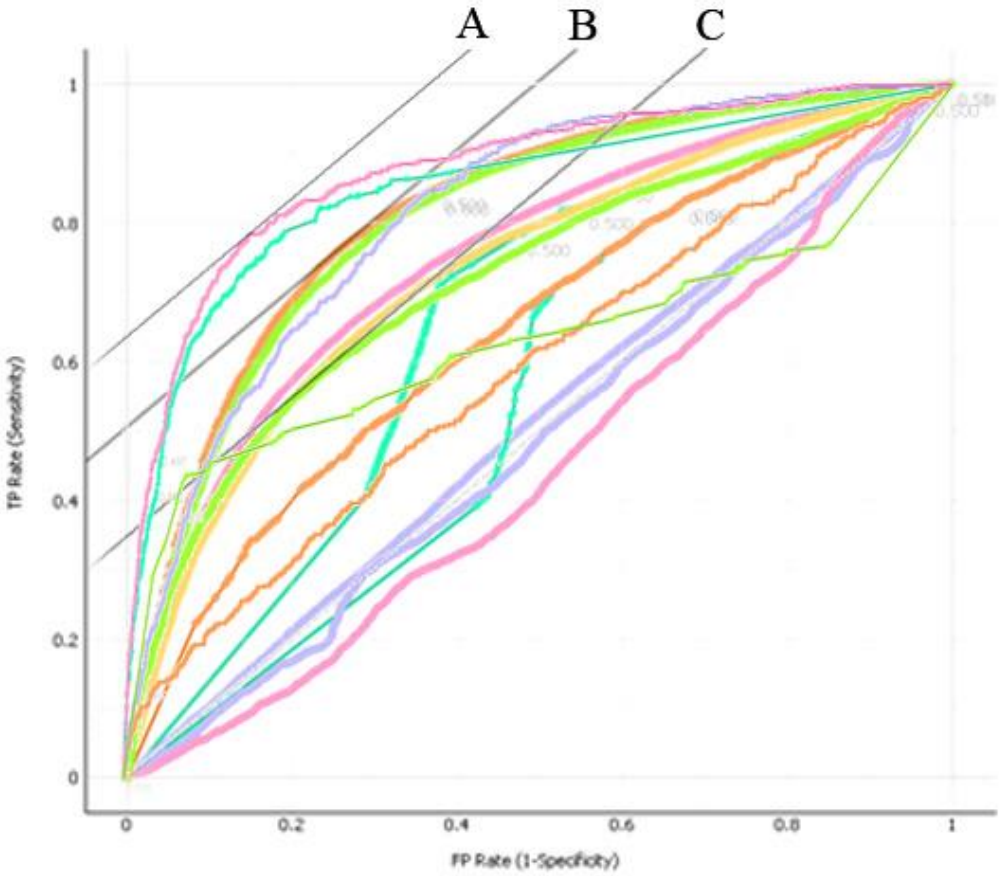
در مطالعه حاضر، بیش از ۴۸۰۰۰ برچسب فشارخون در شرکت‌کنندگان TCGS توسط پنج مدل مختلف یادگیری ماشینی و یادگیری عمیق مبتنی بر CNN ارزیابی و پیش‌بینی شد که نتیجه آن مقدار AUC ۰.۸۸۵ برای بهترین مدل یادگیری ماشین و مقدار ۰.۸۷۷ برای CNN را بدون در نظر گرفتن عوامل خطر اضافی نشان می‌دهد. در مقایسه با سایرین، علاوه بر اینکه این مدل بر اساس AUC از کیفیت بهتری برخوردار است، به دلیل عدم استفاده از سایر پارامترهای بیولوژیکی، ابزار ساده‌تری برای پیش‌بینی خطر است.

بالاترین AUC گزارش‌شده در ادبیات موضوع برای تشخیص بالینی فشارخون بالابر اساس عوامل بالینی فشارخون ۰.۷۵۷ است. مطالعات دیگر شامل عوامل خطر ژنتیکی و بالینی فشارخون بر اساس GRS است که خطر فشارخون را با AUC ۰.۷۸ و ۰.۸۵۴ پیش‌بینی می‌کند (Li *et al.*, 2020; Wu *et al.*, 2020; Niu *et al.*, 2021). بنابراین، یافته‌های این تحقیق عملکرد بهتری را به‌ویژه پس از در نظر گرفتن نشانگرهای ژنومی، جنسیت و سن نشان می‌دهد. همچنین با مقایسه نتایج مدل‌ها نشان داده می‌شود که SVM برای مدل‌سازی این داده‌ها نامناسب است و جنسیت تأثیر قابل‌توجهی بر فشارخون بالا ندارد.

امروزه، مهم‌ترین دستاورد عملی در عصر ژنومی، ایجاد ابزاری برای تشخیص زودهنگام خطر فشارخون بر اساس نشانگرهای ژنومی برای بهبود پیشگیری و درمان در افراد پرخطر است. در این تحقیق ابزاری اصلاح‌شده برای پیش‌بینی خطر فشارخون بالا در سنین پایین‌تر ارائه‌شده است.

۵-۲- مقایسه نتایج و معنی‌داری بهبود

بر اساس (Pepe *et al.*, 2013) برای نشان دادن معنی‌دار بودن بهبود در مدل‌های پیش‌بینی، بر اساس نمودار ROC مقایسه انجام می‌گردد و اگر در تمامی نقاط مساحت زیر نمودار بهتر باشد، نشان‌دهنده معنی‌داری بهبود دقت برای آن مدل است. در شکل ۵-۱ نمودارهای ROC مدل‌های CNN (A) و یادگیری ماشین مبتنی بر کاندید (B) و مبتنی بر امتیاز خطر ژنتیکی (C) آمده است. همان‌طور که در شکل نشان داده‌شده، مدل CNN در تمامی نقاط دارای عملکرد بهتری هست که نشان‌دهنده معنی‌داری بهبود هست.



شکل ۱-۵ مقایسه نمودارهای ROC

۳-۵- بررسی نتایج مدل‌سازی درخت تصمیم

با توجه به مدل‌سازی درخت تصمیم (شکل ۴-۱۵) بر روی‌داده‌ها نکات ذیل به دست می‌آید.

۱- اولین انشعاب در سن ۵۰ سالگی هست که نشان‌دهنده مرز مشخص ۵۰ سال برای افزایش خطر بیماری پرفشاری خون هست.

۲- در شاخه کم‌خطر، انشعاب بعدی در سن ۱۸ سالگی هست که مرز اولیه خطر پرفشاری خون را نشان می‌دهند.

۳- در شاخه پرخطر، انشعاب بعدی در سن ۵۹ سال است که مرز نهایی خطر پرفشاری خون است.

۴- تا سطح چهارم انشعاب بر روی سن هست که نشان می‌دهد سن مهم‌ترین عامل در بروز بیماری پرفشاری خون است.

۵- پس از سن عوامل ژنومی تأثیر بالاتری از جنسیت دارند و می‌توان گفت جنسیت تأثیر مستقیمی در بروز بیماری پرفشاری خون ندارد. این مسئله که جنسیت تأثیر مستقیمی ندارد در مقاله (Hengel, Benitah and Wenzel, 2022) نیز بیان‌شده است.

۶- تفاوت معنی‌داری در ژنوم های مؤثر بیماری پرفشاری خون در بین زنان و مردان نیست. این مسئله در جدول ۴-۵ آمده است که ۱۰ SNP بااهمیت بر روی کروموزوم‌های غیر جنسیتی هستند.

۵-٤- بررسی نتایج حاصل از رتبه‌بندی SNP

در میان نشانگرهای قبلی مرتبط با فشارخون، ده SNP برتر در این مطالعه رتبه‌بندی شده و نتایج نشان داد که SNP های برتر بالاترین p-value را ندارند.

یافته‌های مبتنی بر رتبه‌بندی نشان می‌دهد که برخلاف نتایج قبلی بر اساس p-value، باید نشانگرها را بر اساس ارتباط رتبه‌بندی کنیم (Sullivan and Feinn, 2012). البته لازم به ذکر است که SNPهایی را باید در نظر گرفت که دارای سطح معنی‌داری خوبی با فشارخون بالا هستند.

۵-۵- خطاهای مؤثر در تحقیق

با توجه به اینکه تحقیق و مدل‌سازی انجام‌شده بر مبنای داده هست، ممکن است خطاهای مختلفی در حین مدل‌سازی وجود داشته باشد.

۱- خطای ناشی از اندازه‌گیری رخ‌نمود فشارخون: این خطا در هنگام اندازه‌گیری فشارخون ممکن است رخ دهد. با توجه به توضیحات نحوه اندازه‌گیری این رخ‌نمود، اندازه فشارخون توسط دستگاه فشارسنج جیوه‌ای و پس از استراحت کافی هست که در دو مرحله این مقدار سنجیده شده و درنهایت میانگین آن گزارش‌شده است که دارای حداقل خطا هست.

۲- خطای ناشی از شناسایی SNP درروش NGS: با توجه به تحقیق (Petrackova *et al.*, 2019) خطای فرایند NGS کمتر از ۱ درصد هست (مابین ۰.۱ درصد الی ۱ درصد). این خطا را می‌توان با افزایش حداقل عمق پوشش در فرایند NGS کاهش داد.

۳- خطای ناشی از جنس داده و نوع مدل: در مدل‌سازی‌های انجام‌شده مدل SVM برخلاف تحقیقات مختلف دارای رتبه پایین‌تری نسبت به شبکه عصبی است. دلایل این کاهش کیفیت شامل: بزرگ بودن داده و نداشتن تابع فعال‌سازی مناسب، نویزی بودن داده. در این تحقیق توابع فعال‌سازی مختلفی برای SVM پیاده گردید اما نتیجه مناسب نبود و البته وجود نویز بالا در داده‌ها نیز مزید بر علت بود. این نویز در داده‌های ژنومی با توجه به نرخ خطای فرایند NGS ایجادشده است.

۵-٦- پاسخ به سؤالات تحقیق

سؤالات اصلی تحقیق به‌صورت زیر هست:

- ✓ آیا می‌توان از روش‌های یادگیری ماشین در پیش‌بینی خطر پرفشاری خون مبتنی بر SNP به شکل مؤثرتر از روش‌های موجود بهره برد؟ بله. با توجه به روش‌های استفاده‌شده نشان داده‌شده است که یادگیری ماشین می‌تواند باعث افزایش دقت و کیفیت پیش‌بینی خطر پرفشاری خون باشد.

سؤال فرعی تحقیق نیز به‌صورت زیر هست:

- ✓ آیا می‌توان اطلاعات جمعیت شناختی و داده‌های طولی رخ‌نمود را در بهبود مدل پیش‌بینی کننده استفاده کرد؟ بر اساس داده‌های استفاده‌شده که در شش دوره هست، این امکان ایجاد شد که بتوانیم مقایسه

مناسبی ازلحاظ تأثیر سنی و همچنین وضعیت فرد در سال‌های بعد داشته باشیم که این امر در مدل‌سازی استفاده‌شده است.

- ✓ با توجه به مدل ارائه‌شده، آیا می‌توان SNPهای مهم ژنومی را استخراج کرد؟ روش یادگیری ماشین به‌صورت جعبه سیاه هست و تشخیص SNP بااهمیت دچار مشکل است و بر این اساس از روش‌های رتبه‌بندی به جهت سنجش اهمیت برای SNPهایی که معنی‌داری ارتباط با HTN آن‌ها توسط GWAS بیان‌شده بود، استفاده گردید که نتایج آن به‌صورت یک ابزار پیش‌بینی خطر پیاده‌سازی گردید.

۵-۷- نوآوری‌های تحقیق

در این تحقیق با توجه به مدل‌سازی‌های انجام‌شده می‌توان نوآوری‌های به‌دست‌آمده را به‌صورت ذیل بیان کرد:

۱- با توجه به اهمیت بیماری پرفشاری خون که در حال حاضر بیش از یک میلیارد نفر را در جهان درگیر کرده است، ایجاد یک ابزار که بتواند در سنین جوانی خطر بیماری را پیش‌بینی کند و پیش‌آگهی بدهد ضروری است که این امر باعث افزایش مراقبت‌ها و سبک زندگی افرادی است که دارای خطر بالایی می‌باشند. در این تحقیق بر اساس مدل‌های یادگیری ماشین و تنها داده ژنومی، قبل از بروز عوامل دیگر مانند چاقی، پیش‌آگهی مناسبی داده می‌شود که می‌تواند توسط پزشکان استفاده گردد.

۲- فرد می‌تواند قبل از آنکه دچار بیماری پرفشاری خون شود با سنجش خطر در سنین پایین، سبک زندگی مناسب‌تری را انتخاب کند که مانع از درگیری با بیماری پرفشاری خون می‌شود.

۳- نوع مدل‌سازی انجام‌شده که حاصل ادغام داده‌های طولی در یک مدل یادگیری ماشین هست نوآوری مهم این تحقیق است. البته ایجاد تصویر از داده‌های ژنومی جهت اجرا در مدل CNN نیز یکی از نوآوری‌هاست که باعث کاهش حجم پَروَنجا ژنوم نیز می‌گردد.

۴- به جهت استفاده، ابزاری اولیه ایجادشده است که جزو نوآوری‌های تحقیق است که در بخش ۴-۸ اشاره گردید. این ابزار به‌سادگی می‌تواند با تعداد محدود SNP مؤثر، خطر بیماری را در سنین مختلف محاسبه نماید.

۵- یکی از نتایج تحقیق عدم تأثیرگذاری جنسیت درخطر بیماری پرفشاری خون است که در تحقیقات جدید نیز به این مسئله اشاره‌شده است (Hengel, Benitah and Wenzel, 2022). آنچه باعث افزایش خطر می‌شود، افزایش استرس است که معمولاً در مردان بیشتر است و افزایش استرس نیز عوامل مختلفی دارد من‌جمله عوامل ژنومی.

۶- یکی از خطاهای مهم درروش امتیاز خطر چندژنی، خطای تأثیر کوچک[1] هست. در این خطا به دلیل اینکه تک‌تک نوکلئوتیدها دارای p-value کمتر از حد تعریف‌شده هستند از مدل‌سازی خارج می‌شود. این مسئله در صورتی است که تأثیر چند تک نوکلئوتید باهم دیگر می‌تواند عامل مؤثر در بروز رخ‌نمود باشد. در این تحقیق با توجه به در نظر گرفتن تمامی ژنوم، این مسئله مرتفع شده است.

[1] Small Effect

۵-۸- محدودیت‌ها و پیشنهادهای آتی

این تحقیق به دلیل استفاده از اعتبارسنجی داده‌های داخلی دارای محدودیت‌هایی است. نتایج این مطالعه خارج از داده‌های TCGS برای اعتبار سنجی خارجی استفاده‌نشده است و همچنین داده‌های مورداستفاده برای ارزیابی محدود به داده‌های جمع‌آوری‌شده در TCGS است که ترکیب قومی کمی دارد.

برای پیشنهادهای آتی، می‌توان متغیرهای بالینی مؤثر مانند نژاد را وارد مدل کرد. البته باید دقت کرد که تنها متغیرهایی وارد مدل شوند که وابستگی به سن نداشته باشند.

همچنین می‌توان برای مدل‌سازی بهتر فشارخون به‌صورت مجزا مدل‌سازی IDH و ISH را نیز انجام داد.

برای مدل‌سازی بهتر می‌توان از الگوریتم‌های LightGBM و XGBoost و Ensembleهای دیگر نیز استفاده و نتایج آن را با نتایج مدل‌های ارائه‌شده مقایسه کرد. همچنین می‌توان وضعیت فرد در دوره قبلی را به‌عنوان یکی از ورودی‌های دوره جدید قرار داد و بر اساس آن با پنجره زمانی مدل‌سازی برای پیش‌آگهی انجام داد.

یکی از پیشنهادهای مهم می‌تواند بررسی الگوی ژنومی فشارخون بر اساس داده‌های بومی باشد که در این تحقیق نسخه اولیه آن انجام شد و در ادامه می‌توان آن را تکمیل و گزارش کرد.

۵-۹- خلاصه فصل

در این فصل به بررسی یافته‌های تحقیق، نتیجه‌گیری و جمع‌بندی نتایج پرداختیم. علاوه بر این محدودیت‌ها و پیشنهادهای پژوهشی آینده بیان گردید.

فهرست منابع

Adams, F. (1922) ‘The Genuine Works of Hippocrates’, *Southern Medical Journal*. Southern Medical Association, 15(6), p. 519. doi: 10.1097/00007611-192206000-00025.

Alizadehsani, R. *et al.* (2019) ‘Machine learning-based coronary artery disease diagnosis: A comprehensive review’, *Computers in Biology and Medicine*. Elsevier Ltd, 111(May), p. 103346. doi: 10.1016/j.compbiomed.2019.103346.

Anis, A., Fahiem, M. A. and Tauseef, H. (2016) ‘A Classification Approach Based on Genetic-Data-Structuring for the Prediction of Hypertension’, 4(4), pp. 236–242.

Aulchenko, Y. S., De Koning, D.-J. J. and Haley, C. (2007) ‘Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis’, *Genetics*. Oxford University Press (), 177(1), pp. 577–585. doi: 10.1534/genetics.107.075614.

Azizi, F. *et al.* (2009) ‘Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase {II}’, *Trials*. Springer Science and Business Media , 10(1). doi: 10.1186/1745-6215-10-5.

de Bakker, P. I. W. *et al.* (2008) ‘Practical aspects of imputation-driven meta-analysis of genome-wide association studies’, *Human Molecular Genetics*. Oxford University Press (), 17(R2), pp. R122--R128. doi: 10.1093/hmg/ddn288.

Basile, J. *et al.* (2019) ‘Overview of hypertension in adults - UpToDate’, *Up To Date*, pp. 1–56. Available at: https://www.uptodate.com/contents/overview-of-hypertension-in-adults?search=hypertension&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1%0Ahttps://www-uptodate-com.puce.idm.oclc.org/contents/overview-of-hypertension-in-adults?se.

Bauer, C. *et al.* (2015) ‘{BioMiner}: Paving the Way for Personalized Medicine’, *Cancer Informatics*. {SAGE} Publications, 14, p. CIN.S20910. doi: 10.4137/cin.s20910.

Blondel, V. D. *et al.* (2008) ‘Fast unfolding of communities in large networks’, *Journal of statistical mechanics: theory and experiment*. IOP Publishing, 2008(10), p. P10008.

Breiman, L. (2000) ‘Randomizing outputs to increase prediction accuracy’, *Machine Learning*. Kluwer Academic Publishers, 40(3), pp. 229–242. doi: 10.1023/A:1007682208299.

Bush, W. S. (2019) ‘Genome-Wide Association Studies’, in *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, pp. 235–241. doi: 10.1016/B978-0-12-809633-8.20232-X.

Çalışkan, M., Erol, O. and Öz, G. C. (2020) *The Recent Topics in Genetic Polymorphisms*. Edited by M. Çalışkan, O. Erol, and G. Cevahir Öz. Rijeka: IntechOpen. doi: 10.5772/intechopen.77777.

Cardon, L. R. and Bell, J. I. (2001) ‘Association study designs for complex diseases’, *Nature Reviews Genetics*. Springer Science and Business Media , 2(2), pp. 91–99. doi: 10.1038/35052543.

Cebamanos, L. *et al.* (2014) ‘Regional heritability advanced complex trait analysis for {GPU} and traditional parallel architectures’, *Bioinformatics*. Oxford University Press (), 30(8), pp. 1177–1179. doi: 10.1093/bioinformatics/btt754.

Chowdhury, B. and Garai, G. (2017) ‘A review on multiple sequence alignment from the perspective of genetic algorithm’, *Genomics*. Academic Press, 109(5–6), pp. 419–431. doi: 10.1016/J.YGENO.2017.06.007.

Chowdhury, M. Z. I. *et al.* (2022) ‘Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis’, *PLOS ONE*. Public Library of Science, 17(4), p. e0266334. doi: 10.1371/JOURNAL.PONE.0266334.

Cios, K. J. and William Moore, G. (2002) ‘Uniqueness of medical data mining’, *Artificial Intelligence in Medicine*, 26(1–2), pp. 1–24. doi: 10.1016/S0933-3657(02)00049-0.

Daneshpour, M. S. *et al.* (2017) ‘Rationale and Design of a Genetic Study on Cardiometabolic Risk Factors: Protocol for the Tehran Cardiometabolic Genetic Study (TCGS)’, *JMIR Research Protocols*. {JMIR} Publications Inc., 6(2), p. e28. doi: 10.2196/resprot.6050.

Davies, G. *et al.* (2011) ‘Genome-wide association studies establish that human intelligence is highly heritable and polygenic’, *Molecular Psychiatry*. Springer Science and Business Media , 16(10), pp. 996–1005. doi: 10.1038/mp.2011.85.
Diamandis, E. P. (2012) ‘Biomarker validation is still the bottleneck in biomarker research’, *Journal of Internal Medicine*. Wiley, 272(6), p. 620. doi: 10.1111/j.1365-2796.2012.02579.x.
Eshmurodova Dilrabo Khudayberdi kizi, Ergasheva Iroda Akbarali kizi, K. F. N. and Siab (2023) ‘ASSESSMENT OF POLYGENIC RISK OF ARTERIAL HYPERTENSION’, in *SCIENTIFIC APPROACH TO THE MODERN EDUCATION SYSTEM*, pp. 39–43. Available at: www.interonconf.com.
Esteva, A. *et al.* (2019) ‘A guide to deep learning in healthcare’, *Nature Medicine*. Springer Science and Business Media , 25(1), pp. 24–29. doi: 10.1038/s41591-018-0316-z.
Evangelou, E. and Ioannidis, J. P. A. A. (2013) ‘Meta-analysis methods for genome-wide association studies and beyond’, *Nature Reviews Genetics*. Springer Science and Business Media , 14(6), pp. 379–389. doi: 10.1038/nrg3472.
Farris, J. *et al.* (2021) ‘Genomic analyses of glycine decarboxylase neurogenic mutations yield a large-scale prediction model for prenatal disease’, *PLOS Genetics*. Edited by G. S. Barsh, 17(2), p. e1009307. doi: 10.1371/journal.pgen.1009307.
Fawcett, T. (2006) ‘An introduction to {ROC} analysis’, *Pattern Recognition Letters*. Elsevier , 27(8), pp. 861–874. doi: 10.1016/j.patrec.2005.10.010.
Filshtein, T. J. *et al.* (2019) ‘Reserve and Alzheimer’s disease genetic risk: Effects on hospitalization and mortality’, *Alzheimer’s and Dementia*. Elsevier Inc., 15(7), pp. 907–916. doi: 10.1016/j.jalz.2019.04.005.
Ganesalingam, J. and Bowser, R. (2010) ‘The application of biomarkers in clinical trials for motor neuron disease’, *Biomarkers in Medicine*. Future Medicine Ltd, 4(2), pp. 281–297. doi: 10.2217/bmm.09.71.
Garimella, P. S. *et al.* (2023) ‘A genomic deep field view of hypertension’, *Kidney International*, 103(1), pp. 42–52. doi: 10.1016/j.kint.2022.09.029.
Gibson, G. (2010) ‘Hints of hidden heritability in {GWAS}’, *Nature Genetics*. Springer Science and Business Media , 42(7), pp. 558–560. doi: 10.1038/ng0710-558.
Gilmour, A. R., Thompson, R. and Cullis, B. R. (1995) ‘Average Information {REML}: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models’, *Biometrics*. JSTOR, 51(4), p. 1440. doi: 10.2307/2533274.
Ginsburg, G. S., Donahue, M. P. and Newby, L. K. (2005) ‘Prospects for Personalized Cardiovascular Medicine’, *Journal of the American College of Cardiology*. Elsevier , 46(9), pp. 1615–1627. doi: 10.1016/j.jacc.2005.06.075.
Ginsburg, G. S. and Willard, H. F. (2009) ‘Genomic and personalized medicine: foundations and applications’, *Translational Research*. Elsevier , 154(6), pp. 277–287. doi: 10.1016/j.trsl.2009.09.005.
Giontella, A. *et al.* (2021) ‘Clinical Evaluation of the Polygenetic Background of Blood Pressure in the Population-Based Setting’, *Hypertension*, 77(1), pp. 169–177. doi: 10.1161/HYPERTENSIONAHA.120.15449.
Gray, A., Stewart, I. and Tenesa, A. (2012) ‘Advanced Complex Trait Analysis’, *Bioinformatics*. Oxford University Press (), 28(23), pp. 3134–3136. doi: 10.1093/bioinformatics/bts571.
Guns, R., Liu, Y. X. and Mahbuba, D. (2011) ‘Q-measures and betweenness centrality in a collaboration network: a case study of the field of informetrics’, *Scientometrics*. Springer, 87(1), pp. 133–147.
He, Q. (1999) ‘Knowledge discovery through co-word analysis’. Graduate School of Library and Information Science. University of Illinois~….
Hebbring, S. (2019) ‘Genomic and Phenomic Research in the 21st Century’, *Trends in Genetics*. Elsevier , 35(1), pp. 29–41. doi: 10.1016/j.tig.2018.09.007.
Hengel, F. E., Benitah, J.-P. and Wenzel, U. O. (2022) ‘Mosaic theory revised: inflammation and salt play central roles in arterial hypertension’, *Cellular & Molecular Immunology*. Springer

US, 19(5), pp. 561–576. doi: 10.1038/s41423-022-00851-8.
Heo, B. M. (2018) ‘Prediction of Prehypertenison and Hypertension Based on Anthropometry , Blood Parameters , and Spirometry’. doi: 10.3390/ijerph15112571.
Hernández-Rodríguez, S. *et al.* (2016) ‘A recommender system applied to the indirect materials selection process (RS-IMSP) for producing automobile spare parts’, *Computers in Industry*. Elsevier B.V., 82, pp. 233–244. doi: 10.1016/j.compind.2016.07.004.
Hirschhorn, J. N. and Daly, M. J. (2005) ‘Genome-wide association studies for common diseases and complex traits’, *Nature Reviews Genetics*. Springer Science and Business Media , 6(2), pp. 95–108. doi: 10.1038/nrg1521.
Ho, D. S. W. *et al.* (2019) ‘Machine learning SNP based prediction for precision medicine’, *Frontiers in Genetics*, 10(MAR), pp. 1–10. doi: 10.3389/fgene.2019.00267.
Hordri, N. F. *et al.* (2017) ‘A systematic literature review on features of deep learning in big data analytics’, *International Journal of Advances in Soft Computing and its Applications*, 9(1), pp. 32–49.
Jain, K. K. (2006) ‘A critical review of the Royal Society{\textquotesingle}s report on personalized medicine’, *Drug Discovery Today*. Elsevier , 11(13–14), pp. 573–575. doi: 10.1016/j.drudis.2006.05.005.
Khan, A. *et al.* (2020) ‘A survey of the recent architectures of deep convolutional neural networks’, *Artificial Intelligence Review*. Springer Science and Business Media , 53(8), pp. 5455–5516. doi: 10.1007/s10462-020-09825-6.
Kheradmand, M. *et al.* (2015) *Population Based Cohort Studies in Iran: A Review Article*, *Journal of Mazandaran University of Medical Sciences*. Journal of Mazandaran University of Medical Sciences. Available at: http://jmums.mazums.ac.ir/article-1-5714-en.html (Accessed: 14 February 2021).
Khorraminezhad, L. *et al.* (2020) ‘Statistical and Machine-Learning Analyses in Nutritional Genomics Studies’, *Nutrients*, 12(10), p. 3140. doi: 10.3390/nu12103140.
Kim, H. and Andrade, F. C. D. (2019) ‘Diagnostic status and age at diagnosis of hypertension on adherence to lifestyle recommendations’, *Preventive Medicine Reports*. Elsevier, 13(October 2018), pp. 52–56. doi: 10.1016/j.pmedr.2018.11.005.
Kolifarhood, G., Daneshpour, M. S., *et al.* (2019) ‘Generality of genomic findings on blood pressure traits and its usefulness in precision medicine in diverse populations: A systematic review’, *Clinical Genetics*, 96(1), pp. 17–27. doi: 10.1111/cge.13527.
Kolifarhood, G., Daneshpour, M., *et al.* (2019) ‘Heritability of blood pressure traits in diverse populations: a systematic review and meta-analysis’, *Journal of Human Hypertension*. Springer US, 33(11), pp. 775–785. doi: 10.1038/s41371-019-0253-4.
Kolifarhood, G. *et al.* (2021) ‘Genome-wide association study on blood pressure traits in the Iranian population suggests ZBED9 as a new locus for hypertension’, *Scientific Reports*. Nature Publishing Group UK, 11(1), pp. 1–13. doi: 10.1038/s41598-021-90925-w.
Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017) ‘Imagenet classification with deep convolutional neural networks’, *Communications of the ACM*. AcM New York, NY, USA, 60(6), pp. 84–90.
Kurniansyah, N. *et al.* (2022) ‘A multi-ethnic polygenic risk score is associated with hypertension prevalence and progression throughout adulthood’, *Nature Communications*. Springer US, 13(1), p. 3549. doi: 10.1038/s41467-022-31080-2.
Kurtz, T. W. and Spence, M. A. A. (1993) ‘Genetics of essential hypertension’, *The American Journal of Medicine*. Elsevier , 94(1), pp. 77–84. doi: 10.1016/0002-9343(93)90124-8.
Lajevardi, S Ali *et al.* (2022) ‘A CNN deep learning model to improve SNP-based hypertension risk prediction accuracy’.
Lajevardi, S. Ali *et al.* (2022) ‘Hypertension risk prediction based on SNPs by machine learning models’, *Current Bioinformatics*. Bentham Science Publishers Ltd., 17. doi: 10.2174/1574893617666221011093322.
Lajevardy, S. A. and Kargari, M. (2022) ‘Developing new genetic algorithm based on integer

programming for multiple sequence alignment', *Soft Computing 2022 26:8*. Springer, 26(8), pp. 3863–3870. doi: 10.1007/S00500-022-06790-W.
Lecun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*. Springer Science and Business Media , 521(7553), pp. 436–444. doi: 10.1038/nature14539.
Li, C. *et al.* (2019) 'A prediction model of essential hypertension based on genetic and environmental risk factors in northern han chinese', *International Journal of Medical Sciences*, 16(6), pp. 793–799. doi: 10.7150/ijms.33967.
Li, S. *et al.* (2020) 'Genetic risk scores to predict the prognosis of chronic heart failure patients in Chinese Han', *Journal of Cellular and Molecular Medicine*, 24(1), pp. 285–293. doi: 10.1111/jcmm.14722.
Libbrecht, M. W. and Noble, W. S. (2015) 'Machine learning applications in genetics and genomics', *Nature Reviews Genetics*. Springer Science and Business Media , 16(6), pp. 321–332. doi: 10.1038/nrg3920.
Ligon, A. H. *et al.* (2000) 'Identification of female carriers for Duchenne and Becker muscular dystrophies using a {FISH}-based approach', *European Journal of Human Genetics*. Springer Science and Business Media , 8(4), pp. 293–298. doi: 10.1038/sj.ejhg.5200450.
Loganathan, L. *et al.* (2019) 'Computational and Pharmacogenomic Insights on Hypertension Treatment: Rational Drug Design and Optimization Strategies', *Current Drug Targets*, 21(1), pp. 18–33. doi: 10.2174/1389450120666190808101356.
Mahajan, S. *et al.* (2019) 'Prevalence, Awareness, and Treatment of Isolated Diastolic Hypertension: Insights From the China PEACE Million Persons Project', *Journal of the American Heart Association*, 8(19), pp. 1–17. doi: 10.1161/JAHA.119.012954.
Maher, B. (2008) 'Personal genomes: The case of the missing heritability', *Nature*. Springer Science and Business Media , 456(7218), pp. 18–21. doi: 10.1038/456018a.
Mahrokh, L. *et al.* (2017) 'An efficient microbial fuel cell using a {CNT}{\textendash}{RTIL} based nanocomposite', *Journal of Materials Chemistry A*. Royal Society of Chemistry ({RSC}), 5(17), pp. 7979–7991. doi: 10.1039/c6ta09766a.
Maj, C. *et al.* (2022) 'Dissecting the Polygenic Basis of Primary Hypertension: Identification of Key Pathway-Specific Components', *Frontiers in Cardiovascular Medicine*, 9(February), pp. 1–10. doi: 10.3389/fcvm.2022.814502.
Maltby, D. (2011) 'Big Data Analytics', *ASIST*.
Marcos-Zambrano, L. J. *et al.* (2021) 'Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment', *Frontiers in Microbiology*, 12(February). doi: 10.3389/fmicb.2021.634511.
Martinez-Ríos, E. *et al.* (2021) 'A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data', *Biomedical Signal Processing and Control*. Elsevier Ltd, 68(March), p. 102813. doi: 10.1016/j.bspc.2021.102813.
Mattoo, T. K. (2019) 'Definition and diagnosis of hypertension in children and adolescents - UpToDate', *UpToDate*, (Cv), pp. 1–34. Available at: https://www.uptodate.com/contents/definition-and-diagnosis-of-hypertension-in-children-and-adolescents?search=tension arterial&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1#H12 (Accessed: 14 February 2021).
McCarthy, M. I. *et al.* (2008) 'Genome-wide association studies for complex traits: Consensus, uncertainty and challenges', *Nature Reviews Genetics*. Springer Science and Business Media , 9(5), pp. 356–369. doi: 10.1038/nrg2344.
McKusicK, V. A. *et al.* (1960) 'Medical genetics 1959', *Journal of Chronic Diseases*. Elsevier, 12(1), pp. 1–202. doi: 10.1016/0021-9681(60)90134-X.
McKuSICK, V. A. (1960) 'Genetics and the Nature of Essential Hypertension', *CircuaIation*, XXII.
Montañez, C. A. C., Fergus, P. and Chalmers, C. (2018) 'Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs', in *2018 International Joint*

Conference on Neural Networks (IJCNN).
Mosley, J. D. *et al.* (2020) 'Predictive Accuracy of a Polygenic Risk Score Compared with a Clinical Risk Score for Incident Coronary Heart Disease', *JAMA - Journal of the American Medical Association*, 323(7), pp. 627–635. doi: 10.1001/jama.2019.21782.
Ng, H. T., Goh, W. B. and Low, K. L. (1997) 'Feature selection, perceptron learning, and a usability case study for text categorization', *{ACM} {SIGIR} Forum*. Association for Computing Machinery ({ACM}), 31(SI), pp. 67–73. doi: 10.1145/278459.258537.
Niu, M. *et al.* (2021) 'Identifying the predictive effectiveness of a genetic risk score for incident hypertension using machine learning methods among populations in rural China', *Hypertension Research*. Springer US. doi: 10.1038/s41440-021-00738-7.
Nour, M. and Polat, K. (2020) 'Automatic Classification of Hypertension Types Based On Personal Features By Machine Learning Algorithm', *Mathematical Problems in Engineering*. Hindawi Limited, 2020, pp. 1–13. doi: 10.1155/2020/2742781.
Noyons, E. C. M. (1999) 'Bibliometric mapping as a science policy and research management tool'. DSWO Press.
Padmanabhan, S. and Dominiczak, A. F. (2021) 'Genomics of hypertension: the road to precision medicine', *Nature Reviews Cardiology*. Springer US, 18(4), pp. 235–250. doi: 10.1038/s41569-020-00466-4.
Pattin, K. A. and Moore, J. H. (2009) 'Role for protein{\textendash}protein interaction databases in human genetics', *Expert Review of Proteomics*. Informa {UK} Limited, 6(6), pp. 647–659. doi: 10.1586/epr.09.86.
Pepe, M. S. *et al.* (2013) 'Testing for improvement in prediction model performance', *Statistics in Medicine*, 32(9), pp. 1467–1482. doi: 10.1002/sim.5727.
Pérez-Enciso, M. *et al.* (2019) 'A Guide for Using Deep Learning for Complex Trait Genomic Prediction', *Genes*. MDPI AG, 10(7), p. 553. doi: 10.3390/genes10070553.
Petrackova, A. *et al.* (2019) 'Standardization of Sequencing Coverage Depth in NGS: Recommendation for Detection of Clonal and Subclonal Mutations in Cancer Diagnostics', *Frontiers in Oncology*, 9(September), pp. 1–6. doi: 10.3389/fonc.2019.00851.
Phulka, J. S. *et al.* (2023) 'Current State and Future of Polygenic Risk Scores in Cardiometabolic Disease: A Scoping Review', *Circulation: Genomic and Precision Medicine*. doi: 10.1161/CIRCGEN.122.003834.
Powers, D. M. W. and Ailab (2011) 'EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION', *International Journal of Machine Learning Technology*, pp. 37–63.
Quinlan, J. R. (1986) 'Induction of decision trees', *Machine Learning*. Springer Science and Business Media , 1(1), pp. 81–106. doi: 10.1007/bf00116251.
Rafiei, A. and Amjadi, O. (2013) 'Personalized medicine; a bridge between current medicine and the future healthcare', *Journal of Clinical Excellence*, 1(2), pp. 47–68.
Ramezankhani, A. *et al.* (2016) 'Classi fi cation-based data mining for identi fi cation of risk patterns associated with hypertension in Middle Eastern population', 0(June).
Reddy, B. and Fields, R. (2022) 'Multiple Sequence Alignment Algorithms in Bioinformatics', *Lecture Notes in Networks and Systems*. Springer Science and Business Media Deutschland GmbH, 286, pp. 89–98. doi: 10.1007/978-981-16-4016-2_9/COVER.
Rip, A. and Courtial, J. (1984) 'Co-word maps of biotechnology: An example of cognitive scientometrics', *Scientometrics*. Akadémiai Kiadó, co-published with Springer Science+ Business Media BV~…, 6(6), pp. 381–400.
Risch, N. and Merikangas, K. (1996) 'The Future of Genetic Studies of Complex Human Diseases', *Science*. American Association for the Advancement of Science ({AAAS}), 273(5281), pp. 1516–1517. doi: 10.1126/science.273.5281.1516.
Ruiz, M. E. and Srinivasan, P. (1999) 'Hierarchical neural networks for text categorization (poster abstract)', in *Proceedings of the 22nd annual international {ACM} {SIGIR} conference on Research and development in information retrieval - {SIGIR} {\textquotesingle}99*. {ACM}

Press. doi: 10.1145/312624.312700.
Runcie, D. E. and Crawford, L. (2018) ‘Fast and flexible linear mixed models for genome-wide genetics’, *PLoS Genetics*. Cold Spring Harbor Laboratory, 15(2), pp. 1–24. doi: 10.1101/373902.
Russakovsky, O. *et al.* (2015) ‘{ImageNet} Large Scale Visual Recognition Challenge’, *International Journal of Computer Vision*. Springer Science and Business Media , 115(3), pp. 211–252. doi: 10.1007/s11263-015-0816-y.
Russom, P. (2018) ‘Introduction to Big Data Analytics’, in *Smart Grid Technology*. Cambridge University Press, pp. 38–48. doi: 10.1017/9781108566506.005.
Satyanarayana, N. *et al.* (2020) ‘An intelli AFM: An intelligent association based fuzzy rule miner to predict high blood pressure using bio-psychological factors’, *Intelligent Decision Technologies*. Edited by K. Chrysafiadi, 14(2), pp. 227–237. doi: 10.3233/IDT-190156.
Shah, S. *et al.* (2015) ‘Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations’, *American Journal of Human Genetics*. Cell Press, 97(1), pp. 75–85. doi: 10.1016/j.ajhg.2015.05.014.
Shrestha, A. and Mahmood, A. (2019) ‘Review of deep learning algorithms and architectures’, *IEEE access*. IEEE, 7, pp. 53040–53065.
Silver, D. *et al.* (2016) ‘Mastering the game of Go with deep neural networks and tree search’, *Nature*. Springer Science and Business Media , 529(7587), pp. 484–489. doi: 10.1038/nature16961.
Singh, D., Verma, S. and Singla, J. (2020) ‘A Comprehensive Review of Intelligent Medical Diagnostic Systems’, in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. IEEE, pp. 977–981. doi: 10.1109/ICOEI48184.2020.9143043.
Singh, M. *et al.* (2016) ‘Molecular genetics of essential hypertension’, 1963(April). doi: 10.3109/10641963.2015.1116543.
Skol, A. D. *et al.* (2007) ‘Optimal designs for two-stage genome-wide association studies’, *Genetic Epidemiology*. Wiley, 31(7), pp. 776–788. doi: 10.1002/gepi.20240.
Stehman, S. V (1997) ‘Selecting and interpreting measures of thematic classification accuracy’, *Remote Sensing of Environment*. Elsevier , 62(1), pp. 77–89. doi: 10.1016/s0034-4257(97)00083-7.
Strachan, T. and Read, A. P. (2018) ‘Human population genetics’, in *Human Molecular Genetics*. Garland Science, pp. 397–418. doi: 10.1201/9780429448362-12.
Sullivan, G. M. and Feinn, R. (2012) ‘Using Effect Size—or Why the P Value Is Not Enough’, *Journal of Graduate Medical Education*, 4(3), pp. 279–282. doi: 10.4300/JGME-D-12-00156.1.
Sun, Y. *et al.* (2019) ‘Identification of 12 cancer types through genome deep learning’, *Scientific Reports*. Cold Spring Harbor Laboratory, 9(1). doi: 10.1101/528216.
Szymczak, S. *et al.* (2009) ‘Machine learning in genome-wide association studies’, *Genetic Epidemiology*. Wiley, 33(S1), pp. S51--S57. doi: 10.1002/gepi.20473.
Tohidi, M. *et al.* (2012) ‘Triglycerides and triglycerides to high-density lipoprotein cholesterol ratio are strong predictors of incident hypertension in Middle Eastern women’, *Journal of Human Hypertension*. Nature Publishing Group, 26(9), pp. 525–532. doi: 10.1038/jhh.2011.70.
Torkamani, A., Wineinger, N. E. and Topol, E. J. (2018) ‘The personal and clinical utility of polygenic risk scores’, *Nature Reviews Genetics*. Springer US, 19(9), pp. 581–590. doi: 10.1038/s41576-018-0018-x.
Trerotola, M. *et al.* (2015) ‘Epigenetic inheritance and the missing heritability’, *Human Genomics*. Springer Science and Business Media , 9(1). doi: 10.1186/s40246-015-0041-3.
Visscher, P. *et al.* (2012) ‘Five Years of {GWAS} Discovery’, *The American Journal of Human Genetics*. Elsevier , 90(1), pp. 7–24. doi: 10.1016/j.ajhg.2011.11.029.
Visscher, P. M. *et al.* (2014) ‘Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using {SNP} Data in Unrelated Samples’, *{PLoS} Genetics*. Edited by G. S. Barsh. Public

Library of Science ({PLoS}), 10(4), p. e1004269. doi: 10.1371/journal.pgen.1004269.
Waezizadeh, T. *et al.* (2018) 'Mathematical models for the effects of hypertension and stress on kidney and their uncertainty', *Mathematical Biosciences*. Elsevier , 305, pp. 77–95. doi: 10.1016/j.mbs.2018.08.013.
Wang, B. *et al.* (2020) 'Prediction model and assessment of probability of incident hypertension: the Rural Chinese Cohort Study', *Journal of Human Hypertension*. Springer US. doi: 10.1038/s41371-020-0314-8.
WATSON, J. D. and CRICK, F. H. C. (1953) 'Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid', *Nature*. Springer Science and Business Media , 171(4356), pp. 737–738. doi: 10.1038/171737a0.
Willer, C. J., Li, Y. and Abecasis, G. R. (2010) '{METAL}: fast and efficient meta-analysis of genomewide association scans', *Bioinformatics*. Oxford University Press (), 26(17), pp. 2190–2191. doi: 10.1093/bioinformatics/btq340.
Wu, X. *et al.* (2020) 'Value of a Machine Learning Approach for Predicting Clinical Outcomes in Young Patients With Hypertension', *Hypertension*. Ovid Technologies (Wolters Kluwer Health), 75(5), pp. 1271–1278. doi: 10.1161/hypertensionaha.119.13404.
Xie, J. and Szymanski, B. K. (2011) 'Community detection using a neighborhood strength driven label propagation algorithm', in *2011 IEEE Network Science Workshop*, pp. 188–195.
Xu, J. *et al.* (2019) 'Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives', *Human Genetics*. Springer Science and Business Media , 138(2), pp. 109–124. doi: 10.1007/s00439-019-01970-5.
Yang, J. *et al.* (2010) 'Common {SNPs} explain a large proportion of the heritability for human height', *Nature Genetics*. Springer Science and Business Media , 42(7), pp. 565–569. doi: 10.1038/ng.608.
Yang, J. *et al.* (2011) 'Genome partitioning of genetic variation for complex traits using common {SNPs}', *Nature Genetics*. Springer Science and Business Media , 43(6), pp. 519–525. doi: 10.1038/ng.823.
Yang, J. *et al.* (2013) 'Genome-Wide Complex Trait Analysis ({GCTA}): Methods, Data Analyses, and Interpretations', in *Methods in Molecular Biology*. Humana Press, pp. 215–236. doi: 10.1007/978-1-62703-447-0_9.
Yang, J. *et al.* (2017) 'Concepts, estimation and interpretation of {SNP}-based heritability', *Nature Genetics*. Springer Science and Business Media , 49(9), pp. 1304–1310. doi: 10.1038/ng.3941.
Yin, B. *et al.* (2019) 'Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype', *Bioinformatics*. Oxford University Press (), 35(14), pp. i538--i547. doi: 10.1093/bioinformatics/btz369.
Zhang, S. *et al.* (2018) 'Learning for Personalized Medicine: A Comprehensive Review from a Deep Learning Perspective', *IEEE Reviews in Biomedical Engineering*. IEEE, 12(August), pp. 194–208. doi: 10.1109/RBME.2018.2864254.
Zhao, J. *et al.* (2018) 'Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction', *Scientific Reports*. Cold Spring Harbor Laboratory, 9(1), pp. 1–10. doi: 10.1101/366682.
Zhu, D. *et al.* (2018) 'A machine learning approach for air quality prediction: Model regularization and optimization', *Big Data and Cognitive Computing*, 2(1), pp. 1–15. doi: 10.3390/bdcc2010005.
Ziegler, A. *et al.* (2012) 'Personalized medicine using {DNA} biomarkers: a review', *Human Genetics*. Springer Science and Business Media , 131(10), pp. 1627–1638. doi: 10.1007/s00439-012-1188-9.
Zou, J. *et al.* (2019) 'A primer on deep learning in genomics', *Nature Genetics*. Springer US, 51(1), pp. 12–18. doi: 10.1038/s41588-018-0295-5.
تیموری, م., ابراهیمی, ا .و علوی نیا, س. م. (۱۳۹۴) 'مقایسه روش‌های مختلف یادگیری ماشین در تشخیص پرفشاری خون در بیماران دیابتی با و بدون درنظرگرفتن هزینه‌ها', *مجله تخصصی اپیدمیولوژی ایران*.

دانشپور, م. ا' (1395) .مطالعه گسترده ژنومی در طرح ژنتیک کاردیومتابولیک تهران بستر مناسبی جهت گسترش پزشکی فرادقیق در ایران', *پژوهنده*, ۱۱۲.

عزیزی, ف., هدایتی, م .و دانشپور, م. ا. (۱۳۹۶) 'طراحی و تکمیل نخستین نسخه ژنوم مرجع ایرانیان', *غدد درون ریز و متابولیسم ایران*, ۵.

لاجوردی, س' (1399) .بهبود دقت پیش‌بینی خطر بیماری پرفشارخون بر مبنای تغییرات ژنتیکی با استفاده از روش‌های یادگیری ماشین in ,'*هفدهمین کنفرانس بین المللی مهندسی صنایع*.

لاجوردی, س., علوی, س .و کارگری, م. (۱۳۹۹) 'تحلیل ساختارعلمی موضوع امتیاز خطر ژنتیکی با استفاده از روش تحلیل هم رخدادی واژگان مبتنی بر مقالات پایگاه ScienceDirect', in *ششمین کنفرانس بین المللی مهندسی صنایع و سیستم‌ها ICISE*۲۰۲۰.

واژه نامه فارسی به انگلیسی

Pooling	ادغام
Genome wide Rapid Association Using Mixed Model and Regression	ارتباط سریع ژنومی با استفاده از مدل ترکیبی و رگرسیون
Positive Prediction Value	ارزش پیش‌بینی مثبت
Negative Predictive Value	ارزش های پیشگویانه منفی
Missing	ازدست‌رفته
Cross Validation	اعتبار سنجی متقابل
Risk Scroe	امتیاز خطر
Polygenic Risk Score	امتیاز خطر چند ژنی
Effect size	اندازه اثر
Resequencing	باز توالی
Binding	بخش بندی مقید
Lossless	بدون خطا
Dynamic Programming	برنامه نویسی پویا
Iteratively	بصورت مکرر
Overfitting	بیش‌ازحد برازش
Overfit	بیش برازش
Hypertension	پرفشاری خون
Human Genome Project	پروژه ژنوم انسانی
Iranian Reference Genome Project	پروژه ژنوم مرجع ایران
Electronic Health Record	پرونده الکترونیک سلامت
Personalized Medicine	پزشکی شخصی
back propagation	پس انتشار
Single Nucleotide Polymorphism	چند ریختی تک نوکلئوتیدی
Restriction fragment length polymorphisms	چند ریختی طول قطعه محدود
Wrapper	پوشانه
Convolution	پیچشی
Genetic linkage	پیوند ژنتیکی
loss function	تابع خطا
Advanced Complex Trait Analysis	تجزیه و تحلیل صفات پیچیده پیشرفته
Genome-wide Complex Trait Analysis	تجزیه و تحلیل صفات پیچیده در سطح ژنوم
omic-data-based complex trait analysis	تجزیه و تحلیل صفات پیچیده مبتنی بر داده‌های امیک
Principal Component Analysis	تجزیه و تحلیل مؤلفه های اصلی
Co-word analysis	تحلیل هم واژه
Sequence Alignment	تراز توالی
Pairwise Sequence Alignment	تراز توالی دوبه‌دو
Chip	تراشه
relief	تسکین
ascertainment	تشخیص
Community detection	تشخیص جوامع

Variable number of tandem repeats	تعداد متغیر تکرارهای پشت سر هم
Copy number variation	تغییرات تعداد کپی
short tandem repeats	تکرارهای پشت سر هم کوتاه
Variety	تنوع
Single nucleotide variation	تنوع تک نوکلئوتیدی
Whole genome sequencing	توالی یابی کل ژنوم
feed forward	جلوبرنده
Founder populations	جمعیت های بنیانگذار
hesitation margin	حاشیه تردید
Restricted Maximum Likelihood	حداکثر احتمال محدود
Deletion	حذف
indel	حذف - اضافه
Sensitivity	حساسیت
Relative risk	خطر نسبی
Insertion	درج
Decision Tree	درخت تصمیم
Accuracy	صحت
bimodality	دو وجهی
Phenotype	رخ‌نمود
Longitudinal Phenotype	رخ‌نمود طولی
Classification	دسته‌بندی
Auto encoder	رمزگذار خودکار
Genotype	ژن‌نمود
historical control genotypes	ژن‌نمودهای کنترل تاریخی
Velocity	سرعت
Body Mass Index	شاخص توده بدن
gini index	شاخص جینی
Residual network	شبکه باقی مانده
Convolutional Neural Network	شبکه عصبی پیچیشی
radial basis function neural network	شبکه عصبی تابع پایه شعاعی
Feed forward neural network	شبکه عصبی جلوبرنده
Kohonen self organizing neural network	شبکه عصبی خود سازماندهی کوهونن
Deep Neural Network	شبکه عصبی عمیق
Recurrent neural network	شبکه عصبی گردشی
Fully Connected Network	شبکه کاملاً متصل
Social networks	شبکه‌های اجتماعی
Adversarial networks	شبکه‌های متخاصم
Individual	شخصی
Precision	دقت
Complex Trait	صفت پیچیده
Information gain	بهره اطلاعات
Impact Factor	ضریب تأثیر
Meta-Analysis	فراتحلیل

Recall	بازخوانی
Blood Pressure	فشارخون
Diastolic Blood Pressure	فشارخون دیاستولیک
Systolic Blood Pressure	فشارخون سیستولیک
propensity	گرایش
Genetic Relation Matrix	ماتریس رابطه ژنتیکی
Confusion Matrix	ماتریس مختلط
Support Vector Machine	ماشین بردار پشتیبانی
Mini satellites	ماهواره های کوچک
Positive	مثبت/صحیح
Data Set	دادگان
fuzzy set	مجموعه فازی
Mixed Linear Model	مدل خطی مختلط
Fuzzy Boundry	مرز فازی
Phenome-wide association study	مطالعه ارتباط گسترده پدیده
Genome Wide Association Study	مطالعه ارتباط گسترده ژنوم
Missed Value	مقادیر ازدست‌رفته
Regularization	قاعده‌مند سازی
Negative	منفی / غلط
Area Under Curve	ناحیه زیر منحنی
True Positive Rate	نرخ مثبت واقعی
True Negative Rate	نرخ منفی واقعی
gain ratio	نسبت بهره
Biomarker	نشانگر زیستی
Evidence Theory	نظریه شواهد
Linkage Mapping	نقشه پیوند
Random Sub-sampling	نمونه گیری تصادفی فرعی
Polymerase chain reactions	واکنش های زنجیره‌ای پلیمراز
Inheritance	وراثت
Missing Heritability	وراثت پذیری گمشده
Regional heritability advanced complex trait analysis	وراثت پذیری منطقه ای تجزیه و تحلیل صفات پیچیده پیشرفته
Specificity	ویژگی
Receiver Operating Characteristic	ویژگی عملکرد گیرنده
Multiple Sequence Alignment	هم ترازی چندگانه
Deep learning	یادگیری عمیق
Genome Deep Learning	یادگیری عمیق ژنوم
Machine Learning	یادگیری ماشین

واژه نامه انگلیسی به فارسی

Accuracy	صحت
Advanced Complex Trait Analysis	تجزیه و تحلیل صفات پیچیده پیشرفته
Adversarial networks	شبکه‌های متخاصم
angiotensinogen	آنژیوتانسینوژن
Area Under Curve	ناحیه زیر منحنی
ascertainment	تشخیص
Auto encoder	رمزگذار خودکار
back propagation	پس انتشار
bimodality	دو وجهی
Binding	بخش بندی مقید
Biomarker	نشانگر زیستی
Blood Pressure	فشارخون
Body Mass Index	شاخص توده بدن
Chip	تراشه
Classification	دسته‌بندی
Community detection	تشخیص جوامع
Complex Trait	صفت پیچیده
Confusion Matrix	ماتریس مختلط
Convolution	پیچش
Convolutional Neural Network	شبکه عصبی پیچیده
Copy number variation	تغییرات تعداد کپی
Co-word analysis	تحلیل هم واژه
Cross Validation	اعتبار سنجی متقابل
Data Set	دادگان
Decision Tree	درخت تصمیم
Deep learning	یادگیری عمیق
Deep Neural Network	شبکه عصبی عمیق
Deletion	حذف
Diastolic Blood Pressure	فشارخون دیاستولیک
Dynamic Programming	برنامه نویسی پویا
Effect size	اندازه اثر
Electronic Health Record	پرونده الکترونیک سلامت
Evidence Theory	نظریه شواهد
feed forward	جلوبرنده
Feed forward neural network	شبکه عصبی جلوبرنده
Founder populations	جمعیت های بنیانگذار
Fully Connected Network	شبکه کاملاً متصل
Fuzzy Boundry	مرز فازی
fuzzy set	مجموعه فازی
gain ratio	نسبت بهره
Genetic linkage	پیوند ژنتیکی

Genetic Relation Matrix	ماتریس رابطه ژنتیکی
Genome Deep Learning	یادگیری عمیق ژنوم
Genome Wide Association Study	مطالعه ارتباط گسترده ژنوم
Genome wide Rapid Association Using Mixed Model and Regression	ارتباط سریع ژنومی با استفاده از مدل ترکیبی و رگرسیون
Genome-wide Complex Trait Analysis	تجزیه و تحلیل صفات پیچیده در سطح ژنوم
Genotype	ژن‌نمود
gini index	شاخص جینی
hesitation margin	حاشیه تردید
historical control genotypes	ژن‌نمودهای کنترل تاریخی
Human Genome Project	پروژه ژنوم انسانی
Hypertension	پرفشاری خون
Impact Factor	ضریب تأثیر
indel	حذف - اضافه
Individual	شخصی
Information gain	بهره اطلاعات
Inheritance	وراثت
Insertion	درج
Iranian Reference Genome Project	پروژه ژنوم مرجع ایران
Iteratively	بصورت مکرر
Kohonen self organizing neural network	شبکه عصبی خود سازماندهی کوهونن
Linkage Mapping	نقشه پیوند
Longitudinal Phenotype	رخ‌نمود طولی
loss function	تابع خطا
Lossless	بدون خطا
Machine Learning	یادگیری ماشین
Meta-Analysis	فراتحلیل
Mini satellites	ماهواره های کوچک
Missed Value	مقادیر ازدست‌رفته
Missing	ازدست‌رفته
Missing Heritability	وراثت پذیری گمشده
Mixed Linear Model	مدل خطی مختلط
Multiple Sequence Alignment	هم ترازی چندگانه
Negative	منفی / غلط
Negative Predictive Value	ارزش های پیشگویانه منفی
omic-data-based complex trait analysis	تجزیه و تحلیل صفات پیچیده مبتنی بر داده‌های امیک
Overfit	بیش برازش
Overfitting	بیش‌ازحد برازش
Pairwise Sequence Alignment	تراز توالی دوبه‌دو
Personalized Medicine	پزشکی شخصی
Phenome-wide association study	مطالعه ارتباط گسترده پدیده
Phenotype	رخ‌نمود

Polygenic Risk Score	امتیاز خطر چند ژنی
Polymerase chain reactions	واکنش های زنجیره‌ای پلیمراز
Pooling	ادغام
Positive	مثبت/صحیح
Positive Prediction Value	ارزش پیش‌بینی مثبت
Precision	دقت
Principal Component Analysis	تجزیه و تحلیل مؤلفه های اصلی
propensity	گرایش
radial basis function neural network	شبکه عصبی تابع پایه شعاعی
Random Sub-sampling	نمونه گیری تصادفی فرعی
Recall	بازخوانی
Receiver Operating Characteristic	ویژگی عملکرد گیرنده
Recurrent neural network	شبکه عصبی گردشی
Regional heritability advanced complex trait analysis	وراثت پذیری منطقه ای تجزیه و تحلیل صفات پیچیده پیشرفته
Regularization	قاعده‌مند سازی
Relative risk	خطر نسبی
relief	تسکین
Resequencing	باز توالی
Residual network	شبکه باقی مانده
Restricted Maximum Likelihood	حداکثر احتمال محدود
Restriction fragment length polymorphisms	پلی مورفیسم های طول قطعه محدود
Risk Scroe	امتیاز خطر
Sensitivity	حساسیت
Sequence Alignment	تراز توالی
short tandem repeats	تکرارهای پشت سر هم کوتاه
Single Nucleotide Polymorphism	پلی مورفیسم تک نوکلئوتیدی
Single nucleotide variation	تنوع تک نوکلئوتیدی
Social networks	شبکه‌های اجتماعی
Specificity	ویژگی
Support Vector Machine	ماشین بردار پشتیبانی
Systolic Blood Pressure	فشارخون سیستولیک
True Negative Rate	نرخ منفی واقعی
True Positive Rate	نرخ مثبت واقعی
Variable number of tandem repeats	تعداد متغیر تکرارهای پشت سر هم
Variety	تنوع
Velocity	سرعت
Whole genome sequencing	توالی یابی کل ژنوم
Wrapper	پوشانه

Abstract

Hypertension is one of the most significant underlying ailments of cardiovascular disease; hence, methods that can accurately reveal the risk of hypertension at an early age are essential. Also, one of the most critical personal health objectives is to improve disease prediction accuracy by examining genetic variants.

Therefore, various clinical and genetically based methods are used to predict the disease; however, the critical issue with these methods is the high number of input variables as genetic markers with small samples. One approach that can be used to solve this problem is machine learning.

This study was conducted on the participants' genetic markers in the 20-year research of cardiometabolic genetics in Tehran (TCGS). Various machine learning methods were used, including linear regression, neural network, random forest, decision tree, and support vector machine. The top ten genetic markers were identified using importance-based ranking methods, including information gain, gain ratio, Gini index, χ^2, relief, and FCBF. A model based on a neural network with AUC of 89% was presented. This model has an accuracy and an f-measure of 0.89, which shows the quality. The final results indicate the success of the machine learning approach.

The initial processing method and data modeling have innovation, especially in the deep learning network. The research results in the two machine learning and deep learning models have AUC of 0.87 and 0.89, respectively. This modeling provides the possibility of prognosis at young ages with appropriate accuracy.

Keywords: Deep Learning, Genetic Markers, Hypertension, Disease Risk.

Department of Information Technology
Faculty of Industrial and Systems Engineering

Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Ph.D)

Design and presenting an intelligent model for predicting SNP-based hypertension in uncertainty

By:
Seyed Ali Lajevardy

Supervisor:
Dr. Kargari

Co-Supervisor:
Dr. Daneshpour

Advisor:
Dr. Akbarzadeh

June 2023