6TH INTERNATIONAL CONFERENCE ON APPLIED RESEARCH IN COMPUTER, ELECTRICAL AND INFORMATION TECHNOLOGY
March 6, 2022
Tbilisi - Georgia
CINVU
Kosar University
Avicenna International Community College LLC

ارائه مدل یادگیری عمیق CNN با هدف بهبود دقت پیش بینی خطر بیماری پرفشاری خون بر مبنای SNP

سیدعلی لاجوردی[۱]، مهرداد کارگری[۲]، مریم السادات دانشپور[۳]، مهدی اکبرزاده[۴]

۱-دانشگاه تربیت مدرس - دانشکده صنایع و سیستم ها

۲-دانشگاه تربیت مدرس - دانشکده صنایع و سیستم ها

۳-دانشگاه علوم پزشکی شهید بهشتی - پژوهشکده علوم غدد درون ریز و متابولیسم

۴-دانشگاه علوم پزشکی شهید بهشتی - پژوهشکده علوم غدد درون ریز و متابولیسم

چکیده

فشارخون عامل قابل اصلاح برای بیماری‌های قلبی عروقی مانند بیماری اسکمیک قلب است که از دلایلی اصلی مرگ در سراسر جهان است که با نام کشنده خاموش معرفی می شود. از این رو، روش هایی که می توانند به طور دقیق خطر ابتلا به فشار خون بالا را در سنین پایین آشکار کنند، ضروری هستند. مدل‌های معمول پیش بینی خطر بیماری در درجه اول مبتنی بر عامل‌های شیوه زندگی است. اخیراً، در نظر گرفتن خطر عامل‌های ژنتیکی ازجمله SNPهای مرتبط با بیماری همراه با مدل سازی خطر، دقت پیش‌بینی فردی بیماری‌ها را بهبود بخشیده است. SNP یک تغییر ژنتیکی کوچک در DNA است و رایج‌ترین نوع تنوع ژنتیکی هست که در میان افراد رخ می‌دهد. مدل های مختلفی که از نشانگر های ژنومی جهت پیش بینی بیماری فشار خون استفاده می کنند شامل چهار رویکرد آماری، فرا تحلیل، یادگیری ماشین و مدل بالینی هستند. مهم ترین مساله در این مدل ها تعداد بالای SNP های ورودی و ارتباط این SNP ها با یکدیگر است. در مطالعه حاضر از روش مبتنی بر یادگیری عمیق با رویکرد CNN برای استفاده از SNPهای متعدد فشار خون در یک مطالعه کوهورت طولی استفاده و برای مقایسه نیز PRS با استفاده از دو نرم افزار plink و gcta64 محاسبه شده است. در ابتدا دادهای ژنومی تبدیل به تصویر شده و وارد مدل CNN می شود که لایه های آن شامل لایه پیچش، لایه ادغام و لایه کاملاً متصل و در نهایت لایه خروجی می باشد. داده‌های تحقیق شامل سه بخش داده ژنومی، سن و داده‌های طولی فنوتیپ فشارخون مبتنی بر مطالعه ژنتیک قلبی- متابولیک تهران است. برای مقایسه کارایی مدل از AUC استفاده شده است که مدل CNN با مقدار AUC ۰.۸۷۷، عملکرد بهتری نسبت به PRS و همچنین آخرین مدل های ارائه شده در ادبیات موضوع را نشان می دهد.

کلمات کلیدی: نمره خطر پلی ژنتیک، یادگیری ماشین، بیماری فشار خون، مارکر ژنتیکی ، شبکه عصبی پیچشی

مقدمه

کشنده خاموش[1] که عامل اصلی در بیماری‌های قلبی عروقی مانند بیماری اسکمیک قلب و سکته مغزی است، فشارخون است که از علل مهم مرگ و بیماری در سراسر جهان می باشد. فشارخون و پرفشاری خون؛ هر دو از صفات چندژنی به‌حساب می‌آیند، که تعداد زیادی از مسیرهای متابولیکی را شامل می شوند (Martinez-Ríos et al. 2021; K. T. Mills, Stefanescu, and He 2020; Zhou et al. 2017). از این رو، روش هایی که می توانند به طور دقیق خطر ابتلا به فشار خون بالا را در سنین پایین آشکار کنند، ضروری هستند.

از طرف دیگر یکی از مهم‌ترین اهداف سلامت شخصی، بهبود دقت پیش‌بینی بیماری با بررسی انواع ژنتیکی است. بنابراین، روش های مختلف بالینی و ژنتیکی برای پیش بینی بیماری استفاده می شود. با این حال، مسئله مهم با این روش ها، تعداد بالای متغیرهای ورودی به عنوان نشانگرهای ژنتیکی با تعداد کمی نمونه است. روشی که می تواند بر این مشکل غلبه کند، یادگیری ماشینی است.

ژنوم انسان و تفاوت ژنومی

ژنوم انسان شامل تمام اطلاعات موردنیاز برای حیات انسان است، که از ۳ میلیارد جفت باز در ۲۳ جفت کروموزوم به نام DNA تشکیل شده است. این مجموعه حاوی تمام اطلاعات لازم برای تعیین صفاتی است که فرد را می‌سازد. کدگذاری اختصاصی درون ژن ها از طریق چکونگی ترکیب چهار باز خاص به نام های آدنین، تیمین، گوانین و سیتوزین (A,T,G,C) انجام شده است.

در هنگام مقایسه ژنوم دو فرد، تفاوت‌های بسیاری (به لحاظ تعداد) دیده می شود که این تفاوت‌ها حاصل تغییرات ژنومی مانند چند پلی مورفیسم تک نوکلئوتیدی[2] و تغییرات بزرگ‌تر مانند حذف[3]، جایگزینی[4] و تغییر در تعداد کپی‌های ژنی[5] هستند. هرکدام از این تغییرات همراه با عوامل محیطی می‌توانند منجر به بروز یک فنوتیپ متفاوت در فرد شوند.

SNP یک تغییر ژنتیکی کوچک در DNA است و رایج‌ترین نوع تنوع ژنتیکی هست که در میان افراد رخ می‌دهد. SNP زمانی رخ می‌دهد که باز یک نوکلئوتید به‌عنوان‌مثال نوکلئوتید دارای باز A جایگزین یکی دیگر از سه باز نوکلئوتید (مثلاً G ,C و یا T) می شود. البته این تغییرات می‌توانند: بدون اثر (بدون تغییر در فنوتیپ)، اثر غیر بیماری زا (تغییر در فنوتیپ)، اثر ضد سلامت (دیابت، سرطان، بیماری‌های قلبی، بیماری‌هانتینگتون و هموفیلی) و اثر نهفته (فرضا تغییرات در ژنوم 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 (Filshtein et al. 2019).

[1] Silent killer
[2] Single nucleotide variation
[3] Deletion
[4] Insertion
[5] Copy number variation

مطالعات گسترده ارتباط ژنومی[6] (GWAS) و امتیاز خطر چند ژنی[7] (PRS)

جهت برر سی ارتباط بین ژنوتیپ و فنوتیپ از برر سی توارث[8] در خانواده از آزمون های لینکاژ یا پیو ستگی ژنتیکی[9] استفاده می شد. این روش‌ها در بررسی بیماری‌هایی که یک ژن عامل بروز آن‌ها بود، بسیار موفق بودند. اما این یافته‌ها در بیماری‌های چندعاملی پیچیده تعمیم‌پذیر نمی باشند. در مطالعات ارتباط ژنوم گسترده کشف نشانگرهای زیستی، شناسایی افراد دارای ژن بیماری‌های خاص و شنا سایی نوع واکنش به درمان به‌طور و سیع‌تری صورت گرفته ا ست (M. C. Mills and Rahal 2019).

این مطالعات اغلب از نوع مورد شاهدی انجام می‌گردد؛ بدین معنا که گروهی از بیماران در مقابل افراد سالم انتخاب می شوند و تمامی این افراد برای قسمت عمده چند مورفیسم‌های تک نوکلئوتیدی شناخته شده تعیین ژنوتیپ می شوند، سپس میزان فراوانی اللی هرکدام از این نشانگرها در گروه‌های مورد و شاهد مقایسه می‌شود. اساس این آزمون ها محاسبه میزان اثر[10] بر مبنای محا سبه نسبت شانس[11] ا ست بدین معنا که حضور یا عدم حضور یک آلل چقدر به حضور یا عدم حضور فنوتیپ مورد بررسی در یک جامعه‌ی مشخص و معلوم، مربوط است (Visscher et al. 2012).

امتیاز چند ژنی[12] (PGS) یا امتیاز خطر چند ژنی تخمینی از م سئولیت ژنتیکی یک فرد در برابر یک صفت یا بیماری است که با توجه به مشخصات ژنوتیپ و داده‌های مطالعه ارتباط گسترده ژنوم محاسبه می شود. در حالی که PRS های فعلی معمولاً تنها بخش کوچکی از واریانس صفت را توضیح می دهند، ارتباط آنها با بزرگترین عامل ایجاد تنوع فنوتیپی منجر به کاربرد معمول PRS ها در تحقیقات زیست پزشکی شده است (Choi, Mak, and O'Reilly 2020).

بیماری پرفشاری خون[13] (HTN)

ف شارخون در دو نوبت سنجیده می شود. عدد بالاتر که در زمان انقباض قلب محا سبه می شود، مرتبط با ف شار سیستولی[14] (SBP) است که معمولاً به فشارخونی که در سرخرگ‌هاست گفته می شود. فشارخون دیاستولی[15] (DBP) به فشارخون ورید گفته می‌شود که در زمان استراحت قلب مابین دو نبض است. علل بیماری پرفشاری خون هنوز شناخته نشده ا ست. پر فشاری خون بر ا ساس SBP بزرگتر از ۱۴۰ میلیمتر جیوه و DBP بزرگتر از ۹۰ میلیمتر جیوه تعریف می شود (K. T. Mills et al. 2016; Niu et al. 2021).

مدل های خطر بیماری

مدل‌های معمول خطر بیماری بر ا ساس همه‌گیر شنا سی (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

[6] Genome Wide Association Study (GWAS)
[7] Polygenic Risk Score (PRS)
[8] Inheritance
[9] Genetic linkage
[10] Effect size
[11] Odd ratio
[12] Genetic Risk Score
[13] Hypertension
[14] Systolic Blood Pressure
[15] Diastolic Blood Pressure

مرتبط با بیماری و یا مربوط به فنوتیپ همراه با مدل سازی خطر، دقت پیش‌بینی فردی بیماری‌ها را بهبود بخ شیده ا ست (Mosley et al. 2020).

در حال حا ضر، تو سعه مدل‌های خطر ژنتیکی بر روی د ستیابی به قدرت پیش‌بینی دقیق برای شناخت افراد در خطر به شیوه‌ای قدرتمند تمرکز دارد. مطالعات GWA، SNPها را با توجه به ارتباط آن‌ها با یک بیماری/ فنوتیپ در سطح جمعیت تعریف می‌کند (Abraham and Inouye 2015).

در این تحقیق ابتدا مدل های ژنومی پیش بینی کننده صفات برر سی و در ادامه مدل های مختلف پیش بینی بیماری فشار خون مرور می گردد. این روش ها شامل چهار رویکرد آماری، فرا تحلیل، یادگیری ماشین و مدل بالینی می باشند.

رویکرد آماری

رویکرد آماری، رویکرد غالب در تحلیل ارتباطات گ سترده ژنومی ه ست. اولین مدل آماری GRAMMAR[16] می با شد (Aulchenko, De Koning, and Haley 2007). در این مدل داده ها تحت مدل ترکیبی[17] مورد تجزیه‌وتحلیل قرار می‌گیرند.

پس از مدل GRAMMAR، مدل GCTA ارائه گرد ید (Yang et al. 2017). در روش GCTA برخلاف تجزیه‌وتحلیل روابط تک SNP، به اثرات همه SNPsها به‌عنوان اثرات ت صادفی با یک مدل خطی آمیخته[18] نگاه می شود. افزایش عملکرد این مدل در مدل ACTA ارائه شده ا ست (Gray, Stewart, and Tenesa 2012). با ظهور GPU های سریع، الگوریتم جهت اجرای موازی تغییر کرده و مدل REACTA ارائه گردید (Cebamanos et al. 2014). بر ا ساس SNP های معرفی شده در GWAS های مرتبط با فشار خون، مدل PRS دارای AUC ۰.۸۰۴ می باشد (Vaura et al. 2021).

رویکرد فراتحلیل

در روش‌های فرا تحلیل[19] از مقادیر مختلفی مانند p-vlaue و اندازه نمونه و ادغام مطالعات GWAS با یک فنوتیپ ا ستفاده می شود. مهم ترین مدل در این رویکرد مدل METAL (Willer, Li, and Abecasis 2010) می با شد که امکان ادغام مطالعات مختلف در حوزه GWAS را بر ا ساس اندازه نمونه، واریانس داده، مقدار p-value و اندازه تاثیر ایجاد می کند. برای این رویکرد بسته های نرم افزاری مختلفی مانند GWAMA و MetABEL ارائه شده است (Evangelou and Ioannidis 2013).

رویکرد یادگیری ماشین

در برر سی اثر م شترک چند متغیر ژنتیکی و محیطی، چالش‌های متعددی وجود دارد. اول، در GWAS، ژنوتیپ‌ها تا یک‌میلیون SNP در چند هزار نفر تعیین می شود، که منجر به م شکل کوچک بودن نمونه و بزرگ بودن تعداد بعد

[16] Genome wide Rapid Association Using Mixed Model and Regression
[17] Mixed Model
[18] Mixed Liner Model (MLM)
[19] Meta-Analysis

می شود. دوم، هنگامی که تعداد زیادی از SNPهای ژنوتیپ در یک مقیاس ژنومی هستند، باید عدم تعادل ارتباط[20] بین SNPها (که منجر به متغیرهای همبسته شود، موردتوجه قرار گیرد (Niu et al. 2021; Szymczak et al. 2009).

رویکرد بالینی

در رویکرد بالینی علاوه بر عوامل ژنتیکی، عوامل بالینی نیز در بیماری فشار خون در نظر گرفته می شوند. عوامل مختلفی که در تحقیقات بالینی در نظر گرفته می شود شامل: سابقه خانوادگی، سن، جنس، مصرف نمک، مصرف الکل، مصرف دخانیات، کم تحرکی و چاقی از عوامل خطر بیماری فشاری خون بالا به شمار می رود.

وو و همکاران بر مبنای ملاک های سنتی پیش بینی فشار خون تحقیقی را مبتنی بر روش های یادگیری ماشین ارائه داده اند که دارای AUC ۰.۷۵۷ برای پیش بینی فشار خون است (Wu et al. 2020). همچنین در تحقیق دیگری با استفاده عوامل بالینی سن، مصرف سیگار، سابقه خانوادگی، فعالیت فیزیکی و شاخص جرمی بدن[21] دقت ۰.۸۵۴ و با اضافه کردن پارامترهای ژنتیکی دقت ۰.۸۷۱ بدست آمده است (Niu et al. 2021).

اهداف تحقیق

چالش های مختلفی در مدل سازی خطر مبتنی بر ژنوم و عوامل بالینی وجود دارد. اول، در GWAS، حدود یک میلیون ژنوتیپ SNP در چند هزار نفر خطای اندازه نمونه کوچک نسبت به تعداد صفات را ایجاد می کند. دوم، زمانی که تعداد زیادی از SNP های ژنوتیپ در یک منطقه ژنومی یک سان هستند، عدم تعادل پیوندی[22] (LD) بین SNP ها (که منجر به متغیرهای همبسته می شود) باید در نظر گرفته شود. سوم، افزایش دقت پیش‌بینی در سنین پایین‌تر، حیاتی‌ترین موضوع است (Wu et al. 2020).

در نتیجه، مطالعه حاضر از روش‌های مبتنی بر یادگیری عمیق با رویکرد CNN برای استفاده از SNPهای متعدد فشار خون در یک مطالعه کوهورت طولی استفاده می‌کند.

[20] Linkage Disequilibrium (LD)
[21] Body Mass Index (BMI)
[22] Linkage Disequilibrium

روش شناسی تحقیق

در این مقاله ابتدا با روش امتیاز خطر چند ژنی[۲۳] (PRS) و در ادامه با استفاده از شبکه عصبی پیچشی[24] (CNN) خطر پرفشاری خون مدل سازی می شود و با توجه به اینکه پیش بینی رخداد بیماری در سنین پایین مهم است صرفا جنسیت و سن به همراه SNP و بدون علائم بالینی وارد مدل شده است.

امتیاز خطر چندژنی

از آنجایی که ترکیب ژنتیکی یک فرد از بدو تولد تا حد زیادی ثابت است، اطلاعات ژنتیکی این پتانسیل را دارد که به عنوان یک پیش‌بینی کننده اولیه خطر عمل کند. بیماری فشارخون تحت تأثیر چندین گونه ژنتیکی با اندازه‌های اثر کوچک[۲۵] قرار می‌گیرد، بنابراین پیش‌بینی خطر معنی‌دار، بررسی تأثیر انبوه این گونه‌های چندگانه از طریق محاسبه یک متریک واحد که نشان‌دهنده خطر ژنتیکی کلی یک فرد است، ضروری است. ابتدا یک امتیاز خطر ژنتیکی[۲۶] (GRS) ساده محاسبه شد که یک مجموع ساده از تعداد آلل‌های خطر (معمولاً از چند SNP از GWAS، که گاهی بر اساس اندازه‌های اثر وزن هستند) است که در هر فرد وجود دارد(Padmanabhan and Dominiczak 2020). اخیراً، با توجه به اینکه SNPهایی که آستانه اهمیت بسیار سختگیرانه برای ارتباط گسترده ژنومی را برآورده نمی‌کنند نیز می‌توانند پیش‌بینی‌کننده بیماری باشند، طیف وسیع‌تری از SNPها، از هزاران تا میلیون‌ها SNP برای ایجاد یک GRS بهبود یافته به‌عنوان PRS استفاده شده‌اند (Torkamani, Wineinger, and Topol 2018). باید تأکید کرد اندازه خطر ارائه شده توسط PRS یک محدوده احتمالی است و این با اطلاعات خطر از نشانگرهای ژنتیکی اختلالات تک ژنی متفاوت است.

برای محاسبه PRS از دو نرم افزار plink و gcta64 استفاده می گردد. کد اجرا به صورت شکل ۱ می باشد. لازم به ذکر است اجرای کد روی فایل های به فرمت ped که داده های ژنومی را ذخیره کرده است، در بستر لینوکس اجرا می گردد.

```
plink.exe --ped test.ped --map test.map --noweb --make-bed --out test
gcta64 --bfile test --make-grm --out test --thread-num 10 --reml --pheno test.phen
gcta64 --reml --grm test --pheno test.phen --reml-pred-rand --out test
gcta64  --bfile test --blup-snp test.indi.blp --out test
plink --score test.snp.blp --noweb --bfile test --pheno test.phen --out test
gcta64 --reml --grm test --pheno test.phen --out test --cvblup
```

شکل(۱) : کد shell **محاسبه** PRS

ضمنا برای اعتبارسنجی مدل ساخته شده، داده ها به دو دسته آزمایش[۲۷] و یادگیرنده[۲۸] تقسیم می شوند و اندازه PRS به

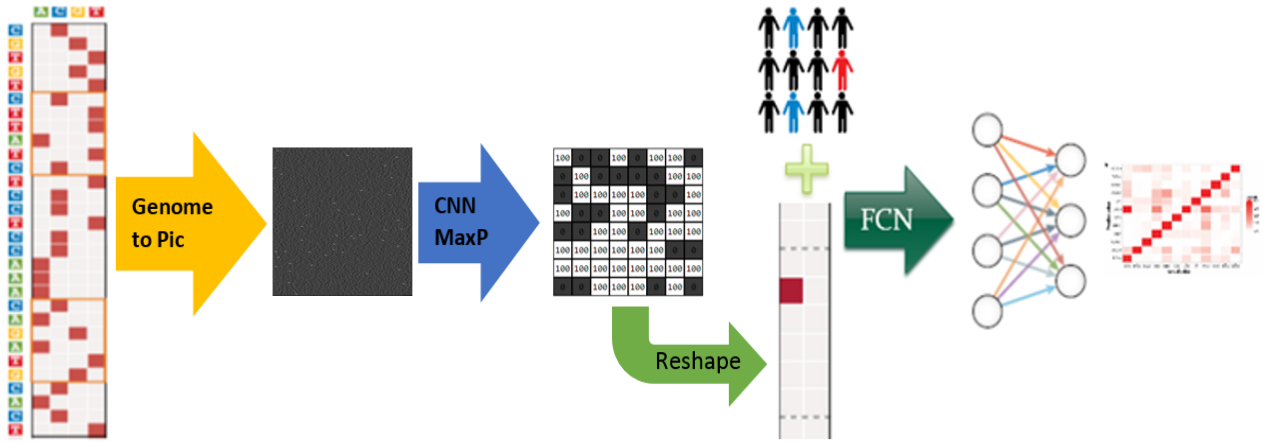
[23] Polygenic Risk Score
[24] Convolutional Neural Network
[25] Small effect
[26] Genetic Risk Score
[27] Test
[28] Train

صورت جداگانه محاسبه می شود و در نهایت در مدلسازی پیش بینی کننده بیماری استفاده می گردد.

شبکه عصبی پیچشی

شبکه عصبی پیچشی برای پردازش داده هایی با دو ویژگی مناسب است: (۱) ویژگی ها از نظر فضایی ارتباط دارند. به عنوان مثال، تصویری که در آن پیکسل ها در یک آرایه مرتب شده اند و یا ویدئو، که تصاویر پشت سر هم می باشد. (۲) ویژگی ها در تمام صفحه همگن هستند. ساختار اصلی CNN از لایه های ورودی به شکل صفحه تصویر تشکیل شده است.

لایه های دیگر شامل لایه پیچش، لایه ادغام[۲۹] و لایه کاملاً متصل[۳۰] (FCN) و لایه خروجی می باشد (Luo et al. 2018). لایه پیچش ویژگی های محلی را بر اساس وزن نورون ها با در هم پیچیدن تمام بخش های تصویر استخراج می کند. لایه ادغام عملیات مستقلی را بر روی هر نقشه ویژگی انجام می دهد، مانند ادغام متوسط[۳۱] یا حداکثر ادغام[۳۲]، که می تواند به طور موثر و ضوح ویژگی را کاهش دهد و تعداد پارامترهای شبکه مورد نیاز برای بهینه سازی کاهش دهد. CNN این ساختارهای اساسی را روی هم می گذارد و خروجی لایه اول را به عنوان ورودی لایه دوم قرار می دهد که این خاصیت توانایی یادگیری عمیق دارد. مراحل اجرای CNN در تصویر ۲ آمده است.

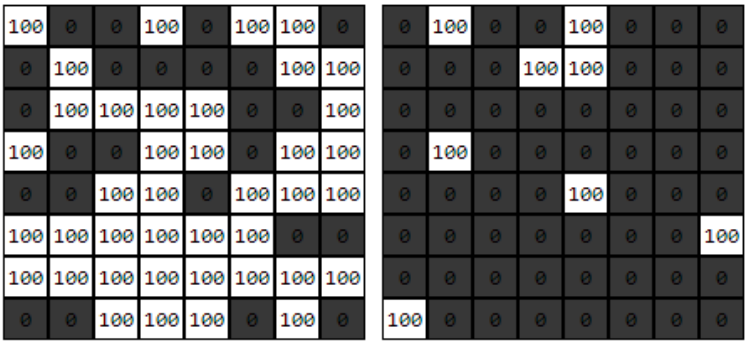


شکل(۲) : مراحل اجرای CNN

در اولین مرحله داده ژنومی که در به صورت فایل ped می باشد به فایل تصویری تبدیل می شود. برای اینکار در هر ۴ پیکسل پشت سر هم تنها یک خانه بر اساس نوع ژنوم مشکی می شود و بقیه خانه ها به رنگ سفید می مانند. برای ساخت بهترین ورودی، طول و عرض تصویر برابر مجذور تعداد ژنوم که برابر با ۶۰۰ هزار SNP است در نظر گرفته می شود. این روش باعث کاهش حجم داده ژنومی با ضریب فشرده سازی ۱۳ درصد می شود. در مرحله بعدی شبکه عمیق مبتنی بر لایه پیچش با ۲۴ فیلتر تشکیل می گردد. فیلترها به صورت ۴*۴ است که در هر فیلتر در هر ردیف تنها یک ستون

[29] Pooling layer
[30] Fully Connected Network
[31] Average pool
[32] Maximum pool

دارای مقدار یک و مابقی صفر می با شد و پس از اعمال فیلتر، تمامی نتایج در همدیگر ادغام می شود. تعداد لایه های CNN وابسته به پیچیدگی تغییرات می با شد که در داده های این تحقیق ۶ لایه منا سب می با شد. بعد از هر لایه پیچش، لایه ادغام بی شینه به صورت ۲*۲ قرار دارد که باعث کاهش بعد ۱/۲ در طول و ارتفاع ت صویر ا ست و در نتیجه در هر مرحله تصویر ۱/۴ می شود. به دلیل ورودی های متفاوت، پس از ۶ لایه داده ها نرمال می شود و در نهایت تصویر تبدیل به آرایه شده و لایه FCN و لایه خروجی قرار گرفته است. نمونه تصویر پس از ۶ لایه در شکل ۳ آمده است.



شکل(۳) : نمونه تصویر ایجاد شده قبل از لایه FCN

داده های تحقیق

داده‌های تحقیق شامل سه بخش داده ژنومی، سن و داده‌های طولی فنوتیپ ف شارخون می با شد. این داده ها از مطالعه ژنتیک قلبی- متابولیک تهران[۳۳] (TCGS) انتخاب شده اند (Daneshpour et al. 2017). TCGS مطالعه کوهورت ژنتیکی برای تعیین عوامل خطر اختلالات غیرواگیر عمده در مطالعه قند و لیپید تهران[۳۴] (TLGS) می با شد (Azizi et al. 2009).

این تحقیق با کد IR.SBMU.ENDOCRINE.REC.1400.008 تو سط کمیته ملی اخلاق در تحقیقات زی ست پز شکی در اردیبه شت ۱۴۰۰ تایید شده ا ست. تمامی شرکت‌کنندگان یک نمونه خون پایه کافی برای آنالیز پلا سما و DNA ارائه کرده اند و برای آنالیزهای مبتنی بر خون و پیگیری طولانی‌مدت ر ضایت کتبی دادند. تمام روش‌های این مطالعه با رعایت د ستورالعمل‌های منا سب و مرتبط انجام شد و در این مطالعه، همه شرکت‌کنندگان به‌جز افراد مبتلا به ف شار خون دیا ستولیک ایزوله (DBP بی شتر از ۹۰ و SBP کمتر از ۱۴۰ میلی‌متر جیوه) وارد شده اند (Mahajan et al. 2019).

برای اندازه گیری ف شار خون، ابتدا شرکت کنندگان به مدت ۱۵ دقیقه روی صندلی می مانند و پس از آن پز شک متخصص فشار خون را دو بار با استفاده از فشارسنج جیوه ای استاندارد اندازه گیری می کند. حداقل یک فاصله ۳۰ ثانیه ای بین این دو اندازه گیری جداگانه وجود دارد و میانگین دو اندازه گیری به عنوان ف شار خون شرکت کننده ثبت می شود (Tohidi et al. 2012).

[33] Tehran Cardio-metabolic Genetic Study
[34] Tehran Lipid and Glucose Study

افرادی که در هر دوره دارای SBP بالاتر از ۱۴۰، یا DBP بیش از ۹۰ با شد یا شرکت کننده در حال م صرف داروی کنترل فشار خون باشد، برچسب فشار خون بالا داده شد و در غیر این صورت، نرمال در نظر گرفته می شوند.

نمونه ژنوم از گلبول‌های سفید با روش ا ستاندارد پروتئیناز K ا ستخراج شده با ترا شه HumanOmniExpress-24-v1-0 توسط شرکت deCODE مطابق با دستگاه Illumina ژنوتیپ شدند (Kolifarhood et al. 2021). داده های ژنومی برای هر فرد دارای ۶۰۰ هزار SNP می باشند که در فایل ped ذخیره سازی شده است.

بر مبنای (Mattoo 2019) جهت بهبود نتایج و کنترل کیفی، دسته‌بندی بر اساس سن (بالای ۱۸ سال و پایین‌تر از ۱۸ سال) و بخش‌بندی SNP از منظر نادر و رایج بودن انجام شده است. بر همین اساس، SNPهایی که دارای مقادیر گمشده[۳۵] کمتر از ۵ درصد و میانگین فراوانی آللی[۳۶] (MAF) بالاتر از ۵ درصد هستند، در نظر گرفته شده است.

اعتبارسنجی و ارزیابی نتایج

معیارهای مختلفی در ادبیات مو ضوع جهت اعتبار سنجی و ارزیابی نتایج مدل‌سازی بیان شده ا ست که در جدول ۱ آمده است. در این جدول TP, TF, FP, FN به ترتیب مقدار مثبت درست، منفی درست، مثبت غلط و منفی غلط می باشد. این معیارها کارایی مدل ساخته شده را در داده های تحقیق بیان می کنند.

معیار	نحوه محاسبه
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall/sensitivity/true positive rate	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
False Positive Rate	$\frac{FP}{FP+TN}$
F1-Score	$2\frac{(Precision)(Recall)}{Precision+Recall}$

جدول (۱) : معیارهای ارزیابی مدل سازی

علاوه بر این معیارها، یکی از معیارهای پرکاربرد در مقایسه بین مدل ها، معیار سطح زیر منحنی مشخصه عملکرد گیرنده[۳۷] (AUC) می باشد.یک طبقه بند کامل دارای AUC برابر با ۱ خواهد و یک طبقه بندی ضعیف دارای AUC ۰.۵ ا ست. یک طبقه بندی کننده خوب دارای AUC بی شتر از ۰.۵ خواهد بود و با ر سیدن به ۱ عالی در نظر گرفته می شود (Martinez-Ríos et al. 2021).

برای اعتبار سنجی مدل سازی نیز از روش 10-fold ا ستفاده شده ا ست که با تقسیم داده ها به ۱۰ قسمت مساوی بدون تکرار و ایجاد مدل های متفاوت، نتیجه نهایی را بر اساس متوسط نتایج بدست آمده اعلام می کند.

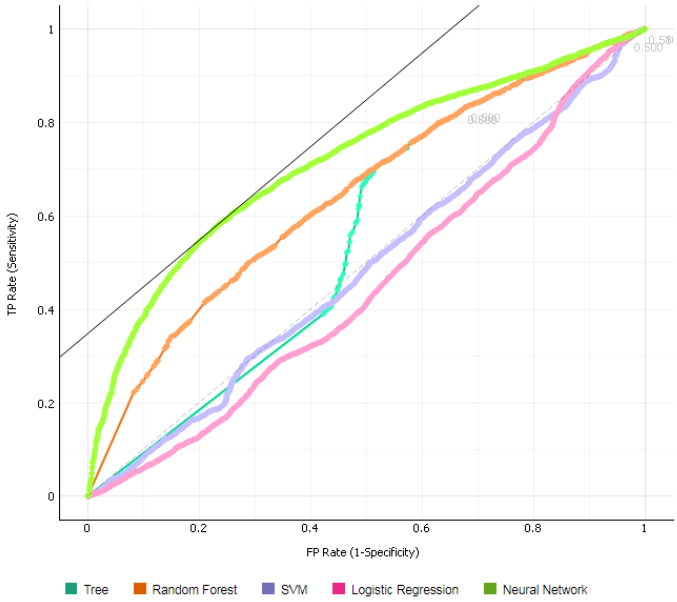
[35] missed value
[36] Mean Allele Frequency
[37] Area Under the receiver operating characteristic Curve

نتایج

پس از محاسبه PRS برای به دست آوردن نتایج طبقه بندی، از مدل های شبکه عصبی ساده[38]، جنگل تصادفی[39]، درخت تصمیم[40]، رگرسیون لجستیک[41] و ماشین بردار پشتیبان[42] (SVM) استفاده گردید. نتایج مدل سازی PRS در جدول ۲ و نمودار ROC در شکل ۴ آمده است.

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

جدول (۲) : نتایج مدل سازی PRS



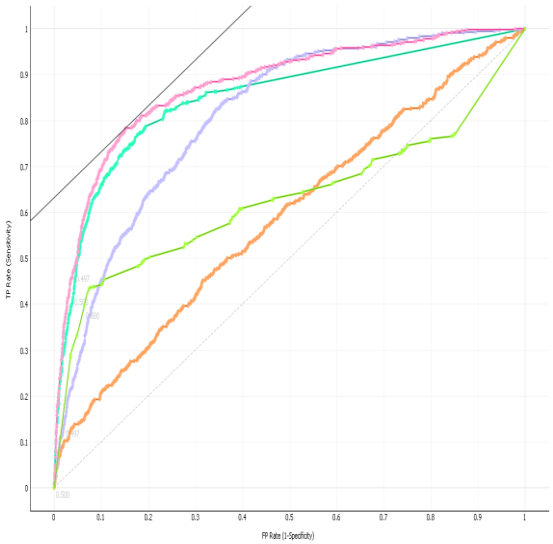
شکل (٤) : نمودار ROC مدل سازی PRS

در مدل سازی CNN برای لایه FCN می توان روش های مختلف طبقه بندی را در نظر گرفت که برای اینکه بتوان با روش قبلی مقایسه انجام داد، تمامی روش های گفته شده، استفاده گردید. نتایج در جدول ۳ و نمودار ROC در شکل ۵ آمده است.

[38] Neural Network
[39] Random Forest
[40] Decision Tree
[41] Logistic Regression
[42] Support Vector Machine

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

جدول (۳) : نتایج مدل سازی CNN



شکل (۵) : نمودار ROC مدل سازی CNN

نتایج نهایی مقایسه ای PRS و مدل سازی CNN در جدول ۴ آمده است.

	AUC	ACC	F1	PRECISION	RECALL
CNN	**۰.۸۷۷**	۰.۹۰۱	۰.۸۹۶	۰.۸۹۳	۰.۹۰۱
PRS	**۰.۷۱**	۰.۷۴	۰.۶۵	۰.۶۵	۰.۷۴

جدول (٤) : مقایسه بهترین نتایج مدل سازی CNN و PRS

بحث و نتیجه‌گیری

در مطالعه حاضر، بیش از ۴۵۰۰۰ نمونه فشار خون در شرکت‌کنندگان TCGS توسط دو مدل PRS و CNN ارزیابی و پیش‌بینی گردید و بنابر نتایج مدلسازی CNN مقدار AUC ۰.۸۷۷ را بدون در نظر گرفتن هیچ عامل خطر اضافی نشان می‌دهد. با توجه به مقادیر AUC که در ادبیات موضوع در مدل سازی های مختلف مبتنی بر عوامل ژنومی برابر ۰.۸۰۴ (Vaura et al. 2021) و مبتنی بر عوامل ژنومی و بالینی برابر ۰.۸۷۱ (Niu et al. 2021) بدست آمده است، کارایی روش CNN کاملا نشان داده می شود و یافته های ما عملکرد بهتری را نشان می دهد. ضمنا در عصر ژنومی، مهمترین دستاورد عملی، ایجاد ابزاری برای تشخیص زودهنگام خطر فشار خون بر اساس نشانگرهای ژنومی برای بهبود پیشگیری و درمان در افراد پرخطر است که رویکرد ما ابزار ارتقاء یافته ای را برای پیش بینی خطر فشار خون بالا در سنین جوانتر ارائه می دهد.

تحقیق ما به دلیل استفاده از اعتبارسنجی داده های داخلی دارای محدودیت هایی است. در این مطالعه، نتایج خارج از داده های TCGS برای اعتبار سنجی خارجی استفاده نشده است که این داده های نیز دارای ترکیب قومی کمی است.

هدف اصلی این مطالعه ارتقای دقت پیش بینی خطر پرفشاری خون با ارائه راهکاری مبتنی بر شبکه عمیق CNN بود که در این مقاله با موفقیت این رویکرد در مقایسه با PRS ارائه گردید. برای کار آینده، می توان با اضافه کردن متغیرهای بالینی موثر، دقت به طور قابل توجهی افزایش داد.

مراجع

Abraham, Gad, and Michael Inouye. 2015. "Genomic Risk Prediction of Complex Human Disease and Its Clinical Application." *Current Opinion in Genetics and Development* 33(Cvd): 10–16. http://dx.doi.org/10.1016/j.gde.2015.06.005.

Aulchenko, Yurii S., Dirk Jan De Koning, and Chris Haley. 2007. "Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method for Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis." *Genetics* 177(1): 577–85.

Azizi, Fereidoun et al. 2009. "Prevention of Non-Communicable Disease in a Population in Nutrition Transition: Tehran Lipid and Glucose Study Phase II." *Trials 2009 10:1* 10(1): 1–15. https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-10-5 (August 7, 2021).

Cebamanos, L., A. Gray, I. Stewart, and A. Tenesa. 2014. "Regional Heritability Advanced Complex Trait Analysis for GPU and Traditional Parallel Architectures." *Bioinformatics* 30(8): 1177–79.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F. O'Reilly. 2020. "Tutorial: A Guide to Performing Polygenic Risk Score Analyses." *Nature Protocols* 15(9): 2759–72. http://dx.doi.org/10.1038/s41596-020-0353-1.

Daneshpour, Maryam 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* 6(2): e28. https://doi.org/10.2196%2Fresprot.6050.

Evangelou, Evangelos, and John P.A. Ioannidis. 2013. "Meta-Analysis Methods for Genome-Wide Association Studies and Beyond." *Nature Reviews Genetics* 14(6): 379–89. http://dx.doi.org/10.1038/nrg3472.

Filshtein, Teresa Jenica et al. 2019. "Reserve and Alzheimer's Disease Genetic Risk: Effects on Hospitalization and Mortality." *Alzheimer's and Dementia* 15(7): 907–16.

Gray, A., I. Stewart, and A. Tenesa. 2012. "Advanced Complex Trait Analysis." *Bioinformatics* 28(23): 3134–36.

Kolifarhood, Goodarz 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* 11(1): 1–14. https://doi.org/10.1038/s41598-021-90925-w.

Luo, Yue et al. 2018. "The Prediction of Hypertension Based on Convolution Neural Network." *2018 IEEE 4th International Conference on Computer and Communications, ICCC 2018*: 2122–27.

Mahajan, Shiwani 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): 1–17. https://www.ahajournals.org/doi/10.1161/JAHA.119.012954.

Martinez-Ríos, Erick, Luis Montesinos, Mariel Alfaro-Ponce, and Leandro Pecchia. 2021. "A Review of Machine Learning in Hypertension Detection and Blood Pressure Estimation Based on Clinical and Physiological Data." *Biomedical Signal Processing and Control* 68(March): 102813. https://doi.org/10.1016/j.bspc.2021.102813.

Mattoo, Tej K. 2019. "Definition and Diagnosis of Hypertension in Children and Adolescents - UpToDate." *UpToDate* (Cv): 1–34. 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.

Mills, Katherine T. et al. 2016. "Global Disparities of Hypertension Prevalence and Control." *Circulation* 134(6): 441–50. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.115.018912.

Mills, Katherine T., Andrei Stefanescu, and Jiang He. 2020. "The Global Epidemiology of Hypertension." *Nature Reviews Nephrology* 16(4): 223–37. http://dx.doi.org/10.1038/s41581-019-0244-2.

Mills, Melinda C., and Charles Rahal. 2019. "A Scientometric Review of Genome-Wide Association


Studies." *Communications Biology* 2(1): 9. http://dx.doi.org/10.1038/s42003-018-0261-x.

Mosley, Jonathan 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): 627–35.

Niu, Miaomiao 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*. http://dx.doi.org/10.1038/s41440-021-00738-7.

Padmanabhan, Sandosh, and Anna F. Dominiczak. 2020. "Genomics of Hypertension: The Road to Precision Medicine." *Nature Reviews Cardiology*. http://dx.doi.org/10.1038/s41569-020-00466-4.

Szymczak, Silke et al. 2009. "Machine Learning in Genome-Wide Association Studies." *Genetic Epidemiology* 33(SUPPL. 1): 51–57.

Tohidi, M., M. Hatami, F. Hadaegh, and F. Azizi. 2012. "Triglycerides and Triglycerides to High-Density Lipoprotein Cholesterol Ratio Are Strong Predictors of Incident Hypertension in Middle Eastern Women." *Journal of Human Hypertension* 26(9): 525–32.

Torkamani, Ali, Nathan E. Wineinger, and Eric J. Topol. 2018. "The Personal and Clinical Utility of Polygenic Risk Scores." *Nature Reviews Genetics* 19(9): 581–90. http://dx.doi.org/10.1038/s41576-018-0018-x.

Vaura, Felix et al. 2021. "Polygenic Risk Scores Predict Hypertension Onset and Cardiovascular Risk." *Hypertension* 77(4): 1119–27. https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.120.16471.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of GWAS Discovery." *American Journal of Human Genetics* 90(1): 7–24. http://dx.doi.org/10.1016/j.ajhg.2011.11.029.

Willer, Cristen J., Yun Li, and Gonçalo R. Abecasis. 2010. "METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans." *Bioinformatics* 26(17): 2190–91.

Wu, Xueyi et al. 2020. "Value of a Machine Learning Approach for Predicting Clinical Outcomes in Young Patients With Hypertension." *Hypertension* 75(5): 1271–78. https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.119.13404.

Yang, Jian et al. 2017. "Concepts, Estimation and Interpretation of SNP-Based Heritability." *Nature Genetics* 49(9): 1304–10.

Zhou, Bin et al. 2017. "Worldwide Trends in Blood Pressure from 1975 to 2015: A Pooled Analysis of 1479 Population-Based Measurement Studies with 19·1 Million Participants." *The Lancet* 389(10064): 37–55.