روشی برای باز‌هدف‌گذاری داروها با در نظر گرفتن عوارض‌جانبی

نام و نام‌خانوادگي نويسنده اول1، نام و نام‌خانوادگي نويسنده دوم2، نام و نام‌خانوادگي نويسنده سوم3

1 رتبه علمی نويسنده در صورت تمايل، گروه آموزشی يا واحد سازمانی مربوطه، نام سازمان، شهر،

آدرس پست الکترونيکی

2 رتبه علمی نويسنده در صورت تمايل، گروه آموزشی يا واحد سازمانی مربوطه، نام سازمان، شهر

آدرس پست الکترونيکی

3 رتبه علمی نويسنده در صورت تمايل، گروه آموزشی يا واحد سازمانی مربوطه، نام سازمان، شهر

آدرس پست الکترونيکی

چكيده

تحقیق و توسعه در زمینه تولید و کشف داروهای جدید پروسه‌ای طولانی و پر‌هزینه است. گزارش‌ها نشان می‌دهد تولید یک داروی جدید از مرحله اول تا رساندن آن به بازار مصرف بین 10 تا 15 سال و با صرف هزینه‌ای بالغ بر 0.8 تا 1.5 میلیارد دلار صورت می‌پذیرد. از این‌رو یافتن یک استراتژی جدید برای کشف دارو امری ضروری به نظر می‌رسد. بازهدف‌گذاری دارو به معنی شناسایی کاربردهای جدید، خارج از محدوده کاربردهای اصلی، برای داروهای تأیید شده است. این کار از طریق شناسایی ویژگی‌های داروها بر پایه محاسبات کامپیوتری و استفاده از قابلیت یادگیری ماشین و بهره‌گیری از کلان‌داده‌ها صورت پذیرفته و باعث کاهش هزینه و ریسک فرآیند تحقیق و توسعه می‌شود.

این پژوهش روشی را برای باز‌هدف‌گذاری دارو مبتنی بر ساخت شبکه‌ای یکپارچه از داده‌های متنوع و ناهمگون و پیشنهاددهی مجدد داروها به روش فاکتورگیری ماتریسی با لحاظ کردن عوارض جانبی داروها ارائه می‌دهد. روش پیشنهادی به ترتیب معیار‌های AUC و AUPR را 1.13 و 14.23 درصد نسبت به روش‌های پیشین بهبود داده است. در این روش ترکیب ویژگی‌های مختلف دارو-بیماری منجر به کسب بیشترین دقت نتیجه شده است که خود کارایی نحوه خلاصه‌سازی و کاهش بعد داده، یکپارچه‌سازی و تجمیع شباهت‌ها را در این روش اثبات می‌کند.

كلمات كليدي

بازهدف‌گذاری دارو، پیش‌بینی پیوند، فاکتورگیری ماتریسی، عوارض جانبی

# مقدمه

تحقیق و توسعه در زمینه تولید و کشف داروهای جدید پروسه‌ای طولانی و پر‌هزینه است. گزارش‌ها نشان می‌دهد تولید یک داروی جدید از مرحله اول تا رساندن آن به بازار مصرف بین 10 تا 15 سال و با صرف هزینه‌ای بالغ بر 0.8 تا 1.5 میلیارد دلار صورت می پذیرد. با این حال علی‌رغم تمام سرمایه‌گذاری‌ها و پیشرفت‌های انجام‌شده در 20 سال اخیر، تعداد داروهای جدیدی که مورد تایید انجمن غذا و دارو آمریکا (FDA) قرار گرفته، چشمگیر نبوده و سالانه نزدیک به 90 درصد داروهای ارزیابی شده توسط این انجمن رد شده و از عرضه‌شان برای مصارف درمانی جلوگیری می‌گردد. طولانی‌بودن پروسه تولید دارو به علت حساسیت و اهمیت تکمیل تمام مراحل آزمایشگاهی می‌باشد[1].

بازهدف‌گذاری دارو به معنی شناسایی کاربردهای جدید، خارج از محدوده کاربرد اصلی برای دارو‌های از قبل تأیید‌شده است. این استراتژی نسبت به کشف داروی جدید برای درمان یک بیماری خاص مزایای مختلفی دارد. اول و شاید مهم‌تر از همه ریسک شکست پایین‌تر این استراتژی است. زیرا در صورت تکمیل مراحل اولیه آزمایش در مدل‌های بالینی و آزمایش‌های انسانی، داروی مدنظر به اندازه کافی ایمن بوده، و احتمالاً از نظر ایمنی در آزمایش‌های بعدی نیاز به اثبات نخواهد داشت. دوم اینکه، با این روش می‌توان مدت زمان تولید دارو را کاهش داد، زیرا بیشتر آزمایش های بالینی، ارزیابی ایمنی و در برخی موارد، تدوین فرمولاسیون از قبل تکمیل شده است. و سوم، سرمایه‌گذاری کمتری مورد نیاز است. این مزایا می‌توانند منجر به بازگشت کم‌خطرتر و سریع‌تر سرمایه‌گذاری در تولید داروهای بازهدف‌گذاری شده شوند[2].

همچنین در مواردی سرعت شیوع یک بیماری جدی مثل کرونا بسیار بالاست و نمی‌توان تا زمان ساخت داروی جدید منتظر ماند. یا در سناریوی دیگری یک بیماری نادر ولی خطرناک شناسایی می‌شود و به دلیل نادر‌بودن بیماری و تعداد کم افراد مبتلا شده، صرف هزینه چند میلیارد دلاری برای شرکت‌های داروسازی توجیه اقتصادی نخواهد داشت لذا پزشکان و محققان باید به بررسی اثربخشی داروهای موجود برای درمان این بیماری بپردازند.

درحال حاضر رشد انفجاری انواع داده‌های زیستی در ابعاد بزرگ و با دقت بالا موجب شکل‌گیری زمینه تحقیقاتی با نام داروشناسی محاسباتی و ایجاد فرصتی برای تحلیل و آنالیز سیستماتیک داده‌های مختلف شده است و تحلیل این داده‌ها جهت بهبود و کاهش ریسک فرآیند توسعه داروها می‌تواند مفید واقع گردد.

سیلدانفیل با نام تجاری ویاگرا نمونه‌ای موفق از بازهدف‌گذاری دارو می‌باشد. سیلدنافیل ابتدا در اواخر دهه 80 میلادی برای درمان بیماری درد قفسه سینه تجویز شد. دارو در طول آزمایشات بالینی فاقد کارآیی لازم بر روی این بیماری شناخته‌شد و روند توسعه آن متوقف گردید تا اینکه بیماران اثرات‌جانبی غیر‌معمول از جمله نعوظ طولانی‌مدت را گزارش کردند. بعد از این بود که محققان با انجام تحقیقات مجدد دارو را برای درمان اختلال نعوظ بازهدف‌گذاری کردند.

مساله پیدا‌کردن مصرف جدید درمانی برای داروها را می‌توان در قالب مساله کلی پیش‌بینی‌پیوند در شبکه و سیستم‌های توصیه‌گر مدل‌سازی کرد. در سیستم‌هاي توصیه‌گر، دو مجموعه از کاربران و اقلام وجود دارد که ارتباطات بین اعضای این دو مجموعه را می‌توان در قالب یک گراف دو‌بخشی مدل کرد. تعاملات و پیوندها در شبکه‌های زیستی و به طور خاص شبکه دارو-بیماری نیز بر همین پایه مدل‌سازی می‌شود.

شبکه‌های اجتماعی، زیستی و ‌‌‌‌‌اطّلاعاتی ‌می‌توانند توسط شبکه‌های پیچیده و پویا توصیف شوند. در این شبکه‌ها، گره‌ها نشان‌دهنده موجودیت‌ها و پیوند‌ها نشان‌دهنده ارتباطات بین موجودیت‌ها هستند. هنگام ایجاد تقریب توپولوژیکی از یک سیستم پیچیده، به علّت محدودیت زمان و فضا، وجود بعضی خطا‌ها یا پیوند‌های زائد اجتناب‌ناپذیر است؛ همزمان بعضی پیوند‌های بالقوّه کشف‌نشده نیز وجود خواهد داشت. بنابراین در این شبکه‌ها امکان وجود پیوند میان دو گره، براساس ویژگی‌های گره‌ها و دیگر پیوند‌های مشاهده‌شده در شبکه بررسی می‌گردد. پیوند‌هایی که ‌‌پیش‌بینی ‌می‌شوند، ‌می‌توانند پیوند‌های گمشده یا پیوند‌های آینده باشند[3].

یكی از روش‌های مهم استفاده شده در سیستم‌های پیشنهاددهنده، روش مبتنی بر فاکتورگیری ماتریس است. در این روش، با استفاده از امتیازات داده شده‌ی کاربران به برخی از آیتم‌ها، برای هر کاربر و هر آیتم، یک بردار ویژگی استخراج می‌شود که طول این بردار برابر با ویژگی‌های نهان کاربر یا آیتم است. ویژگی‌های نهان آیتم‌ها علاوه بر مقادیر به‌دست آمده از ماتریس امتیازها به محتوای آیتم‌ها نیز وابسته است.

توصیه‌گرهايی که مبتنی بر فاکتورگیري ماتریسی هستند با تبدیل کاربران و اقلام به فضاي ویژگی‌هاي نهفته، هرکدام از درایه‌هاي نامعلوم در ماتریس امتیازات را با استفاده از ضرب‌داخلی بردارهاي ویژگی متناظر که از امتیازات معلوم استنتاج شده‌اند، تعیین می‌کند. راهکار پیشنهادي برای مقداردهی اولیه استفاده از ماتریس شباهت بین کاربران و اقلام در مقدار‌دهی اولیه به جای استفاده از مقادیر تصادفی است. نتیجه هر بار ضرب ماتریس‌های ویژگی کاربر و اقلام در یکدیگر باید به ماتریس اصلی امتیازات نزدیک‌‌تر شود. برای این‌کار یک تابع هزینه تعریف می‌شود که هر بار این فاصله را بوسیله کسر ماتریس بدست‌آمده حاصل‌ضرب از ماتریس اولیه اندازه می‌گیرد. با توجه به فاصله بدست‌آمده هر بار ماتریس‌های اولیه‌ی ویژگی با توجه به قوانین بروزرسانی، آپدیت می‌گردند و دوباره عمل ضرب مذکور تکرار می‌گردد. با توجه به اینکه پیدا کردن کمینه عمومی براي این مسئله غیرممکن است، ضرب مذکور تا زمانی که حاصل تفریق و فاصله ماتریس بدست آمده با ماتریس اولیه در دو گام پشت سر‌هم از مقدارتعیین شده‌ای کمتر گردد، ادامه می‌یابد. دقت و همگرایی در این روش معمولاً خیلی وابسته به مقداردهی اولیه ماتریس‌های ویژگی است. اگر مقداردهی اولیه به‌صورت درست انجام نشود، همگرایی به کندي صورت می‌گیرد و حتی در بعضی موارد ممکن است به راه‌حل نادرست بیانجامد[4].

ادامه این مقاله بصورت زیر سازمان‌دهی شده است. در بخش 2 پژوهش‌های پیشین انجام‌شده در زمینه بازهدف‌گذاری دارو و رویکرد‌های مختلف و منابع داده‌ای معرفی می‌گردد. در بخش 3 جزئیات روش پیشنهادی بازهدف‌گذاری دارو با استفاده از اطلاعات عوارض جانبی، شرح داده شده و روش ارائه شده شامل استفاده از تکنیک‌های گشت تصادفی برای ترکیب ویژگی‌های مختلف دارویی و بیماری، روش فاکتور‌گیری ماتریسی شرح داده می‌شود. بخش 4 به بررسی نتایج حاصل از ارزیابی روش پیشنهادی پرداخته و نتایج آن با روش‌های پیشین مقایسه می‌شود و روی برخی از ارتباطات پیش‌بینی شده توسط روش پیشنهادی، بررسی موردی انجام می‌گیرد. در پایان، بخش 5 به بحث و نتیجه‌گیری درخصوص دستاوردهای بدست آمده می‌پردازد و جهت کارهای آتی پیشنهاداتی ارائه می‌دهد.

# پژوهش‌های پیشین

پژوهش‌های صورت گرفته در این خصوص را می‌توان از دو منظر نوع داده مورد استفاده قرار داده شده و نیز از منظر رویکرد محاسباتی به کار گرفته شده، مورد بررسی قرار داد.

با توسعه سریع تکنیک‌های مرتبط با محاسبات زیستی، بانک‌های اطلاعاتی مختلفی از دارو مثل ChemBank, OMIM, Pubmed, DrugBank و بانک‌های اطلاعاتی ژنومی گسترده‌ای نیز مانند MIPS, GenBank و ... ایجاد شده‌اند. پایگاه‌هایی که هر کدام اختصاص به یک یا چند دسته از داده‌های مرتبط با دارو، پروتئین، بیماری و ژن دارند[5].

روش‌هایی که از داده‌های مرتبط با دارو استفاده می‌کنند از اصل شباهت ویژگی این‌چنین بهره می‌گیرند که داروهایی با ساختار شیمیایی مشابه فعالیت‌های بیولوژیکال یکسانی را انجام می‌دهند. روش‌های مبتنی بر اثرانگشت مولکولی و روش‌های مبتنی بر لیگاند در این دسته قرار می‌گیرند. در روش‌های مبتنی بر اثرانگشت مولکولی هر دارو در یک بردار دودویی با مقادیر 0 و 1 کدگذاری می‌شود که صفر و یک به ترتیب به معنی وجود/عدم وجود یک زیرساختار شیمیایی خاص در ترکیب دارو است. در مقاله وانگ و همکاران روشی برای بازهدف‌گذاری دارو براساس اندازه‌گیری شباهت دو بردار دارو ارائه شده است[6].

بیشتر مولکول‌های کوچک زیست‌فعال اثرات خود را با بر‌هم‌کنش با پروتئین‌ها اعمال می‌کنند. این بر‌هم‌کنش را می‌توان با استفاده از مدل‌سازی کامپیوتری ساختار سه‌بعدی پروتئین هدف و دارو تحلیل کرد. این روش موسوم به داکینگ مولکولی بوده و در راهبردهای کشف دارو به کار می‌رود. اروچی و همکاران در پژوهشی در سال 2019 با استفاده از این روش به بازهدف‌گذاری دارویی برای هدف پروتئینی Pim-1 پرداختندکه به عنوان عامل سرطان پروستات و لوسمی حاد مغز استخوان شناخته‌‌شده است[7].

از آن جایی که بیماری‌های مشابه با روش مشابه، درمان می‌شوند، توصیف بهتر ارتباط موجود بین بیماری‌زایی می‌تواند فرضیه‌های بازهدف‌گذاری دارو را پایه‌ریزی کند. اولین کار صورت گرفته با استفاده از این متد پژوهش اگروال و همکاران در سال 2009 است که از طریق بررسی مسیر بیولوژیکی ژن‌های بیماری‌زا به بررسی الگوهای مشترک بیولوژیکی بیماری‌ها پرداخته است[8].

ارگان‌های زنده موجودات را می‌توان با مشاهده عملکرد بیان ژنی آن‌ها در یک حالت خاص درک کرد. بر اساس حالت ارگان زنده، ژن‌های مشخصی بیشتر و یا کمتر بیان می‌شوند و این ژن‌ها از طریق بررسی تعداد نسبی مولکول‌های mRNA رونویسی شده، قابل شناسایی هستند. تمایز بیان ژن را می‌توان دلیلی بر اثر مولکولی خاص دانست که تحت عنوان امضای بیان ژن شناخته می‌شود. یکی از موفق‌ترین نمونه‌های این روش پایگاه‌داده CMap است. این پایگاه که نتیجه تحقیق لمب و همکاران بوده در سال 2006 بوجود آمده و بر اساس این ایده که عملکرد دارو را می‌توان با بررسی پروفایل بیان ژن حاصل از مصرف آن دارو بر روی یک سیستم بیولوژیکی بدست آورد، کار می‌کند[9].

از آنجایی که داده‌های بیولوژیک از طرفی در برخی موارد ناکامل و محدود هستند و از طرف دیگر مفاهیم مکملی را بیان می‌کنند لذا می‌توان از ترکیب منابع‌داده‌ای مختلف بهره گرفت که علاوه بر بالا‌رفتن دقت فرآیند باعث ارائه پیش‌بینی‌هایی منحصر به فرد نیز بشوند. روش‌هایی که تا اینجا مورد بحث قرار گرفتند، هر کدام مبتنی بر یکی از مفاهیم بیولوژیکی خاص بودند. حال از منظر رویکرد محاسباتی مورد استفاده قرار گرفته در بازهدف‌گذاری به بررسی انواع روش‌ها می‌پردازیم.

استفاده از الگوریتم‌هایی مانند ماشین بردار پشتیان، کمترین مربعات و رگرسیون لجستیک جزء قدیمی‌ترین روش‌های بکار گرفته شده در حوزه یادگیری ماشین به‌شمار می‌آیند. در پژوهشی مربوط به سال 2017 پنگ و همکاران روشی به نام NDTISE توسعه دادند که در آن ابتدا به بررسی عمیق‌تر نمونه‌های بدون برچسب (زوج‌هایی که ارتباط بین آن‌ها گزارش نشده) پرداختند[10]. این نمونه‌ها به دو دسته تقسیم شد و نمونه‌های مشکوک از نمونه‌هایی که با قطعیت بیشتر می‌توان درباره عدم ارتباط آن‌ها نظر داد جدا شد. در مرحله بعد تنها از مجموعه دوم نمونه‌هایی به عنوان نمونه‌های منفی به مدل داده شد و در نهایت یک طبقه‌بند پشتیبان وظیفه تصمیم‌گیری را انجام داد.

تحلیل شبکه یکی از استراتژی‌هایی که به طور گسترده در سال‌های اخیر برای بازهدف‌گذاری مورد استفاده قرار گرفته است. مهم‌ترین تفاوت این رویکرد با رویکرد پیشین در نظر گرفتن ویژگی‌های توپولوژیکی است. امکان استفاده از معیارهای شباهت محلی و سراسری وجود دارد که این امر باعث حفظ اطلاعات کمی و کیفی شبکه شده و یکپارچه‌سازی شباهت‌ها را بطور موثرتری اجرا می‌کند. در این رویکرد معمولاً یک شبکه نامتجانس با چند سطح تعریف می‌شود که در هر سطح آن ویژگی‌های یک موجودیت خاص(دارو، بیماری و پروتئین) وجود دارد. در پژوهش لو و همکاران که در سال 2016 برای شناسایی داروی جدید جهت بهبود درمان نوعی خاصی از سرطان ریه انجام شده از خوشه بندی k-means استفاده شده است[11]. در این پژوهش ابتدا گروه داروهایی با ساختار شیمیایی یکسان با داروهای از پیش تایید شده برای درمان سرطان ریه که روی ژن‌های این بیماری اثرگذاری دارند شناسایی گردید و سپس ادامه مطالعه بر روی آن گروه پیگیری گردید. همچنین برای چالش خوشه‌هایی که روی هم قرار گرفته‌اند و با هم گره‌های مشترک دارند نیز راه‌حلی ارائه شده است.

در پژوهش کدیا و همکاران در سال 2020 روشی بر پایه فاکتورگیری ماتریسی غیرمنفی ارائه شده که در آن روی ماتریس‌های تجزیه‌شده‌ی ویژگی در حین عملیات ضرب و بروزرسانی محدودیت غیرمنفی بودن اعمال شده است[12]. علاوه بر آن در این روش جهت مقدار‌دهی اولیه ماتریس‌های ویژگی علاوه بر ماتریس شباهت، کوتاه‌ترین مسیر بین گره‌ها در هر سطح از شبکه بطور مجزا حساب شده و تاثیر آن در مقداردهی اعمال می‌شود.

یو و همکاران در پژوهشی در سال 2020 بسته جامعی به نام BioNEV طراحی کردند که با استفاده از آن دو دسته وظیفه پیش‌بینی پیوند و طبقه‌بندی چند کلاسه در شبکه‌های زیستی قابل اجراست[13]. در بخش پیش‌بینی پیوند سه شبکه مورد علاقه پژوهشگران یعنی دارو-بیماری، برهم‌کنش‌های پروتئینی، تداخلات دارویی و در بخش طبقه‌بندی نیز پیش‌بینی کارکرد پروتئین و طبقه‌بندی انواع عبارات پزشکی پیاده‌سازی شده است.

با استفاده از روش‌های یادگیری عمیق ویژگی‌های ساختاری یک شبکه و وضعیت توپولوژیکال یک گره در سطح شبکه را بعنوان یک ویژگی استخراج کرد و از آن برای محاسبه شباهت دو گره همسان بهره برد. در این روش می‌توان ویژگی‌های هر نمونه دارو، بیماری، ژن، پروتئین و ... بعنوان یک بردار دودویی به صورت خام و بدون دادن هیچ اطلاعات اضافی‌ای به مدل داد و یا ابتدا از روش‌های دیگر بازنمایی اطلاعات مفید را استخراج و سپس به عنوان ورودی به مدل داد. deepDR نام مدلی است که ژنگ و همکاران در سال 2019 برای بررسی بازهدف‌گذاری داروها ارائه دادند[14]. در این مدل ابتدا از گشت تصادفی برای بازنمایی شبکه‌ها استفاده می‌شود و سپس محققین از یک خودرمزگذار برای یکپارچه‌سازی و فشرده‌سازی داده‌ها بهره بردند. مدل در نهایت برای پیش‌بینی پیوند‌های جدید از cVAE استفاده کرده و میزان AUC=0.90 و AUPR=0.92 می‌گردد.

مقالات، کتب و بطور کلی متون منتشرشده در حوزه‌های پزشکی و داروسازی منبع غنی و عظیمی از اطلاعات درخصوص دارو و بیماری به حساب آمده و زمینه مناسبی را جهت داده‌کاوی آن فراهم کرده است. در پژوهشی که در سال 2019 برای کشف دارویی جهت درمان ریزش موی ناشی از شیمی‌درمانی بوسیله ژانگ و همکاران صورت گرفت ابتدا با انجام یک سری جستجو با واژه‌های کلیدی مرتبط مجموعه متونی از پایگاه‌داده‌های زیستی گردآوری شد. با استفاده از واژه‌نامه تمامی ژن‌های مرتبط شناسایی و ارتباط‌شان نیز در قالب شبکه بازنمایی شد. گره‌های بزرگتر تعداد تکرار بیشتری در متون داشتند و یال‌های ضخیم‌تر نیز بواسطه دفعات تکرار مرتبط بودن ژن‌ها در عبارات نشان‌دهنده وزن بیشتر پیوند بودند. بعد از پیداشدن ژن‌های بیماری‌زای موثر و شبیه به یکدیگر، تمامی داروهای مرتبط از پایگاه داده DGIdb استخراج و مطالعات بعدی صرفاً روی این دسته انجام گردید[15].

# مجموعه داده و روش پیشنهادی

روش ارائه شده در این مطالعه شامل استفاده از تکنیک‌های گشت تصادفی برای ترکیب ویژگی‌های مختلف دارویی و بیماری، تجزیه مؤلفه انتشار (DCA) برای استخراج ویژگی‌های مهم‌تر بوده و در نهایت از روش فاکتور‌گیری ماتریسی برای پیش‌بینی نهایی استفاده می‌کند. روش پیشنهادی در گام‌های زیر خلاصه می‌شود:

* جمع‌آوری ویژگی‌های مختلف
* محاسبه ماتریس شباهت از روی ویژگی‌ها با استفاده از تابع جاکارد
* محاسبه ماتریس احتمالات گشت تصادفی برای هر ماتریس شباهت
* خلاصه‌سازی ماتریس‌های حاصل از گشت‌تصادفی و ایجاد یک ماتریس ترکیبی با ابعاد پایین برای دو مجموعه‌ی دارو و بیماری
* استفاده از فاکتورگیری ماتریس برای پیش‌بینی احتمال تعاملات دارو و بیماری

شکل ‏3‑1 نمای کلی این روش ‏را نشان می‌دهد.

## مجموعه داده

بخشی از داده‌های استفاده شده در این پژوهش، از مطالعه لیانگ و همکاران دریافت شده است[16]. محققین این پژوهش داده‌های مورد نیاز را از پایگاه‌داده‌های معتبری چون PubChem، InterPro، UniProt و DincRNA دریافت و پاک‌سازی کرده‌اند. این مجموعه ‌داده شامل سه نوع ماتریس ویژگی برای داروها، یک ماتریس شباهت تجمیع شده و ماتریس تعاملات دارو-بیماری می‌باشد:

* ویژگی ساختار شیمیایی که از پایگاه داده PubChem دریافت شده است[17].
* ویژگی هستان‌شناسی ژن‌های مرتبط با پروتئین‌های هدف که از پایگاه داده InterPro دریافت شده است[18].
* ویژگی ساختار سوم و کروی کمپلکس پروتئین‌های هدف که از پایگاه داده UniProt ‌دریافت شده است[19].
* ویژگی‌ ژن‌های مرتبط هر بیماری که از پایگاه‌داده DincRNA دریافت شده است[20].
* ویژگی معنایی بیماری‌ها که بوسیله متن‌کاوی مقالات موجود در وب و مرتبط با هر بیماری استخراج شده است.
* ماتریس تعاملات دارو-بیماری

در پژوهش ما علاوه بر مجموعه داده بالا و در جهت بهبود عملکرد فرآیند بازهدف‌گذاری، از ویژگی چهارمی نیز به نام عوارض‌جانبی دارو استفاده گردید. ابتدا مجموعه کلی داروهای قرار داده شده در پایگاه‌داده SIDER دریافت شد و در گام بعدی صرفاً ویژگی‌ مربوط به داروهای موجود از میان داده‌ها استخراج گردید و به فرمت قابل بهره‌برداری تبدیل گردید[21].

## محاسبه ماتریس‌های شباهت

ویژگی‌های گفته شده در بالا، با استفاده از تابع جاکارد با فرمول زیر به ماتریس شباهت تبدیل می‌شوند.

|  |  |
| --- | --- |
| *(1)* |  |

دو بردار ویژگی A و B را در نظر بگیرید که هر کدام شامل n درایه با مقادیر ۰ با ۱ هستند.

* به معنای تعداد درایه‌هایی است که در هر دو بردار A و B مقدار ۱ دارد.
* به معنای تعداد درایه‌هایی است که در بردار A برابر ۰ و در B برابر ۱ است.
* به معنای تعداد درایه‌هایی است که در A برابر ۱ و در B برابر ۰ است.

تابع جاکارد برای هر جفت دارو، بردار دودویی ویژگی مربوط به آن دارو‌ها را دریافت می‌کند و با استفاده از فرمول بالا میزان شباهت آن‌ها را محاسبه می‌کند. به عنوان مثال برای محاسبه شباهت دو دارو و بر اساس ویژگی پروتئین‌های هدف، سطر ام و سطر ام ماتریس ویژگی پروتئینی به عنوان ورودی به تابع جاکارد داده می‌شود.

## من خودمم نه خاطرهایجاد ویژگی‌های خلاصه‌شده برای دارو و بیماری

شکل ‏3‑1 نمای کلی روش پیشنهادی

به محاسبه ویژگی‌های با ابعاد کمتر از روی ویژگی‌‌های اولیه، یادگیری ویژگی‌های خلاصه می‌گویند. روش پیشنهادی از تجزیه و DCA[22] برای یادگیری و ایجاد نمایش‌های برداری با ابعاد کوچکتر از ویژگی‌های اصلی ماتریس‌های شباهت، استفاده می‌کند. این الگوریتم که اخیراً توسعه یافته، ترکیبی از روش انتشار شبکه (گشت تصادفی با راه‌اندازی مجدد) و روش کاهش ابعاد بوده و قادر است خصوصیات توپولوژیکی یک شبکه را نیز استخراج کند[23].

### گشت تصادفی با راه‌اندازی مجدد

گشت تصادفی با راه‌اندازی مجدد (RWR)، یک الگوریتم انتشار شبکه است که به‌طور گسترده برای تجزیه و تحلیل داده‌های پیچیده‌ی شبکه بیولوژیکی استفاده شده است[24]–[28]. برخلاف روش‌های گشت تصادفی معمولی، RWR یک احتمال از پیش تعریف شده‌ای برای راه‌اندازی مجدد در گره اولیه در هر تکرار را معرفی می‌کند که می‌تواند الگوی اتصال توپولوژیکی محلی و سراسری در شبکه را در نظر بگیرد تا به طور کامل از روابط مستقیم یا غیرمستقیم زمینه‌ای بین گره‌ها بهره ببرد. به بیان دقیق‌تر ، فرض کنید ماتریس مجاورت یک شبکه شباهت داروها (یا بیماری‌ها) باشد. ما نیز ماتریس دیگری را تعریف می‌کنیم، که در آن هر عنصر احتمال انتقال از گره‌ به گره را نشان می‌دهد، یعنی:

|  |  |
| --- | --- |
| *(2)* |  |

گشت تصادفی برای هر 5 ماتریس شباهت دارویی اجرا می‌شود و نتایج حاصل از گشت تصادفی که در این مرحله شامل اطلاعات ساختاری شبکه نیز هست، به‌دست می‌آید. ابعاد هر یک از این ماتریس‌ها، مشابه ابعاد ماتریس‌های اولیه یعنی تعداد دارو در تعداد دارو و یا تعداد بیماری در تعداد بیماری است.

### چارچوب کاهش ابعاد

حالت‌های انتشار ناشی از فرآیند RWR فوق الذکر، ممکن است به دلیل کیفیت نسبتاً پایین و ابعاد زیاد داده‌های زیستی کاملاً دقیق نباشد. فقدان اطلاعات یا اطلاعات اشتباه در شبکه، حتی به مقدار کم هم می‌تواند به طرز قابل توجهی بر نتایج فرآیند انتشار اثر بگذارد[29]. علاوه بر این، به طور کلی استفاده از داده با ابعاد بالا به عنوان ویژگی‌های ورودی کار مناسبی نیست. برای پرداختن به این مسئله، روش پیشنهادی از یک تکنیک کاهش ابعاد، به نام تجزیه و تحلیل مؤلفه انتشار DCA، استفاده می‌کند تا ابعاد فضای ویژگی‌ها را کاهش داده و ویژگی‌های مهم توپولوژیکی حالت انتشار را به‌دست آورد.

بطور خاص، احتمال اختصاص داده شده به گره در حالت انتشار گره اکنون به طور زیر مدل می‌شود.

|  |  |
| --- | --- |
| *(3)* |  |

که به ویژگی زمینه گره و به ویژگی گره گفته می‌شود و هر دو، ویژگی‌های توپولوژیکی شبکه را توصیف می‌کنند. DCA مجموعه‌ای از حالت‌های مشاهده شده را به عنوان ورودی در نظر می‌گیرد. در این روش با استفاده از آنتروپی نسبی (واگرایی کولبک-لیبرKL) به دنبال کاهش ورودی به برداری با ابعاد کمتر است به طوری که توزیع ورودی با خروجی دارای کمترین تفاوت باشد. این عمل برای تمام ها و ها توسط کمینه‌سازی تابع زیر انجام می‌شود:

|  |  |
| --- | --- |
| *(4)* |  |

که نشان‌گر واگرایی KL بین دو توزیع است. چارچوب DCA برای بهینه‌سازی این تابع از یک روش شبه-نیوتن استاندارد به نام L-BFGS[30] استفاده می‌کند. شایان ذکر است روابط بکار گرفته شده جهت کاهش بعد، یکپارچه سازی و پیش‌بینی نهایی از پژوهش لو و همکاران برگرفته شده است[31].

### یکپارچه‌سازی اطلاعات ماتریس‌ شباهت‌های مختلف

به منظور در نظر گرفتن تمام ماتریس‌های شباهت در فرآیند کاهش ابعاد، چارچوب کاهش ابعاد فوق برای ادغام چندین ماتریس شباهت دارویی گسترش داده می‌شود. فرض کنیم که ماتریس شباهت دارویی داریم. برای ترکیب شبکه‌ها با اطلاعات متنوع (ماتریس‌های شباهت بر اساس ویژگی‌های مختلف) ابتدا RWR بر روی شبکه‌ها به طور جداگانه اجرا شده و سپس حالت‌‌های انتشار شبکه برای هر نود در شبکه ، از طریق فرمول زیر به‌دست می‌آید:

|  |  |
| --- | --- |
| *(5)* |  |

این فرمول مشابه فرمول (۴) می‌باشد؛ با این تفاوت که به ازای هر متغیر است. در واقع، بردار ویژگی زمینه گره برای هر شبکه با هر گره با و بردارهای ویژگی در بین تمام شبکه‌ها مشترک هستند. در نهایت، DCA تابع هدف زیر را بهینه می‌کند:

|  |  |
| --- | --- |
| *(6)* |  |

در این مطالعه احتمال راه‌اندازی مجدد برای گشت تصادفی برابر با ۰.۵ در نظر گرفته شده است. هم‌چنین با توجه به نتایج آزمایش‌ها تحقیقات پیشین[31] مقادیر ابعاد ویژگی‌های پنهان برای دارو و برای بیماری در نظر گرفته شده است که برابر با 10٪ و 15٪ ابعاد اصلی می‌باشد. علاوه بر این، بیشینه تعداد تکرارهای روش پییشنهادی ۲۰ تکرار تعیین شده است.

### پیش بینی از فضای دارویی بر روی فضای بیماری

ما از بردار‌های ابعاد پایین ویژگی‌های دارویی و بیماری که از یادگیری ویژگی‌های خلاصه به‌دست آمده‌اند، برای پیش‌گویی تعاملات جدید بیماری و دارو استفاده می‌کنیم بطوریکه بردارهای ویژگی دارویی پیش‌گویی شده از لحاظ هندسی به ارتباطات شناخته شده با بیماری مرتبط خود نزدیک هستند.

به زبان ریاضی، ما از برای نشان دادن ماتریس ویژگی‌های دارویی (که هر سطر ماتریس، نشان‌دهنده‌ی ویژگی داروی است) و از () برای نشان دادن ماتریس ویژگی‌های بیماری (که هر سطر ماتریس، نشان‌دهنده‌ی ویژگی بیماری است) استفاده می‌کنیم. لازم به ذکر است که و به ترتیب نشان‌دهنده‌ی تعداد دارو و تعداد بیماری‌هاست. فرض کنید نشان‌دهنده ماتریس ارتباطات دارو و بیماری‌ها باشد؛ به طوری که نشان‌دهنده وجود تعامل بین دارو و بیماری و نشان‌دهنده عدم چنین تعاملی باشد. برای یادگیری ماتریس پیش‌بینی ، یک تابع دوسویه تعریف شده است تا تعاملات بیماری-دارویی ناشناخته در (یعنی آن مقادیر صفر در ماتریس ارتباطات) را پیش‌گویی کند. به بیان دقیق‌تر، تابع دوسویه به صورت زیر تعریف می‌شود:

|  |  |
| --- | --- |
| *(7)* |  |

مخفف ماتریس تعاملات دارو- بیماری است و از مرحله یادگیری ویژگی‌های خلاصه‌شده (به عنوان مثال، انتشار شبکه و فرآیندهای کاهش ابعاد شبکه) و ماتریس طرح است که باید آموزش داده شود. سپس از فرمول زیر استفاده می‌کنیم تا مقدار احتمال تعامل بین جفت دارو و بیماری را محاسبه کنیم:

|  |  |
| --- | --- |
| *(8)* |  |

هر چه عدد بزرگ‌تر باشد، به این معنا است که داروی با احتمال بیشتری با بیماری تعامل دارد. ماتریس طرح‌ دارای ابعاد است، به طور معمول بین آن دسته از بردار‌های دارو یا بیماری‌ که از نظر هندسی در فضا نزدیک هستند، هم‌بستگی قابل‌توجهی وجود دارد که در نتیجه می‌تواند تعداد پارامترهای مؤثر مورد نیاز برای مدل‌سازی اثرات متقابل دارو را کاهش دهد. به همین دلیل، ما یک محدودیت رتبه پایین روی اعمال می‌کنیم تا فقط تعداد کمی از فاکتور‌های پنهان از طریق تجزیه مرتبه پایین که محاسبه شوند. این محدودیت رتبه پایین نه تنها مشکل بیش برازش را حل می‌کند بلکه از نظر محاسباتی روند بهینه‌سازی را بهبود می‌دهد[32].

مسئله‌ی بهینه‌سازی با محدودیت رتبه پایین در ماتریس طرح اولیه ، یک مسئله NP-سخت است. یک آرام‌سازی استاندارد از محدودیت درجه پایین، به حداقل رساندن نرم ردیابی (یعنی مجموع مقادیر منفرد) ماتریس است، که معادل کمینه‌سازی نرم‌های فروبنیوس است. بنابراین، با حل کردن مسئله بهینه‌سازی زیر می‌توان فاکتورگیری Z به G و H را انجام داد:

|  |  |
| --- | --- |
| *(9)* |  |

در اینجا پارامتر ، پارامتر منظم‌سازی است که تعادل بین کمینه‌سازی مربعات خطا در جفت تعاملات شناخته شده و نرم فروبنیوس را کنترل می‌کند. مسئله بهینه‌سازی را می‌توان با کمینه‌سازی متناوب حل کرد[33].

# نتایج و ارزیابی

در این مطالعه، برای ارزیابی مدل از روش اعتبارسنجی متقاطع طبقه‌ای و معیارهای اصلی دسته‌بندی استفاده می‌کنیم. اعتبارسنجی متقاطع طبقه‌ای ۵ دسته‌ای، حالت اصلاح‌شده اعتبارسنجی متقاطع ۵ دسته‌ای است که برای مواردی که داده‌های با برچسب صفر و یک نامتوازن هستند، روش ارزیابی مناسب‌تری است و از بروز بیش‌برازش جلوگیری می‌کند.

در اعتبارسنجی متقاطع 5 دسته‌ای داده‌ها به ۵ دسته مساوی تقسیم می‌شوند. روش پیشنهادی با ۴ دسته از داده‌ها آموزش می‌بیند و با یک دسته دیگر مورد آزمون قرار می‌گیرد. معیار ارزیابی برای دسته آزمون محاسبه می شود. این فرآیند تکرار می‌شود تا جایی که تمام بخش‌ها یک بار به عنوان داده آموزشی لحاظ شوند. در نهایت میانگین، معیارهای محاسبه شده در تمام دسته‌های آزمون، به عنوان عملکرد کلی مدل در نظر گرفته می‌شود. تفاوت اعتبارسنجی متقاطع طبقه‌ای با اعتبارسنجی متقاطع در آن است که در مدل طبقه‌ای، وقتی داده‌ها مثلا به ۵ دسته تقسیم‌ می‌شوند، تقسیم داده‌ها به نحوی صورت می‌گیرد که نسبت داده‌های با برچسب صفر به داده‌های با برچسب یک در تمام دسته‌ها تقریبا برابر باشد.

با‌توجه به این‌که مسئله مورد بررسی در این مطالعه، دسته‌بندی است، از معیارهای رایج دسته‌بندی برای ارزیابی آن استفاده می‌شود. برای محاسبه معیارهای دسته‌بندی از یک جدول درهم‌ریختگی مانند جدول ‏4‑1 استفاده می‌شود. چهار کمیت توسط این جدول معرفی می‌شود:

* موارد مثبت صحیح، تعداد زوج دارو-بیماری‌هایی است که مدل آن‌ها را به عنوان زوج دارای تعامل پیش‌بینی کرده و آن دارو و بیماری واقعا با هم ارتباط داشته‌اند.
* موارد مثبت کاذب، نشان‌دهنده تعداد زوج دارو-بیماری‌هایی است که به عنوان زوج دارای تعامل پیش‌بینی شده‌اند اما واقعا ارتباطی بین این زوج گزارش نشده است.
* موارد منفی صحیح تعداد زوج دارو-بیماری‌هایی است که واقعا ارتباطی بین این زوج گزارش نشده است و مدل هم تعاملی بین آن‌ها پیش‌گویی نکرده است.
* موارد منفی کاذب، تعداد زوج دارو-بیماری‌هایی است که واقعا با هم ارتباط داشته‌اند اما مدل تعاملی بین آن‌ها پیش‌گویی نکرده است.

ذکر این نکته نیز حائز اهمیت است که وقتی تعامل بین یک جفت دارو-بیماری صفر گذاشته می‌شود، به معنای آن است که تاکنون شاهدی برای ارتباط بین آن‌ها وجود نداشته است؛ درحالی‌که ممکن است آن‌ها واقعا به هم مرتبط باشند. بنابراین در این نوع مسئله، نمی‌توان تعداد منفی صحیح و مثبت کاذب را به دقت شمارش کرد. روند آموزش همواره نیاز به هر دو نوع داده صفر و یک دارد، بنابراین برخی از زوج‌هایی که در روند آموزشی مدل صفر در نظر گرفته می‌شوند، ممکن است واقعا صفر نباشند.

**جدول ‏4‑1 جدول درهم‌ریختگی**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **ارتباطات واقعی** | |
|  |  | **تعامل** | **عدم تعامل** |
| **ارتباطات پیش‌گویی‌شده توسط مدل** | **تعامل** | مثبت صحیح | مثبت کاذب |
| **عدم تعامل** | منفی کاذب | منفی صحیح |

با توجه به آ‌ن‌که مقادیر معیارهایی مانند دقت، صحت و حساسیت به مقدار آستانه وابسته است، در ارزیابی مدل‌ها از دو معیار جامع‌تر نیز استفاده می‌شود. یکی AUC که سطح زیر نمودار مشخصه عملکرد است. نمودار مشخصه عملکرد، نموداری است که محور افقی آن نرخ مثبت صحیح و محور عمودی آن نرخ منفی صحیح است که به صورت زیر تعریف می‌شوند.

|  |  |
| --- | --- |
| *(10)* |  |
| *(11)* |  |

معیار دیگر، AUPR است که سطح زیر نمودار دقت-حساسیت است که محور افقی آن دقت و محور عمودی آن حساسیت است. در واقع این دو معیار، عملکرد مدل را مستقل از مقدار حد آستانه ارزیابی می‌کنند. در مواری که نسبت نمونه‌های صفر و یک برابر نباشد، معیار AUPR سنجش عادلانه‌تری برای ارزیابی مدل است.

## نتایج روش پیشنهادی

به منظور بررسی اهمیت هر ماتریس شباهت، روش پیشنهادی بر روی تک تک آن‌ها به صورت واحد اجرا شده است. هم‌چنین روش پیشنهادی بر روی ترکیب‌ ماتریس‌های شباهت نیز اجرا شده است. جدول ‏4‑2 نتایج را برای هر ویژگی نشان می‌دهد. همان‌طور که در جدول ‏4‑2 نشان داده شده است، ماتریس شباهت‌های ارتباطات دارو و بیماری و هستان‌شناسی ژن از دیگر ویژگی‌ها بهتر عمل کرده‌اند. علاوه بر این، ترکیب ویژگی‌های مختلف منجر به بهترین نتیجه شده است و نیز کارایی نحوه خلاصه‌سازی و کاهش بعد بوسیله DCA و یکپارچه‌سازی و تجمیع شباهت‌ها بوسیله‌ RWR را اثبات می‌کند. حالت ترکیب ویژگی‌ها + شباهت بر اساس ماتریس ارتباطات دارو و بیماری منجر به بهترین مقدار AUC و AUPR شده است.

جدول ‏4‑2 عملکرد روش پیشنهادی بر روی ماتریس‌های شباهت

|  |  |  |
| --- | --- | --- |
| **نام روش** | **AUC** | **AUPR** |
| الف) ماتریس شباهت بر اساس ساختار شیمیایی d1 | ۸۳.۶۰ | ۸۷.۲۸ |
| ب) ماتریس شباهت بر اساس پروتئین‌های هدف d2 | ۸۴.۴۲ | ۸۸.۴۵ |
| ج) ماتریس شباهت بر اساس هستان‌شناسی ژن d3 | ۸۵.۳۰ | ۸۹.۱۳ |
| د) ماتریس شباهت بر اساس عوارض جانبی d4 | ۸۰.۵۸ | ۸۵.۴۲ |
| d1+d2+d3+d4 | ۸۸.۴۵ | ۹۱.۳۶ |
| شباهت بر اساس ماتریس ارتباطات دارو و بیماری d5 | ۹۲.۰۷ | ۹۳.۴۷ |
| ترکیب ویژگی‌ها + شباهت بر اساس ماتریس ارتباطات دارو و بیماری Integration | ۹۳.۲۳ | ۹۴.۸۳ |

## مقایسه روش پیشنهادی با مطالعات پیشین

جدول ‏4‑3 عملکرد روش پیشنهادی را در مقایسه با دیگر روش‌های موجود نشان می‌دهد. نتایج نشان می‌دهد که روش پیشنهادی عملکرد بسیار بهتری نسبت به سایر روش‌های پیشین دارد و به خوبی توانسته است ارتباطات دارو-بیماری‌های موجود را پیش‌گویی کند. مقدار بالای هم AUC و هم AUPR در روش پیشنهادی در مقایسه با مقادیر بسیار پایین AUPR و مقادیر نسبتا پایین‌تر AUC نشان می‌دهد که روش پیشنهادی عملکرد بسیار موفق‌تری در یافتن تعاملات دارو-بیماری داشته باشد.

جدول ‏4‑3 مقایسه عملکرد روش‌ پیشنهادی با عملکرد روش‌های پیشین

|  |  |  |
| --- | --- | --- |
| **نام روش** | **AUC** | **AUPR** |
| **روش پیشنهادی** | **۹۳.۲۳** | **۹۴.۸۳** |
| TL\_HGBI 2014 [34] | ۷۲.۷ | ۳.۰ |
| MBiRW 2016 [35] | ۸۵.۱ | ۴.۴ |
| SCMFDD 2018 [36] | ۶۳.۸ | ۶ |
| DisDrugPred 2019 [37] | ۹۲.۰ | ۲۴.۳ |
| BLM 2009 [38] | 74.75 | 76.89 |
| Graph Embedding Matrix Factorization-GraRep 2015 [39] | 75.7 | 69.3 |
| Graph Embedding Random walk based-struc2vec 2017 [40] | 80.4 | 80.6 |
| Graph Embedding Neural network based-SDNE 2016 [41] | 77.4 | 75.2 |

روش پیشنهادی به ترتیب معیار‌های AUC و AUPR را 1.13 و 14.23 درصد نسبت به روش‌های پیشین بهبود داده است. از آن‌جایی که معیار AUPR معیار مناسب‌تری برای ارزیابی مسأله ماست روش پیشنهادی بسیار بهتر از روش‌های گذشته عمل کرده است. اهمیت معیار AUPR به نسبت سایر معیار‌ها در این است که این معیار، سنجه‌ی عادلانه‌تری برای ارزیابی این مسأله بوده و زیاد بودن تعداد صفر‌ها در ماتریس ارتباطات که صفر در آن صرفاً نشان‌دهنده نبود ارتباط شناخته شده می‌باشد، در افزایش درصد‌ها تأثیری ندارد. معیار AUC نسبت به تعداد زیاد صفر‌ها در ماتریس ارتباطات حساس است و زیاد بودن این صفر‌ها می‌تواند منجر به افزایش ناعادلانه این معیار گردد. بهبود معیار AUPR توسط روش پیشنهادی می‌تواند نشان‌دهنده عملکرد خوب مدل در برگرداندن برچسب‌های یک باشد که باعث اعتماد به مدل در پیش‌گویی ارتباطات شناخته نشده می‌گردد. این بهبود نسبت به سایر روش‌ها می‌تواند به دلیل استفاده کردن از انواع مختلف ویژگی‌ها و ترکیب‌ آن‌ها باشد. همانطور که در جدول ‏4‑3 نشان داده شد، افزایش ویژگی‌های مختلف منجر به بهبود نتایج شده است. همچنین روش پیشنهادی در این مطالعه به دلیل استفاده‌کردن از گشت تصادفی و تجزیه ماتریسی با رویکردی کارا ویژگی‌های مختلف را به شکل مناسبی با یکدیگر ترکیب می‌کند که این شیوه ترکیبی به دلیل استفاده از فضای پنهان می‌تواند ویژگی‌های توپولوژیکی را نیز در پیش‌گویی نهایی دخیل کند. تمام این موارد در کنار یکدیگر باعث بهبود 14.23 درصدی در معیار AUPR گردیده است.

## مطالعات موردی

به دلیل این‌که در این مسأله صفر‌های موجود در ماتریس ارتباطات دارو و بیماری صرفاً به معنای ناشناخته‌بودن ارتباط(و نه عدم ارتباط قطعی) است و ممکن است این صفر‌ها در آینده تبدیل به یک شوند، ما در این بخش به مطالعه موردی برخی از مثبت‌های کاذب پیش‌گویی شده توسط روش پیشنهادی می‌پردازیم. جدول ‏4‑4 لیستی از بررسی شواهد مطالعات موردی در مقالات پیشین را نشان می‌دهد.

جدول ‏4‑4 شواهد مطالعات موردی در مقالات پیشین

|  |  |  |
| --- | --- | --- |
| **نام دارو** | **نام بیماری** | **منبع** |
| سفپودوکسیم | عفونت استافیلوکوک | [42] |
| لووستاتین | هیپرلیپیدمی نوع دو | [43] |
| تری‌فلوپرازین | اسکیزوفرنی و پارانویا | [44] |
| سفدی‌نیر | عفونت پوستی | [45] |
| نابیلون | تهوع و استفراغ بعد از بیهوشی | [46] |
| پیپراسیلین | عفونت کلبسیلا | [47] |
| پروکائین آمید | فیبریلاسیون بطنی | [48] |
| فلکاینید | فیبریلاسیون بطنی | [49] |
| ارتاپنم | عفونت استافیلوکوک | [50] |

این موارد از جمله ارتباطاتی هستند که در ماتریس ارتباطات اصلی صفر بودند اما مدل به خوبی آن‌ها را یک پیش‌گویی کرده است. به عنوان مثال ارتباط داروی تری‌فلوپرازین با بیماری اسکیزوفرنی با وجود صفر‌بودن در ماتریس ارتباطات اولیه، به خوبی یک تشخیص داده شده است. تری‌فلوپرازین از گروه فنوتیازین‌هاست که عملکرد اصلی این دارو در مهار گیرنده‌های دوپامینی در سلول‌های مغزی است. این دارو از طریق بستن گیرنده‌های دوپامینی در سیستم اعصاب مرکزی اثرات ناشی از تحریك دوپامین را کنترل می‌كند. هم‌چنین این دارو موجب انسداد گیرنده‌های آلفا می‌شود و فعالیت ناشی از هیستامین و سروتونین را خنثی می‌سازد[44]. مورد دیگر داروی نابیلون برای درمان تهوع است. در مطالعات اخیر، نابیلون به عنوان یک مکمل پیشگیری از تهوع و استفراغ پس از عمل، شناخته شده است. این دارو پیش‌تر اثربخشی خود را به طور بالینی در درمان تهوع و استفراغ مرتبط با شیمی‌درمانی نشان داده است[46].

# بحث و جمع‌بندی

پیش‌بینی تعاملات دارو و بیماری اهمیت زیادی دارد، چرا که برخی از دارو‌های مورد استفاده در بیماری‌های می‌توانند برای بیماری‌های دیگر نیز مفید عمل کنند. این موضوع می‌تواند در کاهش هزینه‌های طراحی و تولید دارو و کاهش زمان نیز تاثیرگذار باشد. به دلیل اهمیت این موضوعات، امروزه بیوانفورماتیک و زیست‌شناسی محاسباتی با بهره‌گیری از علم تحلیل‌داده می‌توانند با بررسی داده‌ها و محاسبات مختلف تعاملات دارو-بیماری را پیش‌بینی کنند. با پذیرش این فرضیه که دارو‌های مشابه می‌توانند عمل‌کرد مشابهی داشته باشند و همچنین بیماری‌های مشابه می‌توانند درمان‌های مشابهی داشته باشند می‌توان از شباهت دارو‌ها و بیماری‌ها برای باز‌هدف‌گذاری دارویی استفاده کرد. در این مطالعه، روشی برای پیش‌بینی تعاملات دارو و بیماری با استفاده از ویژگی‌های مختلف دارویی مثل ساختار شیمیایی، عوارض جانبی، پروتئین‌های هدف و غیره و همچنین یک ماتریس ترکیبی شباهت بیماری‌ها ارائه شد.

روش پیشنهادی در این پژوهش از گشت تصادفی و خلاصه‌سازی ویژگی به همراه فاکتورسازی ماتریس برای پیش‌بینی استفاده می‌کند.

نتایج حاصل توسط ارزیابی تقاطعی 5 دسته‌ای با معیار‌های AUC و AUPR اندازه‌گیری شد. این معیار‌ها به ترتیب برابر با 93.23 و 94.83 درصد هستند. میزان کارایی هر ماتریس شباهت نیز در افزایش دقت پیش‌بینی بررسی شد و از میان شباهت‌ها، شباهت ساختاری و شباهت هستان‌شناسی به ترتیب نتایج بهتری از دیگر ویژگی‌ها کسب کردند. بررسی‌ها نشان داد که استفاده از ترکیب ویژگی‌های مختلف می‌تواند به افزایش دقت بیانجامد و کارایی نحوه خلاصه‌سازی و یکپارچه‌سازی انجام شده را اثبات کرد.

برای تحلیل‌های بیشتر مطالعات موردی بر روی پیش‌بینی‌های به مثبت کاذب انجام شد. ارزیابی‌ها نشان می‌دهند که روش پیشنهادی با در نظر گرفتن عوارض جانبی و دیگر ویژگی‌ها توانسته است به خوبی تعاملات ناشناخته را پیش‌بینی کند.

به عنوان کار‌های آینده می‌توان به پژوهش‌های زیر پرداخت:

* استفاده از ویژگی‌های مختلف دیگر
* ترکیب ویژگی‌های داروی با دیگر روش‌های ترکیب ویژگی
* بررسی آنتروپی ماتریس‌های شباهت قبل از در نظر گرفتن آن‌ها و کنار گذاشتن ماتریس‌هایی که ممکن است دارای نویز بالایی باشند.
* با توجه به اینکه تاکنون شدت عوارض جانبی داروها در جایی ثبت نشده در صورت ایجاد شدن این مجموعه‌داده می‌توان آن را به عنوان عاملی در جهت کاهش خطر بازهدف‌گذاری در نظر گرفت.

مراجع

[1] H. Xue, J. Li, H. Xie, and Y. Wang, “Review of drug repositioning approaches and resources,” *Int. J. Biol. Sci.*, vol. 14, no. 10, p. 1232, 2018.

[2] S. Pushpakom *et al.*, “Drug repurposing: progress, challenges and recommendations,” *Nat. Rev. Drug Discov.*, vol. 18, no. 1, pp. 41–58, 2019.

[3] L. Lü and T. Zhou, “Link prediction in complex networks: A survey,” *Phys. A Stat. Mech. its Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[4] M. Hosseinzadeh Aghdam, M. Analoui, and P. Kabiri, “A novel non-negative matrix factorization method for recommender systems,” *Appl. Math. Inf. Sci.*, vol. 9, pp. 2721–2732, Jan. 2015, doi: 10.12785/amis/090558.

[5] G. Jin and S. T. C. Wong, “Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines,” *Drug Discov. Today*, vol. 19, no. 5, pp. 637–644, 2014.

[6] Y. Wang, S. Chen, N. Deng, and Y. Wang, “Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data,” *PLoS One*, vol. 8, no. 11, p. e78518, 2013.

[7] H. Arrouchi, W. Lakhlili, and A. Ibrahimi, “Re-positioning of known drugs for Pim-1 kinase target using molecular docking analysis,” *Bioinformation*, vol. 15, no. 2, p. 116, 2019.

[8] Y. Li and P. Agarwal, “A pathway-based view of human diseases and disease relationships,” *PLoS One*, vol. 4, no. 2, p. e4346, 2009.

[9] J. Lamb *et al.*, “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease,” *Science (80-. ).*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[10] L. Peng *et al.*, “Screening drug-target interactions with positive-unlabeled learning,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–17, 2017.

[11] J. Lu *et al.*, “Identification of new candidate drugs for lung cancer using chemical–chemical interactions, chemical–protein interactions and a K-means clustering algorithm,” *J. Biomol. Struct. Dyn.*, vol. 34, no. 4, pp. 906–917, 2016.

[12] G. Ceddia, P. Pinoli, S. Ceri, and M. Masseroli, “Matrix Factorization-based Technique for Drug Repurposing Predictions,” *IEEE J. Biomed. Heal. Informatics*, 2020.

[13] X. Yue *et al.*, “Graph embedding on biomedical networks: methods, applications and evaluations,” *Bioinformatics*, vol. 36, no. 4, pp. 1241–1251, 2020.

[14] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, “deepDR: a network-based deep learning approach to in silico drug repositioning,” *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, 2019.

[15] N. Zhang, W. Xu, S. Wang, Y. Qiao, and X. Zhang, “Computational Drug Discovery in Chemotherapy-induced Alopecia via Text Mining and Biomedical Databases,” *Clin. Ther.*, vol. 41, no. 5, pp. 972–980, 2019.

[16] X. Liang *et al.*, “LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning,” *Bioinformatics*, vol. 33, no. 8, pp. 1187–1196, Jan. 2017, doi: 10.1093/bioinformatics/btw770.

[17] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, “PubChem: a public information system for analyzing bioactivities of small molecules,” *Nucleic Acids Res.*, vol. 37, no. suppl\_2, pp. W623–W633, 2009.

[18] A. Mitchell *et al.*, “The InterPro protein families database: the classification resource after 15 years,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D213–D221, 2015.

[19] U. Consortium, “The universal protein resource (UniProt) in 2010,” *Nucleic Acids Res.*, vol. 38, no. suppl\_1, pp. D142–D148, 2010.

[20] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, “DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function,” *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 2018.

[21] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, “The SIDER database of drugs and side effects,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, 2016.

[22] H. Cho, B. Berger, and J. Peng, “Diffusion component analysis: unraveling functional topology in biological networks,” in *International Conference on Research in Computational Molecular Biology*, 2015, pp. 62–64.

[23] H. Cho, B. Berger, and J. Peng, “Compact integration of multi-network topology for functional analysis of genes,” *Cell Syst.*, vol. 3, no. 6, pp. 540–548, 2016.

[24] M. Cao *et al.*, “New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence,” *Bioinformatics*, vol. 30, no. 12, pp. i219–i227, 2014.

[25] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, “Walking the interactome for prioritization of candidate disease genes,” *Am. J. Hum. Genet.*, vol. 82, no. 4, pp. 949–958, 2008.

[26] S. Navlakha and C. Kingsford, “The power of protein interaction networks for associating genes with diseases,” *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, 2010.

[27] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, “IsoRankN: spectral methods for global alignment of multiple protein networks,” *Bioinformatics*, vol. 25, no. 12, pp. i253–i258, 2009.

[28] X. Chen, M. X. Liu, and G. Y. Yan, “Drug-target interaction prediction by random walk on the heterogeneous network,” *Mol. Biosyst.*, vol. 8, no. 7, pp. 1970–1978, 2012, doi: 10.1039/c2mb00002d.

[29] M. Kim and J. Leskovec, “The network completion problem: Inferring missing nodes and edges in networks,” in *Proceedings of the 2011 SIAM International Conference on Data Mining*, 2011, pp. 47–58.

[30] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, “Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization,” *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997.

[31] Y. Luo *et al.*, “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information,” *Nat. Commun.*, vol. 8, no. 1, pp. 1–13, 2017.

[32] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon, “Large-scale multi-label learning with missing labels,” in *International conference on machine learning*, 2014, pp. 593–601.

[33] N. Natarajan and I. S. Dhillon, “Inductive matrix completion for predicting gene–disease associations,” *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 2014.

[34] W. Wang, S. Yang, X. Zhang, and J. Li, “Drug repositioning by integrating target information through a heterogeneous network model,” *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.

[35] H. Luo *et al.*, “Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm,” *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, 2016.

[36] W. Zhang *et al.*, “Predicting drug-disease associations by using similarity constrained matrix factorization,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–12, 2018.

[37] P. Xuan, Y. Cao, T. Zhang, X. Wang, S. Pan, and T. Shen, “Drug repositioning through integration of prior knowledge and projections of drugs and diseases,” *Bioinformatics*, vol. 35, no. 20, pp. 4108–4119, 2019.

[38] K. Bleakley and Y. Yamanishi, “Supervised prediction of drug–target interactions using bipartite local models,” *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[39] S. Cao, W. Lu, and Q. Xu, “Grarep: Learning graph representations with global structural information,” in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 891–900.

[40] L. F. R. Ribeiro, P. H. P. Saverese, and D. R. Figueiredo, “struc2vec: Learning node representations from structural identity,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 385–394.

[41] D. Wang, P. Cui, and W. Zhu, “Structural deep network embedding,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1225–1234.

[42] Y. C. Liu, W. K. Huang, and D. L. Cheng, “Antibacterial activity of cef podoxime in vitro,” *Chemotherapy*, vol. 43, no. 1, pp. 21–26, Jan. 1997, doi: 10.1159/000239530.

[43] R. Carmena, G. Roederer, H. Mailloux, S. Lussier-Cacan, and J. Davignon, “The response to lovastatin treatment in patients with heterozygous familial hypercholesterolemia is modulated by apolipoprotein E polymorphism,” *Metabolism*, vol. 42, no. 7, pp. 895–901, 1993, doi: 10.1016/0026-0495(93)90066-W.

[44] R. Macdonald and T. P. Shields Watts, “Trifluoperazine dihydrochloride (‘stelazine’) in paranoid schizophrenia,” *Br. Med. J.*, vol. 1, no. 5121, pp. 549–550, Feb. 1959, doi: 10.1136/bmj.1.5121.549.

[45] H. S. Sader, J. M. Streit, T. R. Fritsche, and R. N. Jones, “Potency and spectrum reevaluation of cefdinir tested against pathogens causing skin and soft tissue infections: A sample of North American isolates,” *Diagn. Microbiol. Infect. Dis.*, vol. 49, no. 4, pp. 283–287, Aug. 2004, doi: 10.1016/j.diagmicrobio.2004.04.015.

[46] D. N. Levin, Z. Dulberg, A.-W. Chan, G. M. T. Hare, C. D. Mazer, and A. Hong, “A randomized-controlled trial of nabilone for the prevention of acute postoperative nausea and vomiting in elective surgery,” *Can. J. Anesth. Can. d’anesthésie*, vol. 64, no. 4, pp. 385–395, 2017.

[47] T. Pillay, D. G. Pillay, M. Adhikari, and A. W. Sturm, “Piperacillin/tazobactam in the treatment of Klebsiella pneumoniae infections in neonates,” *Am. J. Perinatol.*, vol. 15, no. 1, pp. 47–51, Jan. 1998, doi: 10.1055/s-2007-993898.

[48] D. T. Markel *et al.*, “Procainamide and survival in ventricular fibrillation out-of-hospital cardiac arrest,” *Acad. Emerg. Med.*, vol. 17, no. 6, pp. 617–623, 2010, doi: 10.1111/j.1553-2712.2010.00763.x.

[49] R. H. Falk, “Flecainide-induced ventricular tachycardia and fibrillation in patients treated for atrial fibrillation,” *Ann. Intern. Med.*, vol. 111, no. 2, pp. 107–111, 1989, doi: 10.7326/0003-4819-111-2-107.

[50] R. M. Gesser, K. A. McCarroll, and G. L. Woods, “Efficacy of ertapenem against methicillin-susceptible Staphylococcus aureus in complicated skin/skin structure infections: Results of a double-blind clinical trial versus piperacillin-tazobactam,” *Int. J. Antimicrob. Agents*, vol. 23, no. 3, pp. 235–239, 2004, doi: 10.1016/j.ijantimicag.2003.07.013.