دانشکدگان علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

پیش‌بینی ارتباطات miRNA و بیماری با استفاده از شبکه‌های عصبی گرافی

نگارنده

مهرو حاجی‌مهدی

استاد راهنما: دکتر باقر باباعلی

پایان‌نامه برای دریافت درجه کارشناسی
در رشتهٔ علوم کامپیوتر

مرداد ۱۴۰۳

چکیده

شواهد زیادی نشان داده‌اند که miRNAها نقش مهمی در بسیاری از فرآیندهای زیستی ایفا می‌کنند و با انواعی از بیماری‌های پیچیده انسانی ارتباط دارند. پیش‌بینی ارتباط بین miRNAها و بیماری‌ها[1](MDA) به بخش مهمی از فرآیند شناسایی و درمان بیماری‌ها تبدیل شده است. با این حال، شناسایی این ارتباطات از طریق روش‌های تجربی، پیچیده و زمان‌بر است. همچنین، بیان برخی از miRNAها محدود به انواع خاصی از سلول‌ها یا بافت‌هاست و یا نیازمند شرایط محیطی خاصی هستند که این امر باعث می‌شود شناسایی آن‌ها دشوار شود. بنابراین، روش‌های محاسباتی توسعه یافته‌اند تا روش‌های تجربی را تکمیل کنند.

در این پروژه با الهام از پیشرفت‌های چشمگیر در شبکه‌های عصبی گرافی برای پیش‌بینی MDA، نشان می‌دهیم که ساختار ساده‌ای از شبکه‌های عصبی گرافی کانولوشنی قادر است با دقت بالا تعامل بین miRNA و بیماری‌ها را پیش‌بینی کند. به‌طور خاص، ابتدا معیارهای شباهت miRNA و بیماری از چندین منبع اطلاعاتی به دست آمد و بردارهای ویژگی براساس این اطلاعات ساخته شد. سپس، گرافی با miRNA و بیماری به‌عنوان دو نوع گره برای این گراف ایجاد شد که هر کدام دارای بردارهای ویژگی خاص خود هستند. در نهایت، شبکه عصبی گرافی کانولوشنی با یک پرسپترون چندلایه ترکیب شد تا پیش‌بینی نهایی انجام شود. مقادیر AUC و AUPR در آموزش مدل با اعتبارسنجی متقابل به ترتیب برابر با ۰/۱۵ ± ۹۴/۹۷ و ۰/۱۸ ± ۹۴/۹۶ به دست آمدند. علاوه بر این، مطالعات موردی روی سرطان‌های ریه، پستان و خون، نشان می‌دهد که روش ارائه‌شده در این پژوهش می‌تواند روش مفیدی برای استنتاج روابط بین بیماری‌ها و miRNAها باشد.

کلمات کلیدی: پیش‌بینی ارتباط miRNA و بیماری، شبکهٔ عصبی گرافی، معیارهای شباهت

[1] miRNA-Disease Association

پیشگفتار

میکرو ریبونوکلئیک‌اسیدها (microRNA یا به اختصار miRNA) یک دسته از مولکول‌های RNA تک‌رشته‌ای غیرکدگذار با طول تقریبی ۲۲ نوکلئوتید هستند [۱]. پیش‌بینی می‌شود که miRNAها ۱ تا ۵ درصد از ژنوم انسان را تشکیل دهند و حداقل ۳۰ درصد از ژن‌های کدکننده پروتئین‌ها را تنظیم کنند [۲]. در حال حاضر، تعداد ۱۹۱۷ ژن پیش‌ساز miRNA انسانی شناسایی شده است که به ۲۶۵۶ توالی بالغ پردازش می‌شوند، در حالی که وظایف بسیاری از miRNAها هنوز ناشناخته است. بازهای miRNAها با توالی mRNA هدف[۲] خود به صورت کامل یا ناکامل جفت می‌شوند که منجر به کاهش تولید پروتئین از آن mRNAها می‌شود. الگوریتم‌های بیوانفورماتیکی گوناگون پیش‌بینی می‌کنند که miRNAها از ۳۰٪ [۳] تا حتی ۹۲٪ [۴] ژن‌های انسانی را تنظیم می‌کنند.

تحقیقات زیادی گزارش کرده‌اند که miRNAها نقشی حیاتی در فرآیندهای بیولوژیکی متعدد ایفا می‌کنند؛ مانند تاثیر بر چرخهٔ سلولی، تکثیر و تمایز سلول، آپوپتوز[۳]، متابولیسم و سیگنال‌دهی سلولی [۵]. همچنین، بسیاری از مطالعات نشان داده‌اند که بیان غیرعادی و عدم تنظیم miRNAها با توسعه و پیشرفت بسیاری از بیماری‌های انسان مرتبط است؛ از جمله انواع مختلف سرطان [۶]، اختلالات روانی [۷]، آلزایمر [۸]، پارکینسون [۹] و بیماری‌های تخریب عصبی [۱۰]. بنابراین، شناسایی ارتباط بیماری‌ـmiRNA می‌تواند به فهم مکانیسم بیماری در سطح miRNA و شناسایی نشانگرهای بیماری برای تشخیص، درمان، پیش‌بینی و پیشگیری کمک کند.

در حال حاضر روش‌های بیولوژیکی زیادی برای شناسایی miRNAهای مرتبط با بیماری موجود است [۱۱]. پروفایل‌سازی بیان miRNA بینش مهمی در زیست‌شناسی miRNA فراهم

[۲] target mRNA
[۳] apoptosis

کرده است. مطالعات ریزآرایه[۴] بیان miRNAها را در حالت‌های طبیعی و غیرطبیعی از فرایندهای زیستی، از جمله در بیماری‌ها، اندازه‌گیری کرده‌اند. گرچه تحلیل بیان miRNAها می‌تواند داده‌های نوینی در این زمینه پژوهشی فراهم کند، اما پروفایل‌سازی miRNA از بیش از یک تکنیک به طور همزمان بهره می‌برد. پژوهش‌های قبلی تأیید کرده‌اند که درک الگوهای بیان در سطح سلولی برای مطالعات miRNA در بیماری‌ها بسیار مهم است [۱۲]. هیبریداسیون در محل [۵] دقیقاً مشخص می‌کند که کدام miRNAهای خاص در کجا بیان می‌شوند. بیماری‌های مختلف جمعیت‌های سلولی مختلفی را نیز تحت تأثیر قرار می‌دهند. بنابراین، دانستن اینکه کدام mRNAها در کدام جمعیت‌های سلولی بیان می‌شوند، می‌تواند به شناسایی ژن‌های miRNA مرتبط با بیماری و اهداف mRNA آنها کمک کند.

اگرچه چنین روش‌های بیولوژیکی می‌توانند با دقت زیاد رابطهٔ بین miRNAها و بیماری‌ها را تعیین کنند، اما با محدودیت هایی نیز مواجه هستند. به نظر می‌رسد، miRNAها و پروتئین‌های مرتبط با آن‌ها، از جمله کمپلکس‌های ریبونوکلئوپروتئینی[۶] فراوان در سلول باشند. با این حال، miRNAهایی که بیان آن‌ها محدود به انواع سلولی نادر و یا شرایط محیطی خاص باشد، ممکن است در کلون کردن نادیده گرفته شوند [۱]. از طرفی روش‌های بیولوژیک بسیار زمان‌بر و پرزحمت هستند. بنابراین، رویکردهای محاسباتی برای تکمیل رویکردهای تجربی در شناسایی ژن‌های miRNA توسعه یافته‌اند.

در سال‌های اخیر با پیشرفت در جمع‌آوری و ذخیرهٔ داده‌های بیولوژیکی، توسعه روش‌های محاسباتی برای استنتاج ارتباطات بالقوه بین miRNAها و بیماری‌ها توجه بسیاری از دانشمندان حوزه علوم کامپیوتر را به خود جلب کرده است. بیشتر روش‌های محاسباتی فعلی برای پیش‌بینی miRNAهای مرتبط با بیماری بر این فرض استوارند که miRNAهایی با عملکرد مشابه، با بیماری‌های همسان مرتبط هستند و برعکس. این روش‌ها شامل ساختن نماهای miRNA و بیماری بر اساس ویژگی‌های مختلف و سپس استنتاج وابستگی‌های جدید بیماری–miRNA با استفاده از روابط شناخته شده می‌باشد. در حال حاضر، رویکردهای محاسباتی در این حوزه را می‌توان به سه دسته کلی تقسیم کرد [۱۱]: روش‌های مبتنی بر شباهت، روش‌های مبتنی بر یادگیری ماشین و روش‌های مبتنی بر گراف. جزئیات این روش‌ها و مثال‌هایی از پژوهش‌های انجام شده در هریک ازین حوزه‌ها در بخش بعدی توضیح داده می‌شود.

---

[۴] Microarray

[۵] In situ hybridization (ISH)

[۶] ribonucleoprotein

در این پژوهش قصد داریم کارایی شبکه‌های عصبی گرافی را با استفاده از چندین دیدگاه شباهت[۷] در پیش‌بینی MDA مورد سنجش قرار دهیم. بدین منظور، ابتدا اطلاعات شباهت miRNAها را بر اساس اطلاعات شباهت توالی، شباهت خانوادگی، شباهت عملکرد و کرنل شباهت پروفایل تعامل گاوسی، و اطلاعات شباهت بیماری‌ها را بر اساس شباهت عملکردی و شباهت پروفایل تعامل گاوسی محاسبه کردیم. سپس همهٔ این اطلاعات ادغام شد و بردارهای ویژگی برای miRNA و بیماری بدست آمد. در نهایت این بردارها به عنوان ویژگی هر گره در GNN ساخته شده، مورد استفاده قرار گرفت و پیش‌بینی در لایه آخر با استفاده از یک شبکه پرسپترون چندلایه انجام گرفت. به منظور ارزیابی مدل، روش اعتبارسنجی پنج‌گانه روی نسخه ۳.۲ پایگاه داده HMDD بکار گرفته شد. طبق نتایج بدست آمده AUC و AUPR به ترتیب برابر با ۹۴/۹۷ و ۹۴/۹۶ درصد گزارش شد. در ادامه به منظور بررسی اهمیت استفاده از چندین نوع اطلاعات شباهت، مدل بر روی هفت ترکیب از این اطلاعات سنجیده شد. همچنین برای اثبات کارایی مدل، سه بیماری سرطان پستان، سرطان ریه و سرطان خون انتخاب و نتیجه پیش‌بینی مدل با استفاده از پایگاه‌های مختلف ارزیابی شد. پیش‌بینی مدل روی این سه سرطان در ۵۰، ۴۸ و ۴۹ مورد از ۵۰ مولکول توسط بانک داده‌ها تایید شدند. در ادامه ابتدا به مقدمه‌ای بر مولکول miRNA و اهمیت آن در بیماری‌های انسان و همچنین معرفی شبکه‌های عصبی گرافی می‌پردازیم. سپس جزئیات ساخت بردارهای ویژگی و معماری مدل GNN را توضیح می‌دهیم و در نهایت عملکرد مدل را در حالات مختلف بررسی می‌کنیم.

[۷] views similarity

فهرست مطالب

فصل ۱

مطالعات پیشین

روش‌های مبتنی بر شباهت معمولاً بر یک فرض رایج تکیه دارند؛ miRNAهایی که عملکرد مشابهی دارند با بیماری‌هایی که از لحاظ فنوتیپی مشابه هستند، ارتباط دارند و بالعکس. miRNAها و بیماری‌ها به‌صورت بردارهای ویژگی بر مبنای انواع مختلف داده‌های بیولوژیکی، مانند توالی miRNA، حاشیه‌نویسی[۱] miRNA و حاشیه‌نویسی ژن نمایش داده می‌شوند. این روش‌ها معمولاً توجه زیادی به محاسبهٔ شباهت بین miRNAها و بیماری‌ها دارند. تاکنون، روش‌های مختلفی برای محاسبهٔ شباهت ارائه شده است؛ از جمله شباهت معنایی بیماری[۲] [۱۳]، شباهت عملکردی miRNA و کرنل شباهت پروفایل تعامل گاوسی[۳] [۱۴]. برخی پژوهش‌ها از ویژگی‌های دیگری نیز برای ساخت بردار ویژگی miRNA بهره برده‌اند؛ مثل شباهت خوشه‌ای و شباهت خانوادگی [۴۹].

برای مثال، مدل محاسباتی BNPMDA بر مبنای ارتباطات شناخته‌شده بیماری–miRNA، شباهت miRNA و شباهت تجمعی بیماری عمل می‌کند [۱۵]. همچنین، گروه دیگری از پژوهشگران از حاشیه نویسی ژن‌ها برای شناسایی ژن‌های هدف miRNA استفاده کردند تا به‌طور مستقیم شباهت عملکردی miRNAها را اندازه‌گیری کنند [۱۶]. اگرچه مدل‌های مبتنی بر شباهت، عملکرد خوبی دارند، اما عملکردی خوبی در شناسایی ارتباطات پیچیده و غیرخطی بین miRNAها و بیماری‌ها در شبکه‌های ناهمگن ندارند.

---

[۱] Annotation

[۲] Semantic Similarity

[۳] Gaussian Interaction Profile Kernel Similarity

پیش‌بینی عملکرد با مدل‌های یادگیری ماشین معمولاً به سه عامل بستگی دارد: پایگاه‌های داده، نمایش ویژگی‌ها و مدل‌های پیش‌بینی. در این روش، معمولاً از رویکردهای یادگیری ماشین مانند یادگیری منظم‌سازی شده، گام تصادفی[۴]، ماشین‌های بردار پشتیبان[۵] [۱۷] یا درخت تصمیم‌گیری [۱۸] برای کشف ارتباطات بیماری‌ـmiRNA استفاده می‌شود. برای مثال، مدل CNNMDA با استفاده از دو شبکه عصبی کانولوشنی (CNN) به‌طور همزمان نمایشی اولیه و سراسری از جفت ارتباطات بیماری‌ـmiRNA را آموزش می‌دهد [۱۹]. با این حال، در نهایت این همه این مدل‌ها نمایش‌های گوناگون از miRNAها و بیماری‌ها را با هم ترکیب می‌کنند، بنابراین نمی‌توانند یک نمایش تفکیک‌پذیر تولید کنند. در این بین، روش‌های مبتنی بر گراف به دلیل عملکرد برجسته‌شان توجه بیشتری را به خود جلب کرده‌اند.

برخلاف CNNها که در فضای اقلیدسی کار می‌کنند، شبکه‌های عصبی گرافی [۶] در فضاهای غیر اقلیدسی هستند. CNNها فقط روی داده‌های منظم اقلیدسی مانند تصاویر (شبکه‌های ۲ بعدی) و متون (توالی‌های تک بعدی) عمل می‌کنند، درحالی که این نوع داده‌ها می‌توانند به‌عنوان نمونه‌هایی از گراف‌ها در نظر گرفته شوند. بنابراین، تعمیم CNNها روی گراف‌ها ممکن است.

GNN روش‌های مبتنی بر شبکه‌های عصبی هستند که اطلاعات را از طریق انتقال پیام بین همسایگان‌شان رد و بدل می‌کنند. در سال‌های اخیر انواع مختلفی از این نوع شبکه، مانند شبکه‌های عصبی کانولوشنی گرافی [۷]، شبکه‌های گرافی با مکانیسم توجه [۸] و GraphSAGE عملکرد چشمگیری را در وظایف مختلف یادگیری عمیق، از جمله پیش‌بینی ارتباطات بیماری‌ـmiRNA نشان داده‌اند [۲۰].

روش‌های مبتنی بر گراف برای پیش‌بینی ارتباط بیماری و miRNA، معمولاً یک گراف ناهمگن بر پایه ارتباطات شناخته‌شدهٔ بیماری‌ـmiRNA و معیارهای شباهت miRNAها و بیماری‌ها ایجاد می‌کنند. چنین گرافی توانایی زیادی در نمایش روابط پیچیده بین miRNAها و بیماری‌ها دارد. مدل MMGCN با استفاده از GCN به‌عنوان رمزگذار، ویژگی‌های miRNAها و بیماری‌ها را تحت دیدگاه‌های شباهتی مختلف به دست آورده و فرآیند یادگیری نمایشی[۹] را از طریق مکانیسم‌های توجه چندگانه تقویت کرده است [۲۱]. مدل پیشنهادی MKGAT برای یادگیری ویژگی‌های

[۴] Random Walk
[۵] Support Vector Machines (SVM)
[۶] Graph Neural Networks (GNN)
[۷] Graph Convolutional Neural Networks (GCN)
[۸] Graph Attention Networks (GAT)
[۹] Representation learning

جاسازی[۱۰] miRNAها و بیماری‌ها از گراف کانولوشنی همراه با مکانیسم توجه استفاده کرده که از شبکه شباهت miRNA و شبکه شباهت بیماری در ساخت آن استفاده شده است [۲۲]. مدل AMHMDA نیز از یک گراف ناهمگن و دیدگاه‌های مختلف شباهت همراه با استراتژی توجه برای یادگیری نمایش miRNAها و بیماری‌ها استفاده کرده است [۲۳]. این مثال‌ها اهمیت و کارایی شبکه‌های عصبی گرافی را در پیش‌بینی ارتباطات انواع بیماری‌ها و miRNAها نشان می‌دهند.

[۱۰] embedding features

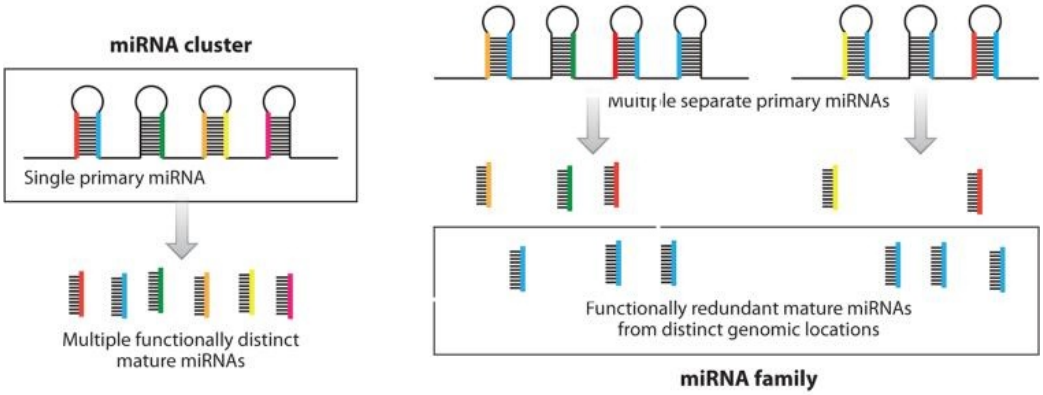
فصل ۲

مفاهیم مقدماتی

۱.۲ miRNA

تا به امروز، بیش از ۲۳۰۰ مولکول miRNA در انسان شناسایی شده است [۲۴]. بسیاری از miRNAهای پستانداران الگوی رونویسی مشابهی با ژن‌های کدگذار پروتئین که در آن قرار دارند، از خود نشان می‌دهند [۲۶]. نسخه‌های اولیه miRNA به نام miRNA اولیه شناخته می‌شوند که ساختار متفاوتی از miRNA بالغ و عملکردی دارد. در سلول، مولکول اولیه miRNA تحت پردازش آنزیم‌های برش دهنده و انواع دیگری از پروتئین‌ها قرار می‌گیرد و ابتدا به miRNA دو رشته‌ای و سپس تک رشته‌ای بالغ با طول تقریبی ۱۸ تا ۲۴ باز تبدیل می‌شود. این مولکول های بالغ می‌توانند با ته صورت کامل یا ناقص با mRNA هدف خود جفت شده و سبب تخریب یا جلوگیری از ترجمه آن به پروتئین شود. به دلیل تطابق ناقص بین miRNA و هدف آن، یک ژن می‌تواند توسط چندین miRNA تنظیم شود و به همین ترتیب، یک miRNA ممکن است بیش از یک ژن هدف داشته باشد [۲۷].

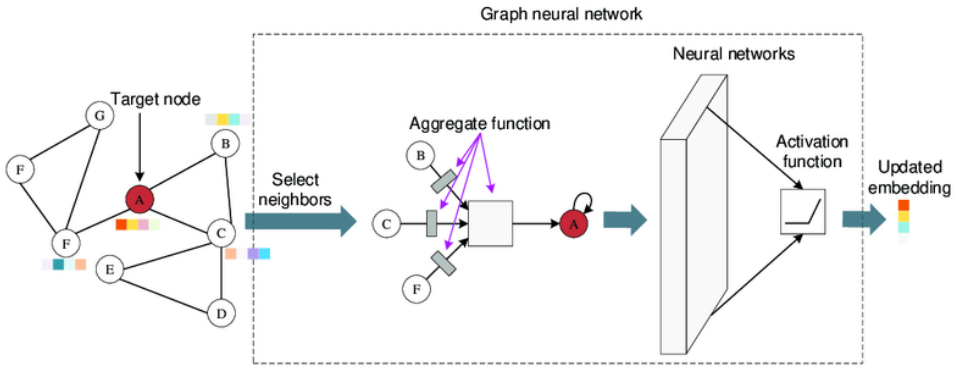
miRNAها نقش مهمی در تعیین سرنوشت سلولی، تکثیر و مرگ سلول ایفا می‌کنند و در بسیاری از فرآیندهای حیاتی سلول دخیل هستند. به عنوان مثال، تاثیر آنها در پاسخ ایمنی، ترشح انسولین، سنتز انتقال‌دهنده‌های عصبی، ریتم شبانه‌روزی و تکثیر ویروس تایید شده است [۲۸]. مطالعات پروفایل بیان ژن نشان داده‌اند که تغییرات در بیان miRNAها در بسیاری از بیماری‌های انسانی وجود دارد. شواهد فراوانی مبنی بر ارتباط سطح بیان ناهنجار این مولکول‌ها با شروع و



شکل ۱.۲: miRNAها به دو صورت خوشه‌ها (شکل سمت چپ) و خانواده‌ها (شکل سمت راست) طبقه‌بندی می‌شوند [۲۹].

پیشرفت بیماری‌های انسانی، اختلالات ژنتیکی و تغییرات در عملکرد سیستم ایمنی وجود دارد [۳۰]. همچنین مشخص شده است که miRNA و در مواردی اعضای خانوادهٔ آنها می‌توانند مسیرهای سرطان‌زا یا سرکوب‌کننده تومور را تنظیم کنند، در حالی که خود miRNAها می‌توانند توسط ژن‌های سرطان‌زا یا سرکوب‌کننده تومور تنظیم شوند [۳۱].

خانوادهٔ miRNA گروهی از این مولکول‌های زیستی هستند که از یک نیای مشترک مشتق می‌شوند [۳۳]. به طور معمول، اعضای یک خانواده miRNA وظایف فیزیولوژیکی مشابهی دارند، اما لزوماً در توالی اولیه یا ساختار ثانویه یکسان نیستند (شکل ۱.۲). این در حالی است که ژن‌های miRNA یک خانواده می‌تواند توالی‌ای از miRNA بالغ عملکردی را به صورت کامل یا جزئی، در خود داشته باشند. بنابراین وجود یک ساختار یا توالی مشترک در آنها می‌تواند باعث عملکرد مشابهی در میان آنها شود. در مقابل miRNAهایی که در ژنوم به صورت مجاور قرار دارند و در یک رونوشت کدگذاری می‌شوند، یک خوشه را تشکیل می‌دهند، با اینکه توالی هر miRNA بالغ یکسان نیست [۳۲]. مشاهده شده که در ژنوم، ژن‌های یک خانواده از miRNAها به‌طور غیرتصادفی در اطراف ژن‌های مؤثر در بیماری‌های عفونی، سیستم ایمنی، سرطان و بیماری‌های تحلیل‌برنده سیستم عصبی سازماندهی شده‌اند [۳۴]. در نتیجه بررسی خانواده miRNA ها، علاوه بر توالی و عملکرد آنها، در ارتباط آنها با فرایندهای زیستی یا بیماری‌ها حائز اهمیت است.



شکل ۲.۲: ساختار و فرآیندهای پردازشی معمول و پایه‌ای شبکه‌های عصبی گرافی [۳۶].

۲.۲ شبکه عصبی گرافی

گراف یک ساختار ریاضیاتی انتزاعی است که برای مدل‌سازی روابط جفتی اشیاء استفاده می‌شود. این ساختار نه تنها قادر به نشان دادن ارتباطات مستقیم بین عناصر است، بلکه ساختارهای پیچیده‌تری مانند خوشه‌ها، مسیرها، دورها و زیرگراف‌ها را نیز شامل می‌شوند. چنین مجموعهٔ کاملی از اطلاعات امکان تحلیل سیستم‌های پیچیده و کشف الگوها یا روندها و یا پیش‌بینی‌هایی را فراهم می‌کند. ابتدا به تعریف گراف با نمادگذاری‌های ریاضی می‌پردازیم و سپس شبکه عصبی گرافی و شبکه کانولوشن گرافی را که روش اصلی این پژوهش است، شرح می‌دهیم.

تعریف ۱.۲. *یک گراف به صورت $G = (V, E, X)$ تعریف می‌شود، به طوری که V مجموعه‌ای از رئوس (گره‌ها) و E مجموعه‌ای از یال‌ها (یا ارتباطات) است که هر یال به صورت یک زوج مرتب نشان داده می‌شود. همچنین $X \in R^{N \times F}$ ماتریس ویژگی رئوس است که N تعداد گره‌ها و F تعداد ویژگی‌های هر گره را مشخص می‌کند* [۳۵].

شبکه‌های عصبی گرافی به یک دسته بزرگ از معماری‌های شبکه عصبی اشاره دارند که به طور خاص برای پردازش و یادگیری از داده‌های ساختاریافته به صورت گراف طراحی شده‌اند. در GNN ها، گره‌ها توسط همسایگان و اتصالات آنها تعریف می‌شوند. اگر همسایگان و اتصالات یک گره را حذف کنیم، گره تمام اطلاعات خود را از دست می‌دهد. بنابراین همسایگان یک گره و اتصالات آن، مفهوم گره را تعیین می‌کنند (شکل ۲.۲). با در نظر گرفتن این موضوع، به

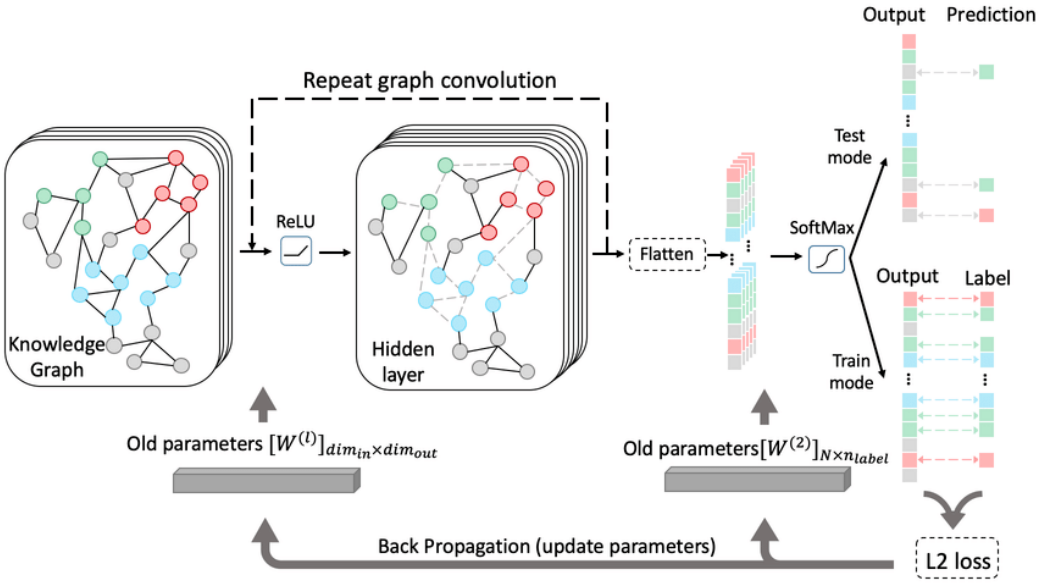
هر گره یک ویژگی اختصاص می‌دهیم تا نماینده مفهوم آن باشد. در حالت کلی، می‌توانیم از یک بردار ویژگی استفاده کنیم. وظیفه تمام مدل‌های GNN تعیین نمایش گره برای هر گره با بررسی اطلاعات همسایگان آن گره است. الگوریتم کلی GNNها را می‌توان به صورت زیر در نظر گرفت:

- **انتقال پیام:** گره‌ها با همسایگان خود از طریق تبادل اطلاعات ارتباط برقرار می‌کنند. این فرآیند که اغلب به عنوان «انتقال پیام» یا «تجمیع همسایگان» شناخته می‌شود، نمایش یک گره را بر اساس ویژگی‌های گره‌های همسایه به‌روزرسانی می‌کند.
- **تجمیع:** توابع تجمیع، پیام‌های ورودی از همسایگان را ترکیب می‌کنند. توابع تجمیع رایج شامل میانگین، جمع، ماکزیمم، یا شبکه‌های عصبی پیچیده‌تر هستند.
- **به‌روزرسانی:** پس از تجمیع، بردارهای ویژگی به‌روزرسانی می‌شود. اینکار معمولاً با استفاده از لایه‌های شبکه عصبی انجام می‌شود.
- **تکرار:** فرآیند انتقال پیام و تجمیع می‌تواند برای چندین لایه یا تکرار انجام شود تا اطلاعات به بخش‌های بزرگتر گراف منتقل شود.

شبکه‌های کانولوشن گرافی نوع خاصی از GNN هستند که به طور خاص برای انجام عملیات کانولوشن بر روی داده‌های ساختاریافته به صورت گراف طراحی شده‌اند. نحوهٔ عملکرد کلی و ساختار این شبکه‌ها در شکل ۳.۲ نشان داده شده است. آنها از عملیات کانولوشن در شبکه‌های عصبی کانولوشنی استاندارد الهام گرفته‌اند اما برای داده‌های غیر اقلیدسی مانند گراف‌ها سازگار شده‌اند [۳۸]. سازوکار کلی GCN ها را می‌توان به صورت زیر فرموله کرد:

$$X^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{S}\tilde{D}^{-1/2}X^{(l)}W^{(l)}) \qquad (1.2)$$

در فرمول بالا $X^{(l+1)}$ ماتریس ویژگی‌های گره‌ها در لایه $(l+1)$ است. این ماتریس نتیجه عملیات کانولوشن در لایه $(l+1)$ بوده و نشان‌دهنده ویژگی‌های به‌روزرسانی شده گره‌ها در این لایه است. $\tilde{D}$ ماتریس قطری از درجات گره‌ها در گراف می‌باشد که در آن هر درایه $\tilde{D}_{ii}$ نمایانگر تعداد یال‌های متصل به گره i است. $\tilde{D}^{-1/2}$ ریشه دوم معکوس این ماتریس است که برای نرمال‌سازی گراف استفاده می‌شود. $\tilde{S}$ ماتریس مجاورت نرمال شده است که اطلاعات یال‌ها و ارتباطات بین



شکل ۳.۲: نحوهٔ عملکرد کلی و ساختار شبکه‌های گرافی کانولوشنی [۳۷].

گره‌ها را در بر دارد. $X^{(l)}$ ماتریس ویژگی گره‌ها در لایه (l) است. $W^{(l)}$ ماتریس وزن لایه (l) است که برای بهینه‌سازی مدل در فرآیند آموزش تنظیم می‌شود. این ماتریس وزن‌ها ویژگی‌های گره‌ها در لایه قبلی را ضرب می‌کند تا ویژگی‌های جدید تولید شود. در نهایت σ تابع فعال‌سازی است که موجب رفتار غیرخطی در مدل می‌شود.

این نوع طراحی شبکه‌های عصبی گرافی اغلب به صورت چند لایه‌ای مورد استفاده قرار می‌گیرند. این طراحی به شبکه اجازه می‌دهد تا روابط پیچیده‌تر و ویژگی‌های سطح بالاتری را در گراف تشخیص دهد به این صورت که خروجی یک لایه GCN به‌عنوان ورودی برای لایه بعدی استفاده می‌شود. این فرآیند برای یک تعداد مشخص از لایه‌ها تکرار می‌گردد تا خروجی نهایی تولید شود. در مدل پیشنهادی از این ساختار برای وظایف مختلفی می‌تواند استفاده شود. در این پژوهش از آن برای یادگیری ارتباط میان miRNA و بیماری استفاده شده است.

فصل ۳

روش پیاده‌سازی‌شده

۱.۳ اندازه‌گیری شباهت‌ها

در زمینه بیوانفورماتیک و زیست‌شناسی محاسباتی، محاسبهٔ شباهت با استفاده از چندین منبع اطلاعاتی مختلف روشی است که برای ادغام چندین نوع داده به منظور ایجاد دید کامل‌تر از روابط و تعاملات در یک سیستم زیستی استفاده می‌شود. با ترکیب داده‌های چندین منبع، محققان می‌توانند اطلاعات جزئی و دقیق‌تری از موجودیت مورد مطالعه بدست آورند. همچنین ادغام چندین منبع می‌تواند تعصبات و محدودیت‌های ذاتی موجود در مجموعه داده‌های مجزا را کاهش دهد. این امر منجر به مدل‌های قابل اعتمادتر و دقیق‌تری می‌شود که می‌تواند پیچیدگی‌های دنیای واقعی را بهتر منعکس کند. چنین رویکردی به‌ویژه در کشف تعاملات عمیق و اغلب غیرآشکار بین موجودیت‌های زیستی بسیار مفید است. به عنوان مثال، miRNA و بیماری‌ها ممکن است از طریق مسیرهای مختلف و مکانیسم‌های تنظیمی مختلفی با هم تعامل داشته باشند. در این بخش، از منابع مختلف داده‌های زیستی برای توصیف کامل معیارهای شباهت miRNAها و بیماری‌ها استفاده می‌کنیم.

۱.۱.۳ کرنل شباهت پروفایل تعامل گاوسی miRNA

پروفایل تعامل گاوسی[۱] یک تکنیک محاسباتی است که برای تحلیل تعاملات بین مولکول‌ها استفاده می شود. این روش اجازه می دهد با استفاده از اطلاعات مرتبط با تعاملات شناخته‌شده، تعاملات جدید و ناشناخته را پیش‌بینی کنیم و معمولا برای فراهم کردن پیش‌بینی‌های دقیق‌تر از تعاملات بین واحدهای زیستی (مثلاً miRNAها و بیماری‌ها) به کار می‌رود. پروفایل تعامل گاوسی بر اساس مدل‌های ریاضی و آماری تعریف می‌شود که اصولاً بر پایه توزیع گاوسی (یا نرمال) بنا شده است. برای هر عنصر، یک پروفایل تعامل ایجاد می‌شود که نشان‌دهنده الگوی تعاملات آن با سایر عناصر است و از یک تابع کرنل گاوسی برای محاسبهٔ شباهت بین پروفایل‌های تعامل استفاده می‌شود.

می‌توان گفت miRNAهایی که از نظر عملکردی مشابه هستند، به احتمال زیاد با بیماری‌هایی که از نظر فنوتیپی مشابه هستند، ارتباط دارند [۴۷]. بر اساس مطالعات گذشته [۵۸]، می‌توان یک miRNA را به عنوان برداری در نظر گرفت که وجود یا عدم وجود ارتباط با هر بیماری در پایگاه‌داده را رمزگذاری می‌کند. بنابراین به سادگی می‌توان این بردار را برابر سطری مربوط به آن miRNA در ماتریس مجاورت بیماری‌ـmiRNA در نظر گرفت. برای دو miRNA مانند m_i و m_j، شباهت گاوسی S_m^{gip} را به صورت زیر تعریف می‌کنیم:

$$S_m^{gip}(m_i, m_j) = e^{-\alpha_m \|IP(m_i) - IP(m_j)\|^2} \qquad (۱.۳)$$

که در آن $IP(m_i)$ و $IP(m_j)$ نمایانگر پروفایل تعامل miRNA و سطر iام و jام ماتریس مجاورت است و N_m نشانگر تعداد miRNAها است. همچنین داریم:

$$\alpha_m = \frac{\alpha'_m}{\frac{1}{N_m} \sum_{i=1}^{N_m} \|IP(m_i)\|^2} \qquad (۲.۳)$$

که این پارامتر اندازهٔ کرنل را تنظیم می‌کند و همچنین α'_m معمولاً مقدار ۱ در نظر گرفته می‌شود [۴۷، ۶۰، ۵۸، ۶۱، ۶۲].

[۱] Gaussian Interaction Profile (GIP)

۲.۱.۳ شباهت توالی miRNA

روش‌های مختلفی برای بررسی شباهت توالی وجود دارد از جمله الگوریتم نیدلمن‌ـوونش[۲] [۴۵] که برای یافتن بهترین هم‌ترازی سراسری بین دو توالی استفاده می‌شود و الگوریتم فاصله لونشتاین[۳] یا کمترین فاصله ویرایش (MED) که یکی از مهم‌ترین الگوریتم‌ها در پردازش زبان طبیعی است [۴۶]. در این مطالعه از الگوریتم MED برای بدست آوردن معیار شباهت توالی miRNAها استفاده شد. این الگوریتم به منظور تعیین میزان تفاوت بین دو رشته (یا کلمه) به کار می‌رود و کمینهٔ تعداد عملیات ویرایشی است که برای تبدیل یک رشته به رشتهٔ دیگر نیاز است. سه نوع عملیات ویرایشی در این الگوریتم تعریف شده است. برای تضمین سازگاری سراسری، رابطهٔ نرمال شدهٔ شباهت توالی $S_m^{seq}(m_i, m_j)$ بین miRNAهای m_i و m_j به صورت زیر بدست می‌آید:

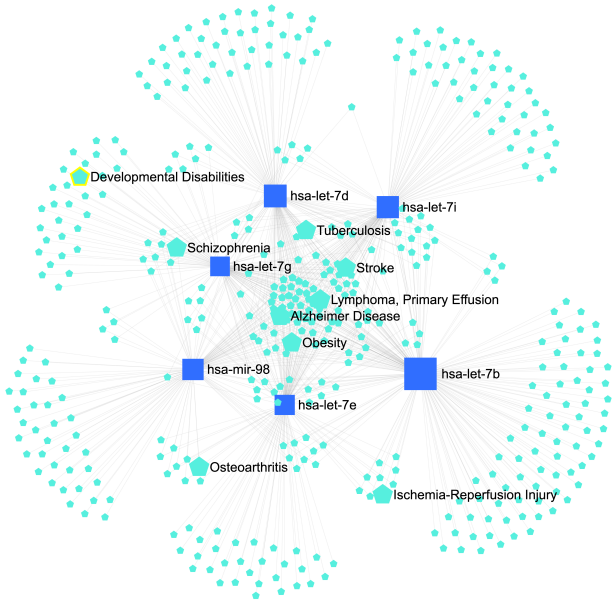
$$S_m^{seq}(m_i, m_j) = ۱ - \frac{MED}{\max(len(m_i), len(m_j))} \qquad (۳.۳)$$

که در آن MED کمترین فاصلهٔ ویرایش بین miRNAها است و $len(m_i)$ و $len(m_j)$ به ترتیب طول m_i و m_j را نشان می‌دهد.

۳.۱.۳ شباهت خانوادگی miRNA

باور بر این است که miRNAهای یک خانواده، بیشتر ممکن است با بیماری‌های مشابه ارتباط داشته باشند [۴۷]. شکل ۱.۳ برخی miRNA عضو خانواده $let-۷$ و بیماری هایی مرتبط با آنها را نشان می دهد. همچنین در گذشته اطلاعات خانواده برای پیش‌بینی ارتباط بیماری و miRNA استفاده شده است [۴۸]. اطلاعات خانواده miRNAها را از پایگاه داده مربوطه (مراجعه شود به بخش ۱.۴) گردآوری و ماتریس شباهت خانواده miRNA به نام S_m^{fam} بدست آمد به طوری که برای m_i و m_j، اگر هر دو به یک خانواده تعلق داشته باشند، $S_m^{fam}(m_i, m_j)$ مقدار ۱ و در غیر این صورت مقدار ۰ اختصاص داده می‌شود.

[۲] Needleman-Wunsch Algorithm
[۳] Levenshtein Distance



شکل ۱.۳: برخی miRNA عضو خانواده $let-۷$ و بیماری هایی مرتبط

۴.۱.۳ شباهت عملکردی miRNA

برای ساخت ماتریس شباهت عملکردی از پژوهش‌های گذشته الهام گرفته شد [۴۹]. ابتدا ژن‌های هدف miRNAها را از پایگاه داده مربوطه (مراجعه شود به بخش ۱.۴) گردآوری و ماتریس تعامل miRNA و ژن به نام GMساخته شد. در این ماتریس سطرها و ستون‌ها به ترتیب miRNAها و ژن‌ها را نشان می‌دهند که اگر تعامل m_i و ژن g_i تایید شده باشد، مقدار آن برابر با ۱ و در غیر این صورت ۰ خواهد بود. بر اساس این ماتریس، پروفایل تعامل گاوسی تعامل miRNA و ژن تعریف می‌شود. در نهایت کرنل پروفایل تعامل گاوسی بین miRNAهای مشابه محاسبه می‌شود و طبق معادلات ۱.۳ و ۲.۳ و در بخش ۱.۱.۳ بدست می‌آید.

۵.۱.۳ کرنل شباهت پروفایل تعامل گاوسی بیماری

فرض ارتباط miRNAها و بیماری‌های مشابه از هر دو سو صدق می‌کند؛ یعنی می‌توان گفت بیماری‌هایی که از لحاظ فنوتیپی مشابه هستند، احتمالاً با miRNAهایی که از نظر عملکردی مشابه هستند مرتبط هستند. بنابراین مشابه بخش ۱.۱.۳، می‌توان کرنل شباهت پروفایل تعامل گاوسی را برای بیماری d_i و بیماری d_j به شرح زیر تشکیل داد:

$$S_d^{gip}(d_i, d_j) = e^{-\alpha_d \|IP(d_i) - IP(d_j)\|^2} \tag{۴.۳}$$

$$\alpha_d = \frac{\alpha'_d}{\frac{1}{N_d}\sum_{i=1}^{N_d} \|IP(d_i)\|^2} \tag{۵.۳}$$

مشابهاً $IP(d_i)$ و $IP(d_j)$ نمایانگر پروفایل تعامل بیماری یا سطر iام و jام ماتریس مجاورت است، N_d تعداد بیماری و پارامتر α_d تنظیم‌گر اندازهٔ کرنل و α'_d که مقدار ۱ در نظر گرفته می‌شود.

۶.۱.۳ شباهت معنایی بیماری‌ها

مانند بسیاری از مقاله‌های گذشته [۵۱، ۵۰، ۵۲، ۴۷، ۵۳، ۵۴] برای توصیف رابطه بین بیماری‌ها، از گراف چرخشی بدون دور (DAG)[۴] برای محاسبه شباهت معنایی آن‌ها استفاده می‌کنیم که در آن گره‌ها نشان‌دهندهٔ بیماری‌ها و یال‌ها نمایانگر رابطه بین بیماری‌ها هستند که یک گره والد را به یک گره فرزند متصل می‌کند. برای بازنمایی شباهت معنایی بیماری‌ها، توصیف‌های MeSH از کتابخانه ملی پزشکی در دسترس هستند. در نتیجه، گراف DAG بیماری d به صورت

$$DAG(d) = (d, T(d), E(d)) \tag{۶.۳}$$

[۴] Directed Acyclic Graph

تعریف می‌شود به طوری که در آن $T(d)$ نشان‌دهنده مجموعه‌ای از بیماری‌هاست که شامل تمامی گره‌های اجداد d به‌علاوه خود d می‌شود و $E(d)$ نشان‌دهنده یال از تمام گره‌های والد به گره‌های فرزند است. مشارکت معنایی $D_d(n)$ بیماری $n \in T(d)$ در بیماری d به صورت زیر قابل محاسبه است:

$$D_d(n) = \begin{cases} ۱ & n = d \\ \max\{\Delta * D_d(n') | n' \in \text{children of } d\} & n \neq d \end{cases} \quad (۷.۳)$$

Δ عامل مشارکت معنایی [۵] نامیده می‌شود و به این معنی است که هرچه فاصله بین بیماری n و اجداد آن بیشتر باشد، سهم معنایی n در بیماری d کمتر است. همچنین معمولا مقدار آن ۰.۵ انتخاب می‌شود [۴۷، ۵۰، ۵۴، ۵۵]. بنابراین مقدار معنایی بیماری d را می‌توان به صورت مجموع مشارکت معنایی تمام گره‌های d تعریف کرد. مقدار معنایی $DV(d)$ را می‌توان به صورت زیر فرموله کرد:

$$DV(d) = \sum_{n \in N_d} D_d(n) \quad (۸.۳)$$

در نهایت شباهت معنایی بیماری d_i و بیماری d_j به صورت زیر تعریف می‌شود:

$$S_d^{sem۱}(d_i, d_j) = \frac{\sum_{n \in T(d_i) \cap T(d_j)} (D_{d_i}(n) + D_{d_j}(n))}{DV(d_i) + DV(d_j)} \quad (۹.۳)$$

به طوری که $D_{d_i}(n)$ و $D_{d_j}(n)$ به ترتیب مشارکت معنایی بیماری n در بیماری d_i و d_j را نشان می‌دهند. ماتریس $S_d^{sem۱} \in R^{N_d \times N_d}$ را اولین معیار شباهت معنایی بیماری در نظر می‌گیریم که در آن N تعداد بیماری‌ها را نشان می‌دهد.

همانطور که ذکر شد، اگر اصطلاحات بیماری در یک لایه از DAG باشند، مشارکت در ارزش معنایی به همان مقدار اختصاص داده می‌شود. با این حال، فراوانی وقوع این بیماری در تمام

[۵] Semantic Contribution Factor

DAGها می‌تواند بسیار متفاوت باشد. یک بیماری خاص که در تعداد کمتری از DAGها وجود دارد، باید سهم بیشتری در مقدار معنایی داشته باشد. برای حل این مشکل می‌توان از شیوه مقالات گذشته استفاده کرد [۴۹، ۴۷، ۵۶، ۵۷]. مشارکت معنایی بیماری n در بیماری d را به صورت زیر در نظر می‌گیریم:

$$C_d(n) = -\log\left(\frac{\text{تعداد DAGهای شامل } n}{N_d}\right) \qquad (۱۰.۳)$$

که در آن N_d تعداد بیماری‌ها را نشان می‌دهد. مقدار معنایی بیماری d را $CV(d)$ می‌نامیم که مانند معادله ۸.۳ قابل محاسبه است. بنابراین معیار دوم شباهت معنایی به صورت زیر محاسبه می‌شود:

$$S_d^{sem۲}(d_i, d_j) = \frac{\sum_{n \in T(d_i) \cap T(d_j)} (D_{d_i}(n) + D_{d_j}(n))}{CV(d_i) + CV(d_j)} \qquad (۱۱.۳)$$

۲.۳ ساخت بردارهای ویژگی

محاسبات و روش بدست آوردن سه معیار شباهت بیماری (شباهت‌های معنایی و کرنل شباهت GIP) و چهار معیار شباهت miRNA (شباهت توالی، شباهت عملکردی، شباهت خانوادگی و کرنل شباهت GIP) در بخش قبل بررسی شد. با ادغام این اطلاعات چندبعدی، می‌توان بردارهای ویژگی برای نمایش miRNAها و بیماری‌ها ساخت.

۱.۲.۳ بردار ویژگی بیماری

در بخش قبل (۶.۱.۳) دو معیار شباهت معنایی بدست آمد. طبق روش مورد استفاده در مطالعات گذشته [۴۹، ۵۷]. این دو معیار تلفیق شدند و معیار نهایی شباهت معنایی بیماری‌ها به صورت زیر بدست آمد:

$$S_d^{sem}(d_i, d_j) = \frac{۱}{۲}(S_d^{sem۱}(d_i, d_j) + S_d^{sem۲}(d_i, d_j)) \qquad (۱۲.۳)$$

معیار شباهت معنایی بیماری‌ها کامل نیست زیرا برخی از بیماری‌ها اطلاعات MeSH ندارند. برای رفع این مشکل، معیار شباهت معنایی بیماری S_d^{sem} را با معیار شباهت گاوسی بیماری S_d^{gip} ترکیب کردیم. بردار ویژگی نهایی بیماری‌ها S_d^{sem} به شرح زیر ساخته می‌شود:

$$S_d^{sem}(d_i, d_j) = \begin{cases} S_d^{gip}(d_i, d_j) & S_d^{sem}(d_i, d_j) = ۰ \\ \frac{۱}{۲}(S_d^{sem}(d_i, d_j) + S_d^{gip}(d_i, d_j)) & S_d^{sem}(d_i, d_j) \neq ۰ \end{cases} \quad (۱۳.۳)$$

برای بیماری d_i، بردار ویژگی توسط iامین ردیف در ماتریس S_d^{sem} نمایش داده می‌شود.

۲.۲.۳ بردار ویژگی miRNA

بسیاری از داده‌های خانوادگی miRNAها در دسترس نیستند زیرا هنوز ارتباط بین آن‌ها کشف نشده است. این موضوع با بررسی ماتریس شباهت S_m^{fam} تایید می‌شود؛ نسبت داده‌های غیر صفر (وجود ارتباط خانوادگی) به صفر حدود ۰/۰۰۴۸ است. بنابراین از معیار شباهت گاوسی miRNA برای حل این محدودیت استفاده می‌کنیم. معیار نهایی شباهت بیماری‌ها S_m^{fam} به شرح زیر ساخته می‌شود:

$$S_m^{fam}(m_i, m_j) = \begin{cases} S_m^{gip}(m_i, m_j) & S_m^{fam}(m_i, m_j) = ۰ \\ \frac{۱}{۲}(S_m^{fam}(m_i, m_j) + S_m^{gip}(m_i, m_j)) & S_m^{fam}(m_i, m_j) \neq ۰ \end{cases} \quad (۱۴.۳)$$

در نهایت ماتریس شباهت عملکردی miRNA به نام S_m^{seq}، ماتریس شباهت توالی S_m^{fam} و ماتریس شباهت خانوادگی S_m^{fam} را با چسباندن آنها به یکدیگری در بعد اول یکپارچه کردیم.

۳.۳ ساختار مدل

ساخت گراف برای پیش‌بینی ارتباط بین miRNA و بیماری‌ها با استفاده از کتابخانه DGL و PyTorch Geometric در پایتون انجام شد. DGL برای تسهیل پیاده‌سازی یادگیری عمیق

بر روی داده‌های گرافی طراحی شده است و یک چارچوب انعطاف‌پذیر و کارآمد برای ایجاد، دستکاری و آموزش شبکه‌های عصبی گراف فراهم می‌کند. تعداد کل گره‌های گراف ۱۷۰۹ یعنی مجموع تعداد بیماری‌ها و miRNAهاست و نوع هر گره مشخص است. این کتابخانه اجازه می‌دهد تا ویژگی‌ها و برچسب‌ها را برای هر گره مشخص کنیم. بنابراین به آسانی بردارهای ویژگی آماده شده در مراحل قبل به این گره‌ها تخصیص داده می‌شود تا آموزش با استفاده از آنها انجام گیرد.

۱.۳.۳ توازن دادگان

در مسائل دسته‌بندی نکتهٔ حائز اهمیت متوازن بودن داده‌های در دسترس است. به صورت کلی مدل‌ها بیشتر به یادگیری ویژگی‌های کلاس غالب (کلاسی که تعداد نمونه‌های بیشتری دارد) می‌پردازد و کلاس‌های کمتر نمایان شده یا نادیده گرفته می‌شوند. بنابراین دقت مدل در پیش‌بینی نمونه‌های با نمونه کمتر، یعنی وجود ارتباط بین دو عنصر مورد بحث، کاهش می‌یابد. این نکته به خوبی اهمیت انتخاب معیارهای ارزیابی مناسب را نشان می‌دهد؛ به عنوان مثال معیار ارزیابی دقت از جمله معیارهای گمراه‌کننده است چرا که در یک دسته‌بندی نامتوازن، مدل می‌تواند با پیش‌بینی همه نمونه‌ها به عنوان کلاس غالب به دقت بالایی دست یابد، حتی اگر در تشخیص کلاس‌های کوچک عملکرد ضعیفی داشته باشد.

بررسی داده‌ها نشان داد که میان miRNA و بیماری‌های موجود ۷۱۱۷۱۴ منفی و تنها ۱۴۵۵۰ ارتباط مثبت وجود دارد به این معنا که تعداد ارتباطات ناشناخته تقریبا دو برابر ارتباطات شناخته شده است. بنابراین داده‌های مربوط به ارتباط بیماری‌ها و miRNAها نامتوازن هستند. برای حل این مشکل از نمونه‌برداری منفی برای مقابله با داده‌های نامتوازن استفاده شد. بدین منظور به صورت تصادفی یک زیرمجموعه از ارتباطات ناشناخته (کلاس منفی) انتخاب شد تا تعداد آن‌ها با تعداد ارتباطات شناخته‌شده (کلاس مثبت) متعادل شود. سپس نمونه‌های مثبت و منفی را ترکیب و به صورت تصادفی مرتب کردیم. از این داده‌های متعادل شده برای برقراری یال‌ها بین گره‌های مرتبط و انتساب برچسب‌های مربوطه (۱ برای وجود ارتباط و ۰ برای عدم وجود ارتباط بین miRNA و بیماری) استفاده کردیم. به این صورت گراف حاصل و ساختارهای داده مرتبط آماده برای تغذیه به مدل شبکه عصبی گراف برای یادگیری و پیش‌بینی هستند.

۲.۳.۳ معماری شبکه گرافی

شبکه طراحی شده شامل دو جزء اصلی است:

۱. لایهٔ کانولوشن گرافی (GCN)

۲. شبکهٔ چندلایه پرسپترون (MLP)

چنین ترکیبی به مدل اجازه می‌دهد تا اطلاعات توپولوژیکی گراف را به خوبی به کار گیرد و از قدرت یادگیری شبکه MLP برای انجام پیش‌بینی‌های دقیق‌تر بهره‌برداری کند. ویژگی‌های اولیه استخراج می‌شوند و فرآیند تولید ویژگی‌های جدید توسط لایهٔ کانولوشنی گرافی چندین بار انجام می‌پذیرد و نتایج به بردار اولیهٔ ویژگی اضافه می‌شود. در ادامه به بررسی عملکرد هر یک از این دو جزء می‌پردازیم.

لایهٔ کانولوشن گراف

این لایه با اعمال کانولوشن بر روی ویژگی‌های گراف، ویژگی‌های جدید را استخراج می‌کند. به طور کلی ماتریس ویژگی یال و وزن یال‌ها محاسبه می‌شود. سپس درجه هر گره محاسبه و با استفاده از توان ۰/۵ − نرمال‌سازی می‌شود. این نرمال‌سازی باعث ایجاد معکوس **ریشه مربع** درجهٔ گره‌ها می‌شود که در فرمول کانولوشن گراف مؤثر است. وزن‌ها یال‌ها با استفاده از درجه‌های نرمال‌سازی شده محاسبه می‌شوند. در مرحله بعد، پیام‌ها نرمال‌سازی می‌شوند و برخی پیام‌ها (اگر شبکه در حال آموزش باشند) به منظور جلوگیری از بیش‌برازش حذف می‌شوند. در نهایت پیام‌رسانی انجام می‌شود که وظیفهٔ آن به‌روزرسانی ویژگی‌های گره‌ها با استفاده از اطلاعات گره‌های همسایه است.

شبکه چندلایه پرسپترون

پس از به‌روزرسانی ویژگی‌های گره‌ها توسط لایهٔ کانولوشن، این ویژگی‌های جدید به MLP داده می‌شوند. MLP قادر است با استفاده از لایه‌های مخفی مختلف، ویژگی‌های پیچیده‌تر را یاد بگیرد و پیش‌بینی نهایی را انجام دهد. پس از تولید ویژگی‌های جدید، تابع فعال‌ساز واحد یکسو شده‌ی خطی (ReLU) اعمال شده و این فرآیند چندین مرتبه انجام می‌گیرد. ویژگی‌های نهایی با یک لایهٔ

تماماً متصل [۶] (FC) استخراج و با استفاده از یک تابع فعال سازی سیگموئید خروجی حاصل می‌شود. همچنین از تابع BCE به عنوان تابع خطا استفاده شد که تفاوت بین پیش‌بینی مدل ($\hat{y}$) و برچسب واقعی (y) را با استفاده از فرمول زیر محاسبه می‌کند:

$$L(y, \hat{y}) = -\frac{۱}{N} \sum_{i=۱}^{N} [y_i \log(\hat{y}_i) + (۱ - y_i) \log(۱ - \hat{y}_i)]$$

[۶] Fully Connected Layer

فصل ۴

نتایج آزمایش‌ها

۱.۴ جمع‌آوری و پردازش دادگان

بانک‌داده HMDD (بانک اطلاعات بیماری miRNA انسانی) به‌طور گسترده‌ای برای پیش‌بینی ارتباطات بیماری‌ـmiRNA مورد استفاده قرار گرفته است. در این مطالعه از نسخه ۳.۲ که شامل ۳۵۵۴۷ ارتباط تایید شده تجربی بین ۱۲۰۶ مولکول miRNA و ۸۹۳ بیماری می‌باشد، استفاده شد [۳۹]. پس از حذف موجودیت‌های تکراری و تعدادی که در بانک‌داده miRBase اطلاعات قابل اعتمادی نداشتند [۴۰]، ۷۱۱۷۱۷ ارتباط شامل ۹۱۷ مولکول miRNA و ۷۹۲ بیماری به‌عنوان مجموعهٔ اصلی مورد استفاده در این پژوهش بدست آمد.

برای ساخت ماتریس‌های شباهت بیماری و miRNA از پایگاه‌داده‌های گوناگونی استفاده شد. برای نمایش شباهت معنایی بیماری‌ها، توصیفگرهای سرعنوان‌های موضوعی پزشکی (MeSH) [۱] از کتابخانه ملی پزشکی ایالات متحده آمریکا دانلود شدند. MeSH یک سیستم طبقه‌بندی سلسله مراتبی و کنترل شده از اصطلاحات پزشکی است که توسط کتابخانه ملی پزشکی ایالات متحده ایجاد و نگهداری می‌شود. این سیستم برای فهرست‌نویسی، جستجو و بازیابی اطلاعات در حوزه علوم زیستی و پزشکی استفاده می‌شود. توصیفگرهای MeSH شامل مجموعه‌ای از اصطلاحات استاندارد هستند که برای توصیف محتوای مقالات، کتاب‌ها و سایر منابع علمی در زمینه پزشکی به

[۱] Medical Subject Heading

کار می‌روند. استفاده از این توصیفگرها امکان مقایسه و تحلیل شباهت‌های معنایی بین بیماری‌های مختلف را فراهم می‌کند. دسترسی به این داده‌ها از طریق وب‌سایت رسمی کتابخانه ملی پزشکی امکان‌پذیر است.

برای ساخت ماتریس شباهت توالی miRNA - miRNA و ماتریس شباهت خانوادگی، اطلاعات توالی و خانواده مولکول‌ها از miRBase دانلود شد [۴۰]. miRBase یک پایگاه داده آنلاین است که اطلاعات جامعی درباره miRNAها ارائه می‌دهد. این بانک اطلاعاتی به‌طور خاص برای ذخیره‌سازی و دسترسی آسان به توالی‌های miRNA و اطلاعات مرتبط با آنها طراحی شده است.

همچنین برای بدست آوردن ماتریس شباهت عملکردی miRNA، ژن‌های هدف miRNAها از پایگاه داده miRTarBase که یک بانک داده جامع است، گردآوری شد [۴۱]. این پایگاه داده به صورت تخصصی به جمع‌آوری تعاملات miRNA و ژن هدف می‌پردازد. با استفاده از این پایگاه داده، می‌توان در رابطه با فرآیندهای تنظیمی و تأثیر miRNAها بر ژن‌ها و بیماری‌ها اطلاعات زیادی بدست آورد.

در بخش مطالعه موردی از ۴ پایگاه داده بیماری‌ـ miRNA برای تایید miRNAهای پیش‌بینی شده توسط مدل استفاده شد. به دلیل اهمیت زیستی miRNAها، تعدادی منابع آنلاین برای ذخیره‌سازی داده‌ها و تحلیل عملکردی MDA توسعه داده شده‌اند. جدول ۱.۴ بانک‌های داده مورد استفاده برای تایید مدل در بخش ۴.۴ و توضیحات کوتاهی درباره هر کدام را نشان می‌دهد. هر یک از این منابع برای جمع‌آوری و ارائه اطلاعات در زمینه‌های مختلفی مورد استفاده قرار می‌گیرند و هر کدام تمرکز و هدف خاص خود را دارد. این موضوع می‌تواند منجر به وجود یا عدم وجود برخی اطلاعات خاص در برخی از آنها شود. همچنین سرعت به‌روز رسانی و نحوه نگهداری اطلاعات نیز ممکن است بین بانک‌های داده متفاوت باشد. بنابراین استفاده از چندین نوع پایگاه داده کمک می‌کند تا اطمینان بیشتری به نتایج داشته باشیم.

۲.۴ معیارهای ارزیابی

برای ارزیابی عملکرد مدل از روش اعتبارسنجی پنج‌گانه متقابل[۲] استفاده شد. تمام نمونه‌ها به‌طور تصادفی به پنج بخش تقریباً مساوی تقسیم شدند. هر کدام از این قسمت‌ها را به نوبت به‌عنوان

[۲] 5 Fold Cross-Validation

نام پایگاه داده	توضیحات
HMDD 4.0	۵۳۵۳۰ داده از ارتباطات miRNA-بیماری شامل ۱۸۱۷ ژن miRNA انسانی، ۷۹ miRNA ویروسی و ۲۳۶۰ بیماری از ۳۷۰۹۰ مقاله [۳۹]
miR2Disease	۳۲۷۳ ارتباط بین ۳۴۹ miRNA و ۱۶۳ بیماری [۴۲]
miRCancer	۹۰۸۰ ارتباط میان ۵۷۹۸۴ miRNA و ۱۹۶ نوع سرطان [۴۳]
dbDEMC	تعداد ۵۶۶۵۵ ارتباط که شامل ۳۲۶۸ miRNA، ۴۰ نوع سرطان و ۱۴۹ زیر نوع سرطان [۴۴]

جدول ۱.۴: پایگاه داده‌های تعامل miRNA و بیماری

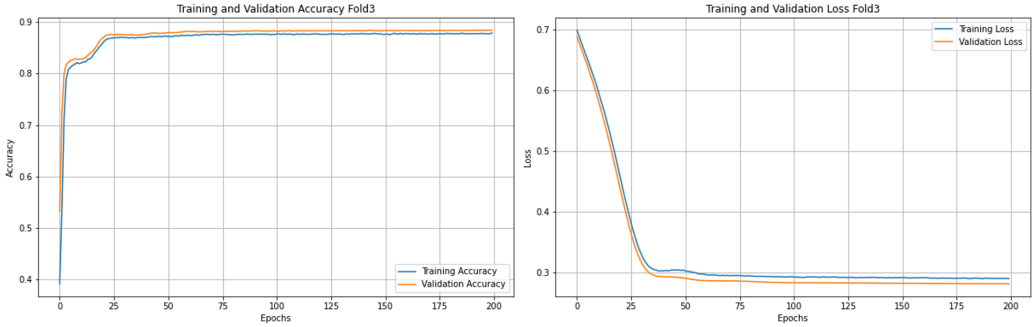
مجموعه تست در نظر گرفته شد، در حالی که قسمت‌های دیگر به‌عنوان مجموعه آموزش استفاده شدند. عملکرد مدل با استفاده از چندین معیار سنجیده شد: صحت (ACC)، دقت (PPV)، نرخ مثبت حقیقی (TPR)، نرخ منفی حقیقی (TNR) ، ضریب همبستگی ماتیوس (MCC)، معیار F۱، مساحت زیر منحنی مشخصه عملکرد (AUC) و مساحت زیر منحنی دقت-بازیابی (AUPR). فرمول‌های معیارهای استفاده شده به شرح زیر هستند:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{۱.۴}$$

$$TNR = \frac{TN}{TN + FP} \tag{۲.۴}$$

$$TPR = \frac{TP}{TP + FN} \tag{۳.۴}$$

$$PPV = \frac{TP}{TP + FP} \tag{۴.۴}$$



شکل ۱.۴: نمودارهای خطای BCE (راست) و دقت (چپ) برای داده‌های آموزش و آزمایش.

$$F۱ = \frac{۲ \times TP}{۲ \times TP + FP + FN} = ۲ \times \frac{PPV \times TPR}{PPV + TPR} \qquad (۵.۴)$$

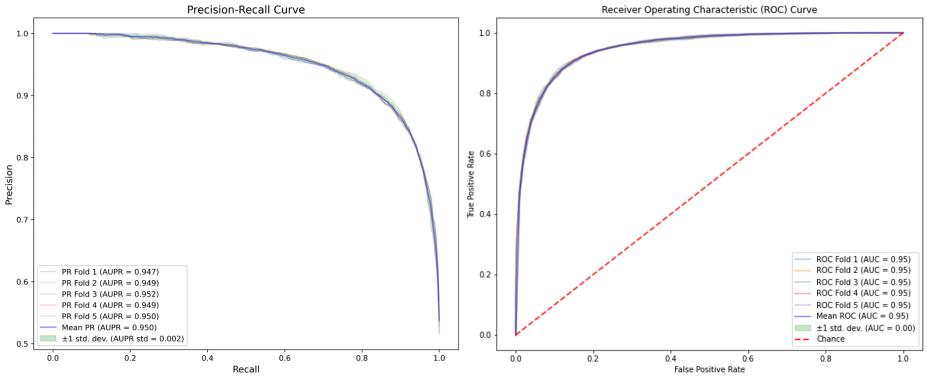
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (۶.۴)$$

در فرمول‌های بالا، TP تعداد نمونه‌های مثبت حقیقی، TN تعداد نمونه‌های منفی حقیقی، FP تعداد نمونه‌های مثبت کاذب و FN تعداد نمونه‌های منفی کاذب را نشان می‌دهد.

۳.۴ ارزیابی عملکرد مدل

برای اندازه‌گیری عملکرد در پیش‌بینی MDA، مدل با در نظر گرفتن تمام معیارهای شباهت برای بردارهای ویژگی بر روی نسخه ۳.۲ پایگاه داده HMDD سنجیده شد. جزئیات هایپرپارامترها و نتایج اعتبارسنجی پنج‌گانه متقابل در جدول ۱.۶ و جدول قابل مشاهده است. شکل ۱.۴ نمودارهای خطا و دقت را برای داده‌های آموزش و آزمایش در طول مدت آموزش نشان می‌دهد.

دقت متوسط مدل بر روی نسخه ۳.۲ پایگاه داده HMDD برابر با ۸۸٫۰۱٪ اندازه‌گیری شد. حساسیت مدل که نسبت نمونه‌های مثبت حقیقی را اندازه‌گیری می‌کند، ۸۸٫۲۷٪ بود. نتایج صحت



شکل ۲.۴: منحنی‌های ROC (راست) و PR (چپ) اعتبارسنجی متقابل.

و نرخ منفی حقیقی متوسط به ترتیب ۸۷/۸۲٪ و ۸۷/۷۵٪ بودند. همچنین میانگین معیار F۱ و ضریب همبستگی ماتیوس به ترتیب ۸۸/۰۴٪ و ۷۶/۰۳٪ اندازه‌گیری شد.

شکل ۲.۴ منحنی ROC و منحنی دقت‌ـ‌فراخوان AUPR را برای روش پیشنهادی نشان می‌دهد. منحنی ROC یک ابزار گرافیکی برای ارزیابی عملکرد مدل‌های طبقه‌بندی است. این منحنی نشان‌دهنده تعادل بین نرخ مثبت کاذب FPR و نرخ مثبت واقعی TPR است. در این منحنی، محور افقی نشان‌دهنده FPR و محور عمودی نشان‌دهنده TPR است. هرچه منحنی به گوشهٔ بالای چپ نزدیک‌تر باشد، نشان‌دهنده عملکرد بهتر مدل است که در شکل راست مشاهده می‌کنیم. منحنی PR نشان‌دهنده تعادل بین دقت و بازیابی است. در این منحنی، محور افقی نشان‌دهنده PPV و محور عمودی نشان‌دهنده TPR است. هرچه منحنی به گوشهٔ بالای راست نزدیک‌تر باشد، نشان‌دهنده عملکرد بهتر مدل است که در شکل چپ مشاهده می‌شود. AUC و AUPR همواره عددی بین صفر و یک هستند که هرچه به ۱ نزدیک‌تر باشد، نشان‌دهندهٔ عملکرد بهتر مدل است. معمولاً AUC و AUPR بالاتر از ۰/۷ به عنوان کارایی خوب و مقدار بالاتر از ۰/۸ کارایی بسیار خوب تلقی می‌شود. متوسط این دو معیار به ترتیب برابر با ۹۴/۹۷٪ و ۹۴/۹۶٪ است که مقادیر بالایی هستند. این نتایج نشان‌دهنده توانایی بسیار خوب مدل پیشنهادی در پیش‌بینی MDA است.

لایه	ACC	PPV	TPR	TNR	F۱	AUC	AUPR	MCC
۱	۸۷٫۹۷	۸۷٫۹۱	۸۸٫۰۰	۸۷٫۹۵	۸۷٫۹۵	۹۴٫۷۰	۹۴٫۶۷	۷۵٫۹۵
۲	۸۷٫۹۲	۸۸٫۰۰	۸۷٫۶۰	۸۸٫۲۳	۸۷٫۸۰	۹۴٫۹۷	۹۴٫۹۳	۷۵٫۸۴
۳	۸۸٫۴۳	۸۸٫۳۶	۸۸٫۸۰	۸۸٫۰۵	۸۸٫۵۸	۹۵٫۱۹	۹۵٫۲۴	۷۶٫۸۵
۴	۸۷٫۹۸	۸۷٫۷۸	۸۸٫۲۱	۸۷٫۷۵	۸۷٫۹۹	۹۴٫۹۶	۹۴٫۹۳	۷۵٫۹۶
۵	۸۷٫۷۶	۸۷٫۰۴	۸۸٫۷۳	۸۶٫۷۹	۸۷٫۸۸	۹۵٫۰۱	۹۵٫۰۲	۷۵٫۵۳
میانگین	**۸۸٫۰۱**	**۸۷٫۸۲**	**۸۸٫۲۷**	**۸۷٫۷۵**	**۸۸٫۰۴**	**۹۴٫۹۷**	**۹۴٫۹۶**	**۷۶٫۰۳**
واریانس	۰٫۲۲	۰٫۴۴	۰٫۴۵	۰٫۵۱	۰٫۲۸	۰٫۱۵	۰٫۱۸	۰٫۴۴

جدول ۲.۴: نتایج ارزیابی مدل روی پایگاه‌داده HMDD. تمام معیارها به درصد (٪) هستند.

۱.۳.۴ ارزیابی عملکرد مدل با بردارهای ویژگی مختلف

همانطور که پیش‌تر اشاره شد، انواع مختلفی از ویژگی‌ها را می‌توان برای بهبود نمایش miRNAها و بیماری‌ها به کار برد. ویژگی‌های مورد استفاده برای miRNAها در این مطالعه شباهت توالی، شباهت خانوادگی و شباهت عملکردی آن‌ها در نظر گرفته شد. برای بررسی تأثیر هر یک از این ویژگی‌ها بر روی عملکرد ساختار پیشنهادی، مدل با ۷ ترکیب مختلف از آن‌ها ارزیابی شد.

در ترکیب‌های ۱، ۲ و ۳، به ترتیب فقط از شباهت توالی، شباهت خانوادگی یا شباهت عملکردی برای ساختن بردار ویژگی miRNAها استفاده شد. در حالت ۴ و ۵ به ترتیب شباهت توالی با شباهت عملکردی و شباهت خانوادگی ترکیب شد و در ترکیب ۶ فقط از شباهت عملکردی و شباهت خانوادگی استفاده شد. در مدل پیشنهادی، ترکیب هر سه نوع شباهت برای نمایش بردارهای ویژگی miRNAها به کار گرفته شد که آن را ترکیب آخر و هفتم در نظر می‌گیریم. توجه داریم که در تمام این ترکیب‌ها برای بردار ویژگی بیماری‌ها از شباهت تجمعی بیماری استفاده شده است. نتایج این ارزیابی در جدول خلاصه شده است.

ترکیب	ACC	PPV	TPR	TNR	F۱	AUC	AUPR	MCC
S_m^{seq}	۸۶/۹۹	۸۶/۸۹	۸۷/۱۱	۸۶/۸۶	۸۷/۰۰	۹۴/۴۴	۹۴/۵۴	۷۳/۹۸
S_m^{fun}	۸۶/۸۵	۸۶/۰۰	۸۸/۰۳	۸۵/۶۸	۸۷/۰۰	۹۴/۳۶	۹۴/۴۸	۷۳/۷۲
S_m^{fam}	۸۸/۰۱	۸۷/۵۹	**۸۸/۵۵**	۸۷/۴۶	۸۸/۰۷	۹۴/۹۶	۹۴/۹۳	۷۶/۰۲
$S_m^{seq} + S_m^{fun}$	۸۷/۰۱	۸۶/۹۴	۸۷/۱۰	۸۶/۹۲	۸۷/۰۲	۹۴/۴۴	۹۴/۵۶	۷۴/۰۲
$S_m^{seq} + S_m^{fam}$	**۸۸/۰۴**	۸۷/۷۹	۸۸/۳۷	۸۷/۷۲	**۸۸/۰۸**	۹۴/۹۶	۹۴/۹۶	**۷۶/۰۸**
$S_m^{fun} + S_m^{fam}$	۸۸/۰۲	۸۷/۶۴	۸۸/۵۱	۸۷/۵۲	۸۸/۰۷	۹۴/۹۵	۹۴/۹۲	۷۶/۰۳
$S_m^{seq} + S_m^{fun} + S_m^{fam}$	۸۸/۰۱	**۸۷/۸۲**	۸۸/۲۷	**۸۷/۷۵**	۸۸/۰۴	**۹۴/۹۷**	**۹۴/۹۶**	۷۶/۰۳

جدول ۳.۴: نتایج ارزیابی مدل با هفت ترکیب مختلف بردارهای ویژگی miRNA روی پایگاه‌داده HMDD. تمام معیارها به درصد (٪) هستند.

بررسی نتایج عملکرد مدل با بردارهای ویژگی مختلف

با نگاه کلی به جدول، متوجه می‌شویم بالاترین دقت، معیار F۱ و ضریب ماتئوس، به ترتیب ۸۸/۰۴٪ و ۸۸/۰۸٪ و ۷۶/۰۸٪ و مربوط به ترکیب ۵ یعنی ادغام شباهت خانوادگی و توالی است. بالاترین نرخ مثبت حقیقی مقدار ۸۸/۵۵٪ و با استفاده از شباهت خانوادگی برای ساخت بردار ویژگی مولکول بدست آمد. همچنین در میان بردارهای تک معیاری (ترکیب ۱ و ۲ و ۳) بهترین عملکرد کلی در استفاده از معیار شباهت خانوادگی دیده می‌شود. بنابراین به نظر می‌رسد شباهت خانواده به تنهایی بهترین نمایش را از miRNAها می‌تواند به مدل نشان دهد.

در مقابل ضعیف ترین نتایج با استفاده از شباهت عملکردی بدست آمد. البته ترکیب این شباهت با دیگری شباهت‌ها اکثرا سبب بهبود عملکرد مدل شده است: برای مثال مقدار TNR در ترکیب شباهت‌های عملکردی و توالی ۸۷/۹۲٪ اندازه گیری شد در حالی که این مقدار در هر یک تنهایی به ترتیب برابر ۸۶/۸۶٪ و ۸۵/۶۸٪ بود.

همانطور که انتظار می‌رفت از بررسی معیارهای ارزیابی می‌توان نتیجه گرفت که به طور کلی ترکیب معیارهای شباهت گوناگون برای نشان دادن موجودیت‌های گراف، می‌تواند توضیف بهتری از آن موجودیت‌ها ارائه دهد. همچنین بنظر می‌رسد معیارهایی که به تنهایی نمایش خوبی از miRNA شکل می‌دهند می‌توانند اثر هم افزایی داشته باشند و ادغام آنها به بهبود قابلیت اعتماد و

دقت پیش‌بینی مدل منجر می‌شود. در نتیجه ادغام هر سه بردار ویژگی شباهت $S_m^{seq} + S_m^{fun} + S_m^{fam}$ بهترین عملکرد و بهبود متعادل را در طیف وسیعی از معیارها ارائه می‌دهد.

۴.۴ مطالعه موردی

برای بررسی توانایی مدل پیشنهادی در استنباط MDA، سه مطالعه موردی بر روی سرطان ریه، سرطان پستان و سرطان خون انجام شد. در هر مورد، تمام ارتباطات شناخته شده از نسخه ۳.۲ HMDD برای ساخت مدل استفاده شد. سپس برای بدست یش بینی احتمال وجود یال میان یک بیماری و miRNA اندیس گره بیماری در گراف مشخض شد و احتمال وجود یال با تمام گره های miRNA گراف بدست آمد. سپس این امتیازات از زیاد به کم مرتب و miRNAهای کاندید بر اساس امتیازات پیش‌بینی‌شان رتبه‌بندی شدند. ۵۰ miRNA برتر انتخاب و با استفاده از ۴ پایگاه داده تایید شدند (مراجعه شود به ۱.۴). برای هر یک از روی سرطان ریه، سرطان پستان و سرطان خون به ترتیب ۵۰، ۴۸ و ۴۹ مورد از ۵۰ مولکول برتر پیش‌بینی شده، تایید شدند. نتایج دقیق ۲۰ مورد برتر برای سرطان پستان، سرطان ریه و سرطان خون به ترتیب در جدول تکمیلی ۲.۶، جدول تکمیلی ۳.۶ و جدول تکمیلی ۴.۶ فهرست شده‌اند.

رشدهای غیرطبیعی بافتی سبب ایجاد تومور در ریه‌ها می‌شود. در میان همهٔ سرطان‌ها، نرخ مرگ و میر ناشی از سرطان‌ها ریه در مردان بالاترین و در زنان در رتبه دوم قرار دارد. تحقیقات متعددی نشان داده‌اند که سرطان ریه با برخی miRNAها مرتبط است. به عنوان مثال، let-7 در پاتوژنز، مهاجرت و متاستاز سرزان ریه نقش دارد [۶۳]. این miRNA توسط مدل پیش‌بینی شد. همچنین سرطان پستان در زنان شایع‌ترین نوع سرطان است. بنابراین، تشخیص زودهنگام این نوع سرطان برای درمان بیماری بسیار حائز اهمیت است. برخی از که در سلول‌های سرطانی پستان شناسایی شده‌اند، شامل miR-9، miR-10b، miR-21، miR-155 miR-210 هستند [۶۵] که همگی توسط مدل پیشنهاد شده اند. پروفایل‌های بیان خاصی از miRNAها در چندین نوع سرطان خونی گزارش شده است. اولین شواهد در مورد دخالت miRNAها در ایجاد این نوع از سرطان‌ها در یک مطالعه مربوط به لوکمی لنفوسیتیک مزمن گزارش شد [۶۴]؛ این مطالعه هیچ ژن کدکننده پروتئینی را شناسایی نکرد، اما چندین نوع مولکول miRNA مثل miR-15a به عنوان عوامل دخیل در این فرایند تایید شد که مدل نیز آنرا پیش‌بینی کرده است.

فصل ۵

جمع‌بندی و پیشنهادات

شناسایی miRNAهای مرتبط با بیماری می‌تواند ما به ما در فهم مکانیسم‌های بیماری‌های انسانی کمک کند. در این پژوهش، مدلی بر پایه شبکه عصبی کانولوشنی گرافی پیشنهاد شد که برای پیش‌بینی MDA آموزش دید. شباهت عملکردی، شباهت توالی و شباهت خانوادگی برای ساخت بردار ویژگی miRNAها و شباهت معنایی در دو سطح برای ساخت بردار ویژگی بیماری‌ها استفاده شد. در ادامه، اهمیت هر یک از معیارهای شباهت در نمایش بهینه موجودیت‌های شبکه بررسی شد. در نهایت، مدل پیشنهادی به منظور پیش‌بینی ارتباطات miRNA-بیماری آموزش داده شد و عملکرد آن با معیارهای مختلفی سنجیده شد. مدل شبکه عصبی گرافی بر روی داده‌های HMDD عملکرد بسیار خوبی نشان داد که نمایانگر کارایی این ساختار برای پیش‌بینی MDA است. نتایج سه مطالعه موردی توانایی بالای مدل را در شناسایی miRNAهای مرتبط با این بیماری‌ها نشان داد. در نتیجه، می‌توان گفت شبکه‌های عصبی گرافی مدل‌های کارامدی برای پیش‌بینی MDA هستند و قادرند اطلاعاتی بسیار ارزشمندی را برای آزمایشات بیولوژیکی در آینده فراهم کند.

با این وجود، مدل پیاده‌سازی شده همچنان جای پیشرفت دارد. آزمایش‌های بیشتری روی بدست آوردن انواع معیارهای شباهت نیاز است. این پژوهش نشان داد که انواع و نحوهٔ ادغام معیارهای مختلف شباهت به نوبه خود بر نتیجهٔ نهایی پیش‌بینی تأثیر می‌گذارد. بنابراین، به خصوص برای بدست آوردن شباهت عملکردی miRNAها، روش‌ها و پایگاه‌های متنوعی نیاز است تا تعاملات پیچیده‌تر میان اجزا را تشخیص دهد و عملکرد شبکه را بهبود بخشد. بررسی روش‌های ادغام این ویژگی‌ها نیر از اهمیت زیادی برخوردار است. در این پژوهش ویژگی‌های هر جزء به

صورت خطی ترکیب شدند در حالی که می توان ابتدا چندین ابرگراف miRNA و بیماری را بر اساس معیارهای شباهتی گوناگون تشکیل داد و سپس با استفاده از کانولوشن ابرگراف، تعاملات مرتبهٔ بالاتر بین گره‌ها را ثبت کرد.

از طرفی با افزودن مکانیسم توجه، مدل می‌تواند اهمیت بخش‌های مختلف ورودی‌ها را در هنگام تصمیم‌گیری تشخیص دهد. این روش تفسیرپذیری پیش‌بینی‌ها را بهبود بخشیده و به شناسایی miRNAها یا بیماری‌های کلیدی که نقش مرکزی در ارتباطات خاص دارند، کمک می‌کند. همچنین استفاده از تکنیک‌های یادگیری تقابلی[۱] در ترکیب با GNNها، به یادگیری نمایش‌های قوی‌تر و قابل تعمیم‌تر از miRNAها یا بیماری‌ها کمک می‌کند؛ به ویژه زمانی که داده‌های برچسب‌دار محدود است. همچنین ترکیب چندین مدل GNN یا ادغام GNNها با سایر رویکردهای یادگیری ماشین منجر به پیش‌بینی‌های قوی‌تر و دقیق‌تر شده است (روش ترکیبی [۲]). در نهایت می توان از GNNهایی که بر روی مجموعه داده‌های زیستی در مقیاس بزرگ از پیش آموزش دیده‌اند، استفاده کرد و آن را برای پیش‌بینی MDA خاص تنظیم نمود.

---

[۱] Contrastive Learning

[۲] Ensemble Methods

فصل ۶

پیوست

۱.۶ دسترسی به داده‌ها و کدها

تمام داده‌ها و کدهای استفاده‌شده در این مطالعه از طریق لینک گیت‌هاب زیر در دسترس هستند. این لینک شامل مجموعهٔ دادگان، کدها و هر گونه ابزار لازم برای بازتولید نتایج ارائه‌شده در این مقاله است: https://github.com/mahroo-hm/MDA-Prediction-Using-GCN

۲.۶ جداول

مشخصات	مقدار
تعداد دور آموزش	۲۰۰
نرخ یادگیری	۰/۰۰۰۱
نرخ کاهش وزن	۰/۰۰۵
تابع خطا	BCE (Binary Cross-Entropy)
بهینه سازی	Adam
تعداد یال‌های آموزش	۴۶۵۶۰
تعداد یال‌های آزمایش	۱۱۶۴۰
تعداد بیماری‌ها	۷۹۲
تعداد miRNAها	۹۱۷
ابعاد ماتریس شباهت بیماری	۱۷۰۹ × ۷۹۲
ابعاد ماتریس شباهت miRNA	۱۷۰۹ × ۲۷۵۱

جدول ۱.۶: اطلاعات پیاده‌سازی

رتبه	miRNA	HMDD4.0	miR2Disease	dbDEMC	miRCancer
۱	hsa-mir-21	تایید شده	تایید شده	تایید شده	تایید شده
۲	hsa-mir-155	تایید شده	تایید شده	تایید شده	تایید شده
۳	hsa-mir-146a	تایید شده	تایید شده	تایید شده	تایید شده
۴	hsa-mir-223	تایید شده	تایید شده	تایید شده	تایید شده
۵	hsa-mir-34a	تایید شده	تایید شده	تایید شده	تایید شده
۶	hsa-mir-126	تایید شده	تایید شده	تایید شده	تایید شده
۷	hsa-mir-17	تایید شده	تایید شده	تایید شده	تایید شده
۸	hsa-mir-29a	تایید شده	تایید شده	تایید شده	تایید شده
۹	hsa-mir-221	تایید شده	تایید شده	تایید شده	تایید شده
۱۰	hsa-mir-145	تایید شده	تایید شده	تایید شده	تایید شده
۱۱	hsa-mir-210	تایید شده	تایید شده	تایید شده	تایید شده
۱۲	hsa-mir-20a	تایید شده	تایید شده	تایید شده	تایید شده
۱۳	hsa-mir-150	تایید شده	-	تایید شده	تایید شده
۱۴	hsa-mir-31	تایید شده	تایید شده	تایید شده	تایید شده
۱۵	hsa-mir-222	تایید شده	تایید شده	تایید شده	تایید شده
۱۶	hsa-mir-143	تایید شده	تایید شده	تایید شده	تایید شده
۱۷	hsa-mir-142	تایید شده	-	-	تایید شده
۱۸	hsa-mir-15a	تایید شده	-	تایید شده	تایید شده
۱۹	hsa-mir-206	تایید شده	تایید شده	تایید شده	تایید شده
۲۰	hsa-mir-200b	تایید شده	تایید شده	تایید شده	تایید شده

جدول ۲.۶: نتایج مطالعه موردی روی سرطان پستان

رتبه	miRNA	HMDD4	miR2Disease	dbDEMC	miRCancer
۱	hsa-mir-155	تایید شده	تایید شده	تایید شده	تایید شده
۲	hsa-mir-150	تایید شده	تایید شده	تایید شده	تایید شده
۳	hsa-mir-15a	تایید شده	تایید شده	تایید شده	تایید شده
۴	hsa-mir-142	تایید شده	-	تایید شده	-
۵	hsa-mir-19a	تایید شده	تایید شده	تایید شده	تایید شده
۶	hsa-mir-145	تایید شده	تایید شده	تایید شده	تایید شده
۷	hsa-mir-195	تایید شده	تایید شده	تایید شده	تایید شده
۸	hsa-let-7b	تایید شده	تایید شده	تایید شده	تایید شده
۹	hsa-mir-122	تایید شده	-	تایید شده	-
۱۰	hsa-mir-132	تایید شده	-	تایید شده	تایید شده
۱۱	hsa-mir-192	تایید شده	تایید شده	تایید شده	تایید شده
۱۲	hsa-mir-182	تایید شده	تایید شده	تایید شده	تایید شده
۱۳	hsa-mir-143	تایید شده	تایید شده	تایید شده	تایید شده
۱۴	hsa-mir-126	تایید شده	تایید شده	تایید شده	تایید شده
۱۵	hsa-mir-15b	تایید شده	تایید شده	تایید شده	تایید شده
۱۶	hsa-mir-93	تایید شده	تایید شده	تایید شده	تایید شده
۱۷	hsa-mir-183	تایید شده	تایید شده	تایید شده	تایید شده
۱۸	hsa-mir-17	تایید شده	تایید شده	تایید شده	تایید شده
۱۹	hsa-mir-20a	تایید شده	-	تایید شده	تایید شده
۲۰	hsa-mir-18a	تایید شده	تایید شده	تایید شده	تایید شده

جدول ۳.۶: نتایج مطالعه موردی روی سرطان ریه

رتبه	miRNA	HMDD4.0	miR2Disease	dbDEMC	miRCancer
۱	hsa-mir-155	تایید شده	تایید شده	تایید شده	تایید شده
۲	hsa-mir-150	تایید شده	تایید شده	تایید شده	تایید شده
۳	hsa-mir-15a	تایید شده	تایید شده	تایید شده	تایید شده
۴	hsa-mir-142	تایید شده	-	تایید شده	-
۵	hsa-mir-19a	تایید شده	تایید شده	تایید شده	تایید شده
۶	hsa-mir-145	تایید شده	تایید شده	تایید شده	-
۷	hsa-mir-122	تایید شده	-	-	تایید شده
۸	hsa-mir-195	تایید شده	تایید شده	تایید شده	تایید شده
۹	hsa-mir-143	تایید شده	تایید شده	تایید شده	تایید شده
۱۰	hsa-let-7b	تایید شده	تایید شده	تایید شده	تایید شده
۱۱	hsa-mir-182	تایید شده	تایید شده	تایید شده	-
۱۲	hsa-mir-192	تایید شده	تایید شده	تایید شده	تایید شده
۱۳	hsa-mir-132	تایید شده	تایید شده	تایید شده	-
۱۴	hsa-mir-126	تایید شده	تایید شده	تایید شده	تایید شده
۱۵	hsa-mir-15b	تایید شده	تایید شده	تایید شده	تایید شده
۱۶	hsa-mir-17	تایید شده	-	تایید شده	تایید شده
۱۷	hsa-mir-183	تایید شده	تایید شده	تایید شده	-
۱۸	hsa-mir-18a	تایید شده	-	تایید شده	تایید شده
۱۹	hsa-mir-20a	تایید شده	تایید شده	تایید شده	تایید شده
۲۰	hsa-mir-93	تایید شده	تایید شده	تایید شده	-

جدول ۴.۶: نتایج مطالعه موردی روی سرطان خون

واژه‌نامهٔ فارسی به انگلیسی

آنتروپی متقاطع دوتایی ..... Binary Cross Entropy (BCE)
ابرگراف ..... Hypergraph
اعتبارسنجی پنج‌گانه متقابل ..... 5 Fold Cross-Validation
الگوریتم‌های طبقه بندی ..... Classification Algorithms
الگوریتم‌های نیدلمن ونش ..... Needleman-Wunsch Algorithm
الگوریتم‌های یادگیری ..... Learning Algorithms
امتیاز لگاریتم درست‌نمایی ..... Log-Likelihood Scores (LLS)
بازخواهی ..... Recall
بردار ویژگی ..... Feature Vector
برچسب ..... Label
بیش‌برازی ..... Overfitting
بیماری‌های تحلیل‌برنده سیستم عصبی ..... Neurodegenerative Diseases
پروفایل تعامل گاوسی ..... Gaussian Interaction Profile (GIP)
پیام‌رسانی ..... Message Passing
تابع سیگموید ..... Sigmoid Function
تابع فعال‌ساز ..... Activation Function
تجمیع ..... Aggregation
تعامل بیماری‌ـmiRNA ..... miRNA-Disease Association (MDA)
حاشیه‌نویسی ..... Annotation
درخت تصمیم‌گیری ..... Decision Tree
دقت ..... Accuracy

دیدگاه‌های شباهت ..... Similarity views
روش‌های ترکیبی ..... Ensemble methods
ژن‌های کدگذار ..... Protein Coding Gens
سرعنوان‌های موضوعی پزشکی ..... Medical Subject Headings (MeSH)
شباهت توالی ..... Sequence Similarity
شباهت خانوادگی ..... Family Similarity
شباهت عملکردی ..... Functional Similarity
شباهت معنایی ..... Semantic Similarity
شبکهٔ عصبی ..... Neural Network
شبکهٔ عصبی کانولوشنی ..... Convolutional Neural Network (CNN)
شبکهٔ گرافی با مکانیسم توجه ..... Graph Attention Networks (GAT)
شبکهٔ گرافی کانولوشنی ..... Graph Convolutional Network (GCN)
شبکهٔ ناهمگن ..... Heterogeneous Network
ضریب همبستگی ماتیوس ..... Matthews Correlation Coefficient (MCC)
فاصله لونشتاین ..... The Levenshtein Distance
فضای اقلیدوسی ..... Euclidean Space
فنوتیپ ..... Phenotype
کدگذار ..... encoder
گام تصادفی ..... Random Walk
گراف بدون جهت بدون دور ..... Directed Acyclic Graph (DAG)
گره ..... Node
لایه ..... Fold
مساحت زیر منحنی ROC ..... Area under the ROC Curve (AUC)
مشارکت معنایی ..... Semantic Contribution
مکانیسم توجه ..... Attention Mechanism
مقدار معنایی ..... Semantic Value
منظم‌سازی ..... Regularization
میانگین بهترین تطابق ..... BMA (Best Match Average)
میکرو ریبونوکلئینک اسید ..... miRNA (micro RiboNeucleinAcid)

میکرو ریبونوکلئینک اسید اولیه ..... pri-miRNA (primary miRNA)
نرخ مثبت حقیقی ..... True Positive Rate (TPR)
نرخ مثبت کاذب ..... True Negative Rate (TNR)
نرخ منفی حقیقی ..... False Positive Rate (FPR)
جاسازی ..... Embedding
واحد خطی اصلاح شده ..... Rectified Linear Unit (ReLU)
وزن‌ها ..... Weights
همترازی سراسری ..... Global Alignment
یادگیری تعاملی گرافی ..... Graph Contrastive Learning
یادگیری ماشین ..... Machine Learning
یادگیری منظم‌سازی شده ..... Regularized Learning
یادگیری نمایشی ..... Representation Learning
یال ..... Edge

مراجع

[1] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004 Jan 23;116(2):281-97. https://doi.org/10.1016/S0092-8674(04)00045-5

[2] Rajewsky, N. L(ou)sy miRNA targets?. Nat Struct Mol Biol 13, 754–755 (2006). https://doi.org/10.1038/nsmb0906-754

[3] Benjamin P. Lewis, I-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, Christopher B. Burge, Prediction of Mammalian MicroRNA Targets, Cell, Volume 115, Issue 7, 2003, Pages 787-798, https://doi.org/10.1016/S0092-8674(03)01018-3

[4] Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., Lim, B., & Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell, 126(6), 1203–1217. https://doi.org/10.1016/j.cell.2006.07.031

[5] Chen, JF., Mandel, E., Thomson, J. et al. The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. Nat Genet 38, 228–233 (2006). https://doi.org/10.1038/ng1725

[6] Meltzer, P. Small RNAs with big impacts. Nature 435, 745–746 (2005). https://doi.org/10.1038/435745a

[7] Forero DA, van der Ven K, Callaerts P, Del-Favero J. miRNA genes and the brain: implications for psychiatric disorders. Hum Mutat. 2010 Nov;31(11):1195-204. https://doi.org/10.1002/humu.21344

[8] Wang WX, Rajeev BW, Stromberg AJ, Ren N, Tang G, Huang Q, Rigoutsos I, Nelson PT. The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1. J Neurosci. 2008 Jan 30;28(5):1213-23. https://doi.org/10.1523/JNEUROSCI.5065-07.2008

[9] Goh, S.Y.; Chao, Y.X.; Dheen, S.T.; Tan, E.-K.; Tay, S.S.-W. Role of MicroRNAs in Parkinson's Disease. Int. J. Mol. Sci. 2019, 20, 5649. https://doi.org/10.3390/ijms20225649

[10] Sabirzhanov B, Faden AI, Aubrecht T, Henry R, Glaser E, Stoica BA. MicroRNA-711-Induced Downregulation of Angiopoietin-1 Mediates Neuronal Cell Death. J Neurotrauma. 2018;35(20):2462-2481. https://doi.org/10.1089/neu.2017.5391

[11] Wei Peng, Zhichen He, Wei Dai, Wei Lan, MHCLMDA: multihypergraph contrastive learning for miRNA–disease association prediction, Briefings in Bioinformatics, Volume 25, Issue 1, January 2024, bbad524, https://doi.org/10.1093/bib/bbad524

[12] Nelson, P. T., Baldwin, D. A., Kloosterman, W. P., Kauppinen, S., Plasterk, R. H., & Mourelatos, Z. (2006). RAKE and LNA-ISH reveal microRNA expression and localization in archival human brain. RNA

(New York, N.Y.), 12(2), 187–191. https://doi.org/10.1261/rna.2258506

[13] Tian, Z., Han, C., Xu, L., Teng, Z., & Song, W. (2024). MGCNSS: miRNA-disease association prediction with multi-layer graph convolution and distance-based negative sample selection strategy. Briefings in bioinformatics, 25(3), bbae168. https://doi.org/10.1093/bib/bbae168

[14] Yan, C., Wang, J., Ni, P., Lan, W., Wu, F. X., & Pan, Y. (2019). DNRLMF-MDA:Predicting microRNA-Disease Associations Based on Similarities of microRNAs and Diseases. IEEE/ACM transactions on computational biology and bioinformatics, 16(1), 233–243. https://doi.org/10.1109/TCBB.2017.2776101

[15] Xing Chen, Di Xie, Lei Wang, Qi Zhao, Zhu-Hong You, Hongsheng Liu, BNPMDA: Bipartite Network Projection for MiRNA–Disease Association prediction, Bioinformatics, Volume 34, Issue 18, September 2018, Pages 3178–3186, https://doi.org/10.1093/bioinformatics/bty333

[16] Yu, S. P., Liang, C., Xiao, Q., Li, G. H., Ding, P. J., & Luo, J. W. (2019). MCLPMDA: A novel method for miRNA-disease association prediction based on matrix completion and label propagation. Journal of cellular and molecular medicine, 23(2), 1427–1438. https://doi.org/10.1111/jcmm.14048

[17] Yujun Yang, Jianping Li and Yimei Yang, "The research of the fast SVM classifier method," 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 2015, pp. 121-124, https://doi.org/10.1109/ICCWAMTIP.2015.7493959

[18] Chen, X., Zhu, C. C., & Yin, J. (2019). Ensemble of decision tree reveals potential miRNA-disease associations. PLoS computational biology, 15(7), e1007209. https://doi.org/10.1371/journal.pcbi.1007209

[19] Xuan, P.; Sun, H.; Wang, X.; Zhang, T.; Pan, S. Inferring the Disease-Associated miRNAs Based on Network Representation Learning and Convolutional Neural Networks. Int. J. Mol. Sci. 2019, 20, 3648. https://doi.org/10.3390/ijms20153648

[20] F. B. Mahmud, M. M. S. Rayhan, M. H. Shuvo, I. Sadia and M. K. Morol, "A comparative analysis of Graph Neural Networks and commonly used machine learning algorithms on fake news detection," 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 2022, pp. 97-102, https://doi.org/10.1109/CDMA54072.2022.00021

[21] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In Proceedings of the 27th ACM International Conference on Multimedia (MM '19). Association for Computing Machinery, New York, NY, USA, 1437–1445. https://doi.org/10.1145/3343031.3351034

[22] Wengang Wang, Hailin Chen, Predicting miRNA-disease associations based on graph attention networks and dual Laplacian regularized least squares, Briefings in Bioinformatics, Volume 23, Issue 5, September 2022, bbac292, https://doi.org/10.1093/bib/bbac292

[23] Ning, Q., Zhao, Y., Gao, J., Chen, C., Li, X., Li, T., & Yin, M. (2023). AMHMDA: attention aware multi-view similarity networks

and hypergraph learning for miRNA-disease associations identification. Briefings in bioinformatics, 24(2), bbad094. https://doi.org/10.1093/bib/bbad094

[24] Sarah W. Burge, Jennifer Daub, Ruth Eberhardt, John Tate, Lars Barquist, Eric P. Nawrocki, Sean R. Eddy, Paul P. Gardner, Alex Bateman, Rfam 11.0: 10 years of RNA families, Nucleic Acids Research, Volume 41, Issue D1, 1 January 2013, Pages D226–D232, https://doi.org/10.1093/nar/gks1005

[25] Lee, Y., Jeon, K., Lee, J. T., Kim, S., & Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. The EMBO journal, 21(17), 4663–4670. https://doi.org/10.1093/emboj/cdf476

[26] Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., & Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. Genome research, 14(10A), 1902–1910. https://doi.org/10.1101/gr.2722704

[27] Dong, H., Lei, J., Ding, L., Wen, Y., Ju, H., and Zhang, X. (2013). "MicroRNA: Function, Detection, and Bioanalysis", Chemical Reviews, 113(8), 6207-6233. https://doi.org/10.1021/cr300362f

[28] Du, T., Zamore, P. Beginning to understand microRNA function. Cell Res 17, 661–663 (2007). https://doi.org/10.1038/cr.2007.67

[29] Greve TS, Judson RL, Blelloch R. microRNA control of mouse and human pluripotent stem cell behavior. Annual Review of Cell and Developmental Biology. 2013 ;29:213-239. https://doi.org/10.1146/annurev-cellbio-101512-122343

[30] Lee, Y. S., & Dutta, A. (2009). MicroRNAs in cancer. Annual review of pathology, 4, 199–227. https://doi.org/10.1146/annurev.pathol.4.110807.092222

[31] Otmani, K., & Lewalle, P. (2021). Tumor Suppressor miRNA in Cancer Cells and the Tumor Microenvironment: Mechanism of Deregulation and Clinical Implications. Frontiers in oncology, 11, 708765. https://doi.org/10.3389/fonc.2021.708765

[32] Komatsu, S., Kitai, H., and Suzuki, H. I. (2023). "Network Regulation of microRNA Biogenesis and Target Interaction", Cells, 12, 306. https://doi.org/10.3390/cells12020306

[33] Guan, Z., Yao, Z., Miao, L., Hu, Y., and Wu, J. (2022). "miRClassify: An advanced web server for microRNA classification and summarization", Computers in Biology and Medicine, 145, 105645. https://doi.org/10.1016/j.compbiomed.2022.105645

[34] Antonov, A., and Carbon, M. (2011). "Large-scale chromosomal mapping of microRNA structural clusters", Nucleic Acids Research, 39(8), 3392-4403. https://doi.org/10.1093/nar/gkq1182

[35] Yang, C., Wang, Z., Dai, K., Wang, Y., Wang, K., Liang, S., Peng, S., Yu, J., and Qing, W. (2021). "MDA-GCNFTG: Identification of microRNA-disease associations based on graph convolutional networks with feature", Briefings in Bioinformatics, 22(6), bbab222. https://doi.org/10.1093/bib/bbab222

[36] Zeng, Y., & Tang, J. (2021). "RLC-GNN: An Improved Deep Architecture for Spatial-Based Graph Neural Network with Application to Fraud Detection", Applied Sciences, 11, 5656. https://doi.org/10.3390/app11125656

[37] Shang, J., Jiang, J., and Sun, Y. (2021). "Bacteriophage Classification for Assembled Contigs Using Graph Convolutional Network", Bacteriophage Classification for Assembled Contigs Using Graph Convolutional Network.

[38] Bharati, S., Sunil, P., and Kalish Gupta, K. (2022). "Graph neural networks: Concepts, methods, challenges, datasets, applications and future directions", Journal of Big Data, 9:18. https://doi.org/10.1186/s40537-022-00573-8

[39] Huang, Z., Feng, Y., Cui, C., and Wu, Y. (2022). "HMDD v4.0: a database for human microRNA-disease associations with experimental support", Nucleic Acids Research, 50(D1), D1373-D1379. https://doi.org/10.1093/nar/gkab1019

[40] Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). "miRBase: from miRNA sequences to function", Nucleic Acids Research, 47(D1), D155-D162. https://doi.org/10.1093/nar/gky1141

[41] Liu, W., Wang, D.-S., Chen, Y., Shang, L., Chen, S., Li, M., Huang, G., and Wei, W. (2022). "miRTarBase 2022 update: an informative resource for experimentally validated miRNA-target interactions", Nucleic Acids Research, 50, D222–D230. https://doi.org/10.1093/nar/gkab1079

[42] Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., Liu, Y. (2009). "miR2Disease: a manually curated database for microRNA deregulation in human disease", Nucleic Acids Research, 37(Database issue), D98-D104. https://doi.org/10.1093/nar/gkn714

[43] Xie, B., Ding, Q., Han, H., Wu, D. (2013). "miRCancer: a microRNA-cancer association database constructed by text mining on literature", Bioinformatics, 29(5), 638-644. https://doi.org/10.1093/bioinformatics/btt014

[44] Xu, F., Wang, Y., Ling, Y., Xu, C., Wang, H., Tschendorff, E. A., Zhao, Y., Zhao, H., He, Y., Zhang, G., Yang, Z. (2022). "dbDEMC 3.0: Functional exploration of differentially expressed miRNAs in human cancers and model organisms", Genomics, Proteomics & Bioinformatics, 20(3), 446-454. https://doi.org/10.1016/j.gpb.2022.01.005

[45] Needleman, Saul B. and Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology, 48(3), 443-453.

[46] Levenshtein, Vladimir I. (1965). "Binary codes capable of correcting deletions, insertions, and reversals", Soviet Physics Doklady, 10, 707-710. https://doi.org/10.1016/0041-5553(65)90818-2

[47] Wang, D., Wang, J., Lu, M., Song, F. and Cui, Q. (2010). "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases", Bioinformatics, 26(13), 1644-1650. https://doi.org/10.1093/bioinformatics/btq241

[48] Zeng, X., X, W. et al. (2021). "Integrated approaches for predicting microRNA function and prioritizing disease-related microRNA using biological information", Briefings in Bioinformatics, 17, 193-203.

[49] Dai, Q., Chu, Y., Li, Z., Xiong, Y., Wei, D-Q. (2022). "MDA-GF: Predicting miRNA-disease associations based on gradient boosting

framework by integrating multi-source information", Computers in Biology and Medicine, 136, 104706. https://doi.org/10.1016/j.compbiomed.2021.104706

[50] Li, M. et al. (2022). "Using GKSNP Feature-Based Sequence Similarity and Graph Neural Network Model to Identify microRNA-Disease Associations", Genes, 13, 1759. https://doi.org/10.3390/genes13101759

[51] Wang, J. et al. (2021). "NMCMDA: Neural multicategory microRNA-disease association prediction", Briefings in Bioinformatics, 22(5), September 2021. https://doi.org/10.1093/bib/bbab074

[52] Sun, F. et al. (2022). "A deep learning method for predicting metabolite-disease associations using graph neural network", Briefings in Bioinformatics, 23(4), July 2022, bbac266. https://doi.org/10.1093/bib/bbac266

[53] Li, Z. et al. (2021). "A graph auto-encoder model for miRNA-disease association prediction", Briefings in Bioinformatics, 22(4), July 2021, bbaa240. https://doi.org/10.1093/bib/bbaa240

[54] Xiao, Q. et al. (2018). "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations", Bioinformatics, 34(2), 239-248. https://doi.org/10.1093/bioinformatics/btx545

[55] Ji, C., Wang, Y., Chen, Y., Zheng, C. and Su, Y. (2022). "Predicting miRNA-Disease Associations Based on Heterogeneous Graph Attention Networks", Frontiers in Genetics, 12:727744. https://doi.org/10.3389/fgene.2021.727744

[56] Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., Liu, Y., Dai, Q., Li, J., Teng, Z., Huang, Y. (2013). "Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors", PLoS One, 8, e70204. https://doi.org/10.1371/journal.pone.0070204

[57] Hu, Z., Bian, Z., Zheng, Y., Peng, E. (2022). "Prediction of miRNA-Disease Associations by Gradient Boosting Model Based on Stacked Auto-Encoder", Molecules, 28, 5013. https://doi.org/10.3390/molecules28135013

[58] van Laarhoven, T., Nabuurs, S.B., Marchiori, E. (2011). "Gaussian interaction profile kernels for predicting drug-target interaction", Bioinformatics, 27(21), 3036-3043. https://doi.org/10.1093/bioinformatics/btr500

[59] Tang, X., Luo, J., Shen, S., Lai, Z. (2021). "Multi-view multichannel attention graph convolutional network for miRNA-disease association prediction", Briefings in Bioinformatics, 22(6), November 2021. https://doi.org/10.1093/bib/bbab174

[60] Chiu, X., Luo, J., Yang, J., Cai, J. and Ding, P. (2018). "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations", Bioinformatics, 34(2), January 2018, 239-248. https://doi.org/10.1093/bioinformatics/btx545

[61] Su, R., Wang, X., Jin, C., Sun, Q., Jiang, Y. (2019). "An integrated framework for identifying potential miRNA-disease associations based on novel negative sample extraction strategy", RNA Biology, 16, 257-269. https://doi.org/10.1080/15476286.2019.1567215

[62] Chen, X., Yan, G-Y. (2013). "Novel human lncRNA-disease association inference based on lncRNA expression profiles", Bioinformatics, 29(20), 2617-2624. https://doi.org/10.1093/bioinformatics/btt426

[63] Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K. L., Brown, D., Slack, F. J. (2005). "RAS is regulated by the let-7 microRNA family", Cell, 120(5), 635-647. https://doi.org/10.1016/j.cell.2005.01.014

[64] Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F., Croce, C. M. (2002). "Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia", Proceedings of the National Academy of Sciences, 99(24), 15524-15529. https://doi.org/10.1073/pnas.242606799

[65] Muñoz, J. P., Pérez-Moreno, P., Pérez, Y., Calaf, G. M. (2023). "The Role of MicroRNAs in Breast Cancer and the Challenges of Their Clinical Application", Diagnostics, 13(19), 3072. https://doi.org/10.3390/diagnostics13193072

Abstract

Numerous pieces of evidence have indicated that microRNA (miRNA) plays a crucial role in a series of significant biological processes and is closely related to complex diseases. Discovering the associations between miRNAs and diseases has become an essential part of disease discovery and treatment. However, uncovering these associations via traditional experimental methods is complicated and time-consuming. Moreover, miRNAs whose expression is restricted to nonabundant cell types or specific environmental conditions could still be missed in cloning efforts. Therefore, computational approaches have been developed to complement experimental methods.

Inspired by the substantial achievements of graph neural networks (GNNs) in miRNA-disease association (MDA) prediction, we show that a simple GNN variant, specifically the Graph Convolutional Network (GCN), is capable of predicting MDA with high accuracy. First, we capture the similarity features of miRNAs and diseases by integrating multi-source information. We then construct a graph where each node (miRNAs or diseases) has its own feature representation. Finally, GCN is combined with a multi-layer perceptron to complete the final prediction. The experimental results show that under 5-fold cross-validation the model is able to achieve AUC and AUPR values of 94.97±0.15 and 94.96±0.18, respectively. Additionally, case studies conducted on lung neoplasms, breast neoplasms, and leukemia, further demonstrate that the method presented in this paper can be a useful approach to predict disease-miRNA associations.

Keywords: disease, miRNA, miRNA-disease association, graph neural networks, similarity measures

University of Tehran
College of Science
School of Mathematics, Statistics, and Computer Science

miRNA-Disease Associations Prediction Using Graph Neural Networks

Mahroo Hajimehdi

Supervisor: Dr. Bagher BabaAli

A thesis submitted in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science

July 2024