A Model for Detection of brain cancer sub-types based on deep random forest and augmented features using genomic data

Fahimeh Fallah, Fatemeh Zamani*

Electrical and Computer Engineering Department, Nooshirvani University of Technology, Babol, Iran
E-mails: zamani@nit.ac.ir; fahimefallah@nit.ac.ir

Short Abstract
Diagnosing the type of cancer, which is called the subtype, is very important in determining the treatment process. This paper focuses on the diagnose of the four subtypes of the brain cancer. Disease subtype diagnosis can be modeled as a classification problem. Due to the significant progress made in bioinformatics in extracting genetic information from the human body, recently this information is widely used in the representing of patients in machine learning. In this paper, three types of genetic information including mRNA, miRNA and DNA methylation are used.
It should be noted that combining different information sources in the form of multimodal data instead of using a single information source increases the accuracy of information classification. To extract more desirable features from the original genetic data, auto-encoder has been used so that the features extracted from auto-encoder are concatenated to the original genetic data.
Random forest has performed well as a classifier in classifying patients based on genetic information. By extending deep methods in neural networks and their good performance, a version of deep random forest with layered structure has been proposed. The deep random forest has the advantage that has a limited number of parameters and lower computational complexity in addition to the optimal performance in information classification. In this paper, deep random forest is used to determine the subtype of a special type of brain cancer. The experiment results show the desired performance of the proposed method.

Keywords
Bioinformatics, classification, autoencoder, deep random forest, multi-omics data, brain cancer.

1- Short Introduction (4-5 lines)

Genomic data have made significant progress in medical activities, including cancer related ones. Cancer diagnosis, cancer subtype Detection, cancer stage prediction and cancer treatment line are examples of For cancer related tasks. GBM is a deadly brain cancer that has subtypes. Diagnosing the subtype of a cancer is effective in its treatment process. In this paper, a method based on machine learning is proposed to detect the subtype of GBM disease using genomic such that the problem of detecting the subtype of GBM is modeled as a classification problem. A classifier is designed for this purpose.

2- Proposed Work and Methodology (including comprision, simulation/experimental results and discusion)

In this paper, a model based on deep forest is presented for the classification of GBM cancer subgroups. For this purpose, three types of genomic data including gene expression, miRNA and DNA methylation have been used. By using three separate autoencoders for each of the mentioned features, with the aim of increasing the classification accuracy of subgroups detection, new features have been extracted. The accuracy of the proposed classifier is compared with the basic methods, deep forest and DFNForest as the most similar method to the proposed method. The results of the proposed method are promising.

3- Conclusion (4-5 lines)

The use of genomic data is very effective in the analysis of cancer disease. Machine learning methods have made great progress in disease diagnosis using genomic data. The recently proposed deep forest learning method has a high performance. In this paper, three genomic features including Gene Expression, miRNA and DNA methylation are used using deep forest to detect GBM cancer subgroup.Also, augmented features extracted from the three autoencoders. Some experiments have been conducted for GBM patients. The results show the effectiveness of the proposed algorithm.

4- References (2-3 references)

J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, H. Dawood, “Open Access A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data”, BMC Bioinformatics, pp. 1–11, 2019.
Z. Zhou, J. Feng, “Deep forest”, national science review, vol. 6, pp. 74–86, 2018.

ارائه مدلی برای تشخیص زیرگروه‌های سرطان مغز مبتنی بر جنگل تصادفی عمیق و ویژگی‌های تقویت شده با استفاده از داده‌های ژنی

فهیمه فلاح
کارشناس ارشد، مهندسی برق و کامپیوتر، دانشگاه صنعتی نوشیروانی، بابل، ایران

فاطمه زمانی
استادیار، مهندسی برق و کامپیوتر، دانشگاه صنعتی نوشیروانی، بابل، ایران

چکیده
تشخیص نوع بیماری سرطان که به ان زیرگروه گفته می‌شود در تعیین روند درمان حایز اهمیت فراوانی است. در این مقاله، هدف تشخیص چهار زیرگروه سرطان مغز می‌باشد. تشخیص زیرگروه بیماری را می توان در قالب یک مسئله طبقه‌بندی مدل کرد. با توجه به پیشرفت‌های چشمگیر صورت گرفته در علم بیوانفورماتیک در استخراج اطلاعات ژنی از بدن انسان، اخیرا از این اطلاعات در توصیف بیماران در یادگیری ماشین استفاده زیادی می‌شود. در این مقاله از سه نوع داده ژنی شامل mRNA، miRNA و متیلاسیون DNA استفاده شده است. ترکیب منابع مختلف اطلاعاتی در قالب داده‌های چندوجهی به جای استفاده از یک منبع اطلاعاتی واحد، به افزایش دقت طبقه‌بندی اطلاعات منجر می‌شود. برای استخراج ویژگیهای مطلوب‌تر از داده‌های ژنی، از خودرمزگذار استفاده شده است بطوریکه ویژگی های استخراج شده از خودرمزگذار، به‌عنوان تقویت کننده در کنار داده‌های ژنی اولیه قرار می‌گیرند. همچنین جنگل تصادفی به‌عنوان یک طبقه‌بندی کننده در طبقه‌بندی بیماران بر مبنای داده‌های ژنی عملکرد مطلوبی داشته است. با گسترش روش‌های عمیق در شبکه‌های عصبی و عملکرد مطلوب آنها، نسخه‌ای از جنگل تصادفی عمیق با ساختار لایه‌ای ارائه شده است. جنگل تصادفی عمیق دارای این مزیت است که در کنار عملکرد مطلوب در طبقه‌بندی اطلاعات، تعداد پارامتر محدودی داشته و پیچیدگی محاسباتی آن پایین‌تر است. در این مقاله از جنگل تصادفی عمیق برای تعیین زیرگروه نوعی از سرطان مغز استفاده شده است. نتایج آزمایش‌ها نشان‌دهنده عملکرد مطلوب روش پیشنهادی است.

کلمات کلیدی
بیوانفورماتیک، طبقه بندی اطلاعات، خودرمزگذار، جنگل تصادفی عمیق، داده‌های چندوجهی، سرطان مغز.

نام نویسنده مسئول: فاطمه زمانی
ایمیل نویسنده مسئول: zamani@nit.ac.ir

تاریخ ارسال مقاله: ۱۴۰۲/۰۳/۰۷
تاریخ(های) اصلاح مقاله: ۱۴۰۲/۰۶/۰۵
تاریخ پذیرش مقاله: ۱۴۰۲/۰۸/۰۶

۱. مقدمه

در سال‌های اخیر با پیشرفتهای صورت گرفته در تجهیزات جمع‌آوری اطلاعات ژن‌ها حجم زیادی از داده‌های ژنی را در دسترس داریم. بیوانفورماتیک دانشی بین رشته‌ای است که سعی می‌کند با استفاده از روش های موجود در علوم کامپیوتر و ریاضیات داده‌های زیستی، داده‌های ژنی را تجزیه و تحلیل کرده و نتایج آن را در کاربردهایی مانند تشخیص بیماری، تشخیص میزان پیشرفت بیماری، تشخیص ژن‌های تاثیرگذار در بیماری و طراحی دارو مورد استفاده قرار دهد[۱]. در دهه گذشته، یادگیری ماشین در پیشرفت چشمگیر تحقیقات بیوانفورماتیک با استفاده از داده‌های ژنی بسیار موثر بوده است [۲].

سرطان به‌عنوان یکی از مهلک‌ترین بیماری‌ها همواره مورد توجه محققین بوده است. هر بیماری سرطان مربوط به یک عضو از بدن انسان، اغلب دارای انواع مختلفی است که به آن زیرگروه گفته می‌شود. تشخیص زیرگروه سرطان در انتخاب روند درمان بسیارحائز اهمیت است [۳]. می‌توان مسئله تشخیص زیرگروه سرطان را در قالب یک مسئله طبقه‌بندی در یادگیری ماشین مدل کرد.

بیماری GBM[1] یک نوع سرطان مغز بسیار تهاجمی با میانگین بقای بسیار کم است که دارای چهارزیر گروه بیماری می‌باشد. تحقیقات زیادی برای بررسی و استخراج اطلاعات کاربردی در رابطه با این بیماری شامل تشخیص، تعیین زیرگروه بیماری و میزان پیشرفت بیماری، با استفاده از روش‌های یادگیری ماشین انجام شده است [۸-۴]. در این تحقیقات اغلب از تصاویر MRI[2] و اطلاعات کلینیکی بیماران استفاده شده است. در این مقاله، هدف تشخیص زیرگروه بیماری در این نوع سرطان می‌باشد.

درخت تصمیم[3] و جنگل تصادفی[4] که در واقع متشکل از تعدادی درخت تصمیم است، دو طبقه‌بند مطرح در یادگیری ماشین هستند که در بسیاری از کاربردهای بیوانفورماتیک از جمله شناسایی بافت‌های بیماری، تعیین زیرگروه بیماری، تشخیص مرحله بیماری و پیش بینی زمان بقای بیماران عملکرد خوبی

[1] GlioBlastoma Multiforme
[2] Magnetic Resonance Imaging
[3] Decision Tree
[4] Random Forest

داشته اند (دربخشهای ۱-۲ و ۲-۲ به تعدادی از تحقیقات مرتبط که از درخت تصمیم و جنگل تصادفی استفاده کرده‌اند، اشاره شده است). از طرفی با مطرح شدن ایده عمیق سازی شبکه عصبی، پیشرفت‌های چشمگیری در کاربردهایی مانند طبقه‌بندی اطلاعات با استفاده از آنها حاصل شده است به‌طوریکه دقت طبقه‌بندی اطلاعات با استفاده از شبکه عصبی عمیق به طور قابل توجهی از سایر الگوریتم‌های یادگیری ماشین بالاتر است. در این راستا یک الگوریتم جنگل تصادفی عمیق مطرح شده است که همانند شبکه عصبی عمیق، یک ساختار لایه لایه در جنگل تصادفی ایجاد کرده و پیشرفت قابل توجهی در طبقه‌بندی اطلاعات را شامل می‌شود [۹]. جنگل تصادفی عمیق در مسائلی مانند طبقه‌بندی تصاویر سنجش از دور [۱۰]، پیش‌بینی تعامل‌های پروتئین-پروتئین[۱۱]، تشخیص بیماری کرونا با استفاده از تصاویر سی‌تی اسکن ریه[۱۲] و طبقه‌بندی سیگنال‌های EEG[5] [۱۳] عملکرد موفقی داشته است.

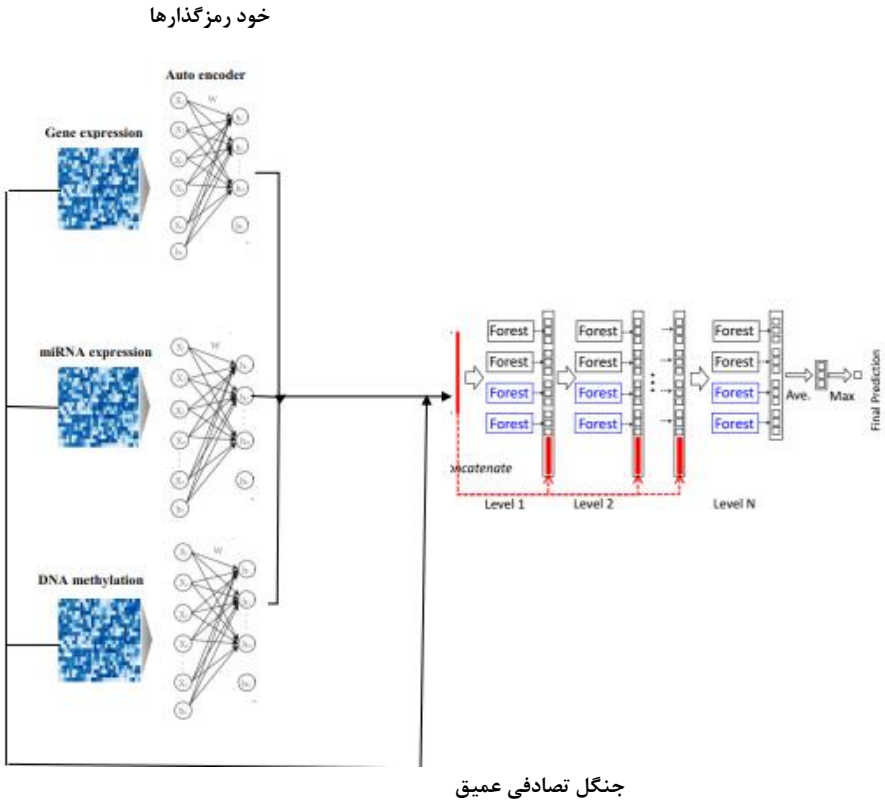
جنگل تصادفی عمیق علاوه بر اینکه کارایی بالایی در طبقه‌بندی اطلاعات دارد نسبت به روش‌های مطرح دیگر مانند شبکه عصبی عمیق دارای دو مزیت زیر است [۱۴]:

- تعداد پارامترهای کمتر
- پیچیدگی محاسباتی پایین‌تر

همان‌طور که در ابتدای این بخش ذکر شد، استفاده از اطلاعات ژنی با توجه به پیشرفت‌های صورت گرفته در استخراج این اطلاعات به صورت گسترده و با دقت بالا اخیراً بسیار مورد توجه قرار گرفته است. در مدل‌های آموزش داده شده با روش‌های یادگیری ماشین، استفاده از داده‌های چندوجهی[6] به صورت هم‌زمان موجب بالا رفتن دقت مدل می‌شود. در این راستا مدل‌هایی با کاربردهای پزشکی مرتبط با سرطان طراحی شده‌اند که از چندین داده ژنی به‌صورت همزمان استفاده کرده‌اند. در [۱۵] از ترکیب اطلاعات ژنی برای تشخیص زیرگروه بیماری، در [۱۶] برای بررسی احتمال زنده ماندن بیمار و در [۱۷] برای پیش‌بینی میزان پیشرفت بیماری مدل‌های یادگیری ماشین بر اساس داده‌های چندوجهی طراحی شده‌اند.

در این مقاله از سه داده ژنی mRNA[7]، miRNA[8] و متیلاسیونDNA[9] مربوط به بیماران مبتلا به بیماری GBM به صورت چندوجهی برای تشخیص زیرگروه بیماری با استفاده از جنگل تصادفی عمیق استفاده شده است. لازم به ذکر است این داده‌های ژنی در اغلب تحقیقات انجام شده مورد استفاده قرار گرفته‌اند [۱۷-۱۵].

به‌منظور بالابردن عملکرد طبقه‌بندی کننده، جنگل تصادفی عمیق، بهتر است داده‌های ژنی مذکور را پردازش کرده و ویژگی‌هایی از آن‌ها استخراج کرد که به بهبود عملکرد طبقه‌بندی کننده کمک کند. در این مقاله از خودرمزگذار[10] برای این منظور استفاده شده است [۱۸]. خودرمزگذار یک روش یادگیری بدون نظارت است که برای استخراج ویژگی از داده‌های ژنی مورد استفاده قرار گرفته‌است [۲۰-۱۹]. در روش پیشنهادی از تعدادی خودرمزگذار برای تقویت و ترکیب ویژگی‌های مذکور استفاده شده است. داده‌های ژنی اولیه به همراه داده‌های حاصل از پردازش خودرمزگذارها، به‌عنوان ورودی به جنگل تصادفی عمیق به منظور تعیین زیرگروه بیماری بیماران داده می‌شوند. در شکل ۱ بلوک دیاگرام روش پیشنهادی مشاهده می‌شود.



شکل ۱- بلوک دیاگرام روش پیشنهادی

[5] Electro Encephalo Gram
[6] Multi Omics
[7] Messenger RiboNucleic Acid
[8] Micro Messenger RiboNucleic Acid
[9] Deoxyribonucleic acid
[10] AutoEncoder

در ادامه در بخش دوم مروری بر پژوهش‌های پیشین در رابطه با روش های مبتنی بر الگوریتم‌های درخت تصمیم، جنگل تصادفی و جنگل تصادفی عمیق در کاربردهای بیوانفورماتیک اشاره خواهد شد. در بخش سوم به معرفی برخی از مفاهیم اولیه پرداخته می‌شود. سپس در بخش چهارم الگوریتم پیشنهادی معرفی شده و در بخش پنجم آزمایش‌هایی که برای ارزیابی الگوریتم انجام شده‌اند توضیح داده خواهد شد. در نهایت مقاله در بخش ششم نتیجه گیری خواهد شد.

۲- مرور پژوهش‌های پیشین

درخت تصمیم یکی از طبقه‌بندی کننده‌های مطرح در یادگیری ماشین می‌باشد که در مسائل مختلفی از آن استفاده شده است[۲۱]. جنگل تصادفی متشکل از تعداد زیادی درخت تصمیم است که به منظور بهبود عملکرد درخت تصمیم پیشنهاد شد و در مسائل مختلفی از آن استفاده شد [۲۳-۲۲]. جنگل تصادفی عمیق، یک جنگل تصادفی با تعداد لایه‌های زیاد می‌باشد (جنگل تصادفی عمیق در بخش ۲-۳ معرفی خواهد شد). در این بخش به مروری بر برخی از مطالعات انجام شده در زمینه روشهای مبتنی بر درخت تصمیم، جنگل تصادفی و جنگل تصادفی عمیق در زمینه‌های مختلف بیوانفورماتیک از جمله پیش بینی زیر گروهها و مراحل پیشرفت سرطانهای مختلف می‌پردازیم.

۲-۱ مرور روش های مبتنی بر درخت تصمیم در پیش بینی بیماری

شرافتیان در سال ۲۰۱۸ برای تشخیص miRNAهای نشانگر بیماری سرطان پستان از درخت تصادفی استفاده کردند[۳]. انصاری وهمکارش در سال ۲۰۱۹ از الگوریتم‌های مختلف درخت تصمیم برای شناسایی پروتئین های موثر و غیرموثر در بیماری استفاده کردند. در نهایت آنها درخت‌های تصمیم مورد استفاده را در قالب جنگل تصادفی تلفیق کردند [۲۴].

ژو و همکارش در سال ۲۰۲۰ از تلفیق درخت تصمیم گرادیان تقویتی با رگرسیون لجستیک برای تشخیص ارتباط بین miRNA ها و بیماری های سرطان روده بزرگ، سرطان معده و سرطان لوزالمعده استفاده کردند [۲۵].

۲-۲ مرور روش های مبتنی بر جنگل تصادفی در پیش بینی بیماری

فراتلو و همکارش از جنگل تصادفی برای تشخیص زیرگروه‌های سرطان پستان استفاده کردند [۲۶]. مجموعه داده مورد استفاده آنها زیر مجموعه‌های از مجموعه داده TCGA[11] بود که حالت نامتوازن داشت. جاگا و همکارش در سال ۲۰۱۴ از جنگل تصادفی برای تشخیص مرحله پیشرفت بیماری سرطان سلول شفاف کلیه بر اساس پروفایل بیان ژن استفاده کردند. آنها برای کاهش ابعاد در فضا از الگوریتم انتخاب ویژگی مبتنی بر همبستگی سریع استفاده کردند [۲۷].

رحیمی و همکارش در سال ۲۰۱۸ از جنگل تصادفی با استفاده از داده های بیان ژن برای طبقه بندی مرحله پیشرفت ۱۸ بیماری از مجموعه داده TCGA استفاده کردند [۲].

دیتما و همکاران در سال ۲۰۱۰ تأثیر و ارتباط هر یک از متغیرهای مختلف در بقای کلی برای بیماران سرطان سر و گردن را بررسی کردند [۲۸]. آنها برای این منظور از جنگل بقای تصادفی استفاده کردند که مجموعه ای از درختان است که از تجزیه و تحلیل بقای سانسور شده داده‌ها به دست می‌آید. در دهه گذشته، تلاش‌های زیادی برای شناسایی مارکرهای زیستی ژن با هدف تولید داروهای مورد تایید سازمان جهانی دارو و غذای امریکا انجام شده است. یکی از مهمترین چالش‌ها در تشخیص مارکر زیستی ژن را می‌توان به عدم تکرارپذیری نتایج حاصل از مطالعات مارکرهای زیستی نسبت داد. کیم و همکارانش در سال ۲۰۲۰ برای ارزیابی توانایی طبقه بندی و تکرارپذیری آماری ابزارهای یادگیری که برای تشخیص مارکرهای زیستی بکار می‌رود از شش مدل یادگیری ماشین در چهار گروه سرطان از مجموعه داده TCGA استفاده کردند که از بین آنها جنگل تصادفی بهترین عملکرد را داشت [۲۹]. لوپز و همکارانش در سال ۲۰۲۰ یک رویکرد مبتنی بر جنگل تصادفی نیمه تمام تکراری برای شناسایی ویژگی امضای miRNA با تجزیه و تحلیل بیان miRNA افتراقی (انجام تجزیه و تحلیل آماری برای کشف تغییرات بیان بین گروه های آزمایش) سرطان سر و گردن پرداختند[۳۰]. چنین امضایی miRNA هایی را شناسایی می‌کند که می‌توانند محرک این نوع سرطان باشند.

۲-۳ مرور روش های مبتنی بر جنگل تصادفی عمیق در پیش بینی بیماری

از جنگل تصادفی عمیق در مسائل زیادی برای طبقه‌بندی اطلاعات استفاده شده است. برای مثال فنگ و همکارانش یک روش مبتنی بر جنگل تصادفی برای تشخیص حرکت دست ارائه کردند [۳۱]. آنها از ترکیب چندین ویژگی برای این منظور استفاده کردند.

مدل استاندارد جنگل تصادفی عمیق ممکن است برای داده‌های زیست شناسی با حجم نمونه کوچک و ابعاد بالا خطر بیش برازش را در آموزش ایجاد کند. عدم تعادل طبقاتی در داده‌های زیست شناسی بسیار رایج است، که مشکلات یادگیری مدل را تشدید می‌کند. برای کاهش این چالش ها جائو و همکاران در سال ۲۰۱۷ یک مدل یادگیری عمیق نام BCDForest[12] (تقویت آبشار جنگل عمیق)، برای طبقه‌بندی زیرگروه‌های سرطان در مجموعه داده‌های زیست شناسی پیشنهاد کردند که می تواند به عنوان اصلاح مدل استاندارد جنگل عمیق در نظر گرفته شود[۳۲]. BCDForest تأکید بر ویژگی‌های مهمتر را در جنگل عمیق پیشنهاد می‌کند به این ترتیب که k ویژگی مهمتر را از هر کدام از جنگل‌ها انتخاب کرده و از انحراف استاندارد این ویژگی ها برای یک خصوصیت جدید استفاده می‌شود. سپس این ویژگی‌های جدید با بردارهای توزیع کلاس خروجی مربوط به آنها ترکیب می‌شود. بردارهای توزیع کلاس تقویت شده در نهایت، در هر لایه با بردار ویژگی اصلی به عنوان ورودی لایه بعدی آبشار به هم متصل می‌شود. آنها روش خود را برای تشخیص زیرگروه بیماری مربوط به چهار بیماری از مجموعه داده TCGA مورد استفاده قرار دادند.

خو و همکارانش در سال ۲۰۱۹ رویکرد مبتنی بر جنگل عمیق آبشاری را پیشنهاد دادند که در هر لایه از جنگل های عمیق انعطاف پذیر استفاده می‌کند. آنها از روش خود برای طبقه‌بندی زیر گروه سرطان در چهار بیماری از مجموعه داده TCGA استفاده کردند [۳۳].

در سال ۲۰۱۹، خو و همکاران یک چارچوب جنگل عصبی عمیق(مبتنی بر شبکه عصبی) انعطاف پذیر با نام DFNForest[13] را پیشنهاد دادند که از ادغام داده‌های ژنی برای تشخیص سه بیماری از مجموعه داده TCGA استفاده می‌کند. آنها از پشته‌ای از خودرمزگذارها برای استخراج ویژگی‌های مهم استفاده کردند [۳۴].

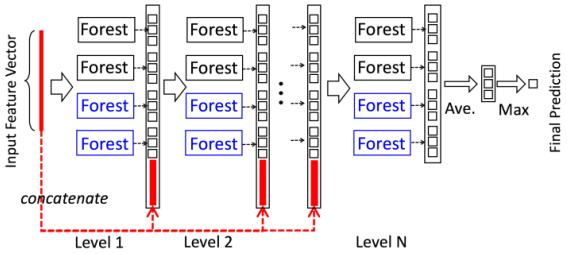
۳- مفاهیم پایه

در این بخش به معرفی مفاهیم پایه به کار رفته در طراحی روش پیشنهادی می‌پردازیم.

۳-۱ جنگل تصادفی عمیق

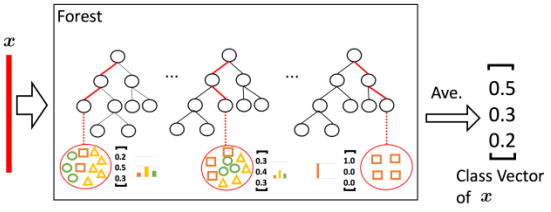
[11] The Cancer Genome Atlas
[12] Boosting Cascade Deep Forest
[13] Deep Flexible Neural Forest

جنگل تصادفی عمیق بر پایه درخت تصمیم طراحی شده که از ساختار لایه‌ای شبکه عصبی عمیق استفاده می‌کند بطوریکه هر سطح اطلاعات خروجی سطح قبل را دریافت کرده نتیجه پردازش آنها به سطح بعدی به‌عنوان ورودی می‌فرستد [۹]. هر لایه از تعدادی جنگل تصادفی تشکیل شده است. برای ایجاد تنوع از و بالاتر بردن کارایی جنگل تصادفی عمیق در طبقه‌بندی اطلاعات از انواع مختلف جنگل‌ها استفاده می‌شود. برای مثال می‌توان از جنگل تصادفی معمولی و جنگل درختی کاملا تصادفی استفاده کرد[۱۵]. هر جنگل درخت کاملاً تصادفی شامل تعدادی درخت کاملاً تصادفی است که با انتخاب تصادفی یک ویژگی در هر گره درخت ایجاد می‌شود. هر درخت رشد میکند تا برگها ایجاد شوند. هر جنگل تصادفی نیز شامل تعدای درخت تصمیم است که در هر درخت تصمیم ویژگیهایی با بالاترین مقدار شاخص جینی انتخاب می‌شوند. شکل۲ ساختار جنگل تصادفی عمیق توضیح داده شده را نشان می‌دهد.



شکل ۲- ساختار جنگل تصادفی عمیق [۱۶]. همانطور که مشاهده می‌شود هر لایه متشکل از دو جنگل تصادفی با درختهای کاملا تصادفی (که با رنگ سیاه نشان داده شده‌اند) و دو جنگل تصادفی معمولی (که با رنگ آبی نشان داده شده‌اند) می‌باشند.

به ازای هر جنگل تصادفی، بردار توزیع هر کلاس محاسبه می‌شود. برای این منظور درصد تعلق داده‌های آموزشی به هر کلاس که توسط درخت‌های آن جنگل تصادفی طبقه‌بندی می‌شوند، محاسبه می‌شود. در نهایت در صد تعلق نمونه‌های آموزشی به کلاسهای مختلف در جنگل مذکور به صورت میانگین درصدهای محاسبه شده در درختان آن جنگل محاسبه می‌شود. در نهایت بردار توزیع کلاس برآورد شده به ازای هر جنگل با بردار ویژگی اصلی برای ورود به لایه بعدی جنگل تصادفی عمیق به هم متصل می‌شوند. برای مثال فرض کنید داده‌ها به سه کلاس تعلق دارند و چهار جنگل در لایه مورد بحث وجود دارد. لایه بعدی، یک بردار ۱۲ تایی از توزیع کلاس‌ها به‌عنوان ویژگی‌های تقویت شده (علاوه بر بردار ویژگی‌های اولیه) دریافت خواهد کرد (شکل ۳). لازم به ذکر است تعداد سطوح آبشار به طور خودکار تعیین می شود.



شکل ۳- در این جنگل تصادفی، فرض شده است داده‌ها به سه کلاس دایره، مربع و مثلث تعلق دارند. همانطور که مشاهده می‌شود بردارهای توزیع هر کلاس برای هر درخت محاسبه شده سپس بردار توزیع کلاس جنگل با میانگین‌گیری بردارهای مربوط به سه درخت محاسبه شده است.

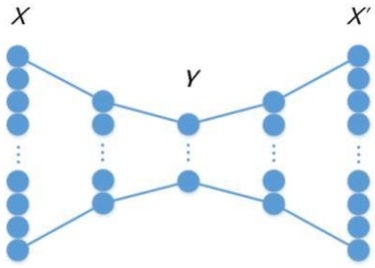
۲-۳ خودرمزگذار

خودرمزگذار، یک روش یادگیری بدون نظارت می‌باشد که در آن از شبکه عصبی برای استخراج ویژگی و کاهش ابعاد استفاده می‌شود [۱۸]. خودرمزگذار از دو بخش، یک رمزگذار و یک رمزگشا تشکیل شده است. مجموعه داده X را با n نمونه و m ویژگی را به‌عنوان ورودی رمزگذار در نظر بگیرید. خروجی رمزگذار را ماتریس Y می‌نامیم که به‌عنوان ورودی به رمزگشا داده می‌شود. اگر خروجی رمزگشا X' باشد که ابعادی مشابه با ابعاد ماتریس X دارد، Y طوری محاسبه می‌شود که تفاوت بین X و X' حداقل باشد. در واقع می‌توان گفت X' بازسازی شده ماتریس X است. ساختار خودرمزگذار در شکل ۴ نشان داده شده است[۱۴]. لازم به ذکر است در ساده‌ترین حالت خودرمزگذار تنها یک لایه پنهان دارد. اگر خودرمزگذار را با یک لایه پنهان در نظر بگیریم، رمزگذار با تابع فعال‌ساز زیر که در حالت کلی می‌تواند غیر خطی باشد، Y را از داده‌های ورودی X محاسبه می‌کند:

$$Y = f(W X + b_X) \qquad (۱)$$

همچنین تابع رمزگشا با تابع فعالساز زیر که در حالت کلی می‌تواند غیرخطی باشد، X' را از روی Y محاسبه می‌کند:

$$X` = g(W`Y + b_Y) \qquad (۲)$$



شکل ۴- ساختار خودرمزگذار در این شکل نمایش داده شده است. X داده‌های ورودی، Y داده‌های میانی با ابعاد کاهش یافته و X' ویژگی‌های استخراج شده با ابعاد مشابه با ابعاد داده های ورودی X می‌باشد [۱۵].

برای آموزش خودرمزگذار باید پارامترهای $\theta = (W, b_X, W', b_Y)$ را طوری محاسبه کند که خطای بازسازی X' از روی X حداقل شود:

$$\Theta = \min_\theta \text{Loss}(X, X`) = \min_\theta \text{Loss}\left(X, g(f(X))\right) \qquad (۳)$$

می‌توان دو حالت برای محاسبه خطای بازسازی در نظر گرفت. حالت خطی که با حداقل سازی مربع خطا پارامترها محاسبه می‌شوند:

$$\text{Loss}_1(\theta) = \sum_{i=1}^{n} ||x_i - x`_i||^2 = \sum_{i=1}^{n} ||x_i - g(f(x_i))||^2 \qquad (۴)$$

و حالت غیرخطی که از آنتروپی متقابل استفاده می‌کند:

$$\text{Loss}_2(\theta) = -\sum_{i=1}^{n} [x_i \log(y_i) + (1 - x_i)\log(1 - y_i)] \qquad (۵)$$

۴- روش پیشنهادی

در این مقاله الگوریتمی برای پیش‌بینی زیرگروه بیماری مبتلایان به بیماری GBM که نوعی سرطان مغز است، پیشنهاد شده است. این بیماری دارای چهار زیرگروه است. بنابراین می‌توان مسئله پیش بینی گروه بیماری GBM را یک

مسئله طبقه‌بندی با چهار کلاس درنظر گرفت.

در این مقاله از سه نوع داده ژنی شامل mRNA، miRNA و متیلاسیونDNA برای توصیف بیماران استفاده شده است که توسط طبقه‌بندی کننده مورد استفاده قرار خواهند گرفت. به منظور تقویت اطلاعات توصیف‌کننده بیماران و در نتیجه افزایش دقت مدل طبقه‌بندی کننده، از سه خودرمزگذار استفاده شده است. خودرمزگذار همانطور که در بخش ۲-۳ توضیح داده شد برای استخراج ویژگی از داده‌های اولیه به کار می‌رود. در نهایت داده‌های ژنی اولیه به همراه ویژگی‌های استخراج شده از خودرمزگذار به مدل طبقه‌بندی کننده داده خواهند شد.

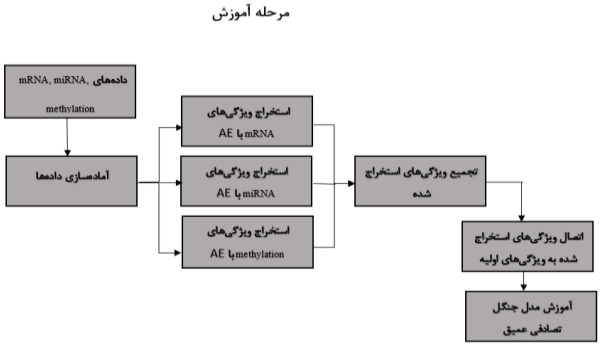
همان‌طور که در بخش ۲ بیان شد، جنگل تصادفی یکی از طبقه‌بندی کننده پرکاربرد هنگام استفاده از داده‌های ژنی می‌باشد. به موازات گسترش شبکه‌های عصبی عمیق که دارای ساختار لایه‌ای با تعداد لایه‌های بالا می‌باشد، جنگل تصادفی عمیق با ساختار لایه‌ای پیشنهاد شد که علاوه بر عملکرد مطلوب در طبقه‌بندی داده‌ها، از جمله داده‌های ژنی(بخش ۳-۲)، دارای دو مزیت نسبت به دیگر روش‌های عمیق می‌باشد: تعداد پارامترهای کمتر، پیچیدگی محاسباتی پایین‌تر. بنابراین در این مقاله از جنگل تصادفی عمیق به عنوان طبقه‌بندی کننده استفاده شد.

در ادامه به بیان جزئیات مرحله‌های آموزش و آزمایش الگوریتم پیشنهادی می‌پردازیم.

۴-۱ مرحله آموزش

در مرحله آموزش، به منظور تقویت ویژگی‌های اولیه، هر یک از سه نوع داده آموزشی به خودرمزگذار سه لایه به عنوان ورودی داده شده پس از آموزش خودرمزگذارها، ویژگی‌های جدید با استفاده از لایه پنهان رمزگذارها استخراج می‌شود.

خروجی‌های حاصل از خودرمزگذارها به همراه ویژگی‌های اولیه تجمیع شده به عنوان بردار ورودی به جنگل تصادفی عمیق به منظور آموزش جنگل تصادفی عمیق ارسال می‌شود. مرحله آموزش روش پیشنهادی در شکل ۵ نشان داده شده‌است.

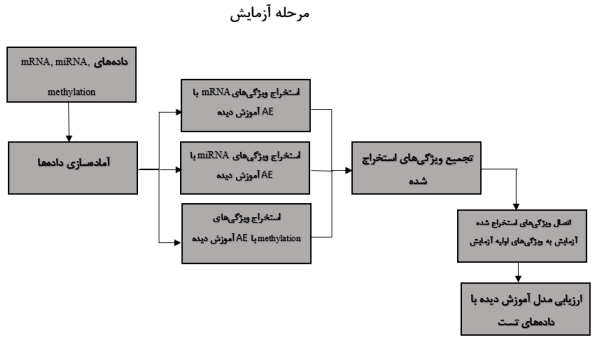


شکل ۵- فلوچارت مرحله آموزش روش پیشنهادی

در مرحله آموزش روش پیشنهادی، ابتدا سه خودرمزگذار که در شکل به صورت خلاصه با AE نشان داده شده‌اند آموزش داده می‌شوند. سپس با تجمع ویژگی‌های استخراج شده از خودرمزگذارها و داده‌های ژنی اولیه، جنگل تصادفی عمیق به صورت بانظارت با استفاده از برچسب داده‌ها آموزش داده می‌شود.

۴-۲ مرحله آزمایش

نمونه‌های آزمایشی شامل سه داده ژنی mRNA، miRNA و متیلاسیون DNA با ویژگی‌های استخراج شده از آنها با استفاده از سه خودرمزگذار آموزش داده شده در فاز آموزش تجمیع می‌شود. سپس زیرگروه بیماری هر نمونه آزمایشی توسط جنگل تصادفی عمیق آموزش داده شده تعیین می‌شود. مرحله آزمایش روش پیشنهادی در شکل ۶ نشان داده شده است.



شکل ۶- فلوچارت مرحله آموزش روش پیشنهادی

۵- نتایج آزمایش‌ها

در این بخش به تشریح آزمایش‌های انجام شده برای ارزیابی مدل پیشنهای برای تشخیص زیرگروه بیماری GBM می‌پردازیم.

۵-۱ پایگاه داده TCGA

در این مقاله از پایگاه داده TCGA به منظور انجام آزمایشها برای ارزیابی روش پیشنهادی استفاده شده است. این پایگاه داده که جمع‌آوری آن از سال ۲۰۰۶ آغاز شد، شامل داده‌های ژنی مانند mRNA، miRNA و متیلاسیونDNA مربوط به سرطان‌های مختلف می‌باشد. این پایگاه داده شامل ۱۱۰۰۰نمونه بافتی بیماران مبتلا به سرطان می‌باشد. در مطالعات بسیاری از داده‌های TCGA برای درک فرآیندهای بیولوژیکی مرتبط با سرطان، شناسایی ژن های متفاوت برای توصیف مولکولی و ژنومی ویژگی های سرطان و در ارائه بینش در مورد درمان بالقوه سرطان را استفاده کردند [۲۹].

ابتدا پیش پردازش داده‌ها شامل حذف داده‌های پرت، انتساب مقادیر از دست رفته و نرمال سازی انجام می شود. لازم به ذکر است مطابق روش به کار رفته در [۳۴] اگر یکی ویژگیهای ذکر شده بیش از ۲۰٪ مقادیر از دست رفته در یک بیمار داشته باشد، داده‌های این بیمار از مجموعه داده‌ها حذف می شود. در غیراینصورت برای مقداردهی داده‌های از دست رفته، از K نزدیکترین همسایه (KNN)[14] برای انتساب استفاده می شود. در نهایت لگاریتم داده‌ها محاسبه می‌شود.

برای ارزیابی روشهای پیشنهادی، از داده های مربوط به بیماری از مجموعه داده TCGA استفاده شده است. داده های مربوط به این بیماری شامل سه مجموعه داده ژنی مربوط به ۲۱۳ بیمار شامل mRNA، miRNA و متیلاسیون DNA است که به ترتیب دارای ابعاد ۱۲۰۴۲، ۵۳۴ و ۱۳۰۵ می‌باشد. GBM دارای چهار زیرگروه می‌باشد که در جدول زیر تعداد افراد مبتلا به هریک از این زیرگروه‌ها از ۲۱۳ بیمار نشان داده شده است.

[14] K Nearest Neighbour

جدول ۱- تعداد افراد مبتلا به چهار زیرگروه بیماری GBM در پایگاه داده TCGA

نام زیرگروه بیماری	تعداد افراد مبتلا
کلاس ۱	۵۶
کلاس ۲	۳۴
کلاس ۳	۵۸
کلاس ۴	۶۵

داده‌ها به نسبت ۷۰ به ۳۰ برای داده‌های آموزشی و آزمایشی درنظرگرفته شد. لازم به ذکر است در تمامی آزمایش‌های انجام شده در بخش‌های ۵-۲ و ۵-۳، خودرمزگذارها با سه‌لایه مورد استفاده قرار گرفته‌اند. تعداد نورون‌های لایه پنهان با استفاده از مجموعه داده اعتبارسنجی در مرحله آموزش تعیین شدند. و تعداد لایه‌های درخت به‌صورت خودکار در مرحله آموزش تعیین شده‌اند. همچنین نتایج گزارش شده تمام آزمایش‌ها، میانگین ۱۰ بار تکرار آن آزمایش می‌باشند.

۵-۲ ارزیابی مدل پیشنهادی با استفاده از داده‌های ژنی به صورت مجزا

در ابتدا هر یک از سه مجموعه داده به صورت مجزا (بدون ترکیب) مورد استفاده قرار گرفتند. برای این منظور دو آزمایش انجام شد.

در آزمایش اول، جنگل تصادفی عمیق با استفاده از داده‌های ژنی و بدون استفاده از ویژگی‌های استخراج شده توسط خودرمزگذار، آموزش داده شد.

در آزمایش دوم، جنگل تصادفی عمیق با استفاده از داده‌های ژنی و ویژگی‌های استخراج شده توسط خودرمزگذار، آموزش داده شد.

برای این منظور سه معیار ارزیابی مورد بررسی قرار گرفت: صحت طبقه‌بندی[15] (جدول ۲)، دقت[16] (جدول ۳) و بازخوانی[17] (جدول ۴).

جدول ۲- مقایسه صحت طبقه‌بندی روش پیشنهادی با جنگل تصادفی عمیق بدون ترکیب داده‌ها.

الگوریتم پیشنهادی بدون ترکیب داده‌ها	جنگل تصادفی عمیق [۹]	داده ژنی مورد استفاده
۰/۸۷	**۰/۸۹**	**mRNA**
۰/۸۱	۰/۷۰	**miRNA**
۰/۸۴	۰/۸۰	**متیلاسیونDNA**

جدول ۳- مقایسه دقت طبقه‌بندی روش پیشنهادی با جنگل تصادفی عمیق بدون ترکیب داده‌ها.

الگوریتم پیشنهادی بدون ترکیب داده‌ها	جنگل تصادفی عمیق [۹]	داده ژنی مورد استفاده
۰/۸۶	**۰/۸۹**	**mRNA**
۰/۷۵	۰/۷۳	**miRNA**
۰/۷۹	۰/۷۲	**متیلاسیونDNA**

جدول ۴- مقایسه بازخوانی طبقه‌بندی روش پیشنهادی با جنگل تصادفی عمیق بدون ترکیب داده‌ها.

الگوریتم پیشنهادی بدون ترکیب داده‌ها	جنگل تصادفی عمیق [۹]	داده ژنی مورد استفاده
۰/۸۷	**۰/۹۱**	**mRNA**
۰/۶۷	۰/۶۶	**miRNA**
۰/۷۳	۰/۶۷	**متیلاسیونDNA**

همان‌طور که در جدول‌های ۲ تا ۴ مشاهده می‌شود، استفاده از خودرمزگذار به‌منظور استخراج ویژگی‌ها و تقویت داده‌های ژنی با آن‌ها در اکثر موارد منجر به بهبود مقادیر مربوط به معیارهای ارزیابی طبقه‌بندی می‌شود.

۵-۳ ارزیابی مدل پیشنهادی با استفاده از داده‌های ژنی به صورت چندوجهی

در ادامه روش پیشنهادی برای تشخیص زیرگروه بیماری GBM با استفاده از هر سه مجموعه داده به صورت ترکیبی مورد ارزیابی قرار گرفته است.

همچنین نتایج با روش های پایه شامل K نزدیکترین همسایه ، ماشین بردار پشتیبان(SVM[18])، جنگل تصادفی (RF[19]) و DFNForest [۳۴] با محاسبه معیار صحت طبقه‌بندی مقایسه شده است (جدول ۵).

جدول ۵- مقایسه صحت طبقه‌بندی روش پیشنهادی با روش‌های دیگر

الگوریتم پیشنهادی	DFN۳۴] [	RF	SVM	KNN	Data
۰/۸۶	۰/۸۶	۰/۸۸	**۰/۸۹**	۰/۷۸	**mRNA**
۰/۷۵	۰/۵۴	**۰/۷۵**	۰/۶۶	۰/۶۰	**miRNA**
۰/۷۹	۰/۵۶	۰/۷۲	۰/۶۷	۰/۵۶	**متیلاسیون DNA**
۰/۹۲	۰/۸۹	۰/۹۰	۰/۸۹	۰/۷۸	**ترکیب داده‌ها (چندوجهی)**

همان‌طور که در جدول ۵ مشاهده می‌شود، صحت طبقه‌بندی روش پیشنهادی نسبت به سایر روش‌ها بالاتر است. همچنین بهترین طبقه‌بندی مربوط به الگوریتم پیشنهادی با استفاده از هر سه داده ژنی می‌باشد که مؤید برتری استفاده از داده‌های چندوجهی به جای استفاده از داده‌ها به صورت مجزا می‌باشد.

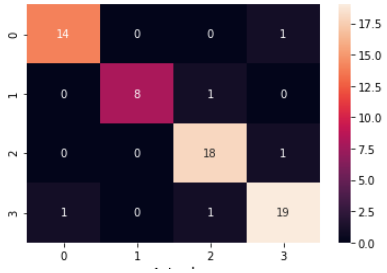
در جدول ۶، صحت، دقت و بازخوانی روش پیشنهادی در طبقه‌بندی اطلاعات با استفاده از داده‌های ژنی به صورت مجزا با ترکیب داده‌ها مقایسه شده است. همانطور که در جدول مشاهده می‌شود، استفاده از داده چندوجهی، عملکرد الگوریتم را در طبقه‌بندی زیرگروه بیماری بهبود داده است.

[15] Accuracy
[16] Precision
[17] Recall
[18] Support Vector Machine
[19] Random Forest

جدول ۶- مقایسه صحت، دقت و بازخوانی طبقه‌بندی روش پیشنهادی با استفاده از داده‌های ژنی به صورت مجزا و به صورت جندوجهی

داده ژنی مورد استفاده	صحت	دقت	بازخوانی
mRNA	۰/۸۷	۰/۸۶	۰/۸۷
miRNA	۰/۸۱	۰/۷۵	۰/۶۷
متیلاسیون DNA	۰/۸۴	۰/۷۹	۰/۷۳
ترکیب داده‌ها (چندوجهی)	۰/۹۳	۰/۹۲	۰/۹۲

افزایش دقت الگوریتم پیشنهادی نسبت به مدلهای دیگر به قدرت مدل‌های یادگیری عمیق است که قادرند ویژگیهای پیچیده‌تری را استخراج کنند.
در شکل ۷ ماتریس درهم ریختگی طبقه‌بندی با استفاده از الگوریتم پیشنهادی نشان داده شده است.



شکل ۷- ماتریس درهم ریختگی طبقه‌بندی داده‌های آزمایشی با اسفاده از الگوریتم پیشنهادی

۶- نتیجه گیری

در سال‌های اخیر، روشهای یادگیری ماشین پیشرفت زیادی در تشخیص بیماری‌ها و استخراج اطلاعات مفید در رابطه با آنها با استفاده از داده‌های ژنی داشته‌اند.

در این مقاله، هدف تشخیص زیرگروه بیماری GBM که نوعی سرطان تهاجمی مغز است، می‌باشد. تشخیص زیرگروه بیماری به پزشکان در تعیین روند درمانی بیمار کمک شایانی می‌کند.

این بیماری دارای چهار زیرگروه است که می‌توان آن را در قالب یک مسئله طبقه‌بندی با چهار کلاس مدل کرد. از بین روشهای مختلف طبقه‌بندی اطلاعات، جنگل تصادفی عملکرد مطلوبی در پردازش داده‌های ژنی داشته است. با گسترش روش‌های عمیق در شبکه‌های عصبی و موفقیت آن‌ها، روش جنگل تصادفی عمیق با همان رویکرد اخیرا پیشنهاد شده است. این روش در کنار عملکرد بالایی که دارد، تعداد پارامترها و پیچیدگی محاسباتی کمی دارد که در مقایسه با روش‌هایی چون شبکه عصبی عمیق، مزین محسوب می‌شود.

همچنین در این مقاله از سه داده ژنی mRNA، miRNA و متیلاسیون DNA استفاده شده است. استفاده از چند داده اطلاعاتی به جای یک داده، به بهبود عملکرد طبقه‌بندی کننده‌ها کمک می‌کند که این مورد در آزمایش‌های انجام شده تایید شد.

همچنین برای استخراج ویژگیهای بهتر از خودرمزگذار استفاده شده است. ویژگی‌های استخراج شده از سه خودرمزگذار در کنار داده‌های ژنی اولیه، داده‌های بهتری برای طبقه‌بندی اطلاعات فراهم می‌آورند که این مورد نیز در آزمایش‌های مجزایی که انجام شد، مورد تایید قرار گرفت.

در نهایت آزمایشهای انجام شده مبتنی بر داده‌های بیماران مبتلا به بیماری GBM با هدف تشخیص زیرگروه این بیماری نشان میدهد تلفیق خودرمزگذار و جنگل تصادفی عمیق نتایج خوبی داشته است.

در پژوهشهای آینده می توان از روشهای مختلف برای ادغام داده‌های ژنی mRNA، miRNA و متیلاسیون DNA استفاده کرد تا ویژگی های تقویت شده بهتری برای طبقه بندی استفاده شود. همچنین به‌جای داده‌های ژنی مورد استفاده، داده‌های ژنی دیگری می‌توان اضافه یا جایگزین کرد که ممکن است تاثیر بیشتری در پیش‌بینی زیرگروه این بیماری داشته باشند که به عنوان کارهای بعدی می توانند مورد مطالعه قرارگیرد.

مراجع

[۱] ب. باباعباسی, «بیوانفورماتیک سلولی و مولکولی»، صفحه ۱-۱۶ ، ۱۳۹۵.

[2] A. Rahimi, and M. Gönen, “Discriminating early- and late-stage cancers using multiple kernel learning on gene sets”, Bioinformatics, vol. 34, no. 13, pp. i412–i421, 2018.

[3] M. Sherafatian, “Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping”, Gene, vol. 677, pp. 111–118, 2018.

[4] W.Y. Cheng, Ch.Ch. Yang, J.H. Kao, Ch.Ch. Shen, Y.Ch. Yang, and M.H. Tsai, "An Intelligent and Prognostic machine learning model for Glioblastoma Multiforme", Research Square, 2023.

[5] P. Sanghani, "Machine Learning Based Overall Survival Prediction of Glioblastoma Multiforme Patients Using Magnetic Resonance Image Derived Features", PhD Dissertation, National University of Singapore, 2018.

[6] S. Bijari, A. Jahanbakhshi, P. Hajishafiezahramini, and P. Abdolmaleki, "Differentiating glioblastoma multiforme from brain metastases using multidimensional radiomics features derived from MRI and multiple machine learning models", BioMed Research International, vol. 2022, 2022.

[7] Y. Kim, K.H. Kim, J. Park, H.I. Yoon, and W. Sung, "Prognosis prediction for glioblastoma multiforme patients using machine learning approaches: Development of the clinically applicable model", Radiotherapy and Oncology, vol. 183, pp. 109617, 2023.

[8] Zh. Ya, L. Ao, H. Jie, and M. Wang, "A novel MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data", IEEE journal of biomedical and health informatics, vol. 24, no. 1 pp. 171-179, 2019.

[9] Z. Zhou, J. Feng, “Deep forest”, national science review, vol. 6, pp.74–86, 2018.

[10] Y.Boualleg, M. Farah, and I.R. Farah, "Remote sensing scene classification using convolutional features and deep forest classifier", IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 12, pp.1944-1948, 2019.

[11] B.Yu, Ch. Chen, X. Wang, Z. Yu, A. Ma, and B. Liu, "Prediction of protein–protein interactions based on elastic net and deep forest", Expert Systems with Applications, vol.176, pp.114876, 2021.

[12] L.Sun, Zh. Mo, F. Yan, L. Xia, F. Shan, Zh. Ding, B. Song, W. Gao, W. Shao, F. Shi, H. Yuan, and H. Jiang, "Adaptive feature selection guided deep forest for covid-19 classification with chest ct", IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2798-2805, 2020.

[13] W.Qin, D. Xu, X. Dong, X. Cui, and S. Zhang, "EEG signal classification based on improved variational mode decomposition and deep forest", Biomedical Signal Processing and Control, vol. 83, pp.104644, 2023.

[14] J. Xia, Z. Ming, and A. Iwasaki, "Multiple sources data fusion via deep forest", In IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1722-1725. IEEE, 2018.

[15] H. Yang, R. Chen, D. Li, and Zh. Wang, "Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data", Bioinformatics, vol. 37, no. 16, pp.2231-2237, 2021.

[16] I. Bichindaritz, G. Liu, and Ch. Bartlett, "Integrative survival analysis of breast cancer with gene expression and DNA methylation data", Bioinformatics, vol. 37, no. 17 pp.2601-2608, 2021.

[17] A. Cheerla, and O. Gevaert,"Deep learning with multimodal representation for pancancer prognosis prediction", Bioinformatics, vol. 35, no. 14, pp.i446-i454, 2019.

[18]Q. Meng, D. Catchpoole, D. Skillicom, P. J. Kennedy, “Relational

autoencoder for feature extraction", *Proc. Int. Jt. Conf. Neural Networks*, pp. 364–371, Proceedings of 2017 International Joint Conference on Neural Networks, 2017.

[19] W. Liu, H. Lin, L. Huang, L. Peng, T. Tang, Q. Zhao, and L. Yang, "Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder", Briefings in Bioinformatics, vol. 23, no. 3 pp. bbac104, 2022.

[20] X. Hu, Y. Zhixiang, Z. Zhiliang, and Y. Peng, "Prediction of miRNA–Disease Associations by Cascade Forest Model Based on Stacked Autoencoder", Molecules, vol. 28, no. 13, pp. 5013, 2023.

[۲۱] مرتضی جهان‌تیغ و مصطفی چرمی، «افزایش صحت طبقه بندی سیگنالهایEEG تصور حرکتی با ترکیب منطقی طبقه‌بندها و با به کارگیری الگوریتم ژنتیک و درختان تصمیم کوچک»، *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۷ ، شماره ۳، صفحه ۹۳۱-۹۳۸، ۱۳۹۶.

[۲۲] فرنوش عارفی و علی نادیان، «تشخیص اجزای بدن انسان در تصاویر RGB-D با استفاده از ویژگی‌های الگوی تغییرات عمق و تفاضل مکانی عمق»، *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۹ ، شماره ۴، صفحه ۱۷۵۵-۱۷۴۵، ۱۳۹۸.

[۲۳] ندا خانبانی و امیرمسعود افتخاری مقدم، «ارائه یک روش تشخیص زبان علامت مبتنی بر رویکردMLRF فازی با استفاده از اطلاعات عمق تصویر»، *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۷ ، شماره ۳، صفحه ۹۷۸-۹۸۷، ۱۳۹۶.

[24] Z. E. Ashari, S. L. Broschat, "T-Tree and t-Forest: Decision Tree and Random Forest Algorithms Including the Relevance Factor with Applications in Bioinformatics", Proceedings of 2019 IEEE International Conference Bioinforma. Biomed, pp. 2779–2783, 2019.

[25] S. Zhou, S. Wang, Q. Wu, R. Azim, W. Li, "Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression", Computational Biology and Chemistry, vol. 85, 2020.

[26] M. Fratello, R. Tagliaferri, "Decision trees and random forests", Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics, vol. 1–3, pp. 374–383, 2018.

[27] Z. Jagga, D. Gupta, "Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms", BMC proceedings, vol. 8, pp. 1–7, 2014.

[28] Datema, Frank R., Ana Moya, Peter Krause, Thomas Bäck, Lars Willmes, Ton Langeveld, Robert J. Baatenburg de Jong, Henk M. Blom., "Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression", Head Neck, vol. 34, no. 1, p. Pages 50-58, 2010.

[29] A. A. Kim, S. Rachid Zaim, V. Subbian, "Assessing reproducibility and veracity across machine learning techniques in biomedicine: A case study using TCGA data", International Journal of Medical Informatics, vol. 141, p. 104148, 2020.

[30] Y. O. Nunez Lopez, B. Victoria, P. Golusinski, W. Golusinski, M. M. Masternak, "Characteristic miRNA expression signature and random forest survival analysis identify potential cancer-driving miRNAs in a broad range of head and neck squamous cell carcinoma subtypes", Reports Pract. Oncol. Radiother, vol. 23, no. 1, pp. 6–20, 2018.

[31] Y. Fang, H. Lu, and H. Liu, "Multi-modality deep forest for hand motion recognition via fusing sEMG and acceleration signals", International Journal of Machine Learning and Cybernetics, vol. 14, no. 4, pp.1119-1131, 2023.

[32] Y. Guo, S. Liu, Z. Li, X. Shang, "BCDForest : a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data", BMC Bioinformatics, vol. 19, no. Suppl 5, pp. 1–13, 2018.

[33] J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, M. M. Khan, "A Novel Deep Flexible Neural Forest Model for Classification of Cancer Subtypes Based on Gene Expression Data", *IEEE Access*, vol. 7, pp. 22086–22095, 2019.

[34] J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, H. Dawood, "Open Access A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data", BMC Bioinformatics, pp. 1–11, 2019.