مقاله علمی پژوهشی
مجله دوفصل نامه
مهندسی ژنتیک و ایمنی زیستی
ISSN 2588-5073 چاپی
ISSN 2588-5081 الکترونیکی
دوره ۱۱، شماره ۱، بهار و تابستان ۱۴۰۱
صفحه ۶۳-۵۴

https://dorl.net/dor/20.1001.1.25885073.1401.11.1.2.9

DOR: 20.1001.1.25885073.1401.11.1.2.9

Genetic Engineering and Biosafety Journal
Volume 11, Number 1
2022

http://gebsj.ir/

https://ecc.isc.ac/showJournal/23064

بررسی پروفایل بیانی پلاکت‌های آموزش‌دیدهٔ توموری به عنوان نشانگر زیستی برای پیش‌آگهی و تشخیص سرطان

Expression profile of tumor educated platelets as biomarkers for diagnosis and early detection of cancer

ساجده باهنر، فهیمه پالیزبان، حسام منتظری*

Sajedeh Bahonar, Fahimeh Palizban, Hesam Montazeri*

گروه بیوانفورماتیک، مرکز تحقیقات بیوشیمی و بیوفیزیک، دانشگاه تهران، ایران

Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Iran

*Corresponding Author, Email:
* نویسنده مسئول مکاتبات، پست الکترونیکی: hesam.montazeri@ut.ac.ir

(تاریخ دریافت: ۱۴۰۱/۱/۴ - تاریخ پذیرش: ۱۴۰۱/۴/۲۱)

چکیده

به این دلیل که بیوپسی مایع نسبت به بیوپسی بافتی ایمن‌تر و کم‌تهاجمی‌تر است، در سال‌های گذشته بررسی زیست‌نشانگرهای موجود در آن برای پیش‌آگهی و تشخیص زودهنگام سرطان حائز اهمیت شده‌است. بررسی تغییر پروفایل بیانی پلاکت‌های آموزش‌دیدهٔ توموری موجود در بیوپسی مایع می‌تواند به عنوان یکی از زیست‌نشانگرها مورد استفاده قرارگیرد. استفاده از مدل‌های یادگیری ماشین دسته‌بندی با توجه به فضای ویژگی برگرفته از داده‌های بیانی این پلاکت‌ها، توانایی پیش‌آگهی و تشخیص زودهنگام سرطان را به ما داده است. در این پژوهش میزان صحت و خطای هفت مدل دسته‌بندی در دو حالت دودسته برای تشخیص نمونه‌های سالم و سرطانی و چنددسته برای تشخیص نمونه‌های سالم و انواع سرطان‌ها از یکدیگر مورد ارزیابی قرارگرفتند. این مدل‌ها روی پروفایل بیانی ۲۰۰۰ ژن پلاکت‌های آموزش دیدهٔ توموری مربوط به بیمارانی با سرطان‌های سینه، کبد، روده، مغز، پانکراس و ریه و همینطور پروفایل بیانی این ژن‌ها در ۵۵ فرد سالم بررسی شدند. داده های مورد استفاده سری GSE68086 هستند که از پایگاه دادهٔ GEO دانلود شدند. همچنین روی این ژن‌ها با روش preranked GSEA تجزیه‌وتحلیل غنی‌سازی مسیر صورت گرفت. نتایج نشان داد مدل ماشین بردارهای پشتیبان با کرنل خطی و غیر خطی از بین مدل‌های دودسته با میانگین خطای ۰/۰۵ و مدل ماشین بردارهای پشتیبان خطی از بین مدل‌های چنددسته با میانگین خطای ۰/۳۳ نرخ خطای کمتری دارند. به طورکلی نتایج حاصل دسته‌بندی پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری و تجزیه‌وتحلیل غنی‌سازی مسیرنشان‌دهندهٔ این است که پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری را می‌توان به عنوان کاندید نشانگر زیستی درنظرگرفت.

واژه‌های کلیدی

پلاکت‌های آموزش‌دیدهٔ توموری،
پیش‌آگهی و تشخیص سرطان،
دسته‌بندی تغییرهای پروفایل بیانی،
مدل‌های یادگیری ماشین،
نشانگر زیستی

Genetic Engineering and Biosafety Journal
Volume 11, Number 1, 2022

Abstract

Since liquid biopsy is less invasive than tissue biopsy, studies on liquid biopsy biomarkers for the early detection of cancer and diagnosis are taken into consideration. Expression profiles of tumor-educated platelets (TEP) in liquid biopsy can be used as one of the biomarkers. The use of classification machine learning models, according to the features space derived from the expression data of TEPs, has given us the ability to predict cancer. In this study, we evaluate different types of classification models namely SVM, LDA, logistic regression, boosting, classification tree, and random forest, on the expression profile of TEPs in 230 patients with breast, liver, colorectal, brain, pancreatic, and lung cancers, as well as the expression profile of these genes in 55 healthy individuals. These models were examined on the expression profile of 2000 high variance selected genes. Also, pathway enrichment analysis was performed on these genes by the GSEA preranked method. The results showed that linear SVM and polynomial SVM models have lower error rates than two-class models and linear SVM models have lower error rates than multi-class models. In general, the results of TEP expression profile classification and pathway enrichment analysis indicate that the expression profile of TEPs can be considered as candidate biomarkers.

Keywords: Tumor educated platelets, cancer diagnosis and early detection, expression profile classification, machine learning models, biomarkers

مقدمه

یکی از راه‌های تشخیص سرطان بررسی نشانگرهای زیستی موجود در بیوپسی مایع است. امروزه تشخیص سرطان به‌طورمعمول با علایم بالینی، رادیولوژی یا تست‌های غربالگری صورت می‌گیرد که همراه‌شدن بیوپسی بافتی با آن‌ها می‌تواند تشخیص دقیق‌تری را به‌دست‌دهد. از آنجایی که بیوپسی مایع نسبت به بیوپسی بافتی ایمن‌تر و کم‌تهاجمی‌تر است، بررسی نشانگرهای زیستی موجود در آن برای تشخیص زودهنگام سرطان و دنبال‌کردن فرآیند پاسخگویی به دارو در بیماران دارای اهمیت بالایی است (Best et al. 2018) . تاکنون چهار دسته نشانگر زیستی شامل اگزوزوم‌ها، سلول‌های تومور در گردش، cell-free nucleotides و پلاکت‌های آموزش‌دیدهٔ توموری، در بیوپسی مایع مورد بررسی قرارگرفته‌اند (Joosse et al. 2015).

امروزه در بیوپسی مایع پلاکت‌های آموزش دیدهٔ توموری (TEP) به‌عنوان یکی از کاندیدهای نشانگر زیستی مورد توجه قرارگرفته‌است. آن‌ها پلاکت‌هایی هستند که از مگاکاریوسیت‌های موجود در مغز استخوان (Noetzli et al. 2019)، ریه‌ها (Lefrançais et al. 2017) یا طحال با فرایند اندوسیتوز و تغییرهای سیتوپلاسمی منشا گرفته‌اند. به عبارت دیگر پلاکت‌ها قطعه‌های سلولی بدون هسته هستند. نقش اصلی پلاکت‌ها شرکت در فرآیند انعقاد خون است. اما دو مشاهدهٔ مهم توسط پژوهشگران در سال‌های ۱۸۶۸ (Trousseau A. 1868) و ۱۸۷۸ (Billroth T. 1878) به عنوان نقطهٔ عطفی برای دنبال‌کردن نقش پلاکت‌ها و ارتباط آن‌ها با تومور شناخته می‌شود. مشاهدهٔ اول وجود خون منعقدشده به‌صورت خودبخودی در بیماران سرطانی و مشاهدهٔ بعدی وجود سلول‌های سرطانی در لخته‌های خون ایجاد شده در این بیماران بود (In T Veld and Wurdinger. 2019) . در مطالعه‌هایی که در سال‌های بعد از آن به صورت دقیق‌تر و مولکولی انجام‌شد مشخص‌شد که سلول‌های توموری مولکول‌های زیستی مختلفی را با ارتباط مستقیم یا با وزیکول‌ها به پلاکت‌ها منتقل می‌کنند تا آن‌ها را تحت آموزش خود دربیاورند و با این روش سبب تغییر در پروفایل بیانی پلاکت‌ها می‌شوند (Schlesinger. 2018). بنابراین بررسی پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری در بیوپسی مایع بیماران سرطانی در مقایسه با افراد سالم می‌تواند نشانگر زیستی خوبی برای تشخیص و دنبال‌کردن روند

درمان سرطان باشد (Varkey and Nicolaides 2021; Sun et al. 2021).

در دهه‌های گذشته یادگیری ماشین به خاطر توانایی ساخت مدل با استفاده از فضای ویژگی برگرفته از داده‌های آموزشی موجود و سپس پیش‌بینی پاسخ داده‌های آزمایشی درشاخه‌های علوم مختلف از جمله زیست‌شناسی و پزشکی مورد اهمیت قرارگرفته‌است (Larrañaga et al. 2006). به‌عنوان نمونه در سال‌های اخیر از یادگیری ماشین برای حل مسئله‌های محاسبه‌ای داده‌های ژنتیکی (Abbasi et al. 2020) از جمله سرطان (Khandezamin et al. 2020; Kourou et al. 2015; Tababaei et al. 2020) استفاده شده است. به‌عنوان یک مثال دقیق تر از مسئله‌های محاسبه‌ای برای سرطان، می‌توان به بررسی پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری برای پیش‌بینی اینکه یک پروفایل بیانی مربوط به یک دادهٔ آزمایشی در دسته سرطانی یا سالم قرارمی‌گیرد، اشاره کرد. در این راستا می‌توان یک مدل دسته‌بندی را با استفاده از یک مجموعه دادهٔ آموزشی تشکیل‌شده از نمونه‌های سالم و سرطانی آموزش داد. در پژوهشی در سال ۲۰۱۵، بست و همکاران با استفاده از پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری و مدل دسته‌بندی ماشین بردار پشتیبان برای پیش‌بینی دسته داده‌های آزمایشی استفاده کردند (Best et al. 2015). همچنین در پژوهش دیگری از روش دسته-بندی هوش ازدحامی (SIEC) که درواقع ترکیبی از روش بهینه سازی ازدحام ذرات و ماشین بردار پشتیبان (SVM) است برای دسته‌بندی داده‌های بیانی پلاکت‌های آموزش دیدهٔ توموری استفاده کردند. در پژوهش مذکور دستورالعمل جدید و مجزایی برای توالی یابی پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری به‌نام thromboSeq ارائه شده‌است (Best et al. 2019) . این پژوهشگران ایدهٔ پژوهش ذکرشده را روی نمونه‌های پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری مربوط به بیماران سرطان ریه و گلیوبلاستوما نیز مورد ارزیابی قرارداده‌اند (Best et al. 2017; Sol et al. 2020) . همچنین در مطالعهٔ دیگری از روش رگرسیون برازش چندگانه با تصمیم بیز برای دسته‌بندی داده‌های بیانی پلاکت‌های آموزش دیدهٔ توموری (Huang et al. 2017) و در پژوهش دیگری نیز از روش imPlatelet که روشی بر پایهٔ تبدیل پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری به تصویر است برای دسته‌بندی نمونه‌ها استفاده شده‌است (Pastuszak et al. 2021) .

تاکنون مدل‌های مختلفی برای دسته‌بندی داده‌ها از جمله مدل ماشین بردار پشتیبان خطی و غیرخطی، رگرسیون لجستیک و مدل‌های مبتنی بر درخت پیشنهاد شده‌است. در این پژوهش هدف بررسی عملکرد مدل‌های دسته‌بندی درختی، ماشین بردار پشتیبان خطی ,ماشین بردار پشتیبان چند جمله‌ای، رگرسیون لجستک لاسو، بوستینگ، جنگل تصادفی و تحلیل تشخیصی خطی در دسته‌بندی پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری مربوط به نمونه‌های سالم و سرطانی به عنوان یک نشانگر زیستی است. همچنین هدف بررسی مجموعه ژن‌هایی است که در تجزیه‌وتحلیل غنی‌سازی مسیر در نمونه‌های بیمار نسبت به سالم غنی می‌شوند.

مواد و روش‌ها

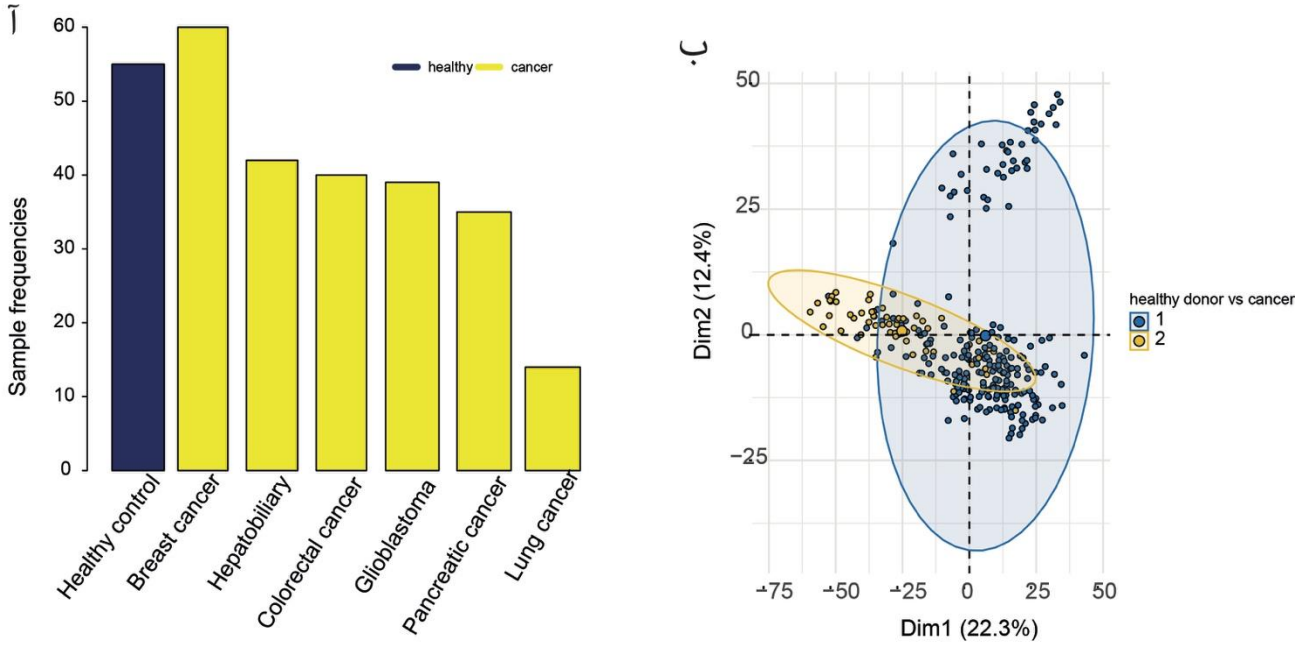
در این پروژه پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری مربوط به ۲۳۰ بیمار سرطانی (شامل ۶۰ نمونهٔ سرطان سینه، ۴۲ نمونهٔ سرطان کبد، ۴۰ نمونهٔ سرطان روده، ۳۹ نمونهٔ گلیوبلاستوما، ۳۵ نمونهٔ سرطان پانکراس، ۱۱ نمونهٔ سرطان ریه) در کنار پروفایل بیانی ۵۵ نمونهٔ بیوپسی مایع از افراد سالم مورد بررسی قرارگرفتند (۱۹ درصد نمونهٔ سالم و ۸۱ درصد نمونهٔ سرطانی). این داده‌ها از پایگاه‌دادهٔ GEO (GSE68086) دانلود شدند. در شکل ۱- (آ) فراوانی این داده‌ها نشان داده شده‌است. سپس داده‌ها با استفاده از بستهٔ edgeR و تابع cpm در R (Robinson et al. 2010) نرمال‌سازی شدند و ماتریس بیان ژن به ابعاد ۲۸۵×۵۷۷۳۶ به دست‌آمد. به دلیل بالابودن ابعاد فضای ویژگی نسبت به نمونه‌های موجود (7~ برابر) ۲۰۰۰ ژن با بیشترین واریانس انتخاب شد و به‌این‌ترتیب فضای ویژگی جدیدی از بیان ژن‌ها در نمونه‌ها برای انجام مراحل بعدی شکل‌گرفت. در شکل ۱- (ب) دو ژن با بیشترین واریانس (PC1, PC2) انتخاب و پراکنش نمونه‌ها بعد از انتخاب ویژگی نشان داده شده‌است.

[DOR: 20.1001.1.25885073.1401.11.1.2.9]

برای دسته‌بندی نمونه‌ها و پیش‌بینی دسته داده‌های آزمایشی در این پژوهش از هفت مدل دسته‌بندی مختلف برای یادگیری ماشین بانظارت استفاده شد. به‌طورکلی در این فرآیند داده‌ها به دو بخش داده‌های آموزشی و داده‌های آزمایشی تقسیم‌بندی شده و از داده‌های آموزشی برای آموزش مدل و تخمین پارامترها استفاده و سپس با استفاده از مدل آموزش‌دیده، دسته داده‌های آزمایشی تخمین زده می‌شود. در این راستا هفت مدل دسته‌بندی مختلف در اعتبارسنجی متقابل ۱۰ تکرار آموزش داده‌شده و صحت و عملکرد آن‌ها در پیش‌بینی دستها مورد ارزیابی قرارگرفت. در هر تکرار ۸۰ درصد داده‌ها به عنوان دادهٔ آموزشی و ۲۰ درصد داده‌ها به عنوان دادهٔ آزمایشی درنظر گرفته‌شدند. از آنجایی که داده‌ها شامل شش دسته مختلف سرطانی و یک دسته سالم است دو روش مختلف دودسته و چنددسته مورد بررسی قرارگرفت. در حالت دودسته تمایز نمونه‌های سالم از سرطانی و در حالت چنددسته، دسته‌بندی نمونه‌های سالم و شش نوع سرطان صورت‌گرفت.

مدل‌های مورد استفاده در این پژوهش شامل رگرسیون لجستیک لاسو، بوستینگ، دسته‌بندی درختی، ماشین بردار پشتیبان خطی ، جنگل تصادفی، ماشین بردار پشتیبان چند جمله‌ای و تحلیل تشخیصی خطی است و در جدول ۱ معادل انگلیسی هر کدام از این مدل‌ها به همراه منبع آن ذکر شده است. این مدل‌ها همگی از دستهٔ مدل‌های پارامتریک هستند که از داده‌های آموزشی برای تخمین پارامترهای آن‌ها استفاده شده و سپس مدل آموزش‌دیده، با پارامترهای تخمین‌زده‌شده می‌تواند دسته داده‌های دیده‌نشده و آزمایشی را تخمین بزند. در بین این مدل‌های دسته‌بندی، مدل‌های تحلیل تشخیصی خطی، ماشین بردار پشتیبان و رگرسیون لجستیک لاسو جزء مدل‌های خطی تعمیم‌یافته هستند و از بین آن‌ها، مدل تحلیل تشخیصی خطی از مدل‌های مولد است که با احتمال پسین کار می‌کند. برای این سه مدل به ترتیب از بسته‌های MASS (Venables and Ripley. 2002) (Friedman et al. 2010) glmnet, e1071 (Dimitriadou et al. 2009) در R استفاده شد. مدل‌های دسته‌بندی درختی، جنگل تصادفی و بوستینگ گروه دیگری از مدل‌های دسته‌بندی هستند که در این پژوهش استفاده شد و در دستهٔ مدل‌های دسته‌بندی مبتنی بر درخت قرار می‌گیرند. این مدل‌ها به‌طورمعمول ساده هستند و عملکرد قوی‌ای ندارند بنابراین با ترکیب‌شدن با روشی مثل بوت استرپ (bootstrapping) عملکرد بهتری پیدا می‌کنند. برای این مدل‌ها نیز به ترتیب از بسته‌های tree (Ripley. 2016) ، randomForest (Liaw and Wiener. 2002) و adabag (Alfaro et al. 2013) در R استفاده شد.

همهٔ مدل‌های ذکرشده به جز مدل ماشین بردار پشتیبان قابلیت دسته‌بندی چنددسته را دارند. بنابراین از همان مدل‌های بوستینگ، جنگل تصادفی، تحلیل تشخیصی خطی و بسته‌های نامبرده در R استفاده شد. به جای مدل رگرسیون لجستیک نیز مدل رگرسیون بیشینه هموار با استفاده از بستهٔ nnet (Venables and Ripley. 2002) در R پیاده سازی شد. در واقع رگرسیون بیشینه هموار یک رگرسیون لجستیک چند جمله‌ای است. برای ماشین بردار پشتیبان خطی و چند جمله‌ای نیز که قادر به تفکیک چنددسته نیستند، دو رویکرد یکی دربرابر یکی یا یکی در مقابل یکی(one-versus-one یا OVO) و یکی دربرابر همه یا یکی در مقابل همه (ove-versus-all یا OVA) پیشنهاد شده‌است. در هر دو راهکار مسئلهٔ دسته‌بندی چنددسته را به چند مسئلهٔ دسته‌بندی دودسته تبدیل می‌کند. در روش یکی در مقابل یکی هر دسته را با یکی از دستهای دیگر به عنوان یک دسته‌بندی دودسته درنظرمی‌گیرد و این کار برای همهٔ دسته‌های دوتایی ممکن تکرار می‌شود. اگر چهار دسته ۱ و ۲ و ۳ و ٤ داشته باشیم در روش یکی در مقابل یکی حالت‌های 1,3 – 1,2 – 3,4 – 2,4 – 2,3 – 1,4 – امکان‌پذیر است. به‌این‌ترتیب می‌توان با اعتبارسنجی متقابل ۱۰ تکرار مدل‌ها را برای هر مجموعهٔ دودسته آموزش داده و پس از آن روی دادهٔ آزمایشی، دستهی که در مجموع بین همهٔ مدل‌ها با بیشترین احتمال پیش‌بینی شده‌است به عنوان دسته آن داده درنظر گرفته‌می‌شود. راهکار یکی در مقابل همه نیز شبیه یکی در مقابل یکی عمل می‌کند با این تفاوت که هر دسته را در برابر سایر دستهای باقیمانده به عنوان یک مسئلهٔ دودسته درنظر می‌گیرد و به‌عنوان‌مثال برای چهار دسته حالت‌های 1,rest – 2,rest 3, rest – 4,rest – درنظر گرفته‌می‌شود .بنابراین از این دو راهکار برای آموزش مدل‌های برپایهٔ ماشین بردار پشتیبان جهت تفکیک چنددسته استفاده شد. در حالت دسته‌بندی چنددسته صحت همهٔ مدل‌ها با اعتبارسنجی متقابل ۱۰ تکرار به‌دست‌آمد.



شکل ۱– (آ) شکل نمودار میله‌ای فراوانی نمونه‌ها در دادهٔ GSE68086 را نشان می‌دهد. به‌طورکلی ۱۹ درصد از کل داده‌ها به نمونه‌های سالم و ۸۱ درصد از آن‌ها به نمونه‌های سرطانی اختصاص دارد. در بین انواع نمونه‌های سرطانی نمونه‌های سرطان سینه (۲۱ درصد) بیشترین فراوانی و نمونه‌های سرطان ریه (۵/۲درصد) کمترین فراوانی را دارند. (ب) این شکل نمودار PCA را نشان می‌دهد که نشاندهندهٔ پراکنش نمونه‌های سالم و سرطانی در دو بعد با بیشترین واریانس بعد از انتخاب ویژگی ۲۰۰۰ ژن است. واریانس در بعد اول ۲۲/۳ درصد و در بعد دوم ۱۲/٤ درصد است.

Figure 1. (A) The bar graph shows the sample frequencies in the GSE68086 data. Overall, 19% of the total data is healthy specimens and 81% of them is cancer specimens. Among the types of cancer samples, breast cancer samples (21%) are the most common and lung cancer samples (5.2%) are the least common. (B) This figure shows the PCA diagram. The PCA plot displays the distribution of health and cancer samples in two dimensions with the most variation after selecting the 2000 gene features. The variance in the first dimension is 22.3 %, and the variance in the second dimension is 12.4 %.

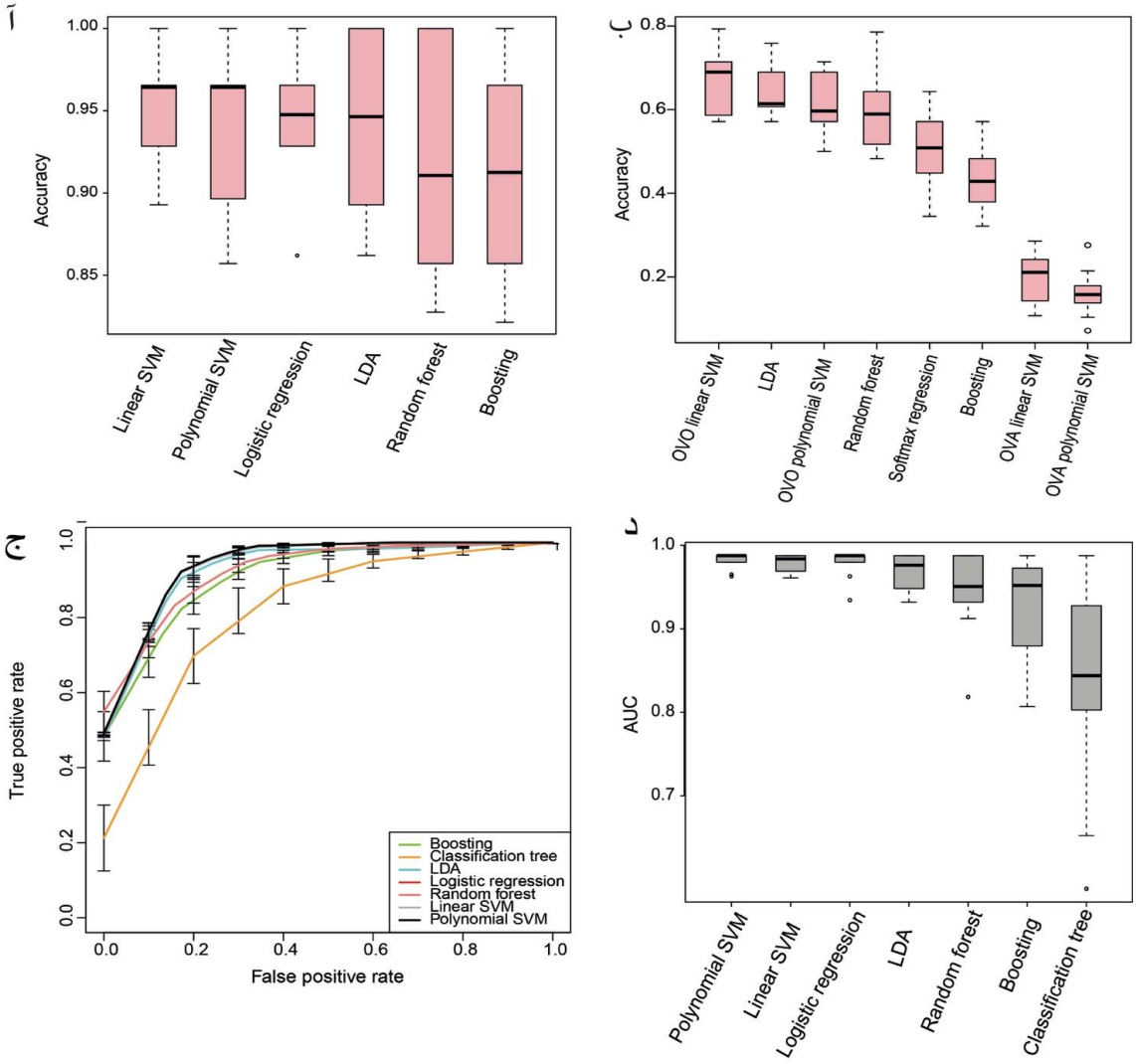
در جهت فهم اینکه ۲۰۰۰ ژن انتخاب‌شده در چه مسیرهای زیستی نقش دارند تجزیه‌وتحلیل غنی‌سازی مسیر با استفاده از روش preranked GSEA صورت‌گرفت. به این منظور ۲۰۰۰ ژن از نظر میزان اهمیت در دسته‌بندی با مدل ماشین بردار پشتیبان چند جمله ای رتبه‌بندی شدند. همچنین مجموعه ژن‌های مربوط به سه ماژول پایگاه‌دادهٔ MsigDB شامل H, C6, C7 دانلود شده و جهت غنی‌سازی مسیر مورد استفاده قرارگرفتند که به ترتیب شامل مجموعه ژن‌های آنکوژنیک، نشان و ایمونولوژیک هستند. پس از آن با استفاده از بسته fgsea (Sergushichev. 2016) در R تجزیه‌وتحلیل غنی‌سازی مسیر صورت‌گرفت. در این پروژه همهٔ تجزیه‌وتحلیل‌ها در R ورژن 4.1.0 انجام شد.

نتایج

در این پژوهش صحت شش مدل دسته‌بندی مختلف با استفاده از اعتبارسنجی متقابل ۱۰ تکرار سنجیده شد. این سنجش در دو حالت تفکیک دودستهٔ دسته‌های سالم و سرطانی و تفکیک چنددستهٔ شش نوع سرطان مختلف و دستهٔ سالم صورت‌گرفت. نتایج نشان داد که در دسته‌بندی دودسته مدل ماشین بردار پشتیبان خطی با میانگین صحت ۰/۹۵ و بعد از آن مدل ماشین بردار پشتیبان چند جمله‌ای با میانگین صحت ۰/۹٤ بهترین کارایی را در دسته‌بندی نمونه‌های سرطانی و سالم دارند. شکل ۲ – (آ) نمودار جعبه‌ای میزان صحت همهٔ مدل‌ها در تفکیک دو دسته با اعتبارسنجی ۱۰ تکرار را نشان می‌دهد. برای تفکیک چند دستهٔ

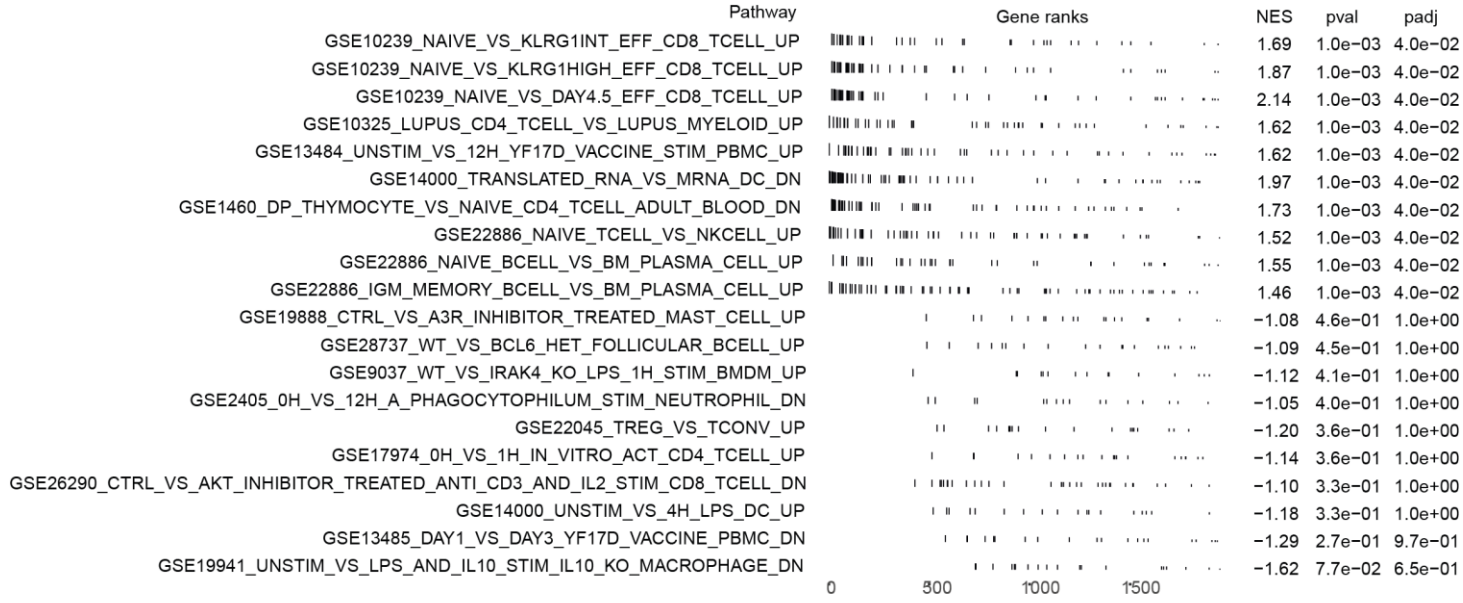
[DOR: 20.1001.1.25885073.1401.11.1.2.9]

مدل‌های جنگل تصادفی، تحلیل تشخیصی خطی و بوستینگ قابلیت انجام این دسته‌بندی را دارند. همچنین مدل رگرسیون بیشینهٔ هموار گسترش یافتهٔ مدل رگرسیون لجستیک برای تفکیک چند دسته است، اما مدل‌های ماشین بردار پشتیبان خطی و چند جمله‌ای قادر به تفکیک چند دسته نیستند. به این منظور از دو راهکار یکی در مقابل یکی و یکی در مقابل همه استفاده شد که در بخش مواد و روش‌ها توضیح داده شده‌است. نتایج حاصل نشان داد که مدل ماشین بردار پشتیبان خطی یکی در مقابل یکی با میانگین صحت ۰/٦۷ بهترین عملکرد را در تفکیک نمونه های شش نوع سرطان مختلف و نمونه های سالم دارد. بعد از آن مدل تحلیل تشخیصی خطی با میانگین صحت ۰/٦٤ در درجهٔ دوم و مدل یکی در مقابل یکی ماشین بردار پشتیبان چند جمله‌ای با میانگین صحت ۰/٦۲ در درجهٔ سوم قرار دارد. شکل ۲- (ب) نشاندهندهٔ صحت مدل‌های مختلف برای دسته‌بندی چنددسته روی دادهٔ مورد بررسی است.



شکل ۲- (آ) نمودار جعبه ای صحت 6 مدل دودسته که با اعتبارسنجی متقابل ۱۰ تکرار به‌دست آمده‌است. در این نمودار مدل ماشین بردار پشتیبان خطی با میانگین ~ 0.95 و میانه 0.96 ~ بهترین عملکرد را در دسته‌بندی دودسته نشان می‌دهد. (ب) مانند قسمت آ نمودار جعبه‌ای صحت مدل‌ها را نشان می‌دهد با این تفاوت که این نمودار برای تفکیک چنددستهٔ مدل‌ها ترسیم شده‌است. در دسته‌بندی چنددسته مدل یکی در مقابل یکی ماشین بردار پشتیبان خطی با میانگین صحت 0.67 ~ و میانه ی 0.69 ~ بهترین عملکرد را دارد. (ج) منحنی راک برای سنجش حساسیت و اختصاصیت مدل‌ها در دسته‌بندی دودسته مورد استفاده قرارگرفت. مدل ماشین بردار پشتیبان چند جمله‌ای با بیشترین سطح زیر منحنی بهترین عملکرد را دارد. (د) نمودار جعبه‌ای برای سطح زیر نمودار (AUC) مدل‌های مختلف را نشان می‌دهد که همانطور که در قسمت ج توضیح داده‌شد ماشین بردار پشتیبان چند جمله‌ای با AUC حدود ۰/۹۹٤ بهترین عملکرد را دارد.

Figure 2. (A) The accuracy of six two-class classification models acquired by cross-validation of ten replicates is depicted in a box plot. Linear SVM model with mean ~ 0.95 and median ~ 0.96 shows high-performance results in two-class classification. (B)0 Similar to part A, the boxplot shows the accuracy of the model, but this plot is drawn for multiclass classification. The OVO linear SVM model performs best with mean accuracy ~ 0.67 and median ~0.69. (C) Roc curves were used to measure the sensitivity and specificity of the two-class classification models. The polynomial SVM model with the largest area under the curve provides the best performance. (D) Box plot of AUC of different two class classification models shows that the polynomial SVM model has the best performance with AUC ~ 0.994.



شکل ۳- در این شکل نتایج تجزیه‌وتحلیل غنی‌سازی مسیر به روش preranked GSEA نشان داده شده است. همانطور که مشخص است ۱۰ مسیر با adjusted pvalue کمتر از ۰/۰۵ به طور معناداری در نمونه‌های سرطانی نسبت به سالم غنی شده‌اند. این ۱۰ مسیر، مسیرهای مختلف زیستی دخیل در فرآیندهای ایمونولوژیک هستند.

Figure 3. In this figure, the results of pathway enrichment analysis by preranked GSEA method are shown. As can be seen, 10 pathways with adjusted pvalue less than 0.05 were significantly enriched in cancer samples compared to healthy ones. These 10 pathways are various biological pathways involved in immunological processes.

برای سنجش حساسیت و اختصاصیت مدل‌های دسته‌بندی دودسته می‌توان از منحنی مشخصهٔ عملکرد (ROC curve) استفاده‌کرد. در این منحنی حساسیت مدل‌ها روی محور عمودی با نرخ مثبت صحیح نشان داده‌می‌شود. حساسیت نشاندهندهٔ نسبت نمونه‌هایی است که دستهٔ آن‌ها توسط مدل به درستی سرطانی تشخیص داده‌شده‌اند به مجموعه نمونه‌هایی که دستهٔ آن‌ها به طور کاذب سالم تشخیص داده‌شده‌است. همچنین محور افقی نشاندهندهٔ نرخ مثبت کاذب است. نرخ مثبت کاذب درصد نمونه‌هایی است که به طور کاذب بیمار تشخیص داده‌شده‌اند و از فرمول specificity – 1 به‌دست می‌آید. بنابراین منحنی مشخصهٔ عملکرد مقایسه‌ای بین نرخ مثبت کاذب و نرخ مثبت صحیح مدل انجام می‌دهد و هرچه سطح زیر منحنی (AUC) به یک نزدیک‌تر باشد به این معناست که مدل عملکرد بهتری دارد چون حساسیت و اختصاصیت بیشتری دارد.

در این پژوهش برای مقایسهٔ مدل‌های دسته‌بندی دودسته منحنی راک ترسیم شد و نتایج نشان داد که مدل ماشین بردار پشتیبان چند جمله‌ای با میانگین AUC ~ 0.994 بهترین عملکرد را دارد. بعد از آن مدل ماشین بردار پشتیبان خطی با میانگین AUC ~ 0.992 در درجهٔ دوم از لحاظ عملکرد قرار دارد. منحنی راک مدل‌ها در شکل ۲- (ج) نشان داده‌شده‌است. این منحنی با اعتبارسنجی متقابل ۱۰ تکرار روی نمونه‌های مورد بررسی به‌دست‌آمد و شکل ۲- (د) نشاندهندهٔ نمودار جعبه‌ای برای AUC مدل‌ها است.

در ادامه، تجزیه‌وتحلیل غنی‌سازی مسیر روی نمونه‌های پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری انجام شد. در این راستا ابتدا ۲۰۰۰ ژن با مدل ماشین بردار پشتیبان چند جمله‌ای رتبه‌بندی شده و سپس با روش preranked GSEA تجزیه‌وتحلیل

[DOR: 20.1001.1.25885073.1401.11.1.2.9]

صورت‌گرفت. نتایج حاصل نشان داد که از بین سه مجموعهٔ ژن‌های آنکوژنیک، نشان و ایمونولوژیک منتخب از پایگاه دادهٔ MsigDB، ۱۰ مجموعه ژن ایمونولوژیک با $padj < 0.05$ به طور معناداری در نمونه‌های سرطانی نسبت به سالم در داده‌های پروفایل بیانی مورد بررسی غنی شده‌اند. نتایج تجزیه‌وتحلیل به‌دست‌آمده در شکل ۳ قابل مشاهده است.

بحث

در دهه‌های گذشته مشاهده‌های جالب توجهی دربارهٔ پلاکت‌های بیماران مبتلا به سرطان به‌دست آمده‌است. از جمله این مشاهده‌ها می‌توان به وجود خون منعقد شدهٔ خودبخودی در بیماران سرطانی و همچنین وجود سلول‌های سرطانی در لخته‌های خون ایجاد شده اشاره کرد (In 'T Veld and Wurdinger. 2019) . پژوهش‌های بعدی نشان داد که سلول‌های سرطانی با ارتباط مستقیم یا به‌وسیلهٔ وزیکول‌ها، پلاکت‌ها را تحت آموزش خود درآورده و سبب تغییر در پروفایل بیانی پلاکت‌ها می‌شوند. ازاین‌رو به آن‌ها پلاکت‌های آموزش‌دیدهٔ توموری گفته می‌شود (Schlesinger. 2018) . تا کنون در چندین پژوهش سعی بر آموزش مدل‌های دسته‌بندی به روش یادگیری ماشین با نظارت و سپس استفاده از آن مدل‌ها برای پیش‌بینی دستهٔ داده‌های آزمایشی داشته‌اند. در این دسته‌بندی تفکیک دسته‌های سالم و سرطانی یا تفکیک بین انواع سرطان‌ها و حالت سالم صورت گرفته‌است. بست و همکاران از مدل ماشین بردار پشتیبان (Best et al. 2015) و در پژوهش دیگری از مدل دسته‌بندی هوش ازدحامی که روشی مشابه مدل ماشین بردار پشتیبان است (Best et al. 2019) برای دسته‌بندی داده‌های پلاکت-های آموزش دیدهٔ توموری استفاده کردند. همچنین در پژوهش‌های دیگر هانگ و همکاران از روش رگرسیون برازش چندگانه با تصمیم بیز (Huang et al. 2017) و پاستوزاک و همکاران از روشی بر پایهٔ تبدیل پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری به تصویر برای دسته‌بندی نمونه‌های سالم و سرطانی استفاده کردند (Pastuszak et al. 2021) .

بر مبنای بررسی‌هایی که روی مقاله‌های موجود در زمینه دسته‌بندی پلاکت‌های آموزش دیدهٔ توموری انجام دادیم، در پژوهش‌های پیشین مقایسه‌ای بین آموزش مدل‌های مختلف دسته‌بندی با استفاده از یادگیری ماشین با نظارت صورت نگرفته است بنابراین در این پروژه هدف اصلی آموزش مدل‌های مختلف دسته‌بندی با استفاده از ۲۳۰ نمونهٔ pan-cancer و ۵۵ نمونهٔ سالم برگرفته از پایگاه‌دادهٔ GEO (GSE68086) بود. در ابتدا ۲۰۰۰ ژن، با بیشترین واریانس بیان در بین نمونه‌ها، انتخاب شدند و بنابراین ابعاد ماتریس بیان ژن دانلود شده از ابعاد ۲۸۵×۵۷۷۳۶ به ۲۸۵×۲۰۰۰ کاهش‌یافت. پس از آن مدل‌های دسته‌بندی مورد بررسی شامل مدل‌های جنگل تصادفی، رگرسیون لجستیک لاسو، ماشین بردار پشتیبان چند جمله‌ای، ماشین بردار پشتیبان خطی، بوستینگ، دسته‌بندی درختی و تحلیل تشخیصی خطی با استفاده از ماتریس بیان ژن با ویژگی‌های انتخاب‌شده آموزش داده‌شدند و صحت و عملکرد آن‌ها با اعتبارسنجی متقابل ۱۰ تکرار سنجیده شد.

نتایج حاصل نشان داد که مدل‌های ماشین بردار پشتیبان خطی و چند جمله‌ای از سایر مدل‌های دسته‌بندی هم در حالت دودسته و هم در حالت چنددسته بالاترین میزان صحت و بالاترین میزان AUC را دارند. میانگین AUC به‌دست‌آمده برای ماشین بردار پشتیبان چند جمله‌ای ۰/۹۹٤ ~ و برای مدل ماشین بردار پشتیبان با کرنل خطی ۰/۹۹۲ ~ به‌دست‌آمد. در پژوهشی که بست و همکاران (Best et al. 2015) انجام دادند میانگین AUC برای مدل ماشین بردار پشتیبان حدود ۰/۹۸٦ به‌دست آمده‌است که در مقایسه با نتایج به‌دست‌آمده در اینجا، مدل ماشین بردار پشتیبان آموزش داده شده در این پژوهش عملکرد بهتری دارد. این تفاوت می‌تواند به دلیل متفاوت بودن روش انتخاب ویژگی در این پژوهش که انتخاب ژن‌هایی با بیشترین واریانس است، باشد. آموزش مدل ماشین بردار پشتیبان در پژوهش مذکور نیز با استفاده از داده‌های GSE68086 صورت گرفته است. با مقایسهٔ دو مدل دسته‌بندی ذکر شده با بهترین عملکرد در حالت دودسته نسبت به چنددسته عملکرد خیلی بهتری دارند و علت این نتیجه می‌تواند مربوط به تعداد نمونه‌های کم موجود در هر دسته سرطانی باشد به‌عنوان‌مثال فقط ۱۵ نمونه (۵ درصد) از نمونه‌ها مربوط به سرطان کبد است. در واقع تعداد کم نمونه‌های موجود در هر دسته سرطانی ممکن است نتواند دانسته‌های جامعی از جامعهٔ آماری بدهد بنابراین سبب بیش‌برازش مدل‌های مورد بررسی می‌شود. از بین مدل‌های مورد بررسی مدل تحلیل تشخیصی خطی در حالت تفکیک چنددسته بعد از مدل

یکی در مقابل یکی ماشین بردار پشتیبان خطی صحت خوبی نسبت به سایر مدل‌ها نشان داد درحالی که در دسته‌بندی دودسته در درجهٔ چهارم صحت دسته‌بندی قرار داشت. از بین روش‌های مبتنی بر درخت روش جنگل تصادفی در هر دو حالت چنددسته و دودسته عملکرد و صحت بهتری داشت که علت آن این است که این روش سعی در کم‌کردن همبستگی بین درخت‌ها دارد.

به طور کلی نتایج این پژوهش نشان داد که پروفایل بیانی پلاکت-های آموزش دیدهٔ توموری موجود در بیوپسی مایع می‌تواند نشانگر زیستی خوبی برای تشخیص سرطان باشد. با آموزش مدل‌های دسته‌بندی روی داده‌های موجود می‌توان دسته سالم یا بیمار یک نمونه پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری را پیش‌بینی کرد. همچنین علاوه بر تشخیص pan-cancer می‌توان با آموزش مدل‌های دسته‌بندی به صورت چنددسته نوع سرطان را نیز برای یک نمونهٔ با دسته نامشخص پیش‌بینی کرد که در این پژوهش به بررسی این موارد پرداخته‌شد. همچنین با دسته‌بندی ساب‌تایپ‌های سرطانی، از پروفایل بیانی پلاکت‌های آموزش دیدهٔ توموری می‌توان برای تشخیص بهتر و بررسی روند درمان به عنوان نشانگر زیستی استفاده کرد و مشابه روند این پژوهش با استفاده از داده‌هایی که دسته ساب‌تایپ‌های سرطانی آن مشخص است مدل‌های دسته‌بندی را آموزش داده و برای پیش‌بینی دسته داده‌های آزمایشی از آن استفاده‌کرد.

جدول ۱- مدل‌های دسته‌بندی مورد استفاده در این پژوهش

Table 1. The classification models used in this research

نام فارسی مدل	نام انگلیسی مدل	منبع
جنگل تصادفی	Random forest	**Breiman. 2001**
رگرسیون لجستیک لاسو	LASSO logistic regression	**DeMaris. 1995**
ماشین بردار پشتیبان خطی	Linear support vector machine (SVM)	**Saunders et al. 1998**
بوستینگ	Boosting	**Freund. 1995**
دسته‌بندی درختی	Classification tree	**Breiman et al. 2017**
تحلیل تشخیصی خطی	Linear discriminant analysis (LDA)	**Keinosuke Fukunaga. 2013**

منابع

Abbasi K, Masoudi-Nejad A. 2020. Design of a bioinformatics model to predict drug compound properties and its application in inhibition of HIV replication and BACE-1. Genetic Engineering and Biosafety Journal. 9(2): 181-193 (In Farsi with English abstract).

Alfaro E, Gáamez M, and García N. **2013**. adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software 54: 1–35.

Best MG, In 't Veld SGJG, Sol N, and Wurdinger T. **2019**. RNA sequencing and swarm intelligence–enhanced classification algorithm development for blood-based disease diagnostics using spliced blood platelet RNA. 14(4): 1206–1234.

Best MG, Sol N, In 't Veld SGJG, Vancura A, Muller M, Niemeijer ALN, et al. and Wurdinger T. **2017**. Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets. Cancer Cell 32(2): 238-252.e9.

Best MG, Sol N, Kooi I, Tannous J, Westerman BA, Rustenburg F, et al., and Wurdinger T. **2015**. RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. Cancer Cell 28(5): 666–676.

Best MG, Wesseling P, and Wurdinger T. **2018**. Tumor-Educated Platelets as a Noninvasive Biomarker Source for Cancer Detection and Progression Monitoring. Cancer research 78(13): 3407–3412.

Billroth T. 1878. Lectures on surgical pathology and therapeutics. New Sydenham society. v 2.

Breiman L. **2001**. Random Forests. Machine Learning 45(1): 5–32.

Breiman L, Friedman JH, Olshen RA, and Stone CJ. **1984**. Classification and regression trees. 1st edition.

[DOR: 20.1001.1.25885073.1401.11.1.2.9]

Champon & Hall. New York. USA. 1-368.

DeMaris A. **1995**. A Tutorial in Logistic Regression. Journal of Marriage and the Family 956–968.

Dimitriadou E, Hornik K, Leisch F, Meyer D, and Maintainer AW. **2009**. The e1071 Package. Misc Functions of Department of Statistics (e1071), TU Wien R package 1: 5-24.

Freund Y. **1995**. Boosting a Weak Learning Algorithm by Majority. Information and Computation 121(2): 256–285.

Friedman J, Hastie T, and Tibshirani R. **2010**. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software 33(1): 1-22.

Huang G, Yuan M, Chen M, Li L, You W, Li H, Cai JJ, and Ji G. **2017**. Integrating multiple fitting regression and Bayes decision for cancer diagnosis with transcriptomic data from tumor-educated blood platelets. Analyst 142(19): 3588–3597.

In 'T Veld SGJG, and Wurdinger T. **2019**. Tumor-educated platelets. Blood 133(22): 2359–2364.

Joosse SA, Pantel K, SA J, and K P. **2015**. Tumor-Educated Platelets as Liquid Biopsy in Cancer Patients. Cancer cell 28(5): 552–554.

Keinosuke Fukunaga. **2013**. Introduction to Statistical Pattern Recognition. 2nd edition. Elsevier.

Khandezamin Z, Naderan Tahan M, Rashti MJ. 2020. Intelligent detection of breast cancer with feature selection based on logistic regression and support vector machine Classification. Journal of Soft Computing and Information Technology 9(2): 115-123 (In Farsi with English abstract).

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. 2015. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13: 8-17.

Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, and Robles V. **2006**. Machine learning in bioinformatics. Briefings in Bioinformatics 7(1): 86–112. (In Farsi with English abstract).

Lefrançais E, Ortiz-Muñoz G, Caudrillier A, Mallavia B, Liu F, Sayah DM, Thornton EE, Headley MB, David T, Coughlin SR, Krummel MF, Leavitt AD, Passegué E, and Looney MR. **2017**. The lung is a site of platelet biogenesis and a reservoir for haematopoietic progenitors. Nature 544(7648): 105–109.

Liaw A, and Wiener M. **2002**. Classification and Regression by randomForest. R news 2(3): 18–22.

Noetzli LJ, French SL, and Machlus KR. **2019**. New insights into the differentiation of megakaryocytes from hematopoietic progenitors. Arteriosclerosis, Thrombosis, and Vascular Biology 39(7): 1288–1300.

Pastuszak K, Supernat A, Best MG, In 't Veld SGJG, Łapińska-Szumczyk S, Łojkowska A, Różański R, Żaczek AJ, Jassem J, Würdinger T, and Stokowy T. **2021**. imPlatelet classifier: image-converted RNA biomarker profiles enable blood-based cancer diagnostics. Molecular oncology 15(10): 2688–2701.

Ripley B, Ripley M B. **2016**. tree: Classification and Regression Trees. R Packag version 1.0: 37.

Robinson MD, McCarthy DJ, and Smyth GK. **2010**. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1): 139–140.

Saunders C, Stitson MO, Weston J, Bottou L, Schölkopf B, and Smola A. **1998**. Support Vector Machine Reference Manual, 1-26.

Schlesinger M. **2018**. Role of platelets and platelet receptors in cancer metastasis. Journal of Hematology & Oncology 11(1): 1–15.

Sergushichev AA. **2016**. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv 060012: 1–9.

Sol N, In 't Veld GJG, Vancura A, Tjerkstra M, Leurs C, Rustenburg FO, et al., and Wurdinger T. **2020**. Tumor-Educated Platelet RNA for the Detection and (Pseudo)progression Monitoring of Glioblastoma. Cell Reports Medicine 1(7): 100101.

Sun K, Wang H, Xu X, Wei X, Su J, Zhu K, Fan J, Calin G, and Grignani F. **2021**. Tumor-Educated Platelet miR-18a-3p as a Novel Liquid-Biopsy Biomarker for Early Diagnosis and Chemotherapy Efficacy Monitoring in Nasopharyngeal Carcinoma. Frontiers in Oncology 11 : 4066.

Tababaei A, Derhami V, Sheikhpour R, Pajoohan M. 2020. Feature selection based on information theory to select effective genes for diagnosis of cancer subtypes using microarray data. Iranian Journal Of Biomedical engineering. 13(4): 351-362 (In Farsi with English abstract).

Trousseau A. trans by Bazire P. V. 1868. Lectures on Clinical Medicine. London. New Sydenham Society.

Varkey J, and Nicolaides T. **2021**. Tumor-Educated Platelets: A Review of Current and Potential Applications in Solid Tumors. Cureus 13(11): e19189.

Venables WN, and Ripley BD. **2002**. Modern Applied Statistics with S. 4th edition. Springer. New York. USA. 251-266.