دانشکده مهندسی برق و کامپیوتر

سمینار کارشناسی ارشد

در رشته مهندسی کامپیوتر گرایش نرم افزار

بررسی روش‌های مبتنی بر یادگیری چند هسته‌ای برای طبقه‌بندی اطلاعات بیماری‌ها با استفاده از مجموعه داده‌های ژنی

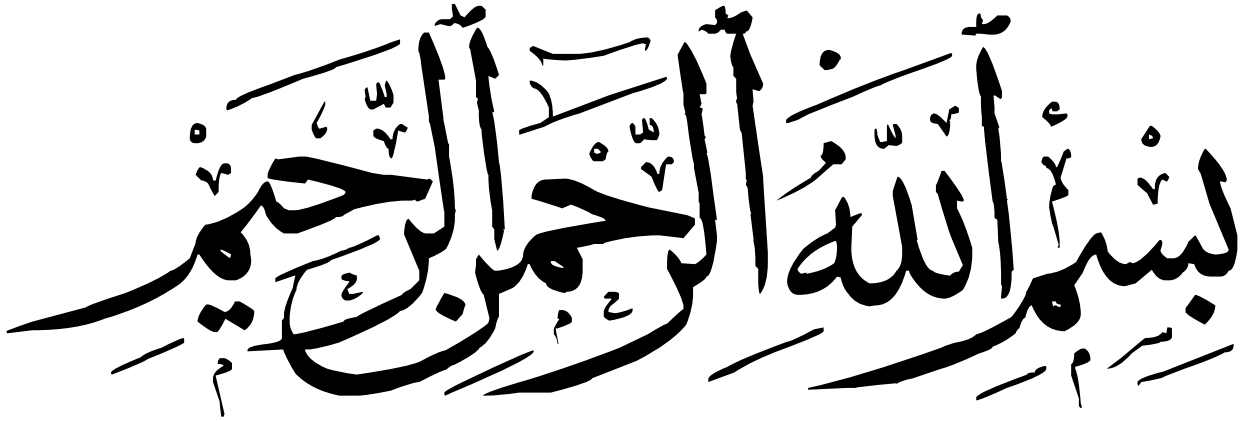
توسط

امیرحسین محمدجانی

استاد راهنما سمینار

دکتر فاطمه زمانی

بهار ۱۴۰۰


بسم الله الرحمن الرحيم

دانشکده مهندسی برق و کامپیوتر

سمینار کارشناسی ارشد

در رشته مهندسی کامپیوتر گرایش نرم افزار

بررسی روش‌های مبتنی بر یادگیری چند هسته‌ای برای طبقه‌بندی اطلاعات بیماری‌ها با استفاده از مجموعه داده‌های ژنی

توسط

امیرحسین محمدجانی

استاد راهنما سمینار

دکتر فاطمه زمانی

بهار ۱۴۰۰

چکیده

بیوانفورماتیک دانشی است که به جنبه‌های ریاضی و محاسباتی زیست شناسی برای فهم و پردازش داده‌های زیستی می پردازد. گسترش روزافزون و حجم عظیم داده‌های ژنومی و نیاز به ذخیره، بازیابی و تحلیل مناسب این داده‌ها موجب پیدایش علم بیوانفورماتیک گردید. با رشد قابل توجه داده های بیولوژیکی کشف دانش از داده‌ها نقش بسزایی در تجزیه و تحلیل داده‌ها و حل مشکلات در حال ظهور خواهد داشت ولی انجام الگوریتم‌های بیوانفورماتیک به صورت دست نویس و غیر خودکار بسیار دشوار است، به این منظور برای به دست آوردن دانش از داده‌های زیستی از ابزارها و روش‌های یادگیری ماشین استفاده می‌شود. یکی از روش‌های یادگیری ماشین که کاربرد بسیاری در بیوانفورماتیک دارد یادگیری چند هسته‌ای می‌باشد. روش‌های هسته‌ای طی دهه گذشته شاهد افزایش محبوبیت زیادی در بیوانفورماتیک بوده‌اند. یکی از دلایل عمده محبوبیت روش‌های هسته در بیوانفورماتیک قدرت آن‌ها در ادغام داده است. دومین مزیت روش‌های هسته این است که می‌توان آن‌ها را به راحتی روی داده‌های ساخت یافته (مثال گراف ها، مجموعه ها، سری های زمانی و رشته ها) اعمال کرد. داده‌های بیوانفورماتیک می‌تواند در بسیاری از زمینه‌های مختلف از جمله تشخیص، پیش‌بینی و درمان بیماری مورد استفاده قرار گیرد. به عنوان مثال برای بیماری سرطان که یک بیماری جدی و تهدید کننده زندگی است که در همه دن‌یا ش‌یوع دارد و ی‌ک‌ی از م‌ه‌م‌ت‌ری‌ن دل‌ای‌ل م‌رگ و م‌ی‌ر در ج‌ه‌ان م‌ح‌س‌وب م‌ی‌ش‌ود، تشخیص بیماری و مراحل پیشرفت آن برای پیشگیری و درمان بسیار مهم است زیرا میزان کشندگی این بیماری هنگامی که در مراحل آخر تشخیص داده شود، بیشتر است. همچنین نوع درمان در مراحل مختلف بیماری نیز متفاوت است. با افزایش در دسترس بودن خصوصیات ژنومی برای نمونه برداری از تومور گرفته شده از بیماران، الگوریتم‌های یادگیری ماشین برای کارهای پیش‌بینی و درمان سرطان اعمال شده است. در این پژوهش به بررسی روش‌های مبتنی بر یادگیری چند هسته‌ای برای طبقه‌بندی اطلاعات بیماری‌ها با استفاده از مجموعه داده‌های ژنی می‌پردازیم. با این هدف که نتایج نسبت به روش‌های موجود بهبود یابد و به دقت بالاتری دست یابیم.

واژه های کلیدی: یادگیری ماشین، ماشین بردار پشتیبان، یادگیری چند هسته‌ای، پیش‌بینی بیماری، بیوانفورماتیک، داده‌های ژنی

فهرست مطالب

فهرست شکل ها

فهرست جدول ها

فصل ۱

مقدمه

۱-۱ مقدمه

بیوانفورماتیک[۱] دانشی است که به جنبه‌های ریاضی و محاسباتی زیست‌شناسی برای فهم و پردازش داده‌های زیستی می‌پردازد. گسترش حجم داده‌های ژنومی و نیاز به ذخیره، بازیابی و تحلیل مناسب این داده‌ها، موجب پیدایش علم بیوانفورماتیک گردید. هدف اولیه بیوانفورماتیک افزایش سطح فهم و درک از فرایندهای زیستی است و تمرکز آن در توسعه و کاربرد تکنیک‌های محاسباتی جامع به منظور کسب این هدف است [1].

با پیشرفت تکنولوژی و افزایش چشمگیر داده‌های زیستی، علاوه بر ذخیره‌سازی و نگهداری، استخراج اطلاعات سودمند از این حجم از داده نیز چالش بزرگی را برای پژوهشگران به وجود آورده است. به این منظور برای به دست آوردن دانش از داده‌های زیستی از ابزارها و روش‌های یادگیری ماشین استفاده می‌شود [2]. با رشد قابل توجه داده های بیولوژیکی، کشف دانش از داده‌ها نقش بسزایی در تجزیه و تحلیل داده‌ها و حل مشکلات در حال ظهور خواهد داشت [3]. پیش از ظهور روش‌های یادگیری ماشین در بیوانفورماتیک، الگوریتم‌های بیوانفورماتیک به صورت دست‌نویس و غیرخودکار برنامه‌نویسی می‌شدند که برای مسائلی مانند پیش‌بینی ساختار پروتئین بسیار دشوار بوده‌است [4].

در ادامه به توضیح برخی مفاهیم اصلی در ارتباط با اسیدهای نوکلئیک[۲] و پروتئین‌ها[۳] که مواد خام علم بیوانفورماتیک محسوب می‌شوند، می‌پردازیم.

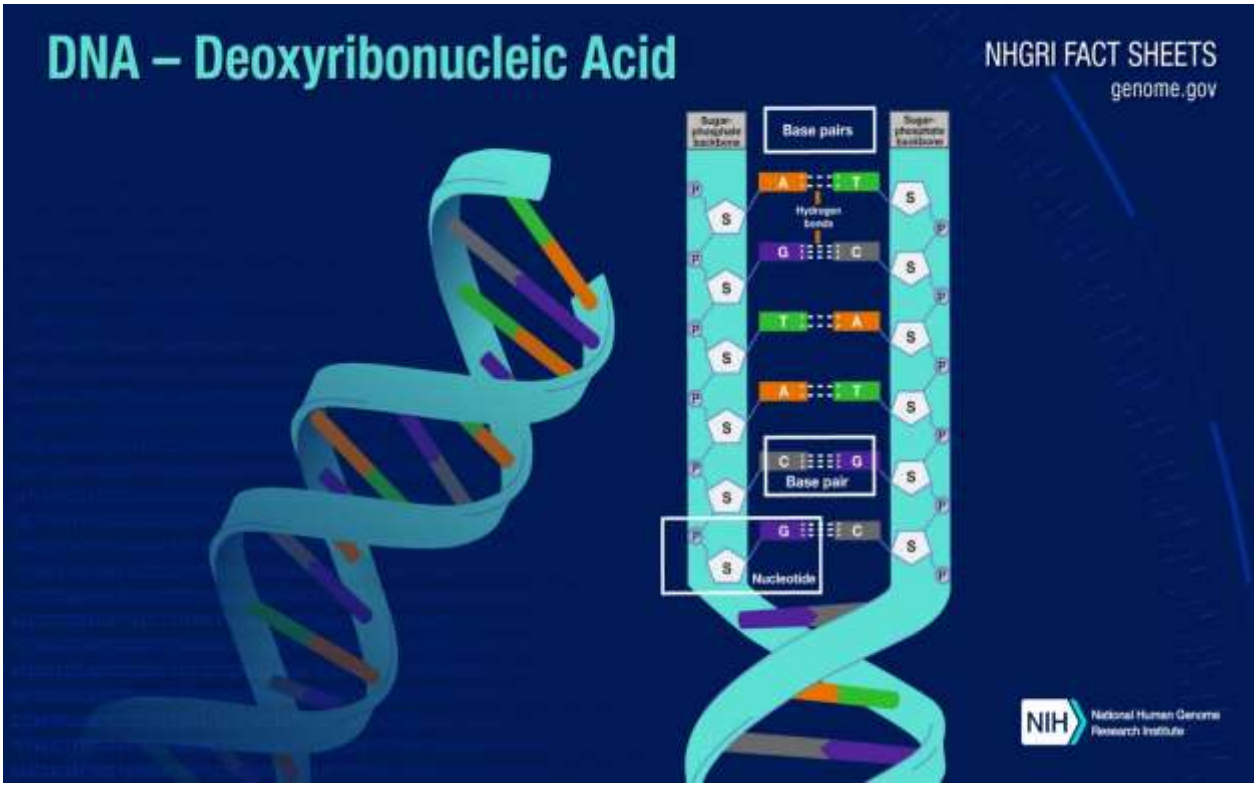
اسیدهای نوکلئیک، نوکلئوتیدها هستند که شامل یک باز متصل به قند همراه با گروه فسفات متصل به قند می‌باشند. قرار گیری نوکلئوتیدها به دنبال هم در یک مولکول DNA[۴] یا RNA[۵] یک توالی نوکلئوتیدی را تشکیل می‌دهد [5].

DNA: مولکول DNA یک توالی نوکلئیک اسید است که تمام کدهای ژنتیکی جانوران، گیاهان و حتی ویروس‌ها را حمل می‌کند که این اطلاعات برای رشد، تکامل، بقا، تولید مثل و سایر عملکردهای موجودات حیاتی است. محل قرارگیری آن در هسته‌ی سلول است که به عنوان مرکز ارسال دستورالعمل‌های مورد نیاز بدن جانداران شناخته می‌شود. زیر واحد اصلی DNA نوکلئوتید می‌باشد که از سه بخش اصلی یعنی قند، فسفات و یکی از چهار باز آلی آدنین[۶](A)، تیمین[۷](T)، سیتوزین[۸](C) و گوانین[۹](G) تشکیل می‌شود (شکل ۱-۱). DNA یک مارپیچ دو رشته‌ای است که رشته‌های آن به دور یک محور مرکزی معمولاً به صورت راست گَرد پیچ می‌خورند. ستون‌های

---

۱ Bioinformatic
۲ Nucleic acid
۳ Protein
۴ Deoxyribonucleic acid
۵ Ribonucleic acid
۶ Adenine
۷ Thymine
۸ Cytosine
۹ Guanine

قند-فسفات همانند نرده‌های پلکان به دو قسمت خارجی بازهای آلی پیچیده و به این ترتیب در معرض محیط آبی داخل سلول هستند و بازهای آلی که خاصیت آبگریزی دارند همانند پله های نردبان در داخل مارپیچ قرار می‌گیرند. در ساختمان DNA هر کدام از دو رشته به صورت موازی و ناهمسو روبروی هم قرار دارند؛ به این ترتیب که اگر یک رشته در ابتدا دارای گروه فسفات (‘۵) و در انتهای رشته هیدروکسیل (‘۳) باشد، وضعیت در رشته مقابل برعکس است؛ به این معنی که در ابتدای رشته دوم، گروه هیدروکسیل و در انتهای آن گروه فسفات قرار دارد [5].



شکل ۱-۱ ساختار DNA

ژن[۱۰] : هر قسمتی از DNA که دارای دستورالعمل یا کد ایجاد یک پروتئین مشخص است و منجر به ایجاد یک ویژگی یا عملکرد خاص می‌شود یک ژن نامیده می‌شود. در بدن انسان تقریبا 30000 ژن وجود دارد [5].

ژنوم[۱۱] : مجموع تمام ژن‌های موجود در یک بافت یا سلول است که علم ژنومیکس[۱۲] به مطالعه آن می‌پردازد [5].

RNA: مولکول RNA از روی مولکول DNA ساخته می‌شود. ساختار شیمیایی RNA به ساختار شیمیایی DNA بسیار شبیه است با دو تفاوت؛ یک این که RNA دارای قند ریبوز است در حالی که DNA دارای قند کمی متفاوت

[۱۰] Gene
[۱۱] Genome
[۱۲] Genomics

تر به نام دئوکسی ریبوز است (گونه‌ای از ریبوز که یک اتم اکسیژن در آن کم است) و دوم این که RNA دارای نوکلئوباز یوراسیل[۱۳] است در حالی که به جای آن DNA دارای تیمین است. برخلاف DNA بیشتر مولکول‌های RNA تک رشته‌ای هستند [6]. انواع بسیار مختلف RNA وجود دارد که به چند مورد اشاره می‌کنیم [6].

- RNA پیام رسان [۱۴](mRNA): یک گروه مهم از RNA هست که یک برنامه ژنتیکی را برای تولید یک محصول پروتئینی، رمز (کدینگ) می‌کند. در سنتز پروتئین، mRNA کدهای ژنتیکی را از DNA در هسته به ریبوزوم‌ها (مکان های ترجمه پروتئین در سیتوپلاسم) حمل می‌کند [6].
- RNA انتقال دهنده [۱۵](tRNA): RNA کوچک غیر رمزگذار (۷۴-۹۳ نوکلئوتید) مورد نیاز ریبوزوم برای تبدیل mRNA به پروتئین می‌باشد [6].
- RNA کوچک [۱۶](miRNA): دو RNA تک رشته تقریباً مکمل ۲۰-۲۵ نوکلئوتیدی که از DNA رونویسی شده‌اند و ترجمه نمی‌شوند. miRNA بیان ژن‌های دیگر را تنظیم می‌کند زیرا مکمل قسمت‌هایی از mRNA است [6].

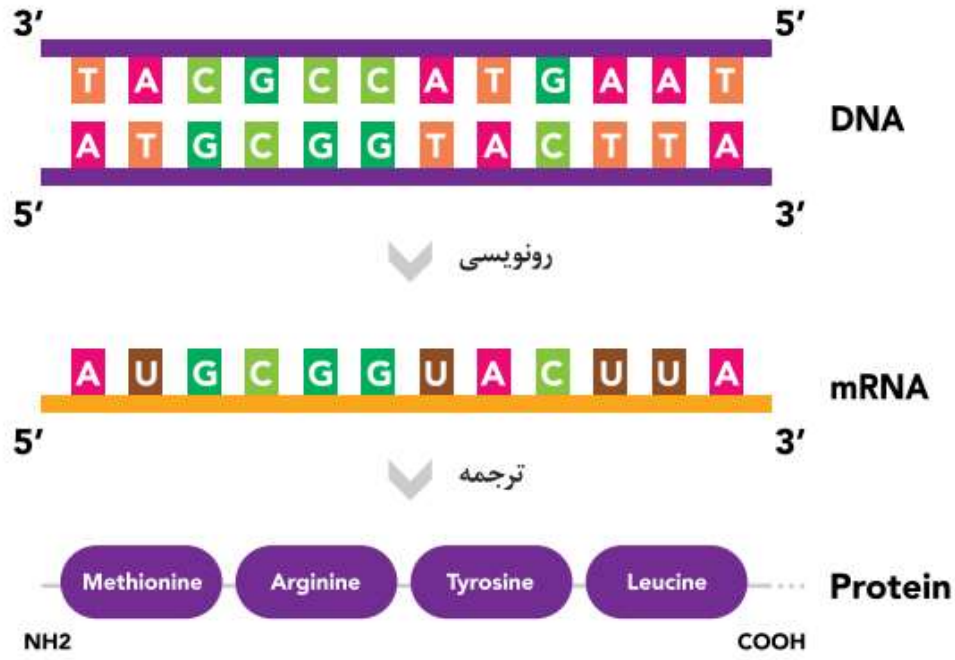
پروتئین: پروتئین‌ها در طی فرایند ترجمه تولید می‌شوند و شامل تبدیل کدون (ترکیبی از سه نوکلئوتید) به پروتئین است. هر ترکیب سه نوکلئوتیدی یک اسید آمینه را تعیین می‌کند و در مجموع ۲۰ اسید آمینه مختلف حروف الفبا پروتئین‌ها را تشکیل می‌دهند [6].

توالی یابی RNA [۱۷]: توالی یابی از نو به مفهوم تولید توالی اولیه از ژنوم یا ترانسکریپتوم (مجموعه ی رونوشت های (mRNA) موجود در یک سلول یا جمعیتی از سلول‌ها) در موجوداتی که اطلاعات ژنتیکی اندکی از آنها در دسترس است و ژنوم آنها هنوز توالی یابی نشده است می‌باشد. این نوع توالی یابی در سطح ترانسکریپتوم به توالی یابی RNA معروف است. توالی یابی RNA با استفاده از توالی نسل بعدی (NGS[۱۸]) حضور و مقدار RNA را در یک نمونه بیولوژیکی در یک لحظه مشخص نشان می‌دهد [5].

بیان ژن[۱۹]: بیان ژن یک فرایند حیاتی است که پلی را بین اطلاعات رمزگذاری شده در داخل یک ژن و یک محصول نهایی ژن عملکردی مانند پروتئین فراهم می‌کند. برای بیان پروتئین، این فرایند شامل مرحله رونویسی و ترجمه است (شکل ۱-۲ را ببینید). توانایی تنظیم بیان ژن به سلول‌ها امکان می‌دهد پروتئین عملکردی را هر زمان که برای عملکرد طبیعی یا زنده ماندن آن‌ها لازم باشد، ارائه دهند [7].

---

[۱۳] Uracil
[۱۴][۱۴] Messenger RNA
[۱۵] Transfer RNA
[۱۶] Micro-RNA
[۱۷] RNA-Seq
[۱۸] Next-generation sequencing
[۱۹] Gene Expression



شکل ۱-۲ نمای شماتیک مرحله رونویسی و ترجمه. ابتدا در طی مرحله رونویسی RNA از روی DNA ساخته می‌شود. سپس در طی فرایند ترجمه پروتئین از RNA ساخته می‌شود [5].

متیلاسیون DNA[۲۰]: اپی ژنتیک را می‌توان به عنوان یک تغییر پایدار در پتانسیل بیان ژن توصیف کرد که در طول تکامل و تکثیر سلولی اتفاق می‌افتد بدون اینکه تغییری در توالی ژن ایجاد شود. متیلاسیون DNA یکی از اتفاقات شایع اپی ژنتیکی است که در ژنوم پستانداران رخ می‌دهد. این تغییر اگرچه وراثتی است اما برگشت پذیر است و آن را به یک هدف درمانی تبدیل می‌کند. متیلاسیون DNA یک تغییر شیمیایی کووالانسی است که منجر به افزودن یک گروه متیل (CH3) در موقعیت کربن پنجم حلقه سیتوزین می‌شود[8] .

تنوع تعداد کپی (CNV[۲۱]): نشان دهنده کپی یا حذف توالی ژنومی بزرگتر از اندازه ۱ کیلو بایت است[9] .

در ادامه در فصل دوم به بیان مفاهیم پایه و در فصل سوم به مرور پژوهش‌های پیشین در زمینه استخراج اطلاعات بیماری با استفاده از یادگیری چند هسته‌ای می‌پردازیم. در فصل چهارم یکی از الگوریتم‌های ارائه شده در پژوهش‌های پیشین را پیاده سازی می‌کنیم و در فصل پنجم به نتیجه گیری می‌پردازیم.

[۲۰] DNA Methylation

[۲۱] Copy Number Variation

فصل ۲

مفاهیم پایه

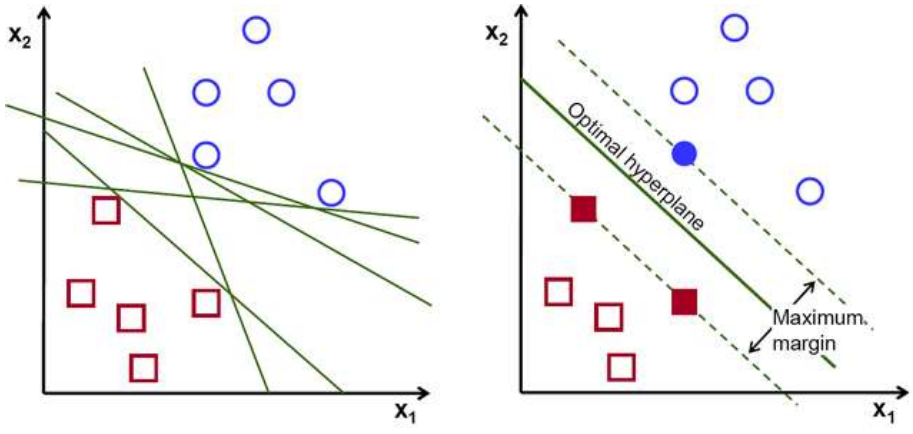
۲-۱ مقدمه

در این قسمت به بیان برخی از مفاهیم یادگیری ماشین می‌پردازیم. ابتدا ماشین بردار پشتیبان[۲۲] (SVM) و یادگیری چند هسته‌ای[۲۳] (MKL) را تعریف می‌کنیم و سپس روش‌های ارزیابی الگوریتم‌ها را بیان می‌کنیم.

۲-۲ ماشین بردار پشتیبان

هدف از الگوریتم ماشین بردار پشتیبان یافتن یک ابرصفحه در یک فضای N بعدی (N - تعداد ویژگی ها) است که به طور مشخص نقاط داده را طبقه بندی می‌کند. برای جدا کردن دو کلاس از نقاط داده بسیار ابرصفحه وجود دارد که می‌توانند انتخاب شوند. هدف ما یافتن ابرصفحه‌ای است که دارای حداکثر حاشیه باشد، یعنی فاصله بین نقاط داده هر دو کلاس حداکثر باشد (شکل ۲-۱ را ببینید). ابرصفحات مرزهای تصمیم گیری هستند که به طبقه‌بندی نقاط داده کمک می‌کنند. نقاط داده‌ای که در دو طرف ابرصفحه قرار دارند را می‌توان به طبقات مختلف نسبت داد. همچنین ابعاد ابرصفحه به تعداد ویژگی‌ها بستگی دارد. اگر تعداد ویژگی‌های ورودی ۲ باشد، ابرصفحه فقط یک خط است. اگر تعداد ویژگی های ورودی ۳ عدد باشد، آنگاه ابرصفحه، به صفحه‌ای دو بعدی تبدیل می‌شود. بردارهای پشتیبانی نقاط داده‌ای هستند که به ابرصفحه نزدیک‌تر هستند و موقعیت و جهت‌گیری ابرصفحه را تحت تأثیر قرار می‌دهند. با استفاده از این بردارهای پشتیبان، حاشیه طبقه‌بندی را به حداکثر می‌رسانیم [10].

بر اساس نظریه Vapnik-Chervonenkis (تئوری VC)، یک SVM از قانون به حداقل رساندن ریسک ساختاری (SRM) پیروی می‌کند که نه تنها خطای آموزش را به حداقل می رساند بلکه پیچیدگی ماشین یادگیری را محدود می‌کند، بنابراین توانایی‌های تعمیم را بهبود می‌بخشد [11].



شکل ۲-۱ ابرصفحه های ممکن برای تفکیک خطی مجموعه داده [10]

[۲۲] Support Vector Machine

[۲۳] Multiple Kernel Learning

می توان یک ابرصفحه را به صورت معادله (۱-۱) بیان کرد.

$$w.x + b = 0 \qquad (۱\ ۱)$$

جایی که . علامت ضرب است. w بردار نرمال است که به ابرصفحه عمود است. b بایاس ابرصفحه جدا کننده است. ما می‌خواهیم w و b را طوری انتخاب کنیم که بیشترین فاصله بین ابر صفحه‌های موازی که داده‌ها را از هم جدا می‌کنند، ایجاد شود. این ابرصفحه‌های موازی با استفاده از رابطه زیر توصیف می‌شوند.

$$w.x + b = 1 \quad \text{و} \quad w.x + b = -1 \qquad (۲\ ۱)$$

۲-۲-۱ حاشیه سخت

اگر داده‌های آموزشی جدایی پذیر خطی باشند، ما می‌توانیم دو ابر صفحه در حاشیه نقاط به‌طوری‌که هیچ نقطه مشترکی نداشته باشند، در نظر بگیریم و سپس سعی کنیم فاصله آن‌ها را حداکثر کنیم. با استفاده از هندسه، فاصله این دو صفحه $\frac{2}{\|w\|}$ است. بنابراین باید $\|w\|$ را مینیمم کنیم [12]. برای اینکه از ورود نقاط به حاشیه جلوگیری کنیم، شرایط زیر را اضافه می‌کنیم:

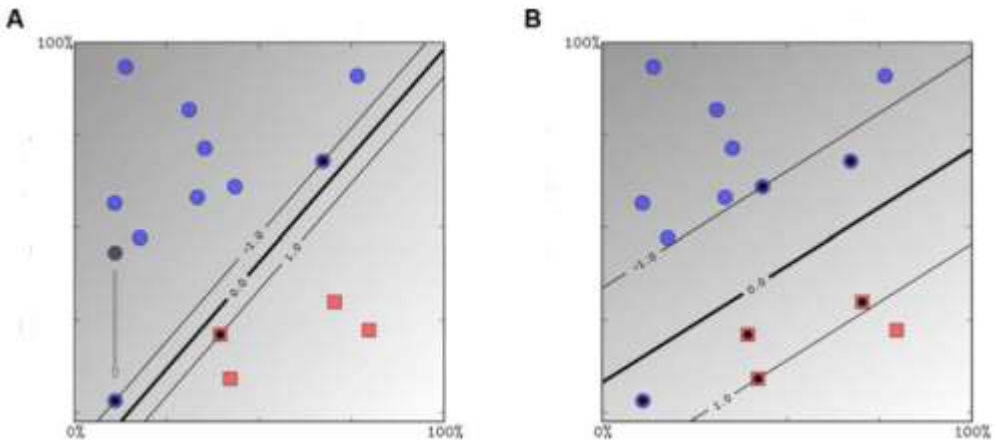
$$y_i(w.x_i + b) \geq 1, \qquad \forall\ 1 \leq i \leq n \qquad (۳\ ۱)$$

با کنار هم قرار دادن این دو یک مسئله بهینه‌سازی به دست می‌آید:

$$min\ \frac{1}{2}\|w\|^2 \quad , y_i(w.x_i + b) \geq 1, \qquad \forall\ 1 \leq i \leq n \qquad (۴\ ۱)$$

۲-۲-۲ حاشیه نرم

داده‌ها اغلب به صورت خطی قابل تفکیک نیستند و حتی اگر قابل تفکیک باشد، می‌توان با اجازه دادن به طبقه‌بندی کننده در طبقه‌بندی نادرست برخی از نقاط، یک حاشیه بزرگتر به دست آورد (شکل ۲-۲ را ببینید). نتایج تجربی نشان می‌دهد که حاشیه بزرگتر عملکرد بهتری نسبت به SVM با حاشیه سخت خواهد داشت [12].



شکل ۲-۲ تفاوت بین SVM با حاشیه سخت (A) و SVM با حاشیه نرم (B) [12]

برای اینکه به SVM اجازه خطا را بدهیم، قید زیر را با قید معادله ۱-۴ جایگزین می‌کنیم.

$$y_i(w.x_i + b) \geq 1 - \xi_i\,, \qquad \forall\, 1 \leq i \leq n \tag{۱-۵}$$

جایی که $\xi_i \geq 0$ یک متغیر slack است که اجازه می‌دهد یک نمونه در حاشیه قرار بگیرد و یا طبقه بندی نادرست داشته باشد. برای جلوگیری از استفاده بیش از حد متغیرهای slack تابع بهینه سازی را به صورت زیر اصلاح می‌کنیم.

$$min\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{۱-۶}$$

$$\text{s.t.} \quad \forall i \qquad y_i(w.x_i + b) \geq 1 - \xi_i \quad and \quad \xi_i \geq 0$$

C یک پارامتر مثبت از پیش تعریف شده بین سادگی مدل و خطای طبقه بندی است. به جای حل این مسئله بهینه سازی اولیه معمولا مسئله بهینه سازی دوگانه مربوطه را حل می‌کنیم. ابتدا تابع لاگرانژ را به صورت زیر می‌نویسیم [12].

$$\mathcal{L} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \; - \sum_{i=1}^{N}\alpha_i\big(y_i(w^T x_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{N}\beta_i\xi_i, \tag{۱-۷}$$

و با توجه به متغیرهای تصمیم گیری مسئله اولیه برای یافتن موارد زیر مشتقات را می‌گیریم.

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \Longrightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \qquad (۸\ ۱)$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + \beta_i \ \ \forall i.$$

سپس این موارد را دوباره به تابع لاگرانژ وصل می‌کنیم تا مقدار عینی مسئله دوگانه را پیدا کنیم که می‌تواند به صورت زیر نوشته شود:

$$minimize \quad -\sum_{i=1}^{N} \alpha_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j}$$

$$s.t. \qquad \forall i, \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \ \ and \ C \geq \alpha_i \geq 0 \qquad (۹\ ۱)$$

۲-۲-۳ماشین بردار پشتیبان براساس هسته

قضیه کاور گزاره‌ای در تئوری یادگیری محاسباتی است و یکی از انگیزه های نظری اولیه برای استفاده از روش‌های هسته غیر خطی در برنامه‌های یادگیری ماشین است. این قضیه بیان می‌کند که با توجه به مجموعه‌ای از داده های آموزشی که به طور خطی قابل تفکیک نیستند می‌توان با پیش بینی آن از طریق برخی تحولات غیرخطی، آن را به یک مجموعه آموزشی تبدیل کرد که به صورت خطی قابل تفکیک است. این قضیه به دلیل نظریه پرداز اطلاعات توماس ام. کاور که آن را در سال ۱۹۶۵ بیان کرد نامگذاری شده است.

داده های واقعی اغلب به طور خطی در فضای ورودی تفکیک نمی‌شوند. برای غلبه بر این مشکل، داده‌ها در یک فضای مشخصه با ابعاد بالا ترسیم می‌شوند که در آن داده‌ها کم و احتمالاً قابل تفکیک هستند. در عمل، نگاشت داده‌ها به صراحت انجام نمی‌شود. در عوض یک تابع هسته برای ساده کردن محاسبه مقدار ضرب داخلی داده‌های تبدیل شده در فضای ویژگی استفاده شده است. یعنی انتخاب یک تابع هسته به معنی تعریف نگاشت از فضای ورودی به فضای ویژگی است. برای این منظور در معادله ۲-۹ عبارت $x_i^T x_j$ را با تابع هسته $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ جایگزین می‌کنیم [13].

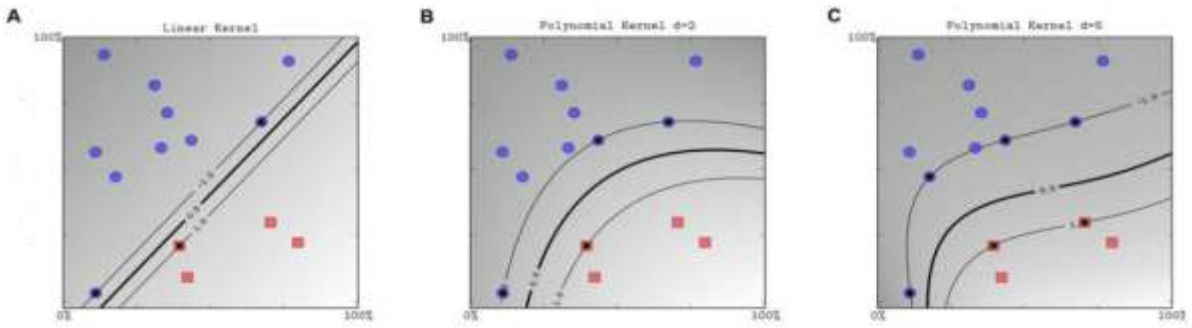
توابع هسته که به طور کلی استفاده می شوند:

- linear: $k(x, x') = \langle x, x' \rangle$;
- polynomial: $k(x, x') = (\langle x, x' \rangle + k)^d$;
- radial basis function (RBF): $k(x, x') = e^{-\gamma \|x - x'\|^2}, \gamma > 0$;
- sigmoid: $k(x, x') = tanh(\sigma \langle x, x' \rangle + r)$

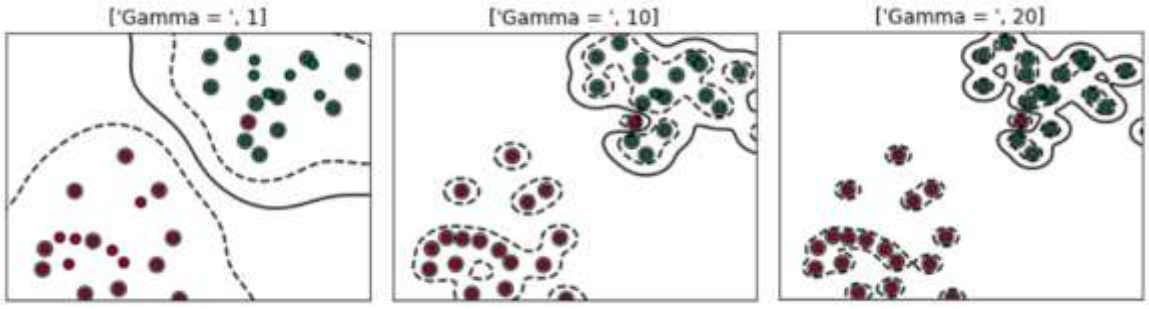
هسته چندجمله‌ای که نشان دهنده شباهت بردارها (نمونه های آموزشی) در یک فضای مشخصه نسبت به چند جمله‌ای‌های متغیرهای اصلی است و به این ترتیب امکان یادگیری مدل غیر خطی را فراهم می‌کند. در هسته چندجمله‌ای k اغلب ۰ (همگن) یا ۱ (ناهمگن) انتخاب می شود و پارامتر d درجه چندجمله‌ای می‌باشد که انعطاف پذیری چندجمله‌ای را کنترل می‌کند(شکل ۲-۳) [12].

هسته تابع پایه شعاعی (RBF) تابعی است که مقدار آن به فاصله از مبدا یا از نقطه ای بستگی دارد. γ پارامتر این هسته می‌باشد و صاف بودن مرز تصمیم گیری هسته شعاعی را کنترل می‌کند که اگر مقدار آن افزایش یابد مدل دچار Overfits می‌شود و اگر مقدار آن کاهش یابد مدل دچار Underfits می‌شود (شکل ۲-۴) [12].

هسته سیگموئید دو پارامتر σ و r می‌گیرد. برای $\sigma > 0$ می‌توانیم σ را به عنوان یک پارامتر مقیاس گذاری داده‌های ورودی و r را به عنوان یک پارامتر تغییر دهنده که آستانه نگاشت را کنترل می‌کند، در نظر بگیریم [14].



شکل ۲-۳ تاثیر پارامتر درجه در هسته چندجمله‌ای [12]



شکل ۲-۴ تاثیر پارامتر گاما در هسته RBF [15]

دانش ذاتی هر کار طبقه بندی با تعریف یک هسته مناسب بین اشیا به دست می‌آید که دو مزیت دارد: (۱) توانایی تولید مرزهای تصمیم گیری غیرخطی با استفاده از روش‌های طراحی شده برای طبقه‌بندی‌های خطی و (۲) امکان استفاده از طبقه‌بندی بر روی داده‌هایی که نمایانگر فضای بردار آشکاری نیستند به عنوان مثال توالی DNA / RNA یا توالی پروتئین و یا ساختارهای پروتئینی [12].

روش‌های هسته‌ای طی دهه گذشته شاهد افزایش محبوبیت زیادی در بیوانفورماتیک بوده اند. یکی از دلایل عمده محبوبیت روش‌های هسته در بیوانفورماتیک قدرت آن‌ها در ادغام داده است. دومین مزیت روش‌های هسته این است که می‌توان آن‌ها را به راحتی روی داده‌های ساخت یافته (مثال گراف ها، مجموعه ها، سری های زمانی و رشته ها) اعمال کرد [16].

۲-۳ یادگیری چند هسته‌ای

عملکرد پیش‌بینی SVM بسیار وابسته به عملکرد هسته استفاده شده است. روش استاندارد انتخاب بهترین عملکرد هسته در میان مجموعه ای از کاندیداها با استفاده از یک استراتژی اعتبارسنجی متقابل است. با این حال به جای انتخاب یک تابع هسته واحد، استفاده از یک ترکیب وزنی از هسته‌های ورودی ممکن است عملکرد پیش‌بینی بهتری داشته باشد که به MKL معروف است. الگوریتم های MKL ممکن است هسته‌های مختلف محاسبه شده در ورودی یکسان را ترکیب کنند یا هسته‌های محاسبه شده در ورودی مختلف را ترکیب کنند [17].

روش های مختلفی وجود دارد که از طریق آن‌ها می‌توان ترکیب هسته‌ها را انجام داد و هرکدام ویژگی‌های پارامتر ترکیبی خاص خود را دارند. ترکیب هسته‌ها را در الگوریتم‌های MKL موجود به سه دسته اساسی دسته بندی می‌کنند: ۱)روش های ترکیبی خطی محبوب‌ترین هستند و دارای دو دسته اساسی هستند: جمع بدون وزن (یعنی استفاده از مجموع یا میانگین هسته به عنوان هسته ترکیبی) و مجموع وزنی. ۲)روش های ترکیبی غیرخطی از توابع غیر خطی هسته ها یعنی ضرب، توان و نمایی استفاده می‌کنند. ۳)روش‌های ترکیبی وابسته به داده ها، وزن هسته خاصی را برای هر نمونه داده اختصاص می‌دهند. با این کار می‌توانند توزیع‌های محلی را در داده‌ها شناسایی کنند و قوانین مناسب ترکیب هسته را برای هر منطقه یاد بگیرند [17].

می‌توان الگوریتم های موجود MKL را از نظر روش آموزشی به دو گروه اصلی تقسیم کرد: ۱)روش‌های یک مرحله‌ای که هم پارامترهای تابع ترکیب و هم پارامترهای یادگیرنده مبتنی بر ترکیب را در یک گام محاسبه می‌کنند. می توان از یک رویکرد ترتیبی یا یک رویکرد همزمان استفاده کرد. در رویکرد پی در پی، ابتدا پارامترهای تابع ترکیب تعیین می‌شوند و سپس یک یادگیرنده مبتنی بر هسته با استفاده از هسته ترکیبی آموزش می‌یابد. در رویکرد همزمان، هر دو مجموعه پارامترها با هم یاد می گیرند. ۲)روش‌های دو مرحله ای از یک رویکرد تکراری استفاده می‌کنند که در هر تکرار ابتدا پارامترهای تابع ترکیب را هنگام تثبیت پارامترهای پایه یادگیرنده به‌روز

می‌کنند و سپس هنگام رفع پارامترهای تابع ترکیب، پارامترهای یادگیرنده پایه را به‌روز می‌کنند. این دو مرحله تا زمان همگرایی تکرار می‌شوند [17].

MKL می تواند دو کاربرد داشته باشد: (الف) هسته های مختلف با مفاهیم مختلفی از شباهت مطابقت دارند و به جای تلاش برای یافتن بهترین روش‌ها یک روش یادگیری برای ما انتخاب می‌کند یا ممکن است از ترکیبی از آن‌ها استفاده کند. استفاده از یک هسته خاص ممکن است منبع بایاس باشد و در صورت انتخاب یک یادگیرنده در بین مجموعه ای از هسته‌ها می‌توان راه حل بهتری پیدا کرد. (ب) هسته های مختلف ممکن است از ورودی‌هایی که از نمایش‌های مختلف (احتمالاً از منابع یا روش‌های مختلف) بدست می‌آیند استفاده کنند. از آنجا که این نمایش‌ها متفاوت هستند اندازه‌گیری‌های متفاوتی از شباهت مربوط به هسته های مختلف دارند. در چنین حالتی ترکیب هسته یکی از راه های ممکن برای ترکیب چندین منبع اطلاعاتی است [17].

۲-۴ ارزیابی

پس از ساخت طبقه‌بندی کننده برای مشخص شدن کارایی آن باید آن را با توجه به نتیجه‌ای که در برابر داده‌های آزمایش و آزمون تولید می‌کند سنجید. معیارهای مختلفی برای سنجش وجود دارد. با فرض اینکه طبقه‌بند برای تشخیص بیمار بودن یک فرد آموزش داده شده باشد، نتیجه‌ای که حاصل می‌شود یکی از چهار حالت زیر می‌شود.

۱. مثبت درست(TP[۲۴]): فرد بیمار باشد و طبقه‌بند بدرستی فرد را بیمار تشخیص دهد.
۲. منفی درست(TN[۲۵]): فرد بیمار نباشد و طبقه‌بند بدرستی فرد را سالم تشخیص دهد.
۳. مثبت نادرست(FP[۲۶]): فرد بیمار نباشد و طبقه‌بند به اشتباه فرد را بیمار تشخیص دهد.
۴. منفی نادرست(FN[۲۷]): فرد بیمار باشد و طبقه‌بند به اشتباه فرد را سالم تشخیص دهد.

با توجه به یک طبقه بندی کننده و مجموعه‌ای از نمونه‌ها (مجموعه آزمایش)، می‌توان یک ماتریس سردرگمی دو به دو ساخت که نمایانگر مجموعه چهار حالت ممکن است. ماتریس سردرگمی یک ماتریس N * N است (N = تعداد کلاس‌های هدف) شامل شماری از مقادیر پیش بینی شده و واقعی است که در ارزیابی عملکرد یک مدل طبقه بندی کمک می‌کند [18].

[۲۴] True Positive
[۲۵] True Negative
[۲۶] False Positive
[۲۷] False Negative

جدول ۲-۱ ماتریس سردرگمی یکی از معیارهای ارزیابی الگوریتم‌های دسته‌بندی می‌باشد.

		برچسب واقعی	
		بیمار	سالم
برچسب پیش‌بینی	بیمار	TP	FP
	سالم	FN	TN

در ادامه تعدادی از معیارهای ارزیابی را توضیح می‌دهیم.

دقت[۲۸] به این معناست که طبقه‌بندی کننده تا چه اندازه خروجی را درست پیش‌بینی می‌کند [19].

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

صحت یا درستی[۲۹] نشان می‌دهد که وقتی طبقه‌بندی کننده نتیجه را مثبت پیش‌بینی می‌کند تا چه اندازه درست است [19].

$$Precision = \frac{TP}{TP + FP}$$

فراخوانی[۳۰] یا حساسیت[۳۱] به احتمال مثبت شدن صحیح نتیجه آزمون وقتی که نمونه بیمار است، اشاره می‌کند که می‌توان آن را به صورت زیر به دست آورد. حساسیت را نرخ مثبت صحیح (TPR[۳۲]) نیز می‌نامند [19].

$$\text{Recall} = Sensitivity = \text{TPR} = \frac{TP}{TP + FN}$$

تشخیص‌پذیری یا ویژگی[۳۳] نیز به احتمال منفی شدن صحیح نتیجه آزمون وقتی که نمونه سالم (فاقد بیماری) است، اشاره می‌کند. تشخیص‌پذیری یا ویژگی را می‌توان به صورت زیر به دست آورد. تشخیص‌پذیری یا ویژگی را نرخ منفی صحیح (TNR[۳۴]) نیز می‌گویند [19].

$$\text{Specificity} = \text{TNR} = \frac{TN}{TN + FP}$$

---

۲۸ Accuracy
۲۹ Precision
۳۰ Recall
۳۱ Sensitivity
۳۲ True Positive Rate
۳۳ Specificity
۳۴ True Negative Rate

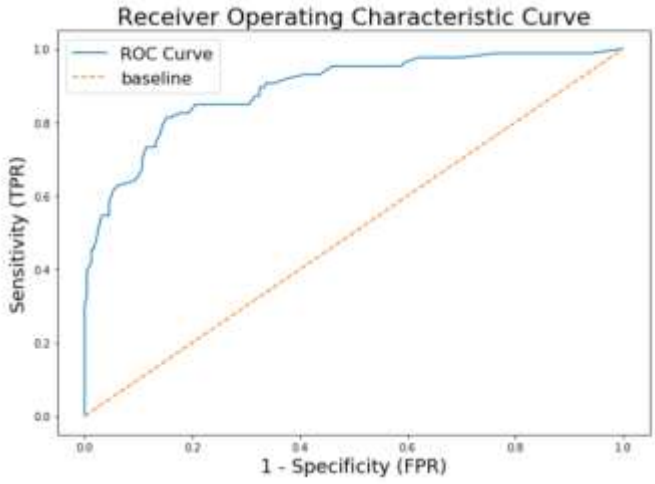
متمم تشخیص‌پذیری یا ویژگی (ویژگی − ۱) می‌باشد. این مفهوم به احتمال مثبت شدن نادرست نتیجه آزمون وقتی که نمونه سالم (فاقد بیماری) است، اشاره می‌کند [19]. این مفهوم را نرخ مثبت غلط و یا به اختصار FPR[۳۵] نیز می‌گویند که مقدار آن از رابطه زیر به دست می‌آید:

$$1 - \text{Specificity} = \text{FPR} = \frac{FP}{TN + FP}$$

ضریب همبستگی متیوس[۳۶] (MCC) خلاصه بهتری از عملکرد الگوریتم‌های طبقه‌بندی را نشان می‌دهد [19].

$$\text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

منحنی مشخصه عملکرد سیستم (ROC[۳۷]) یک تکنیک برای تجسم، سازماندهی و انتخاب طبقه‌بندی‌ها بر اساس عملکرد آن‌ها است. تجزیه و تحلیل ROC برای استفاده در تجسم و تجزیه و تحلیل رفتار سیستم های تشخیصی گسترش یافته است. جامعه تصمیم گیری پزشکی دارای مقالات گسترده ای در مورد استفاده از نمودارهای ROC برای آزمایش تشخیصی است. نمودارهای ROC نمودارهای دو بعدی هستند که در آنها حساسیت بر روی محور Y و متمم ویژگی در محور X رسم می شود. نمودار ROC معاملات نسبی بین مزایا (مثبت درست) و هزینه‌ها (مثبت نادرست) را نشان می‌دهد. اگر نقاط بالای خط نیمساز قرار گیرند برای ما مطلوب است. قرار گرفتن نقاط بر روی خط نیمساز نشانه تصادفی بودن نتایج است. اگر نقاط در رو یا زیر خط نیمساز قرار بگیرند یعنی طبقه‌بند نامناسب است [18].



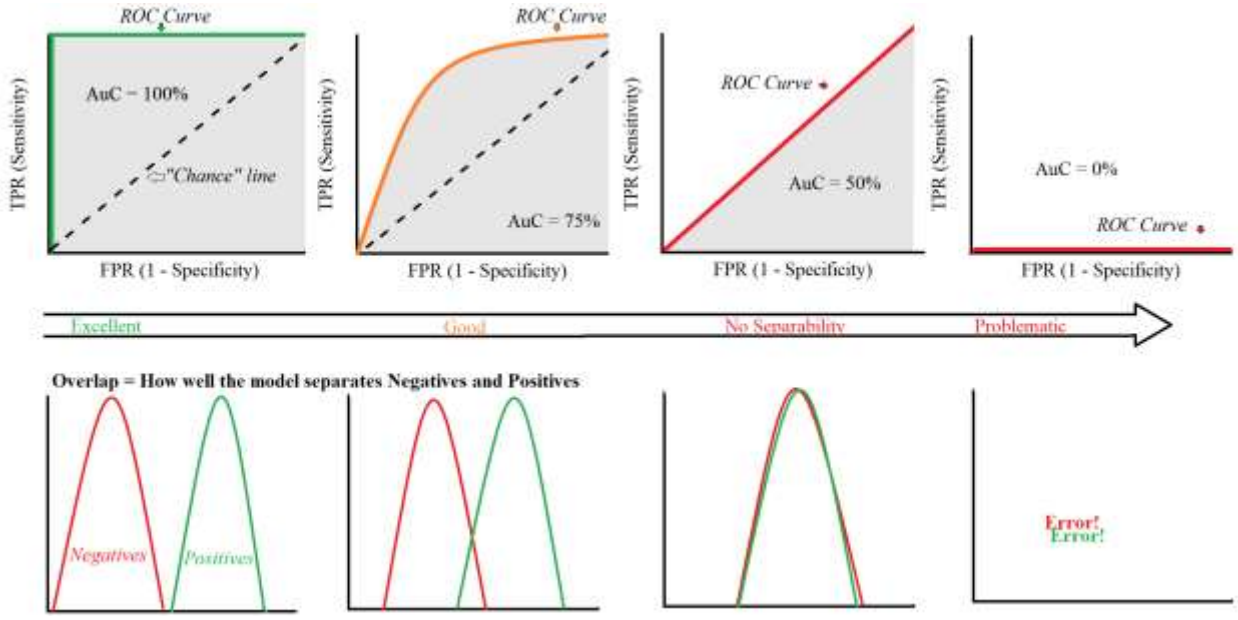
شکل ۲-۵ منحنی مشخصه عملکرد سیستم (ROC) [20]

[۳۵] False Positive Rate

[۳۶] Matthews Correlation Coefficient

[۳۷] Receiver Operating Characteristic

منحنی ROC تصویری دو بعدی از عملکرد طبقه‌بندی کننده است. برای مقایسه طبقه‌بندی‌ها ممکن است بخواهیم عملکرد ROC را به یک مقدار مقیاسی واحد نشان دهیم که عملکرد مورد انتظار را نشان می‌دهد. یک روش معمول برای محاسبه مساحت زیر منحنی ROC، AUC[۳۸] می‌باشد. AUC مقداری بین ۰ تا ۱ می‌گیرد. هر چه مقدار AUC بیشتر باشد یعنی عملکرد طبقه‌بند مناسب‌تر است (شکل ۲-۶) [18].



شکل ۲-۶ معیار ارزیابی سطح زیر منحنی ROC (AUC) [21]

علاوه بر معیارهای ارزیابی نامبرده شده، از معیار دیگری به نام مقدار احتمال(p-value) برای مشخص شدن پایداری نتیجه حاصل از الگوریتم استفاده می‌شود. احتمال رد فرض صفر به شرط آنکه فرض صفر صحیح باشد مقدار احتمال نامیده می‌شود. احتمال مشاهده اثر (E) در هنگام درست بودن فرضیه صفر (H_0) است.

$$p - value = P(E \mid H_0)$$

بنابراین وقتی مقدار p به اندازه کافی (α) کم باشد فرضیه صفر را رد می‌کنیم و نتیجه می‌گیریم که اثر مشاهده شده پایدار است. مقدار p بر اساس فرض درست بودن H0 محاسبه می‌شود، نمی تواند اطلاعاتی راجع به درست بودن H0 ارائه دهد. این استدلال همچنین نشان می‌دهد که اولاً، p نمی تواند احتمال درست بودن فرضیه جایگزین باشد. ثانیاً، مقدار p بسیار به اندازه نمونه بستگی دارد [22].

[۳۸] Area Under the Curve

فصل ۳

مرور پژوهش‌های پیشین

۳-۱ مقدمه

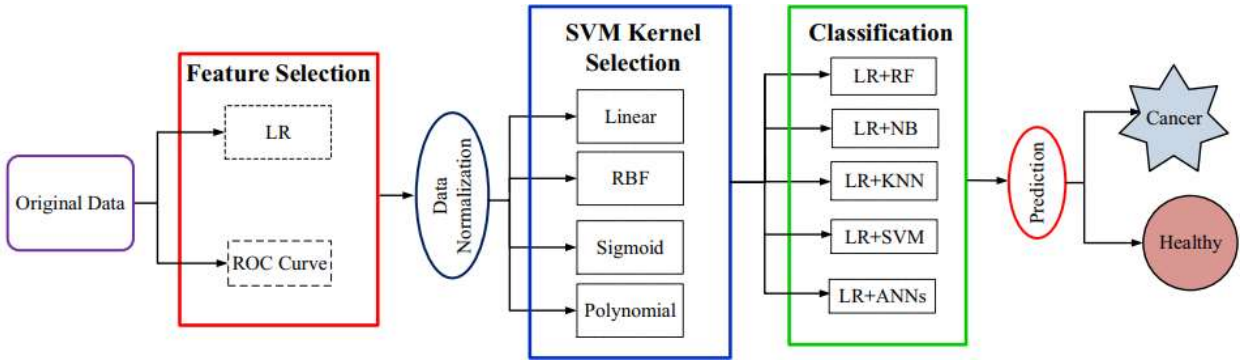
در این فصل مطالعات انجام شده را در دو بخش یادگیری ماشین و یادگیری چند هسته‌ای بیان خواهیم کرد.

۳-۲ مرور مطالعات مبتنی بر ماشین بردار پشتیبان

در این قسمت به بررسی کارهای انجام شده بر اساس ماشین بردار پشتیبان با کاربرد پیش‌بینی بیماری و مرحله آن می‌پردازیم. منظور از مرحله بیماری میزان پیشرفت بیماری می‌باشد.

Zhao و همکارانش در سال ۲۰۱۹ یک مدل یکپارچه طبقه بندی سرطان روده بزرگ با استفاده از ماشین بردار پشتیبان پیشنهاد کردند [23]. شکل ۳-۱ الگوریتم پیشنهاد شده در این مقاله را نشان می‌دهد.

در این مطالعه یک مدل یکپارچه مبتنی بر رگرسیون لجستیک (LR) و ماشین بردار پشتیبانی (SVM) برای طبقه بندی سرطان روده بزرگ (CRC[۳۹]) به نمونه های سرطانی و طبیعی پیشنهاد شد.



شکل ۳-۱ مدل پیشنهادی Zhao و همکارانش برای طبقه بندی سرطان روده بزرگ [23]

در تجزیه و تحلیل عوامل CRC مرتبط، ۸۹ فرد سالم و ۹۲ بیمار CRC را انتخاب کردند. همه شرکت کنندگان ساکنان محلی فرانسه و آلمان بودند. اطلاعات شامل محل، سن و جنسیت، نوع تومور، درجه تومور و DNA کلیه شرکت کنندگان بود.

از رگرسیون لجستیک (یک مدل آماری رگرسیون برای متغیرهای وابسته دوسویی مانند بیماری یا سلامت، مرگ یا زندگی است. روشی برای طبقه بندی بردار ورودی داده شده به یکی از دو کلاس است.) برای یافتن شاخص‌ترین شاخصه های بیماری ($p<0.05$) با غربالگری هر یک از ویژگی‌ها برای کاهش ویژگی‌های زائد با توجه به p-value استفاده شد. احتمالات پیش‌بینی شده توسط رگرسیون لجستیک با استفاده از منحنی ROC تجزیه و تحلیل شد. اگر مقدار AUC (سطح زیر منحنی ROC) کمتر از ۰,۵ باشد، این نشان می‌دهد که هر نمایش کاملا اتفاقی است

[۳۹] Colorectal cancer

و مقداری نزدیک به ۱ نشان می‌دهد که ویژگی به شدت دو گروه را تفکیک می‌کند. پس از اندازه گیری با این روش‌ها، شاخص ترین ویژگی‌ها به عنوان ورودی SVM انتخاب شدند.

گام نرمال سازی داده ها به چند دلیل مورد استفاده قرار می‌گیرد: (۱) برای جلوگیری از غلبه ویژگی ها با محدوده عددی بزرگتر بر ویژگی‌ها با محدوده های عددی کوچکتر، (۲) برای جلوگیری از مشکلات عددی در هنگام محاسبه و (۳) برای اینکه طبقه‌بندی کننده دقت بالاتری به دست آورد. هر ویژگی به محدوده [۰،۱] با روش (۱-۱۰) مقیاس داده شد.

$$V' = \frac{V - MIN}{MAX - MIN} \qquad (۱۰\ ۱)$$

بطوریکه V مقدار اصلی، Min و Max مرز پایین و بالایی مقدار ویژگی و V' مقدار مقیاس شده است.

در این مطالعه از SVM به عنوان طبقه‌بندی کننده استفاده شد. هسته‌های Sigmoid، RBF، Linear و Polynomial با یکدیگر مقایسه شدند و هسته RBF با بهترین عملکرد انتخاب شد (جدول ۳-۱).

جدول ۳-۱ نتایج حاصل از اعتبارسنجی ۵ قسمتی برای انواع مختلف هسته [23]

Prediction models	Positives			Negatives			ACC (%)	MCC (%)
	TP	FN	SN (%)	TN	FP	SP (%)		
Linear	75	17	81.5	80	9	89.9	85.6	65.63
RBF	80	12	87.0	83	6	93.3	90.1	80.30
Sigmoid	80	12	87.0	81	8	91.0	89.0	77.99
Polynomial	75	17	81.5	79	10	88.8	85.1	70.41

در انتها الگوریتم SVM با برخی الگوریتم های رایج طبقه‌بندی از جمله RF، NB، KNN و ANN مقایسه شد و SVM با دقت ۹۰,۱٪ بهترین عملکرد را به دست آورد (جدول ۳-۲).

جدول ۳-۲ نتایج حاصل از اعتبارسنجی 5-fold از مدل‌های پیش‌بینی [23]

Prediction models	Positives			Negatives			ACC (%)	MCC (%)
	TP	FN	SN (%)	TN	FP	SP (%)		
LR + RF	78	14	84.8	80	9	89.9	87.3	74.72
LR + NB	70	22	76.1	76	13	85.4	80.7	63.55
LR + KNN	75	17	81.5	78	11	87.6	84.5	69.24
LR + SVM	80	12	87.0	83	6	93.3	90.1	80.30
LR + ANNs	80	12	87.0	80	9	89.9	88.4	76.84

Bhalla و همکارانش در سال ۲۰۱۷ به دنبال شناسایی ژن‌هایی بودند که به کمک آن‌ها مرحله بیماری سرطان سلول کلیوی را تشخیص دهند [24].

مجموعه داده های بیان RNA-seq از نوع سرطان KIRC[۴۰] یعنی کارسینومای سلول شفاف کلیه از پورتال داده TCGA به دست آمد. داده های بیان RNA-seq برای ۵۲۳ نمونه شامل داده‌های بیان ژن برای ۲۰۵۳۴ ژن در دسترس بود. بیماران مرحله اول و دوم را بیماران مرحله اولیه و بیماران مرحله سوم و چهارم را بیماران مرحله انتهایی تعریف کردند. در این مطالعه از ۸۰٪ نمونه ها (۴۱۹ بیمار) برای آموزش و آزمایش استفاده شد که مجموعه داده های آموزشی نامیده می‌شود. از ۲۰٪ نمونه (۱۰۴ بیمار) برای اعتبارسنجی خارجی استفاده شد که به عنوان مجموعه داده‌های اعتبار سنجی مستقل یا خارجی نامیده می‌شود.

پردازش داده ها. مقادیر بیان ژن دارای دامنه وسیعی از تنوع است. بنابراین مقادیر RSEM[۴۱] (میزان بیان ژن تخمین زده شده از داده‌های RNA-seq) را پس از افزودن ۱ به عنوان یک عدد ثابت با استفاده از log2 تبدیل کردند. قبل از نرمال سازی داده‌ها، ویژگی‌ها با واریانس کمتر از ۰,۲۵ را حذف کردند. پس از حذف ویژگی‌ها با واریانس کم، ۲۰۵۳۴ ویژگی به ۱۹۱۶۶ ویژگی کاهش یافت. پس از آن مقادیر RSEM تبدیل شده log2 را برای هر ژن نرمال کردند و آن را به Z-score تبدیل کردند. برای محاسبه تحول و نرمال سازی از معادلات (۱-۱۱) و (۱-۱۲) استفاده شده است:

$$x = log_2(RSEM + 1) \qquad (۱۱ ۱)$$

$$\text{Z\_score} = \frac{x - \bar{x}}{s} \qquad (۱۲ ۱)$$

بطوریکه Z_score نمره نرمال شده است، x عبارت ژن تبدیل شده به سیستم log، $\bar{x}$ عبارت متوسط بیان ژن در مجموعه داده های آموزشی است و s انحراف معیار ژن در مجموعه داده آموزش است. از میانگین و انحراف معیار ویژگی‌های آموزش برای عادی سازی مجموعه داده‌های اعتبار سنجی استفاده شد.

مدل مبتنی بر آستانه. در این مطالعه از روش ساده‌ای برای تمایز مرحله اولیه و انتهایی سرطان با استفاده از بیان یک ژن استفاده شده است. این روش مبتنی بر این واقعیت است که ژن‌های کمی در مراحل مختلف سرطان به طور متفاوت بیان می‌شوند. در این روش برای هر ژن یک آستانه انتخاب کردند که تعیین می‌کند نمونه با توجه به بیان آن ژن در مرحله اولیه یا مرحله انتهایی باشد. اگر ژن در مرحله اولیه بیش از حد بیان شود یعنی میانگین

[۴۰] Kidney Renal Clear Cell Carcinoma

[۴۱] RNA-Seq by Expectation Maximization

بیان نرمال شده آن در مرحله اولیه بیشتر از مرحله انتهایی در داده‌های آموزشی باشد و برای یک نمونه داده شده بیان نرمال آن بیش از آستانه باشد، آن نمونه را به عنوان مرحله اولیه طبقه‌بندی می‌کنند در غیر این صورت آن را به عنوان مرحله انتهایی طبقه بندی می‌کنند. در حالی که اگر ژن در مرحله انتهایی، بیان بیش از حد داشته باشد یعنی میانگین بیان نرمال شده آن در مرحله اولیه نسبت به مرحله انتهایی در داده های آموزشی کمتر باشد و برای یک نمونه مشخص بیان نرمال شده آن بیش از آستانه باشد آن نمونه را به عنوان مرحله انتهایی طبقه‌بندی می‌کنند در غیر اینصورت آن را به عنوان مرحله اولیه طبقه بندی می‌کنند. علاوه بر این برای بهینه سازی آستانه دستیابی به بهترین عملکرد از روش تکرار استفاده شد که در آن آستانه به طور سیستماتیک برای طیف وسیعی از مقادیر بیان نرمال شده در تمام نمونه‌های یک ژن خاص افزایش یا کاهش می‌یابد. برای هر ژن، آن آستانه‌ای انتخاب شد که حداکثر عملکرد طبقه بندی را از نظر سطح زیر منحنی مشخصه گیرنده (ROC) ارائه می دهد. در این مطالعه، این رویکرد را به عنوان رویکرد آستانه و این مدل‌ها را به عنوان مدل‌های مبتنی بر آستانه می‌نامند.

به منظور رتبه بندی ژن‌ها، یک مدل مبتنی بر آستانه برای هر ژن ایجاد کردند و ۵۰ ژن برتر را بر اساس مقدار ROC مدل آنها شناسایی کردند. سپس ماتریس همبستگی برای ۵۰ ژن را محاسبه کردند و در هر ترکیبی از ژن با همبستگی بیشتر از ۰,۶ ، ژن با ROC کمتر حذف شد و ۲۸ ژن از ۵۰ ژن باقی ماند. از این ۲۸ ژن به عنوان ویژگی ورودی برای ایجاد مدل‌های یادگیری ماشین برای تمایز مرحله اولیه و انتهایی سرطان استفاده کردند. همانطور که در جدول ۳-۳ ذکر شده، مدل SVM در مجموعه داده آموزش با دقت ٪۷۳,۲۷ و ٪۷۸= ROC و همچنین در مجموعه داده اعتبارسنجی با دقت ٪۷۱,۱۵ و ۰,۷۷= ROC بهترین عملکرد را بدست آورد. از هسته RBF برای SVM استفاده کردند.

جدول ۳-۳ عملکرد مدل‌های طبقه بندی بر اساس ژن های انتخابی با استفاده از تکنیک‌های مختلف یادگیری ماشین [24]

Technique	Dataset	Performance Measures				
		Sensitivity	Specificity	Accuracy (%)	MCC	ROC
RF	Training	73.62	72.12	73.03	0.45	0.77
	Validation	73.02	60.98	68.27	0.34	0.74
Naive Bayes	Training	75.98	67.27	72.55	0.43	0.76
	Validation	77.78	60.98	71.15	0.39	0.76
SMO	Training	83.86	55.76	72.79	0.42	0.70
	Validation	80.95	53.66	70.19	0.36	0.67
J48	Training	64.17	66.06	64.92	0.3	0.67
	Validation	68.25	58.54	64.42	0.26	0.67
SVM	Training	75.98	69.09	73.27	0.45	0.78
	Validation	74.6	65.85	71.15	0.4	0.77

در تجزیه و تحلیل بعدی، ۲۸ ژن فوق را به دو گروه جدا کردند. (۱) گروه A حاوی ۱۶ ژن که در مرحله اولیه بیماری بیش از حد بیان شده‌اند و (۲) گروه B شامل ۱۲ ژن که در مرحله انتهایی بیماری بیش از حد بیان می‌شوند. در مرحله بعدی، مدل‌های مبتنی بر آستانه را با استفاده بیش از دو ژن ایجاد کردند و بهترین مجموعه ژن‌ها را از گروه A و B شناسایی کردند. برای این منظور، تجزیه و تحلیل ژن‌های گروه A را انجام دادند، جایی که ژن دارای رتبه برتر بیان با ۱۵ ژن باقیمانده به روشی تکراری بهترین جفت ژن را شناسایی کرد. این بهترین جفت ژن سپس با ژن‌های دیگر یکی یکی ترکیب می‌شود تا سه ژن برتر و غیره شناسایی شود. سرانجام چهار ژن برتر (setA-1) را از ژن‌های گروه A بدست آوردند. همین تمرین برای ژن‌های گروه B که چهار ژن برتر (setB-1) را ارائه می‌دهند نیز تکرار شد. فقط چهار ژن را انتخاب کردند زیرا با افزایش تعداد ژن‌ها عملکرد بیشتر نشد. علاوه بر این انواع مختلفی از مدل‌های پیش‌بینی را با استفاده از setA-1 ایجاد کردند و به ROC = ۰,۷۶ برای مدل های SVM در مجموعه داده های آموزشی دست یافتند. همچنین عملکرد را روی یک مجموعه داده مستقل ارزیابی کردند و عملکرد مشابه ROC =۰,۸ را بدست آوردند. به طور مشابه برای setB-1 حداکثر ROC= ۰,۷۴ را در مجموعه داده آموزش و اعتبارسنجی بدست آوردند. ترکیب setA-1 و setB-1 (Combo-1) با ROC =۰,۷۷ و 0.80 به ترتیب در مجموعه داده آموزش و مجموعه داده اعتبار سنجی خارجی بدست آمد (جدول ۳-۴).

جدول ۳-۴ عملکرد مدل های مبتنی بر ماشین بردار پشتیبانی (SVM) و جنگل تصادفی (RF) با استفاده از مجموعه‌های مختلفی از ویژگی های انتخاب شده در آموزش و مجموعه داده‌های اعتبار سنجی مستقل یا خارجی توسعه یافته است [24]

Features	Dataset	Technique	Performance Measures				
			Sensitivity	Specificity	Accuracy (%)	MCC	ROC
setA-1 (4 genes)	Training	SVM	71.65	70.3	71.12	0.41	0.76
	Validation		68.25	78.05	72.12	0.45	0.80
	Training	RF	70.87	65.45	68.74	0.36	0.69
	Validation		73.02	58.54	67.31	0.32	0.74
setB-1 (4 genes)	Training	SVM	71.26	70.3	70.88	0.41	0.74
	Validation		74.6	68.29	72.12	0.42	0.74
	Training	RF	80.31	49.7	68.26	0.32	0.65
	Validation		82.54	51.22	70.19	0.36	0.68
Combo-1 (8 genes)	Training	SVM	75.20	70.30	73.27	0.45	0.77
	Validation		77.78	68.29	74.04	0.46	0.80
	Training	RF	81.1	55.15	70.88	0.38	0.73
	Validation		82.54	51.22	70.19	0.36	0.74

به جای ترکیب دو ژن با استفاده از میانگین ساده (همانطور که در بخش قبلی نشانگرهای زیستی چندین ژن انجام شد) در اینجا مدل های مبتنی بر SVM را با استفاده از بهترین دو ژن، سه ژن، چهار ژن و غیره توسعه

دادند تا حداقل تعداد ویژگی‌های به دست آوردن بهترین مدل را شناسایی کنند. ابتدا ژن‌های گروه A برای تولید مدل SVM مبتنی بر دو ژن مورد استفاده قرار گرفتند و سپس بهترین جفت ژن را برای توسعه مدل SVM کشف کردند. به همین ترتیب، ژن سوم را با بهترین جفت ژن جستجو کردند که بهترین مدل SVM را ارائه می‌دهد. این فرآیند تا زمان اشباع عملکرد تکرار شد و با استفاده از پنج ژن (setA-2) بهترین مدل SVM را بدست آوردند. روش مشابهی برای ژن‌های گروه B تکرار شد تا پنج ژن برتر (setB-2) را پیدا کنند. مدل SVM با استفاده از بهترین مجموعه setA-2 با حداکثر ROC=۰,۷۵ با دقت ۷۰,۸۸٪ در مجموعه آموزش و دقت ۶۷,۳۱٪ و ROC=۰,۷۵ در مجموعه داده اعتبار سنجی بدست آورد. همانطور که در جدول ۳-۵ ذکر شده است، مدل SVM براساس setB-2 حداکثر دقت ۶۹,۶۹٪ و ROC=۰,۷۶ در مجموعه داده های آموزشی و دقت ۶۴,۴۲٪ و ROC=۰,۷۲ در مجموعه داده های اعتبار سنجی را نشان می‌دهد. به منظور افزایش عملکرد، بهترین ژن‌های گروه A و B را ترکیب کردند و ده ژن برتر را ترکیب کردند (Combo-2). مدل SVM بر اساس این ده ژن دارای حداکثر دقت ۷۲,۶۲٪ و ROC=۰,۷۸ در داده های آموزش و دقت ۷۰,۱۹٪ با ROC=۰,۷۷ در مجموعه داده‌های اعتبارسنجی است (جدول ۳-۵).

جدول ۳-۵ عملکرد مدل های بردار پشتیبانی (SVM) و جنگل تصادفی (RF) با استفاده از مجموعه‌های مختلفی از ویژگی‌های انتخاب شده از طریق روش SVM در آموزش و مجموعه داده‌های اعتبار سنجی مستقل [24]

Features	Dataset	Technique	Performance Measures				
			Sensitivity	Specificity	Accuracy (%)	MCC	ROC
setA-2 (5 genes)	Training	SVM	68.9	73.94	70.88	0.42	0.75
	Validation		65.08	70.73	67.31	0.35	0.75
	Training	RF	81.5	56.97	71.84	0.4	0.73
	Validation		77.78	46.34	65.38	0.25	0.68
setB-2 (5 genes)	Training	SVM	68.9	70.91	69.69	0.39	0.76
	Validation		60.32	70.73	64.42	0.3	0.72
	Training	RF	71.65	64.85	68.97	0.36	0.71
	Validation		69.84	56.1	64.42	0.26	0.70
Combo-2 (10 genes)	Training	SVM	72.44	72.89	72.62	0.45	0.78
	Validation		71.43	68.29	70.19	0.39	0.77
	Training	RF	76.19	65.85	72.12	0.42	0.76
	Validation		70.47	70.3	70.41	0.4	0.76

به طور گسترده‌ای بررسی شده است که تفاوت‌های خاص جنسیتی در ایجاد و بقای تومورهای مختلف وجود دارد. مدل‌های خاص جنسیتی را برای طبقه بندی بیماری توسعه دادند. ابتدا مدل‌های خاص مرد با استفاده از ۸۰٪ نمونه های مرد (۱۵۹ مرحله اولیه و ۱۰۹ مرحله آخر) ساخته شد. برای انتخاب ۶۴ ویژگی از روش انتخاب ویژگی مبتنی بر Weka (بسته نرم‌افزاری) استفاده شد. این ویژگی‌ها برای توسعه مدل‌ها مورد استفاده قرار گرفتند و

حداکثر ROC ٪۸۷ با دقت ٪۸۰,۲۲ در مجموعه داده های آموزشی و ٪۷۷,۱۴ دقت با ROC ٪۸۰,۰ در مجموعه داده های اعتبار سنجی (۴۱ نمونه اولیه و ۲۹ نمونه مرحله آخر) با استفاده از SVM به دست آوردند (جدول ۳-۶). سپس مدل‌های طبقه بندی بر روی نمونه های زن (۹۳ مرحله اولیه و ۵۴ مرحله آخر) که ROC ٪۹۰,۰ با دقت ٪۸۵,۷۱ با استفاده از SVM حاصل شد، آموزش دیدند. دقت ٪۹۵,۷۸ و ROC ٪۸۲,۰ در مجموعه داده‌های اعتبارسنجی زنان (۲۴ نمونه مرحله اولیه و ۱۴ نمونه مرحله آخر) به دست آمد(جدول ۳-۶). این نتایج نشان می‌دهد که مدل‌های خاص جنسیتی می‌توانند بهتر باشند.

جدول ۳-۶ عملکرد مدل‌های بردار پشتیبان (SVM) و مدل‌های جنگل تصادفی (RF) بر اساس جنسیت با استفاده از ژن‌ها/ویژگی‌های منتخب Weka در آموزش و مجموعه داده‌های اعتبار سنجی مستقل [24]

Gender	Technique	Dataset	Performance Measures				
			Sensitivity	Specificity	Accuracy (%)	MCC	ROC
Female	RF	Training	87.1	88.89	87.76	0.75	0.93
		Validation	75	71.43	73.68	0.45	0.76
	SVM	Training	89.25	79.63	85.71	0.69	0.90
		Validation	75	85.71	78.95	0.59	0.82
Male	RF	Training	83.02	73.39	79.10	0.57	0.83
		Validation	75.61	58.62	68.57	0.35	0.72
	SVM	Training	83.02	76.15	80.22	0.59	0.87
		Validation	78.05	75.86	77.14	0.53	0.80

۳-۳ مرور مطالعات مبتنی بر یادگیری چند هسته‌ای

در ادامه به بررسی کارهای انجام شده بر اساس یادگیری چند هسته‌ای با کاربردهای مختلف از جمله پیش‌بینی بیماری، مرحله و زیرگروه‌های آن و همچنین بقای بیمار می‌پردازیم.

Du و همکارانش در سال ۲۰۱۷ یک روش انتخاب ویژگی دو مرحله‌ای با استفاده از یادگیری چند هسته‌ای پیشنهاد کردند و اثر بخشی روش خود را در پیش‌بینی بیماری بیان کردند [25].

روش پیشنهادی Simple MKL-Feature Selection (SMKL-FS) نامیده می‌شود. برای داده‌های بیان مجموعه‌ای از ویژگی‌ها چهار دسته اصلی ویژگی وجود دارد: ویژگی‌های مرتبط[۴۲]، ویژگی‌های زائد[۴۳]، ویژگی‌های بی ربط[۴۴] و ویژگی‌های نویز[۴۵]. بیشتر ویژگی‌ها، ویژگی‌های نامربوطی هستند که ابتدا توسط بسیاری از روش‌های انتخاب ویژگی برای تجزیه و تحلیل داده های بیان حذف می شوند. در مرحله اول روش ارائه شده، ویژگی‌های

[۴۲] Relevant features
[۴۳] Redundant features
[۴۴] Irrelevant features
[۴۵] Noisy features

مرتبط توسط اندازه گیری امتیاز هر ویژگی با استفاده از فرآیند بهینه‌سازی MKL شناسایی می‌شوند. اگر پیچیدگی محاسباتی در نظر گرفته شود می‌توان مجموعه کوچکی از ویژگی‌های مرتبط را در مرحله اول انتخاب کرد. در مرحله دوم، یک طرح انتخاب تعبیه شده[۴۶]، یعنی انتخاب رو به جلو، برای جستجوی زیرمجموعه ویژگی‌های فشرده از مجموعه ویژگی‌های کاندید به دست آمده در مرحله اول اعمال می‌شود.

ابتدا MKL برای انتخاب مجموعه ویژگی‌های مرتبط اعمال می‌شود. برای MKL، روش SimpleMKL انتخاب می‌شود تا ضریب dm از ترکیب هسته بدست آید. SimpleMKL برای انجام بهینه سازی روی پارامترهای SVM (α_i) و ضرایب هسته (d_m) از فرایند گرادیان کاهشی تکرار شونده (یک الگوریتم تکراری مرتبه اول که مینیمم محلی را در یک تابع مشتق پذیر پیدا می‌کند) استفاده می‌کند. تابع هدف بهینه به شرح زیر تعریف می شود:

$$J = \min_{d_m} \max_{\alpha} W(\alpha, d_m) \qquad s.t. \quad \sum_{m=1}^{M} d_m = 1, \qquad d_m \geq 0 \qquad (۱۳\ ۱)$$

با استفاده از SimpleMKL می‌توان مقدار J را برای هر ویژگی از مجموعه ویژگی S در فرآیند بهینه‌سازی $W(\alpha, d_m)$ با $\min_{d_m} \max_{\alpha} W(\alpha, d_m)$ بدست آورد. برای انتخاب مجموعه ویژگی‌های مرتبط، لیست J برای ویژگی‌ها محاسبه می‌شود تا ارتباط بین ویژگی‌ها و نمونه‌ها اندازه گیری شود. در آخر، لیست J را صعودی مرتب می‌کنند و لیست ویژگی‌های رتبه‌بندی شده S_r را بدست می‌آورند. سپس n^* از ویژگی‌های برتر انتخاب شده و مجموعه ویژگی S_{n^*} بدست می‌آید. شکل ۳-۲ فرایند انتخاب ویژگی مرتبط را نشان می‌دهد.

Selecting the Relevant Feature Algorithm :

Input:
Training examples: $X_0 = [x_1, x_2, \dots, x_i, \dots, x_N]^T$
Class label: $y = [y_1, y_2, \dots, y_i, \dots, y_N]^T$
List of features: $S = [1,2, \dots f]$
Relevant feature number: n^*
Initialize:
for each feature *f* of *S* **do**
compute list of J by using SimpleMKL
end for
sort J ascend and get the ranked features list S_r
select top n^* features, obtained feature set S_{n^*}
Output:
Feature set S_{n^*}

[۴۶] Embedded selection scheme

شکل ۳-۲ شبه کد الگوریتم انتخاب ویژگی‌های مرتبط [25]

در مرحله دوم انتخاب رو به جلو برای جستجوی زیر مجموعه ویژگی‌های فشرده از مجموعه ویژگی‌های کاندید به دست آمده در مرحله اول اعمال می‌شود. برای جستجوی زیر مجموعه ویژگی فشرده از مجموعه ویژگی‌های موجود در S_{n^*} از انتخاب رو به جلو استفاده می‌شود. برای این کار از فرمول ۱-۱۴ استفاده می‌کنند.

$$J_Z = \min_{d_m} \max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_{m}^{M} d_m k_m \left(x_i^Z, x_j^Z \right) \right) \qquad (۱۴\ ۱)$$

بطوری که Z مجموعه‌ای است که دارای ویژگی‌های انتخاب شده مانند $Z = \{f_1, f_2, \dots, f_n\}$ می‌باشد. J_Z با استفاده از روش MKL محاسبه می‌شود. برای انتخاب زیرمجموعه‌ای با r ویژگی از S_{n^*} به روش افزایشی از یک فرایند رو به جلو استفاده می شود. در ابتدا ، نمره $J_0 = +\infty$ و زیرمجموعه Z، تهی تنظیم می‌شود. هر ویژگی در زیرمجموعه ویژگی‌ها مانند $f_1, f_2, \dots, f_n$ جستجو می‌شود و توابع هدف $J_{f_1}, J_{f_2}, \dots, J_{f_n}$ با استفاده از روش MKL محاسبه می‌شود. ویژگی f_i که بیشترین کاهش $\Delta J = J_0 - J_{f_i}$ را ایجاد می‌کند، به Z اضافه می‌شود. سپس الگوریتم ویژگی f_j را که بیشترین کاهش ΔJ را از مجموعه $\{S_{n^*} - Z\}$ در Z ایجاد می‌کند، انتخاب می‌کند. روند انتخاب افزایشی تا زمانی که $\Delta J \leq 0$ شود یا به تعداد تکرار داده شده برسد، ادامه می‌یابد. شکل ۳-۳ فرایند انتخاب زیر مجموعه ویژگی فشرده را نشان می‌دهد.

Selecting Compact Feature Subset Algorithm :

Input:
Training examples: $X_0 = [x_1, x_2, \dots, x_i, \dots, x_N]^T$
Class label: $y = [y_1, y_2, \dots, y_i, \dots, y_N]^T$
Relevant Feature Set: S_{n^*}
Initialize:
Feature Subset: $Z = [\,]$
Set score of $J_0 = +\infty$
While $\Delta J \leq 0$ or $length(Z) \leq k$ **do**
Compute $J_Z = \min_{d_m} \max_{\alpha} W(\alpha, d_m)$ of selected features set Z by using SimpleMKL
for each feature $i \in \{S_{n^*} - Z\}$ **do**
compute J_{Z+i} of selected features set $\{Z, i\}$ by using SimpleMKL
compute $\Delta J_i = J_Z - J_{Z+i}$
end for
select feature f which generates the largest ΔJ reduction and $Z = \{Z, f\}$
end while
Output:
Feature ranked list Z.

شکل ۳-۳ شبه کد الگوریتم انتخاب زیر مجموعه ویژگی فشرده [25]

برای اندازه گیری عملکرد MKL از سه هسته خطی $K(x_i,x)=(x_i,x)$، چندجمله‌ای $K(x_i,x)=[(x_i,x)+1]^2$ و RBF $K(x_i,x)=exp\left(-\frac{\|x_i-x\|^2}{2}\right)$ استفاده کردند.

در این مطالعه از سه نوع داده بیان برای اندازه گیری عملکرد روش‌های انتخاب ویژگی استفاده شده است. مجموعه داده های ریزآرایه mRNA از پایگاه داده GEO[۴۷] و مجموعه داده های تعیین توالی mRNA و تعیین توالی miRNA از پایگاه داده TCGA بارگیری شد. داده های بارگیری شده از هر پایگاه داده مربوط به ۸ نوع سرطان هستند. برای تبدیل داده‌ها به قالب تعریف شده در این مطالعه، پیش پردازش روی داده‌ها صورت می‌گیرد. مقادیر از دست رفته هر مجموعه بیان تخمین زده می‌شود. اگر مقادیر از دست رفته یک mRNA (یا miRNA) کمتر از 20% کل نمونه ها باشد، این مقادیر از دست رفته با استفاده از روش کمترین مربعات محلی (LLSimpute[۴۸]) تخمین زده می شود. پس از این فرآیند، این مجموعه داده ها با روش میانه قدرمطلق انحراف (MAD[۴۹]) نرمال می‌شوند تا همه نمونه‌ها دارای زمینه مشابه باشند. در آزمایشات روش MKL نسبت به روش های دیگر که در مقالات قبلی ارائه شده بود، در مجموع نتیجه بهتری بدست آورد (جدول ۳-۷ را ببینید).

الگوریتم LLSimpute نشان دهنده یک ژن هدف است که مقادیر از دست رفته را به عنوان یک ترکیب خطی از ژن‌های مشابه دارد. ژن‌های مشابه توسط k نزدیکترین همسایه یا k ژن منسجم که دارای مقادیر بزرگ قدرمطلق ضرایب همبستگی پیرسون هستند، انتخاب می‌شوند [26].

در روش میانه قدرمطلق انحراف ابتدا فاصله نمونه‌ها با میانه آن‌ها را به دست می‌آورند. سپس میانه قدرمطلق این فاصله‌ها را به عنوان خروجی آزمون اعلام می‌کنند.

جدول ۳-۷ نتایج اثر بخشی روش های انتخاب ویژگی مختلف [25]

Methods	SVM-RFE	SVM-RCE	mRMR	IMRelief	SlimPLS	OSFS	FGM	SMKL-FS
KIDNEY	0.922	0.832	0.987	0.901	0.896	0.893	0.916	**0.994**
BRCA	0.839	0.963	0.979	0.817	0.973	0.893	0.953	*0.990*
LUNG	0.891	0.946	0.979	0.953	0.831	0.945	0.946	**0.980**
HNSC	0.979	0.955	0.991	0.879	0.874	0.920	0.874	**0.994**
LIHC	0.906	0.836	0.911	0.813	0.871	0.789	**0.925**	0.917
PRAD	0.897	0.933	0.930	0.892	0.905	0.794	0.836	*0.946*
STAD	0.855	0.870	0.853	0.790	0.823	0.760	0.827	*0.880*
THCA	0.925	0.901	**0.969**	0.842	0.876	0.878	0.928	0.967
Mean	0.902	0.904	0.950	0.861	0.881	0.859	0.901	**0.958**

[۴۷] Gene Expression Omnibus

[۴۸] Local Least Squares imputation

[۴۹] Median Absolute Deviation

Pfeifer در سال ۲۰۱۵ برای شناسایی زیرگروه‌های سرطان، ادغام داده‌ها با استفاده از یادگیری چند هسته‌ای را پیشنهاد کردند [27].

برای ادغام چندین نوع داده از یادگیری چند هسته‌ای برای چارچوب کاهش ابعاد (MKL-DR) استفاده می‌کنند و رویکرد را گسترش می‌دهند و آن را یادگیری چند هسته‌ای منظم برای کاهش ابعاد (rMKL-DR) می‌نامند. این روش از یک طرف مبتنی بر یادگیری چند هسته‌ای و از سوی دیگر بر اساس چارچوب جاسازی شده گراف[۵۰] برای کاهش ابعاد است.

به طور کلی یادگیری چند هسته‌ای وزن β را بهینه می‌کند که به صورت خطی مجموعه‌ای از ماتریس‌های هسته ورودی را ترکیب می‌کند. در اینجا هر نوع داده ورودی به عنوان یک ماتریس هسته جداگانه نشان داده می‌شود. بنابراین این روش می‌تواند برای داده‌هایی که نمایش ویژگی‌های مختلف دارند استفاده شود.

MKL-DR بر اساس چارچوب جاسازی گراف برای کاهش ابعاد توصیف شده که امکان تلفیق تعداد زیادی از روش‌های کاهش بعد را فراهم می‌کند. در این چارچوب، بردار تصویر[۵۱] v (برای تصویر کردن به زیر فضایی یک بعدی) یا ماتریس تصویر[۵۲] V (برای تصویر کردن به ابعاد بالاتر) بر اساس معیار حفظ گراف بهینه می‌شود.

$$\min_{v} imize \sum_{i,j=1}^{N} \left\| v^T x_i - v^T x_j \right\|^2 w_{ij} \qquad (۱۵\ ۱)$$

$$\text{s.t. } \sum_{i=1}^{N} \|v^T x_i\|^2 d_{ii} = const, \text{ or} \qquad (۱۶\ ۱)$$

$$\sum_{i,j=1}^{N} \left\| v^T x_i - v^T x_j \right\|^2 w'_{ij} = const. \qquad (۱۷\ ۱)$$

W یک ماتریس شباهت با ورودی های W_{ij} و D (یا W) یک ماتریس محدودیت برای جلوگیری از راه حل بی اهمیت است. انتخاب ماتریس‌های W و D (یا Wو W') طرح کاهش ابعاد را تعیین می‌کند که باید اجرا شود.

نسخه هسته‌ای مسئله بهینه سازی با محدودیت (۵-۲۱) را می توان با استفاده از یک نگاشت ضمنی از داده‌ها به یک فضای هیلبرت با ابعاد بالا استخراج کرد. علاوه بر این می توان نشان داد که بردار تصویر بهینه v در دامنه

[۵۰] Graph embedding framework

[۵۱] Pojection vector

[۵۲] Projection matrix

نقاط داده x_i قرار دارد، بنابراین $v = \sum_{n=1}^{N} \alpha_n \phi(x_n)$. مسئله بهینه سازی کامل برای rMKL-DR به این ترتیب است:

$$\min_{\alpha,\beta} \sum_{i,j=1}^{N} \left\| \alpha^T K^i \beta - \alpha^T K^j \beta \right\|^2 w_{ij}$$

$$s.t. \sum_{i,j=1}^{N} \left\| \alpha^T K^i \beta \right\|^2 d_{ij} = const. \qquad (۱۸\ ۱)$$

$$\|\beta\|_1 = 1,\ \beta_m \geq 0, m = 1,2,\dots,M.$$

بطوریکه

$$\alpha = [\alpha_1 \dots \alpha_N]^T \in \mathbb{R}^N,$$

$$\beta = \left[\beta_1 \dots \beta_M\right]^T \in \mathbb{R}^M, \qquad (۱۹\ ۱)$$

$$K^i = \begin{pmatrix} K_1(1,i) & \cdots & K_M(1,i) \\ \vdots & \ddots & \vdots \\ K_N(N,i) & \cdots & K_M(N,i) \end{pmatrix} \in \mathbb{R}^{N \times M}.$$

مسئله بهینه سازی را می‌توان به راحتی برای تصویر کردن به بیش از یک بعد گسترش داد. در آن حالت، یک ماتریس تصویر $A = [\alpha_1 \ \dots \ \alpha_p]$ به جای بردار تصویر α بهینه می‌شود. سپس A با توجه به روش کاهش ابعاد انتخاب شده، همزمان با بردار وزن هسته β بهینه می‌شود. از آنجا که بهینه سازی همزمان این دو متغیر دشوار است از نزول مختصات استفاده می‌شود. به عنوان مثال A و β به صورت متناوب بهینه می‌شوند تا زمانی که همگرایی یا حداکثر تعداد تکرارها حاصل شود. با استفاده از این چارچوب، الگوریتم کاهش ابعاد پیش‌بینی‌های حفظ موقعیت (LPP[۵۳]) را اعمال می‌کنند. این یک روش محلی بدون نظارت است که هدف آن حفظ فاصله هر نمونه تا نزدیکترین همسایگان آن است. همسایگی یک نقطه داده i به عنوان N(i) نشان داده می شود. برای LPP، ماتریس های W و D به این صورت تعریف می شوند:

[۵۳] Locality Preserving Projections

$$w_{ij} = \begin{cases} 1, & if\ i \in N_k(j)\ \vee j \in N_k(i) \\ 0, & else \end{cases}$$

$$d_{ij} = \begin{cases} \sum_{n=1}^{N} w_{in}, & if\ i = j \\ 0, & else \end{cases} \qquad (۲۰\ ۱)$$

رویکرد rMKL-DR با LPP از این پس rMKL-LPP نامیده می‌شود. فرآیند خوشه بندی با استفاده از k-means انجام می‌شود. برای ارزیابی خوشه‌ها از عرض سیلوئت استفاده می‌کنند، معیاری که برای هر نقطه داده نشان می‌دهد چقدر در خوشه خاص خود جای می‌گیرد در مقایسه با اینکه در بهترین خوشه دیگر باشد. وقتی به طور متوسط از تمام نقاط داده به دست می‌آید، میانگین مقدار سیلوئت حاصل اشاره می‌کند که یک خوشه چقدر منسجم است و خوشه ها به چه اندازه از هم جدا شده‌اند.

از rMKL-LPP برای پنج مجموعه داده سرطان استفاده کردند. برای هر مجموعه داده، الگوریتم با هر دو مقدار اولیه قابلیت اجرا دارد، یا با بهینه سازی A و یا با بهینه سازی β شروع می شود. برای هر دو نتیجه کاهش ابعاد، نقاط داده یکپارچه با استفاده از k-means با $k \in \{2,3,\dots,15\}$ خوشه‌بندی شدند. با استفاده از میانگین مقدار سیلوئت نتیجه خوشه بندی، تعداد بهینه خوشه‌ها را انتخاب کردند. سپس از این معیار برای انتخاب بهترین خوشه‌بندی در میان دو مقداردهی اولیه مختلف استفاده شد. در بیشتر موارد، مقداردهی اولیه β منجر به مقادیر کمی بهتر سیلوئت می‌شود.

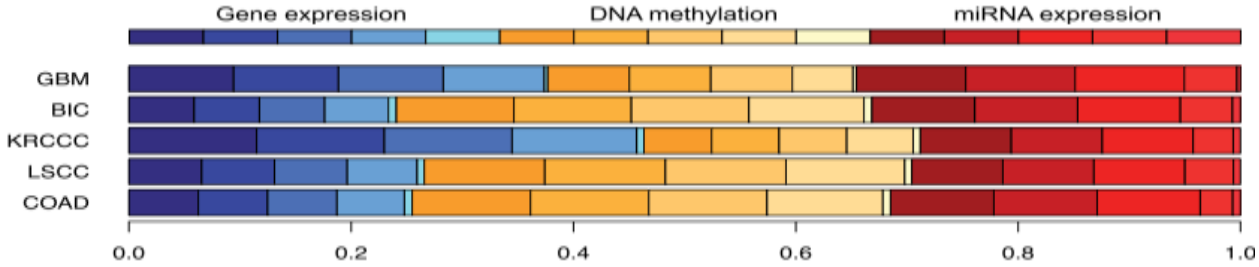
برای هر نوع داده، از تابع هسته گاسین برای محاسبه ماتریس‌های هسته و عادی سازی آنها در فضای ویژگی استفاده کردند. برای بررسی اینکه چقدر روش قادر به مدیریت چندین هسته ورودی برای انواع داده های منفرد است، دو سناریو ایجاد کردند. اولین حاوی یک ماتریس هسته در هر نوع داده بود که γ_1 با توجه به قانون $\gamma = \frac{1}{2d^2}$ انتخاب شد و d تعداد ویژگی های داده است. از آنجا که این منجر به سه هسته می‌شود، سناریو 3K نامیده می‌شود. برای سناریو ۲، با تغییر پارامتر هسته γ به گونه‌ای که $\gamma_n = c_n\gamma_1$ و $c_n \in \{10^{-6}, 10^{-3}, 1, 10^3 10^6\}$ پنج ماتریس هسته تولید کردند. بنابراین این سناریو 15K نامیده می‌شود. خوشه بندی های حاصل را با نتایج SNF در جدول ۳-۸ مقایسه کردند. به طور کلی عملکرد rMKL-LPP با پنج ماتریس هسته بهترین بود.

جدول ۳-۸ تجزیه و تحلیل بقا از نتایج خوشه بندی همجوشی شبکه شباهت (SNF) و rMKL-LPP با یک و پنج هسته در هر نوع داده. اعداد داخل پرانتز تعداد خوشه ها را نشان می دهد. [27]

Cancer type	SNF	rMKL-LPP	
		3K	15K
GBM	2.0E-4 (3)	4.5E-2 (5)	6.5E-6 (6)
BIC	1.1E-3 (5)	3.0E-4 (6)	3.4E-3 (7)
KRCCC	2.9E-2 (3)	0.23 (6)	4.0E-5 (14)
LSCC	2.0E-2 (4)	2.2E-3 (2)	2.4E-4 (6)
COAD	8.8E-4 (3)	2.8E-2 (2)	2.8E-3 (6)

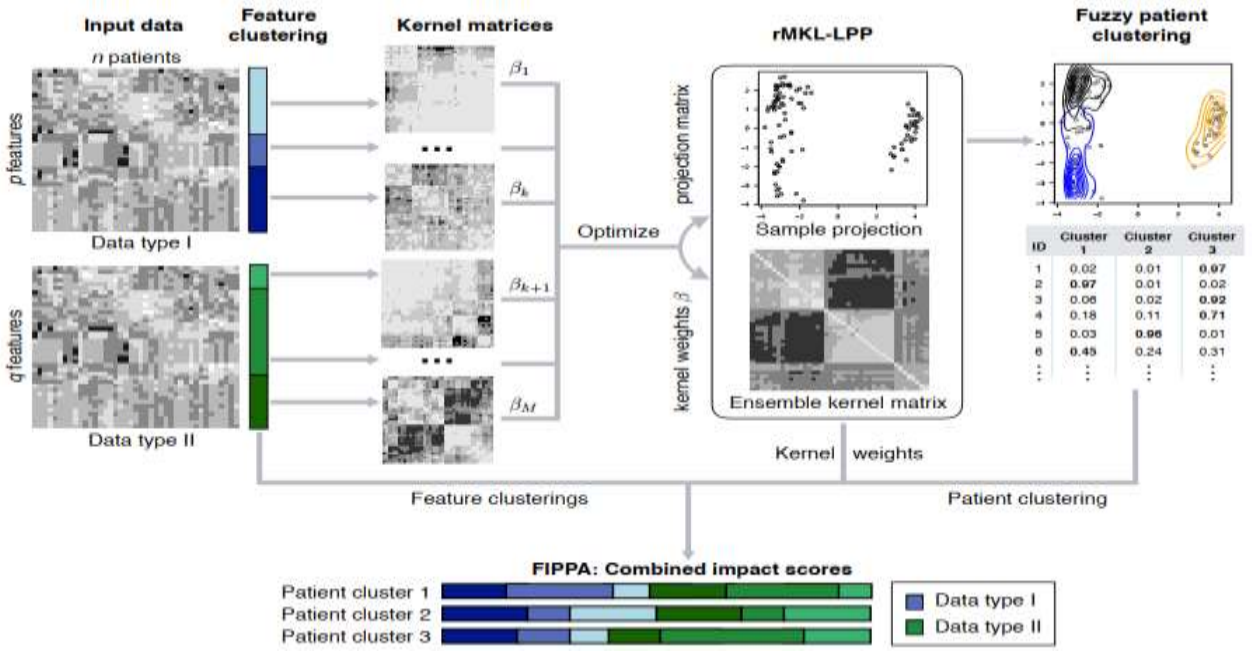
یک مزیت دیگر روش rMKL-LPP با پنج هسته در هر نوع داده این است که فرد نیازی نیست در مورد بهترین اندازه گیری شباهت برای نوع داده از قبل تصمیم بگیرد. علاوه بر این، نتایج نشان می‌دهد که حتی در برخی از سناریوها داشتن بیش از یک ماتریس هسته در هر نوع داده برای گرفتن درجات مختلف شباهت بین نقاط داده (بیماران در این سناریو کاربردی) حتی ممکن است مفید باشد.

برای rMKL-LPP با پنج هسته در هر نوع داده، شکل ۳-۴ تأثیر هر ماتریس هسته بر ماتریس نهایی یکپارچه را نشان می‌دهد. نوار بالا نشان می‌دهد که نمودار برای سهمی برابر در تمام ماتریس های هسته چگونه است. در مقایسه با این، می‌بینید که ماتریس های هسته با استفاده از مقادیر بالا برای پارامتر $\gamma = \gamma_1 * 10^6$، تأثیر بسیار کمی برای انواع سرطان دارند. این نتایج با این قانون موافق است که γ باید به اندازه $\frac{1}{2d^2}$ یا پایین تر انتخاب شود که برای انتخاب γ_1 استفاده شده است. علاوه بر این، همه انواع داده ها به ماتریس هسته ترکیبی کمک می‌کنند و می توان تفاوت هایی را برای انواع مختلف سرطان مشاهده کرد، به عنوان مثال برای BIC، داده های متیلاسیون DNA تأثیر بیشتری دارد در حالی که برای KRCCC اطلاعات بیشتری از داده های بیان ژن گرفته شده است.



شکل ۳-۴ مشارکت ماتریس‌های مختلف هسته به هر ورودی در ماتریس هسته کلی واحد. این سه رنگ بیان ژن (آبی)، متیلاسیون DNA (زرد) و بیان miRNA (قرمز) را نشان می دهند [27].

Speicher و Pfeifer در سال ۲۰۱۸ برای شناسایی زیرگروه‌های سرطانی از خوشه بندی و یادگیری چند هسته‌ای استفاده کردند [28]. شکل ۳-۵ رویکرد پیشنهادی این مقاله را نشان می دهد.



شکل ۳-۵ رویکرد پیشنهادی Speicher و Pfeifer برای شناسایی زیرگروه های سرطانی [28]

در این مطالعه روشی را پیشنهاد می‌دهند که خوشه بندی ویژگی‌ها را با خوشه بندی نمونه‌ای مبتنی بر چندین هسته ترکیب کند و FIPPA را معرفی می‌کنند، نمره ای که تأثیر خوشه ویژگی را بر یک خوشه بیمار اندازه‌گیری می‌کند. همانطور که در شکل ۳-۵ نشان داده شده است ویژگی‌های هر نوع داده را با استفاده از k-means خوشه بندی می‌کنند به گونه‌ای که می‌توانند براساس هر خوشه ویژگی یک ماتریس هسته تولید کنند. ماتریس‌های هسته سپس با استفاده از یک رویکرد یادگیری چند هسته‌ای ادغام می‌شوند. این روش یک وزن برای هر ماتریس هسته و یک ماتریس تصویر بهینه می‌کند که منجر به نمایش نمونه‌ها در ابعاد پایین می‌شود. سپس با استفاده از خوشه‌بندی فازی (یک الگوریتم خوشه‌بندی که امکان تقسیم داده‌ها به بیش از یک خوشه را فراهم می‌کند) نمونه‌ها را خوشه‌بندی می‌کنند. این روش دو مزیت در مقایسه با روش‌های استاندارد فراهم می‌کند. (۱) افزایش همگنی ویژگی‌ها با شناسایی خوشه‌های ویژگی می‌تواند نویز را در هر ماتریس هسته کاهش دهد. (۲) در دسترس بودن خوشه‌های ویژگی و وزن هسته مربوطه امکان تفسیر بیشتر خوشه‌های بیمار را فراهم می‌کند.

رویکرد خود را برای شش مجموعه داده مختلف سرطان تولید شده توسط TCGA که از مرورگر UCSC Xena بارگیری شده اند، اعمال کردند. انواع سرطان تحت پوشش عبارتند از: BRCA[۵۴]، LUAD[۵۵]، HNSC[۵۶]، LGG[۵۷]، THCA[۵۸] و PRAD[۵۹]. برای هر بیمار سرطانی برای خوشه بندی از متیلاسیون DNA، داده‌های بیان ژن، تغییرات

[۵۴] Breast invasive carcinoma
[۵۵] Lung adenocarcinoma
[۵۶] Head and Neck squamous cell carcinoma
[۵۷] Brain Lower Grade Glioma
[۵۸] Thyroid carcinoma
[۵۹] Prostate adenocarcinoma

تعداد کپی و داده‌های بیان miRNA استفاده کردند. برای اعتبار سنجی آزمایشی خود پارامتر هر دو مرحله خوشه بندی را روی یک مقدار ($c \in \{2, \dots, 6\}$) قرار دادند. در یادگیری چندهسته‌ای از هسته گاسین استفاده کردند. پارامتر هسته γ را وابسته به تعداد ویژگی‌های d در مجموعه ویژگی‌های مربوطه بر اساس قاعده انگشت[۶۰] برابر $\gamma = \frac{1}{2d^2}$ انتخاب کردند. با ضرب γ با یک فاکتور $f_\gamma \in \{0.5,1,2\}$، سه هسته در هر نوع داده تولید کردند و فقط از ماتریس یک هسته که بالاترین واریانس را در اولین مولفه های اصلی d ارائه داد استفاده کردند. مقایسه روش پیشنهادی با سایر روش ها در جدول ۳-۹ ذکر شده است.

جدول ۳-۹ مقایسه روش پیشنهادی مقاله FC+ rMKL-LPP با سایر روش ها [28]

Cancer	average kLPP		rMKL-LPP		FC + rMKL-LPP	
	p-value	C	p-value	C	p-value	C
BRCA	3.7E-2	6	7.3E-2	6	5.0E-2	4
HNSC	1.4E-3	6	1.4E-3	6	9.96E-3	5
LGG	<1.0E-16	3:6	<1.0E-16	3:6	<1.0E-16	3:6
LUAD	0.15	2	2.9E-2	2	3.1E-2	6

وزن هر خوشه ویژگی در ترکیب با ماتریس هسته این امکان را می‌دهد تا امتیازات fFIPPA را محاسبه کنند. نمرات fFIPPA محاسبه شده اجازه شناسایی خوشه های ویژگی را می دهد که بیش از حد متوسط به شباهت نمونه ها در یک خوشه نمونه ($fFIPPA^+$، فرمول ۱-۲۱) و عدم شباهت نمونه ها در دو خوشه مختلف ($fFIPPA^-$، فرمول۲-۲۲) کمک می‌کنند، بدین ترتیب پایه و اساس خوشه تولیدی را نشان می‌دهد.

$$fFIPPA^{+}_{c,m} = \frac{1}{|N|^2}\sum_{i,j=1}^{N} p_c(x_i \wedge x_j)\frac{\beta_m K^{+}_m[i,j]}{\mathbb{K}^{+}[i,j]} \qquad (۲۱\ ۱)$$

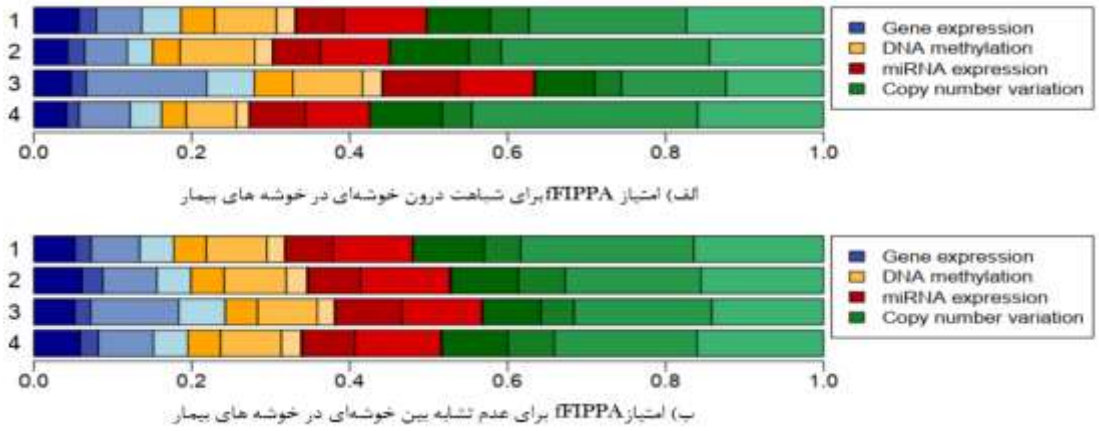
$$fFIPPA^{-}_{c,m} = \frac{1}{|N|^2}\sum_{i,j=1}^{N} p_c(x_i \oplus x_j)\frac{\beta_m K^{-}_m[i,j]}{\mathbb{K}^{-}[i,j]} \qquad (۲۲\ ۱)$$

$$p_c(x_i \oplus x_j) = \left(p_c(x_i) + p_c(x_j) - 2p_c(x_i)p_c(x_j)\right) \quad and \quad p_c(x_i) = p(x_i \in c)$$

بطوریکه m خوشه ویژگی $m \in \{1, \dots, M\}$، c خوشه نمونه $c \in \{1, \dots, C\}$ می‌باشد. $(x_i \wedge x_j)$ برای ایجاد نفوذ بالا برای جفت‌هایی استفاده می‌شود که احتمالاً هر دو مشترک به خوشه c تعلق دارند و $p_c(x_i \oplus x_j)$ منجر به افزایش یک فاکتور برای جفت نمونه هایی می‌شود که دقیقاً یکی از آن‌ها احتمال بالایی برای c دارد.

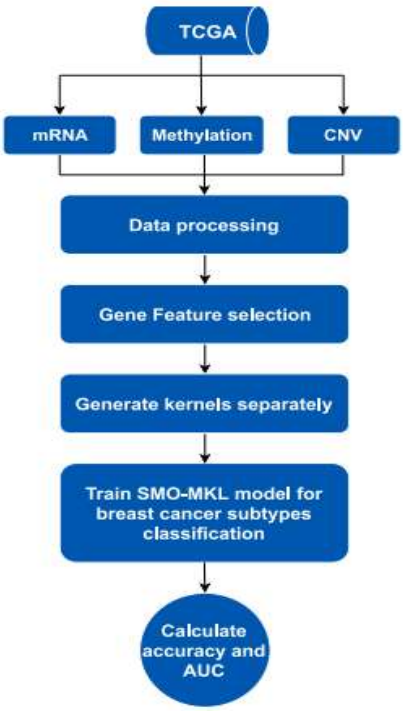
[۶۰] Rule of thumb

به عنوان مثال شکل ۳-۶ نشان می‌دهد که در خوشه‌های ۲ و ۴ تاثیر داده‌های تغییر تعداد کپی، بر شباهت بین خوشه‌ای بیشتر از عدم تشابه بین خوشه‌ای است.



شکل ۳-۶ fFIPPA مثبت و منفی از هر خوشه ویژگی و خوشه بیماران برای سرطان پستان با ۴ خوشه. هر ردیف یک خوشه بیمار را نشان می‌دهد [28].

Tao و همکارانش در سال ۲۰۱۹ به طبقه بندی زیرگروه های سرطان پستان با استفاده از یادگیری چندهسته‌ای پرداختند [29]. شکل ۳-۷ فرایند پیش بینی زیرگروه‌های سرطان پستان را در این مقاله نشان می‌دهد.



شکل ۳-۷ فرایند پیش بینی زیرگروه‌های سرطان پستان توسط Tao و همکارانش [29]

در مرحله اول، داده های متمایز سرطان پستان از جمله داده های mRNA، داده های متیلاسیون DNA و داده‌های تغییر تعداد کپی (CNV) را از TCGA جمع آوری کردند. مجموعه داده حاوی ۶۰۶ نمونه بیمار از سرطان پستان بود که به پنج زیرگروه تقسیم شد که در جدول ۳-۱۰ ذکر شده است.

جدول ۳-۱۰ تعداد زیرگروه‌های مشخص سرطان پستان [29]

Breast Subtypes	Cancer Patients
Luminal A	277
Luminal B	40
TNBC	70
HER2 (+)	11
Unclear	208

علاوه بر این قبل از انجام انتخاب ویژگی داده‌ها را نرمالسازی کردند. برای داده های mRNA از داده‌های اصلی TCGA استفاده کردند و ژن هایی را که بیش از ۲۰۰ نمونه داده از دست رفته داشتند، حذف کردند. در نهایت هر داده omics به عنوان یک ماتریس دو بعدی که در آن هر ردیف نشان دهنده یک نماد ژن و هر ستون نمایانگر یک نمونه با زیرگروه مربوط به خود بود، نشان داده شد.

از آنجا که برخی از ژن ها ممکن است در طبقه بندی زیرگروه های سرطان پستان تأثیر کمی داشته و یا حتی هیچ تأثیری نداشته باشند، با استفاده از تکنیک‌های انتخاب ویژگی برخی ژن‌های قابل توجه را انتخاب کردند. برای انتخاب ویژگی ابتدا از آزمون Wilcoxon rank-sum (در پاراگراف بعد شرح داده خواهد شد) بر روی هر داده omics استفاده کردند تا p-value ژن‌ها را جداگانه بدست آورند. سپس نرخ کشف اشتباه بنیامین-هوخبرگ[۶۱] (در پاراگراف بعد شرح داده خواهد شد) را برای تنظیم این p-valueها انتخاب کردند. در نهایت ژن با p-value کمتر از ۰,۰۵ را به عنوان ژن قابل توجه انتخاب کردند.

آزمون Wilcoxon rank-sum جزء آزمون غیر پارامتری است و برای سنجش تفاوت میان نمونه‌ها به کار می‌رود و فقط در صورت مستقل بودن داده‌ها قابل استفاده است. مراحل انجام آزمون Wilcoxon rank-sum در جدول ۳-۱۱ ذکر شده است [30]. رویه بنیامین-هوخبرگ ابزاری قدرتمند است که نرخ کشف اشتباه را کاهش می‌دهد. تنظیم نرخ به کنترل این واقعیت کمک می‌کند که بعضی اوقات p-value کوچک (کمتر از ۵٪) به طور تصادفی اتفاق می‌افتد که می تواند شما را اشتباهاً به رد کردن فرضیه صفر درست سوق دهد. به عبارت دیگر، روش بنیامین-هوخبرگ به شما کمک می‌کند تا از خطاهای نوع I (مثبت نادرست) جلوگیری کنید [31].

[۶۱] Benjamini-Hochberg False Discovery Rate

جدول ۳-۱۱ مراحل لازم برای انجام آزمون Wilcoxon rank-sum [30]

مرحله	جزئیات
۱	بدون توجه به اینکه از چه گروهی می آیند، همه مشاهدات به ترتیب افزایش درجه بندی می‌شوند. اگر دو مشاهده بدون توجه به گروه، دارای اندازه برابر باشند، به آن‌ها میانگین دو رتبه داده می‌شود.
۲	رتبه‌ها را در گروه‌های کوچکتر از دو گروه جمع می‌شوند. اگر دو گروه از اندازه مساوی برخوردار باشند، می توان یکی را انتخاب کرد.
۳	مقدار P مناسب محاسبه می‌شود.

برای پیش بینی زیرگروه‌های سرطان پستان، هسته MKL را روی داده های mRNA، داده های متیلاسیون و داده های CNV به طور جداگانه تولید کردند. از آنجا که مقیاس داده‌های مختلف omics متفاوت است، بنابراین با استفاده از فرمول (۱-۲۳) این هسته‌ها را نرمالسازی کردند.

$$K_{norm}(X_i, X_j) = \frac{K(X_i, X_j)}{\sqrt{K(X_i, X_i)K(X_j, X_j)}} \qquad (۲۳\ ۱)$$

بطوریکه K نشانگر تابع هسته، X_i نمایانگر i-اُمین نقطه داده و X_j نمایانگر j-اُمین نقطه داده است.

در این مطالعه از هسته خطی، هسته RBF و هسته چندجمله‌ای استفاده کردند. در کاربردهای عملی هسته‌های مختلف را می‌توان با هم ترکیب کرد. اگر داده ها به صورت خطی از هم قابل تفکیک باشند، هسته خطی کافی است و در غیر اینصورت از هسته RBF و هسته چندجمله ای استفاده می شود. از MKL بهبود یافته‌ای که توسط بهینه سازی حداقل متوالی (SMO) با $l_p norm$ تنظیم شده استفاده کردند که به راحتی قابل اجراست و برای مشکلات بزرگ مقیاس کارآمد است.

برای آزمایش از آنجا که SVM یک طبقه بندی کننده دو کلاسه است، ابتدا ACC و AUC طبقه‌بندی هر یک از دو زیرگروه سرطان پستان با سه هسته را بدست آوردند که به ترتیب در جدول ۳ ۱۲ و جدول ۳-۱۳ نشان داده شده است. در بیشتر موارد با استفاده ترکیبی از انواع داده‌ها، دقت و AUC بیشتری حاصل شد.

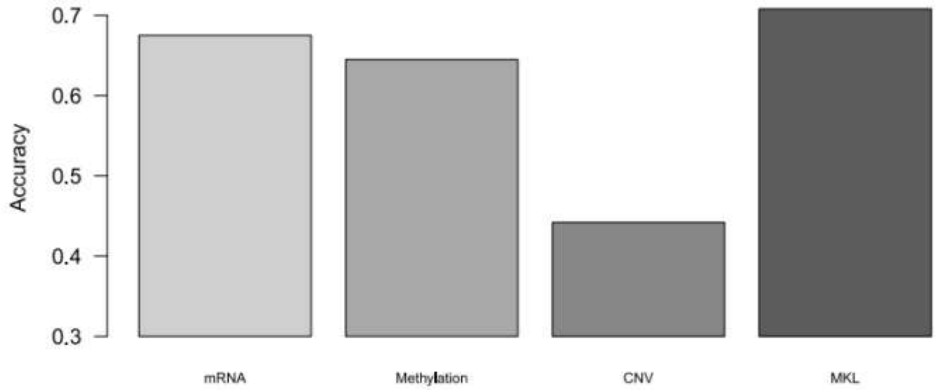
جدول ۳-۱۲ ACC طبقه بندی بین هر دو زیرگروه از سرطان پستان با سه هسته [29]

Breast Cancer Subtypes	mRNA	Methylation	CNV	MKL
Luminal A vs. luminal B	0.436	0.436	0.490	**0.681**
Luminal A vs. HER2 (+)	0.739	0.566	0.739	**0.870**
Luminal A vs. TNBC	**0.868**	0.867	0.604	0.859
Luminal A vs. Unclear	0.760	**0.849**	0.473	0.831
Luminal B vs. HER2 (+)	0.732	0.776	0.485	**0.837**
Luminal B vs. TNBC	0.871	**0.883**	0.855	0.873
Luminal B vs. Unclear	0.696	0.748	**0.770**	0.747
HER2 (+) vs. TNBC	0.5	0.5	0.5	**0.708**
HER2 (+) vs. Unclear	0.495	0.498	0.5	**0.731**
TNBC vs. Unclear	0.806	0.836	0.717	**0.846**
Mean	0.690	0.696	0.613	**0.798**

جدول ۳-۱۳ AUC طبقه بندی بین هر دو زیرگروه از سرطان پستان با سه هسته [29]

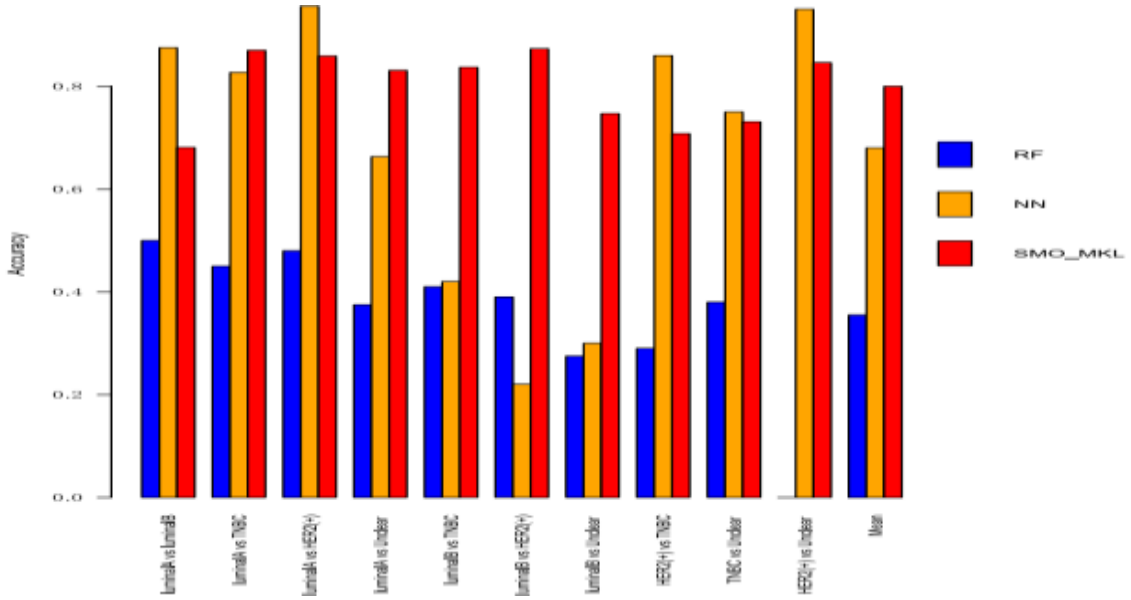
Breast Cancer Subtypes	mRNA	Methylation	CNV	MKL
Luminal A vs. luminal B	0.835	0.632	0.810	**0.848**
Luminal A vs. HER2 (+)	0.973	0.903	0.979	**0.986**
Luminal A vs. TNBC	**0.934**	0.926	0.909	0.930
Luminal A vs. Unclear	0.824	0.878	0.589	**0.901**
Luminal B vs. HER2 (+)	0.843	0.824	0.725	**0.895**
Luminal B vs. TNBC	**0.947**	0.932	0.941	0.945
Luminal B vs. Unclear	0.875	0.808	0.835	**0.896**
HER2 (+) vs. TNBC	0.867	0.778	0.741	**0.869**
HER2 (+) vs. Unclear	0.925	0.873	0.859	**0.962**
TNBC vs. Unclear	0.902	0.918	0.834	**0.929**
Mean	0.893	0.847	0.822	**0.916**

همچنین به منظور آزمایش این الگوریتم برای پیش‌بینی زیرگروه‌های سرطان با ادغام چندین طبقه‌بندی باینری به طبقه‌بند چند کلاسه دست یافتند و از روش یک در مقابل یک (OvO) استفاده کردند. در آزمایش نتایج با اعتبارسنجی 10-fold به دست آمد. ACC طبقه بندی در شکل ۳-۸ نشان داده شده است که MKL بهترین دقت را بدست آورد.

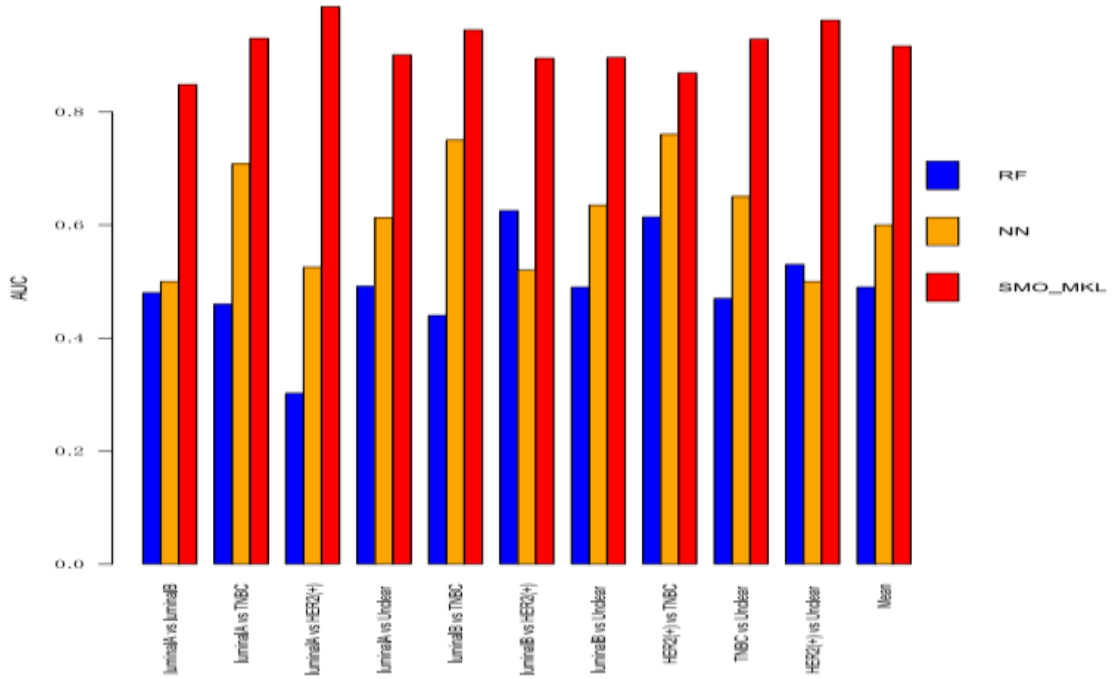


شکل ۳-۸ دقت طبقه بند چند کلاسه در زیرگروه های سرطان پستان [29]

همچنین، SMO-MKL را با جنگل تصادفی و شبکه عصبی در مجموعه های داده مشابه مقایسه کردند. ACC و AUC بدست آمده به ترتیب در شکل ۳-۹ و شکل ۳-۱۰ نشان داده شده است. دلیل اینکه برخی از دقت‌های شبکه عصبی از SMO-MKL بهتر بود این است که برخی از طبقه‌بندی‌های باینری دارای مشکل عدم تعادل هستند و شبکه عصبی به طور کامل یک کلاس را طبقه بندی می‌کند، اما AUC این طبقه‌بندی‌های باینری پایین است.

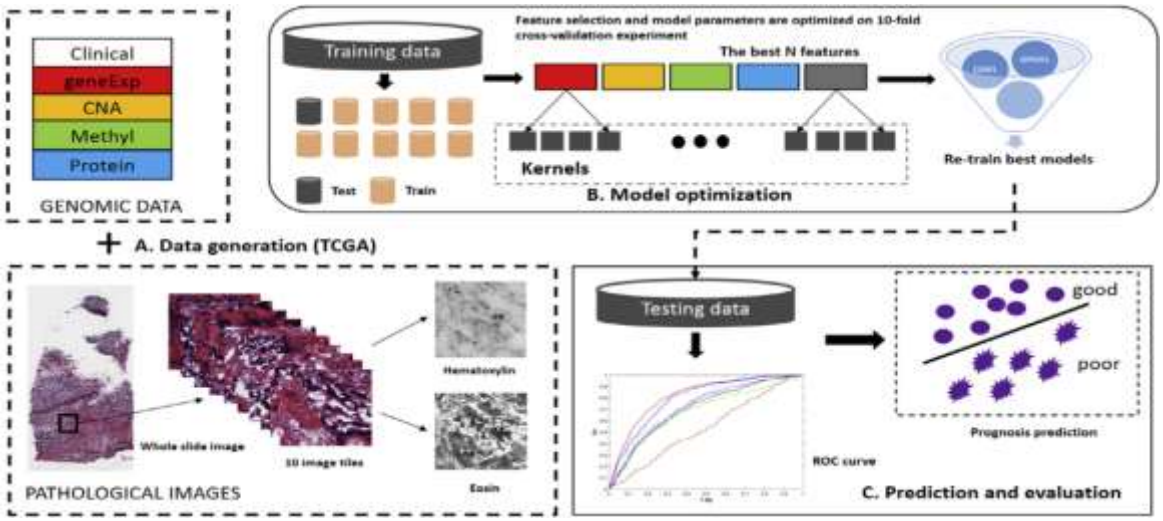


شکل ۳-۹ ACC جنگل تصادفی، شبکه عصبی و SMO-MKL در هر دو زیرگروه سرطان پستان [29]



شکل ۳-۱۰ AUC جنگل تصادفی، شبکه عصبی و SMO-MKL در هر دو زیرگروه سرطان پستان [29]

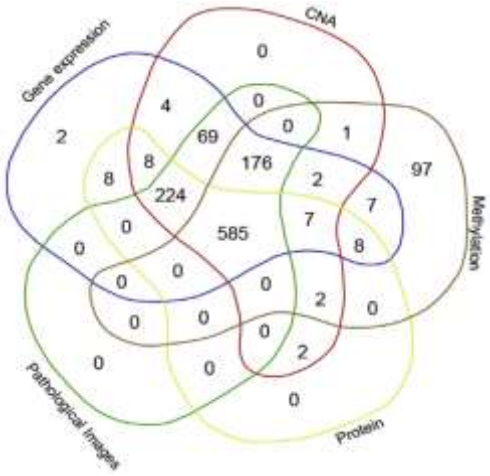
Sun و همکارانش در سال ۲۰۱۸ برای پیش‌بنی بقا بیماری سرطان پستان یک روش قدرتمند و جدید به نام GPMKL ارائه کردند [32]. شکل ۳-۱۱ رویکرد پیشنهادی این مقاله را نشان می‌دهد.



شکل ۳-۱۱ رویکرد پیشنهادی توسط Sun و همکارانش برای پیش‌بینی بقا بیماران دارای سرطان پستان [32]

مجموعه داده‌هایی شامل بیان ژن، تغییرات تعداد کپی، متیلاسیون ژن، بیان پروتئین و تصاویر پاتولوژیک نمونه‌های سرطان پستان را از پورتال TCGA بارگیری کردند. سپس از نمودار Venn (شکل ۳-۱۲) برای نمایش دقیق

جزئیات بیماران در انواع مختلف داده استفاده کردند و در نهایت ۵۸۵ بیمار معتبر را که متشکل از انواع داده ذکر شده بود، بدست آوردند. این مجموعه داده شامل ۵۷۸ بیمار زن و ۷ بیمار مرد بود که ۷ بیمار مرد را حذف کردند.



شکل ۳-۱۲ نمودارهای ون از تقاطع بین انواع داده استفاده شده توسط Sun و همکاران [32]

در این مطالعه پیش بینی بقای سرطان پستان را به عنوان یک مسئله طبقه بندی باینری در نظر گرفتند و بیماران در این مطالعه از نظر زمان زنده ماندن به دو دسته بازماندگان بلند مدت و کوتاه مدت تقسیم شدند. از بین ۵۷۸ بیمار، ۴۴۵ بیمار به عنوان بازمانده کوتاه مدت و ۱۳۳ بیمار به عنوان بازمانده طولانی مدت در نظر گرفته شد. علاوه بر این، بازماندگان کوتاه مدت ۰ و بیماران طولانی مدت ۱ برچسب گذاری شدند.

ابتدا ژن‌هایی با مقادیر از دست رفته (NA) در بیش از ۱۰٪ بیماران برای بیان ژن، تغییرات تعدا کپی، متیلاسیون ژن و بیان پروتئین را حذف کردند. پس از آن، مقادیر از دست رفته باقیمانده در هر نوع داده با استفاده از الگوریتم نزدیکترین همسایگان وزنی[۶۲] (در پاراگراف بعد شرح داده خواهد شد) محاسبه شد. علاوه بر این پروفایل های بیان ژن نرمال شد و به سه دسته تقسیم شدند: بیان کم (۱-) ، بیان بیش از حد (۱) و متوسط (۰). برای ویژگی‌های تغییرات تعداد کپی مستقیماً از داده‌های اصلی با مقادیر نسبی تعداد کپی برای هر ژن استفاده کردند. برای متیلاسیون ژن و بیان پروتئین به طور مستقیم از داده های اصلی استفاده کردند که توسط Zscore (فرمول ۱-۲۴) نرمال شده بود. ویژگی های تصاویر پاتولوژیک را با CellProfiler (یک ابزار منبع باز رایگان است که برای کمک به محققان در اندازه گیری کمی ویژگی‌ها از تصاویر ایجاد شده است) استخراج کردند.

$$Zscore = \frac{value - mean}{standard\ deviation} \qquad (۲۴\ ۱)$$

[۶۲] Weighted Nearest Neighbors Algorithm

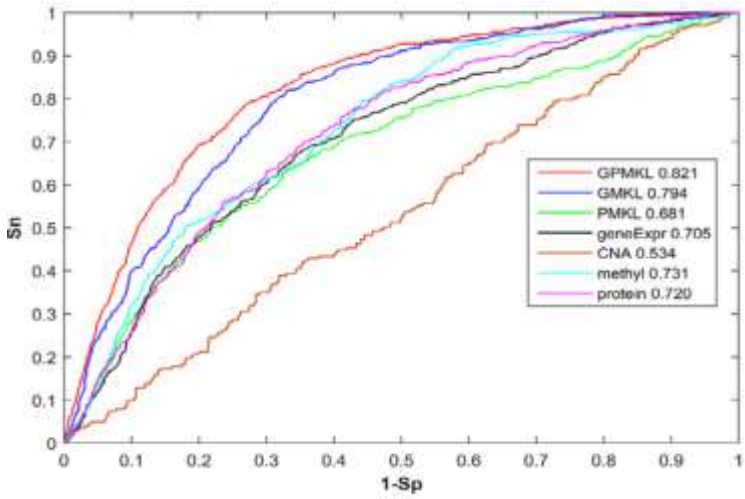
الگوریتم نزدیکترین همسایگان وزنی ژن‌هایی را با پروفایل بیان مشابه ژن مورد نظر برای محاسبه مقادیر از دست رفته انتخاب می‌کند. اگر ژن A را در نظر بگیرید که در آزمایش ۱ دارای یک مقدار از دست رفته است، این روش K ژن دیگر با بیشترین شباهت بیان نسبت به ژن A در سایر آزمایشات پیدا می‌کند که دارای مقدار موجود در آزمایش ۱ هستند. سپس از میانگین وزنی مقادیر در آزمایش ۱ از K نزدیکترین ژن برای محاسبه مقدار از دست رفته ژن A استفاده می‌شود. در میانگین وزنی، سهم هر ژن با شباهت بیان آن با ژن A وزن می‌شود [33].

برای ادغام انواع داده های مختلف در این مطالعه از MKL استفاده کردند زیرا بکارگیری چندین هسته در مقایسه با یک هسته واحد می‌تواند عملکرد تصمیم را قدرتمندتر کرده و عملکرد پیش بینی را افزایش دهد. بنابراین با استفاده از روش simpleMKL مدلی را پیشنهاد کردند که داده‌های ژنومی (بیان ژن، تغییرات تعداد کپی، متیلاسیون ژن و بیان پروتئین) و تصاویر پاتولوژیک را ادغام می‌کند. با توجه به این واقعیت که داده های استفاده شده شامل پنج نوع داده بود، ۵ هسته مختلف را به طور مستقل ساخته و در نهایت آن‌ها را در یک مدل سراسری ادغام کردند. هر هسته با هر نوع داده مستقل مطابقت دارد (بیان ژن، تغییرات تعداد کپی، متیلاسیون ژن، بیان پروتئین و تصاویر پاتولوژیک). همه انواع هسته SVM را گاسین (فرمول ۱-۲۵) انتخاب کردند و تمام دامنه جستجوی پارامتر $\delta = [0.25,0.5,1,2,5,7,10,12,15,17,20]$ بود.

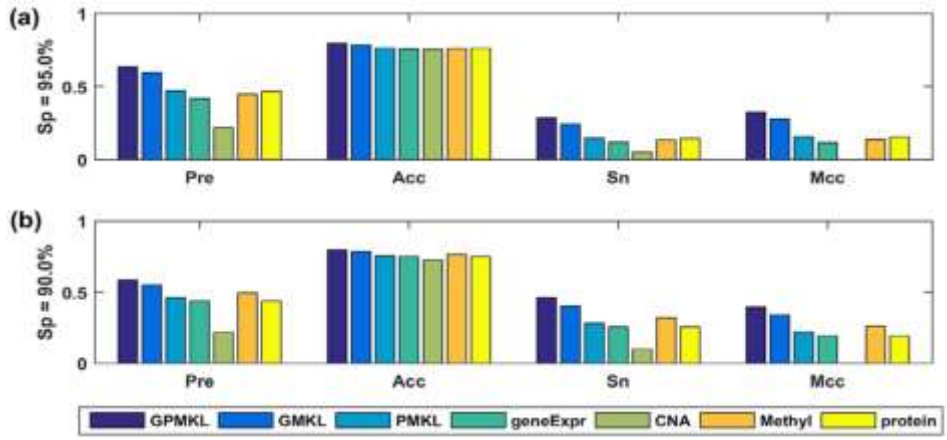
$$K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\delta^2}\right) \qquad (۱-۲۵)$$

simpleMKL همچنین یک مسئله دوگانه است که با نسخه چند هسته‌ای SVM پیاده سازی شده است. روش simpleMKL مبتنی بر روش تنظیم L2-norm است که نسبت به سایر روش‌های طبقه‌بندی کارایی بیشتری دارد، همچنین به جای یادگیری ترکیب هسته از هسته‌های مستقل، به طور مستقیم یک مسئله بهینه‌سازی ماشین بردار پشتیبان را حل می‌کند که هزینه محاسبات را بسیار کاهش می‌دهد و همچنین simpleMKL از الگوریتم گرادیان کاهشی برای یافتن بهترین پارامترها استفاده می‌کند. دو مدل مستقل به نام GMKL و PMKL برای مقایسه توسعه دادند که به ترتیب از داده های ژنومی و تصاویر پاتولوژیک استفاده می‌کنند.

منحنی‌های ROC برای مقایسه عملکرد پیش بینی هفت روش مختلف در هر سطح مشخص ترسیم شدند و در شکل ۳-۱۳ نمایش داده شده اند. علاوه بر منحنی ROC، مقدار AUC مربوطه برای هر روش نیز در شکل ۳-۱۳ نمایش داده می‌شود. همچنین برای هر روش، آستانه‌ای تنظیم شد به گونه‌ای که ویژگی یا تشخیص پذیری هر روش برابر با ۹۰٫۰٪ (متوسط) یا ۹۵٫۰٪ (زیاد) باشد. سپس مقادیر Pre، Acc، Sn و Mcc مربوط به هر روش را محاسبه کردند که در شکل ۳-۱۴ نشان داده شده است.



شکل ۳-۱۳ منحنی ROC برای طبقه بندی بازماندگان بلند مدت و کوتاه مدت از مجموعه داده های سرطان پستان، GPMKL مقدار AUC 0.821 بدست می‌آورد [32].



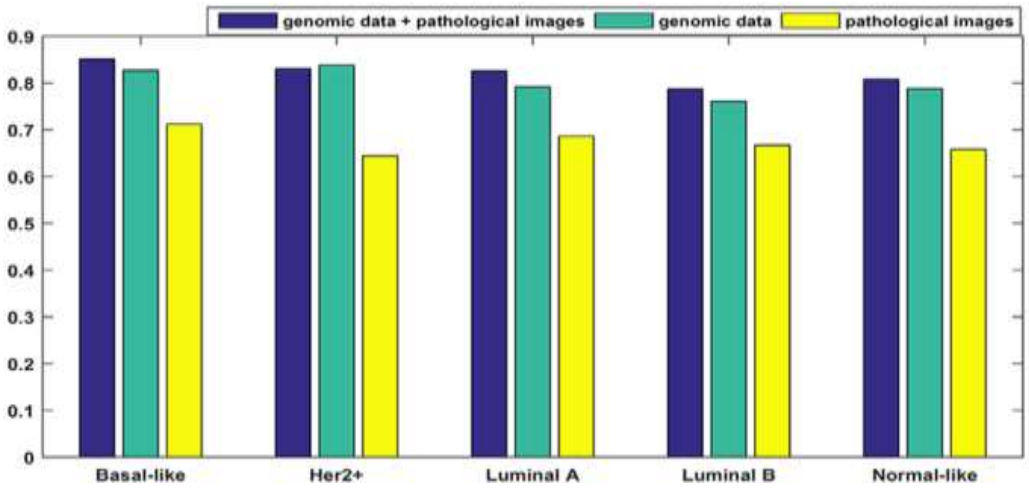
شکل ۳-۱۴ مقایسه عملکرد بین GPMKL و سایر مدل ها در معیارهای مختلف با تشخیص پذیری ۹۵٪ (a) و تشخیص پذیری ۹۰٪ (b) [32].

برای بررسی بیشتر اثر بخشی GPMKL، همچنین آن را با شش تا از محبوب ترین مدل‌های پیش‌بینی بقا مقایسه کردند که نتایج دقیق آن در جدول ۳-۱۴ ذکر شده است.

جدول ۳-۱۴ مقایسه عملکرد روش پیشنهادی با سایر مدل‌های موجود با استفاده از AUC [32]

Methods	Genomic data	Pathological image	Genomic data + pathological image
LASSO-Cox	0.697 ± 0.069	0.655 ± 0.059	0.698 ± 0.060
En-Cox	0.667 ± 0.080	0.649 ± 0.056	0.677 ± 0.067
PCRM	0.620 ± 0.067	0.608 ± 0.055	0.546 ± 0.043
RSF	0.722 ± 0.049	0.620 ± 0.067	0.718 ± 0.055
BoostCI	0.716 ± 0.037	0.622 ± 0.058	0.717 ± 0.037
superPC	0.659 ± 0.069	0.595 ± 0.056	0.698 ± 0.048
GPMKL	**0.802 ± 0.032**	**0.690 ± 0.046**	**0.828 ± 0.034**

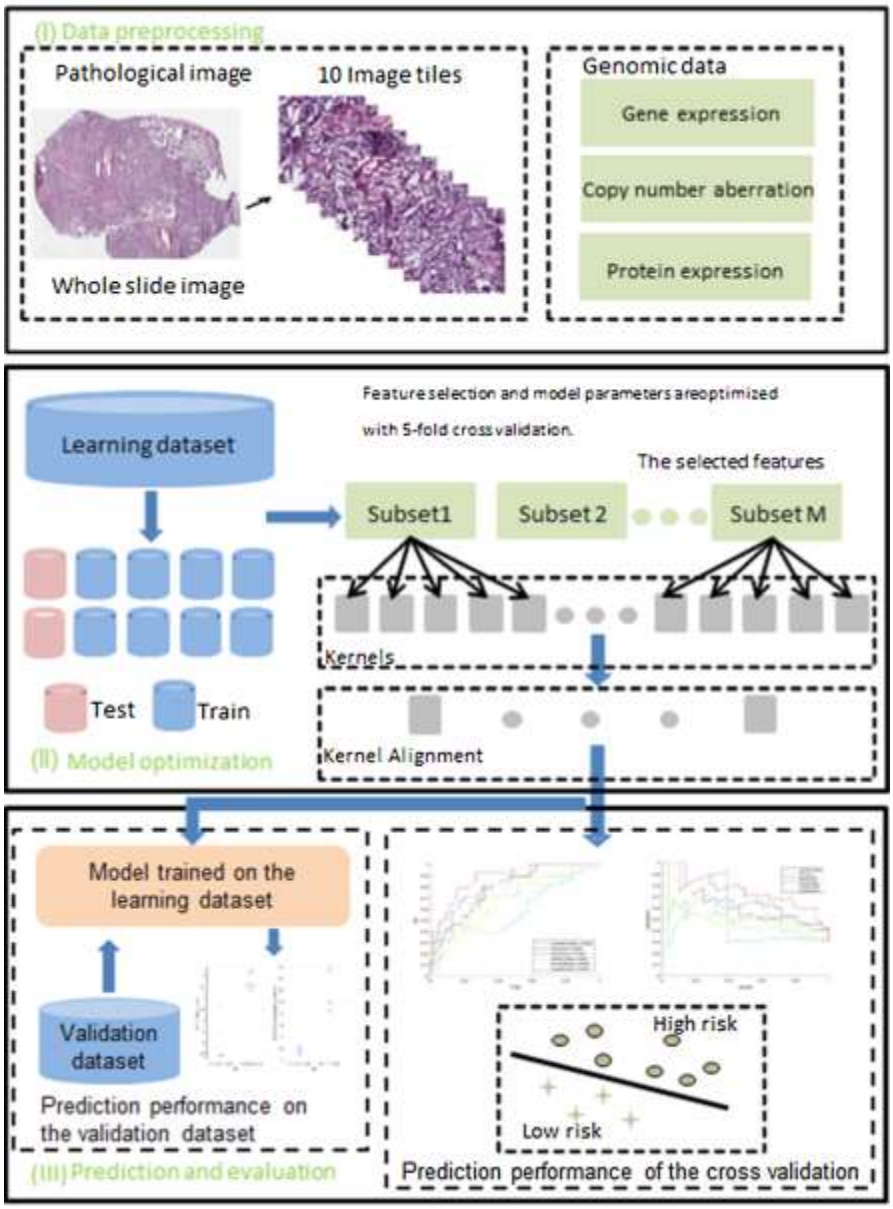
در انتها از روش پیشنهادی برای تعیین تاثیرات استفاده جداگانه از داده های ژنومی و تصاویر پاتولوژیک و یا استفاده ترکیبی از آن‌ها در تشخیص ۵ زیر گروه مختلف سرطان پستان استفاده شده است که نتیجه آن در شکل ۳-۱۵ نشان داده شده است.



شکل ۳-۱۵ زیرگروه های سرطان پستان عملکرد پیش بینی متفاوتی را توسط GPMKL نشان می دهند [32].

Zhang و همکارانش در سال ۲۰۱۹ برای پیش‌بینی بقای بیماران دارای سرطان سلول سنگفرشی ریه که جراحی انجام دادند، یک روش جدید مبتنی بر یادگیری چند هسته‌ای به نام LSCDFS-MKL ارائه کردند [34]. شکل ۳-۱۶ رویکرد پیشنهادی این مقاله را نشان می دهد.

مجموعه داده‌ها را برای بیمارانی دارای سرطان کارسینوما سلول سنگفرشی ریه که انواع داده از جمله تغییرات تعداد کپی، بیان ژن، بیان پروتئین و تصاویر پاتولوژیک را داشتند، از پایگاه داده TCGA بارگیری کردند. در این مطالعه پیش‌بینی را یک مسئله باینری در نظر گرفتند. ۷۰ بیمار به عنوان بازماندگان پر خطر (کمتر از ۵ سال) و ۳۱ بیمار به عنوان بازماندگان کم خطر (بیش از ۵ سال) طبقه بندی شدند. علاوه بر این، بازماندگان پرخطر و کم خطر به ترتیب ۰ ، ۱ برچسب گذاری شدند. ویژگی ها با مقادیر از دست رفته در بیش از ۱۰٪ بیماران را حذف کردند. سپس برای هر نوع داده واحد، از یک الگوریتم نزدیکترین همسایگان وزنی برای تخمین باقی مانده مقادیر از دست رفته استفاده کردند. با پیروی از مطالعه قبلی، پروفایل های بیان ژن را عادی کردند و آن‌ها را به سه دسته بیان بیش از حد (۱) ، خط پایه (۰) و بیان کم (۱-) گسسته کردند. برای مجموعه داده تغییرات تعداد کپی مستقیماً داده های اصلی را که حاوی مقادیر نسبی تعداد کپی خطی است، بازیابی کردند. سپس داده ها با Zscore نرمال شدند.



شکل ۳-۱۶ رویکرد پیشنهادی توسط Zhang و همکارانش برای پیش‌بینی بقا بیماران دارای سرطان سلول سنگفرشی ریه پس از جراحی [۳۱]

پس از انجام عملیات فوق تغییرات تعداد کپی، بیان ژن و بیان پروتئین هنوز به ترتیب از ۲۴۷۷۶، ۱۸۹۶۳ و ۲۱۵ ویژگی تشکیل شده بود. با این حال، با افزایش تعداد ویژگی‌ها ممکن بود عملکرد ضعیف تر باشد زیرا حجم نمونه کم بود و مشکل نفرین ابعاد وجود داشت. بنابراین از همبستگی خطی[۶۳] برای انتخاب تعداد مطلوب ویژگی‌ها

[۶۳] linear correlation

استفاده کردند. به طور خاص بیشترین ویژگی‌های آموزنده با اندازه‌گیری ضرایب همبستگی آن‌ها با برچسب‌ها بدست آمد. ویژگی‌های تصاویر پاتولوژیک را با CellProfiler استخراج کردند.

برای ادغام انواع مختلف داده از یادگیری چند هسته‌ای استفاده کردند چرا که چندین هسته نه تنها می‌توانند عملکرد بهتری نسبت به یک هسته داشته باشند، بلکه تفسیرپذیری را نیز افزایش می دهند. در LSCDFS-MKL، روشی به نام یادگیری چند هسته‌ای زیر مجموعه‌ای برای هر زیر مجموعه ویژگی ارائه کردند که به معنای اندازه گیری شباهت نمونه‌ها با هسته‌های مختلف بود. در این مطالعه، هسته‌های پایه را با تنظیم پهنای باند مختلف هسته گاسین به دست آوردند. بنابراین، با هر مدل یادگیری چند هسته‌ای زیر مجموعه‌ای، شباهت نمونه‌ها در مقیاس‌های مختلف به دست آمد. برای کنترل تبعیض و وضوح هسته‌ها دامنه‌ای از σ را استفاده کردند. هسته پایه زیرمجموعه ویژگی m-اُم را با مقیاس s به صورت زیر تعریف کردند.

$$K_s^{(m)}\left(x_i{}^{(m)}, x_j{}^{(m)}\right) = exp\left(-\frac{\left\|x_i^{(m)} - x_j^{(m)}\right\|^2}{2\sigma_s^2}\right) \qquad (۲۶\ ۱)$$

بطوریکه در آن m شاخص زیر مجموعه ویژگی‌ها است ($m = 1,2, \dots, M$) و M تعداد کل زیر مجموعه‌های ویژگی است. s شاخص مقیاس ها است ($s = 1,2, \dots, S$) و S تعداد کل مقیاس‌ها است. علاوه بر این، $x_i^{(m)}$ نشان دهنده نمونه i-اُم زیر مجموعه ویژگی m-اُم است.

طبق مدل یادگیری چند هسته‌ای زیر مجموعه‌ای، باید S×M هسته‌ی پایه را بسازند که بُعد آنها N×N باشد، جایی که N تعداد موارد آموزش را نشان می دهد. با افزایش موارد آموزش، محاسبه و ذخیره هسته‌ها به سرعت افزایش می‌یابد. برای حل این مشکل هسته پایه را انتخاب کردند که بتواند مجموعه داده را به بهترین وجه از آن هسته‌های اصلی منعکس کند. بنابراین از ترازبندی هسته KA[۶۴] استفاده کردند که نه تنها می تواند در حدود مقدار مورد انتظار متمرکز شود، بلکه بازده محاسباتی را به پیچیدگی زمان $O(n^2)$ کاهش می دهد. KA بین دو ماتریس هسته K_s و K_{ideal} در زیر مجموعه ویژگی یکسان به صورت زیر محاسبه می‌شود.

$$KA(K_s, K_{ideal}) = \frac{\langle K_s, K_{ideal}\rangle_F}{\sqrt{\langle K_s, K_s\rangle_F \langle K_{ideal}, K_{ideal}\rangle_F}}$$
$$\langle K_s, K_{ideal}\rangle_F = \sum_{i=1}^{N}\sum_{j=1}^{N} K_s(x_i, x_j) K_{ideal}(x_i, x_j) \qquad (۲۷\ ۱)$$

[۶۴] Kernel Alignment

بطوریکه K ماتریس هسته (گرم ماتریس) می‌باشد و K_{ideal} ماتریس هسته ایده‌آل می‌باشد که به صورت $K_{ideal} = yy^T$ تعریف می‌شود و y برچسب‌ها هستند. K_{ideal} می تواند به عنوان یک ملاک برای انتخاب هسته‌های متمایز شناخته شود. برای بدست آوردن هسته بهینه از یک سری هسته پایه S با مقیاس های مختلف، از فرمول زیر استفاده کردند.

$$K_{s_{max}}^{(m)}\left(x_i^{(m)}, x_j^{(m)}\right) = \max_s KA\left(K_s^{(m)}\left(x_i^{(m)}, x_j^{(m)}\right), K_{ideal}\right)$$
$$s.t. \qquad s \in [1, S], m \in [1, M] \qquad (۲۸\ ۱)$$

پس از آن فقط M هسته‌ی پایه مربوط به هر زیر مجموعه ویژگی را بدست آوردند. بنابراین از طریق یک ترکیب خطی وزنی از این هسته‌ها توسط فرمول۱-۲۹ هسته بهینه را بدست آوردند.

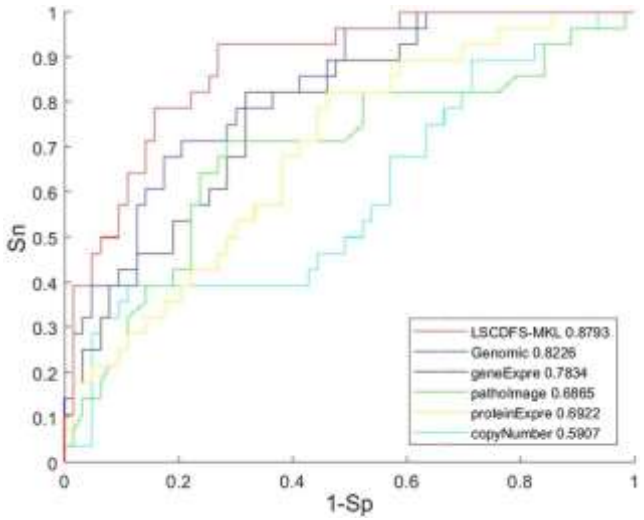
$$K\left(x_i, x_j\right) = \sum_{m=1}^{M} d_m K_{s_{max}}^{(m)}\left(x_i^{(m)}, x_j^{(m)}\right)$$
$$s.t. \quad d_m \geq 0 \; and \; \sum_{m=1}^{M} d_m = 1 \qquad (۲۹\ ۱)$$

الگوریتم: LSCDFS-MKL

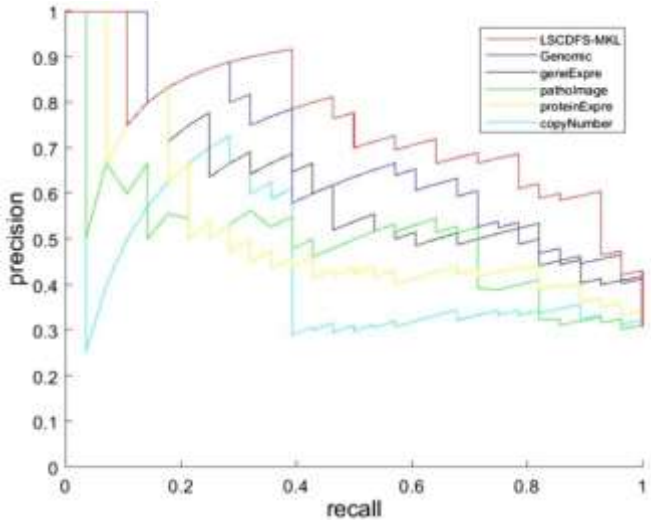
- مقدار دهی اولیه به دامنه مقادیر مقیاس هسته به صورت $[\sigma_{min}, \sigma_{max}]$.
- تعداد S مقیاس نمونه با دامنه قبلی.
- محاسبه‌ی هسته‌های مختلف با مقیاس های مختلف برای هر زیر مجموعه ویژگی
- انتخاب هسته زیرمجموعه توسط KA برای هر زیر مجموعه ویژگی
- یافتن ضریب d_m برای هر هسته زیر مجموعه ویژگی‌ها با حل مسئله بهینه سازی (فرمول ۱-۳۰)
- استفاده از d_m برای حل مسئله طبقه بندی MKL.

$$\min_{w,b,\xi,d} \left\{\frac{1}{2}\sum_{m=1}^{M}\frac{1}{d_m}\|w_m\|^2\right\} + C\sum_{i=1}^{N}\xi_i$$
$$s.t. \begin{cases} y_i\left(\sum_{m=1}^{M}\langle w_m, \phi_m(x_i)\rangle + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \\ \sum_{m=1}^{M} d_m = 1, \quad and \quad d_m \geq 0 \end{cases} \qquad (۳۰\ ۱)$$

برای ارزیابی خصوصیات روش پیشنهادی، منحنی مشخصه عملکرد گیرنده (ROC) و منحنی صحت-فراخوانی[۶۵] (PRC) معرفی شد که معمولاً برای نشان دادن توانایی تشخیص طبقه‌بندی باینری استفاده می‌شود. این دو منحنی به ترتیب در شکل ۳-۱۷ و شکل ۳-۱۸ نشان داده شده است. جدا از منحنی ROC، سطح زیر منحنی (AUC) محاسبه شد. علاوه بر این، ویژگی یا تشخیص پذیری Sp، حساسیت Sn، ضریب همبستگی متیوس MCC، دقت Acc و صحت Pre را محاسبه کردند.



شکل ۳-۱۷ منحنی ROC برای طبقه بندی بازماندگان کم خطر و پرخطر با انواع داده های مختلف توسط [34] LSCDFS-MKL



شکل ۳-۱۸ منحنی PRC برای طبقه بندی بازماندگان کم خطر و پرخطر با انواع داده های مختلف توسط [34] LSCDFS-MKL

[۶۵] Precision Recall Curve

علاوه بر این، برای ارزیابی عملکرد استفاده از انواع داده های مختلف، حد آستانه‌ای با ویژگی یا تشخیص پذیر ۹۰٪ و ۹۵٪ را تنظیم کردند. پس از آن مقادیر Acc، Pre، Sn و Mcc را نیز محاسبه کردند. همانطور که در جدول ۳-۱۵ ذکر شده است، ترکیب انواع مختلف داده بهتر از استفاده از یک نوع داده واحد است.

جدول ۳-۱۵ مقایسه عملکرد انواع داده مختلف در معیارهای مختلف [34]

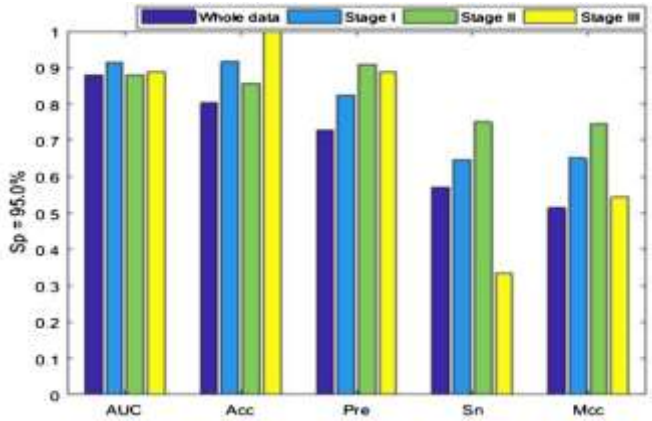
	S_p = 90.0%				S_p = 95.0%			
	Acc	*Pre*	*Sn*	*Mcc*	*Acc*	*Pre*	*Sn*	*Mcc*
LSCDFS-MKL	0.8022	0.8125	0.4643	0.5052	0.8022	0.7273	0.5714	0.5133
Genomic	0.7802	0.7857	0.3929	0.4416	0.7473	0.6471	0.3929	0.3524
geneExpre	0.7363	0.7000	0.2500	0.2987	0.7582	0.6667	0.4286	0.3862
patholmage	0.7033	0.5714	0.1429	0.1650	0.7033	0.5385	0.2500	0.2041
proteinExpre	0.7253	0.6667	0.2143	0.2577	0.7033	0.5385	0.2500	0.2041
copyNumber	0.7473	0.7273	0.2857	0.3371	0.7363	0.6250	0.3571	0.3175

برای ارزیابی بیشتر اثربخشی، LSCDFS-MKL با سایر مدل‌های پیش بینی پیشرفته مقایسه شد و نتیجه در جدول ۳-۱۶ ذکر شده است.

جدول ۳-۱۶ مقایسه عملکرد بین LSCDFS-MKL و مدل‌های دیگر پیش‌بینی [34]

Methods	AUC
LASSO-Cox	0.6179
EN-Cox	0.6332
PCRM	0.5493
RSF	0.6950
BoostCI	0.7347
superPC	0.7160
LSCDFS-MKL	0.8793

برای بررسی اینکه آیا LSCDFS-MKL می‌تواند بازماندگان کم خطر را از بازماندگان پر خطر در هر لایه از مراحل تشخیص دهد مقایسه‌ای بین مراحل مختلف سرطان انجام شد و نتیجه در شکل ۳-۱۹ نشان داده شده است.

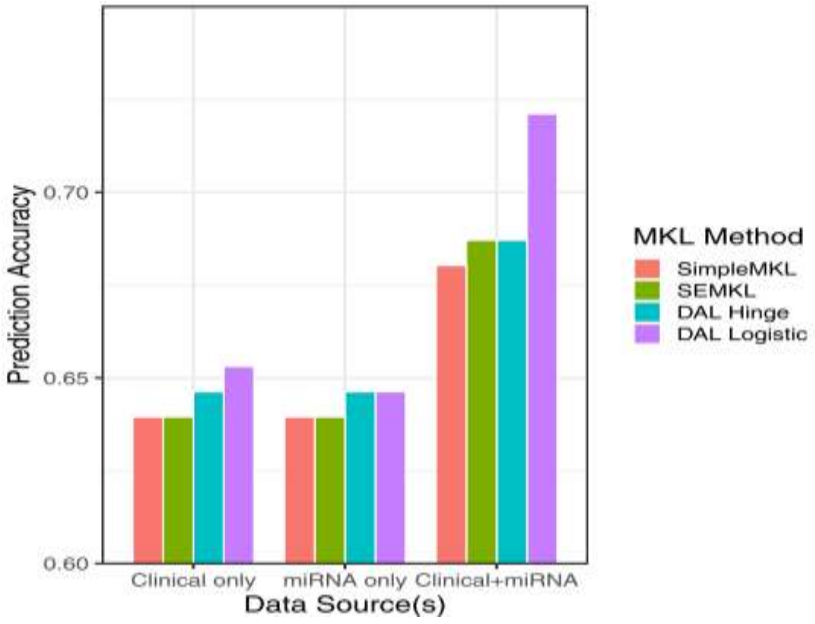


شکل ۳-۱۹ مقایسه عملکرد روش پیشنهادی درون لایه‌های هر مرحله از بیماری در معیارهای مختلف [34]

Wilson و همکارانش در سال ۲۰۱۹ به دنبال اثبات کارایی یادگیری چند هسته‌ای به عنوان ابزار ادغام داده در پیش‌بینی بقای بیمار بودند [35].

برای نشان دادن MKL به عنوان یک ابزار ادغام داده، MKL را بکار گرفتند تا بهترین هسته را برای داده‌های بیان ژن بالینی و miRNA به طور جداگانه پیدا کنند و سپس آن‌ها را در یک تجزیه و تحلیل واحد ترکیب کنند. هدف این بود که پیش بینی کنند آیا یک بیمار دارای سرطان تخمدان بیش از سه سال پس از تشخیص زندگی خواهد کرد؟

از مجموعه داده سرطان تخمدان بارگیری شده از پایگاه داده TCGA که در یکی از مقالات قبلی استفاده شده بود، استفاده کردند. از ۲۸۳ نمونه موجود در این مجموعه داده، از ۷۰٪ (۱۹۸) به عنوان نمونه آموزش و از ۳۰٪ (۸۵) به عنوان مجموعه آزمایش استفاده کردند. برای جلوگیری از نفرین ابعاد، ۶۵ ژن که بر اساس p-value دارای رتبه برتر بودند را انتخاب کردند. از این ۶۵ ژن برای هدایت SVM با اعتبارسنجی 10-fold، برای چندین هسته RBF استفاده کردند ($\sigma = 10^{-10}, \ldots, 10^{10}$) تا دامنه‌ای که منجر به بالاترین دقت پیش‌بینی می‌شود را شناسایی کنند. در نهایت در تجزیه و تحلیل MKL از یک هسته خطی و ۳ هسته RBF با $\sigma = 10^{-4}, 10^{-3}, 10^{-2}$ استفاده کردند. استفاده از داده های miRNA فقط دقت پیش بینی مشابه اطلاعات بالینی را دارد، اما استفاده از هر دو منبع داده منجر به دقت بالاتری از هر یک از منابع داده به صورت فردی می‌شود (شکل ۳-۲۰ را ببینید).



شکل ۳-۲۰ صحت پیش بینی MKL با استفاده از داده های بالینی (Clinical) و miRNA به صورت جداگانه و ترکیبی [35]

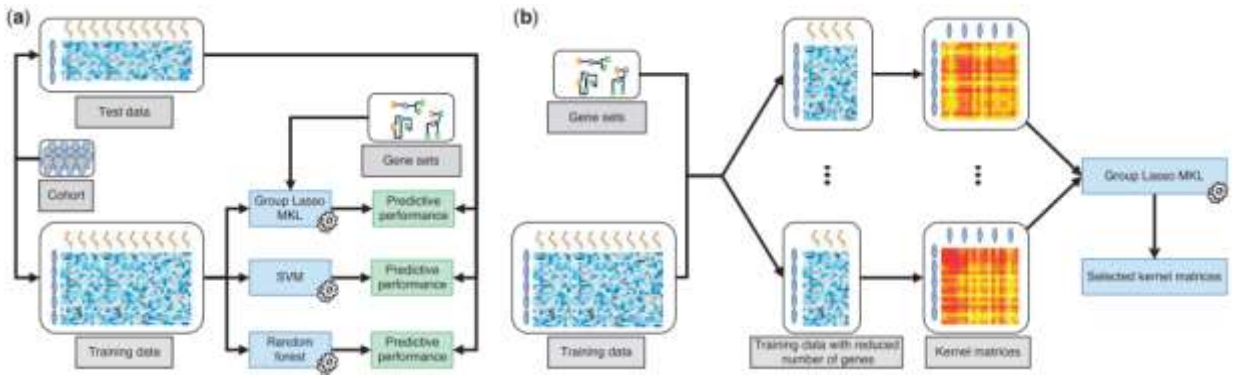
۱۵ نوع سرطان را انتخاب کردند. داده های بیان ژن برای یافتن روابط بین مجموعه‌های ژنی و نتیجه بقای دودویی استفاده شد. در این تجزیه و تحلیل بر روی ۵۰ مجموعه ژن متمرکز شدند که در مجموعه ژن مشخصی است که در یکی از مقالات قبلی معرفی شده است و بیانگر حالات یا فرایندهای بیولوژیکی کاملاً مشخص هستند. نتایج اجرای ۴ روش نامبرده شده روی هر یک از ۱۵ مجموعه داده در جدول ۳-۱۷ ذکر شده است.

جدول ۳-۱۷ دقت پیش بینی برای هر یک از چهار روش و هر نوع سرطان و تعداد مجموعه ژن (هسته) در نظر گرفته شده برای MKL. مقادیر به رنگ قرمز با دقیق ترین روش و مقادیر آبی با کم دقت ترین روش مطابقت دارند [35]

Cancer Type	Number of Gene Sets	SimpleMKL	SEMKL	DAL Hinge	DAL Logistic
BLCA	27	0.635	0.635	0.659	0.635
BRCA	50	0.551	0.573	0.570	0.557
CESC	48	0.676	0.757	0.757	0.811
COAD	40	0.632	0.632	0.596	0.632
GBM	48	0.675	0.725	0.725	0.725
HNSC	50	0.691	0.680	0.619	0.670
KIRC	50	0.729	0.698	0.677	0.698
LGG	50	0.818	0.782	0.800	0.800
LIHC	50	0.667	0.651	0.714	0.683
LUAD	2	0.593	0.630	0.605	0.605
LUSC	27	0.535	0.535	0.576	0.545
OV	9	0.667	0.621	0.636	0.636
SKCM	50	0.573	0.607	0.562	0.618
STAD	4	0.616	0.644	0.589	0.630
UCEC	8	0.719	0.684	0.772	0.702

Rahimi و Gonen در سال 2018 چارچوبی برای تعیین مرحله بیماری سرطان بر مبنای MKL با استفاد از ماشین بردار پشتیبان پیشنهاد کردند [36]. شکل ۳-۲۱ الگوریتم پیشنهاد شده در این مقاله را نشان می‌دهد.

به جای شناسایی لیستی از ویژگی‌های بیان ژن و سپس تغذیه کردن این زیر مجموعه در الگوریتم یادگیری ماشین، پیشنهاد کردند که این دو مرحله را با دانش قبلی در مورد مسیرها / مجموعه ژن‌ها در یک مدل واحد ترکیب کنند. برای این منظور یک ماتریس هسته جداگانه برای هر مجموعه ژن ایجاد کردند و آن‌ها را با استفاده از یک الگوریتم یادگیری چند هسته‌ای، Group Lasso MKL، به صورت وزندار و خطی ترکیب کردند.



شکل ۳-۲۱ الگوریتم پیشنهادی Rahimi و Gonen برای تعیین مرحله بیماری سرطان [36]

$$minimize \quad -\sum_{i=1}^{N}\alpha_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

$$w.r.t. \quad \alpha \in \mathbb{R}^N \qquad (۳۱\ ۱)$$

$$\forall i, \ \text{s.t.} \ \sum_{i=1}^{N}\alpha_i y_i = 0 \quad \mathrm{C} \geq \alpha_i \geq 0$$

یادگیری چند هسته‌ای برای به دست آوردن همزمان وزن هسته و سایر پارامترهای ماشین بردار پشتیبان، مسئله بهینه سازی زیر را حل می‌کند.

$$minimize \quad J(\eta)$$

$$w.r.t. \quad \eta \in R^P \qquad (۳۲\ ۱)$$

$$\forall m, \ \text{s.t.} \ \sum_{m=1}^{P}\eta_m = 1 \quad \eta_m \geq 0$$

بطوریکه η بردار وزن‌های هسته، P تعداد هسته‌های ورودی است. $J(\eta)$ مربوط به مسئله بهینه سازی در رابطه (۵-۱۳) می‌باشد که عبارت $k(x_i, x_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$ با $\sum_{m=1}^{P}\eta_m k_m(x_i, x_j)$ جایگزین می‌شود.

ابتدا الگوریتم وزن هسته‌ها را به صورت مقادیر یکنواخت $\eta_m^{(1)} = 1/P$ در نظر می‌گیرد. در هر تکرار t، مسئله بهینه سازی داخلی (یعنی یک مدل استاندارد SVM) را با استفاده از وزن هسته فعلی η^t حل می‌کند تا ضرایب بردار پشتیبان α^t محاسبه شوند. سپس وزن هر هسته در تکرار بعد t+1 با استفاده ازمقادیر وزن هسته در تکرار t با فرم بسته زیر محاسبه می‌شود:

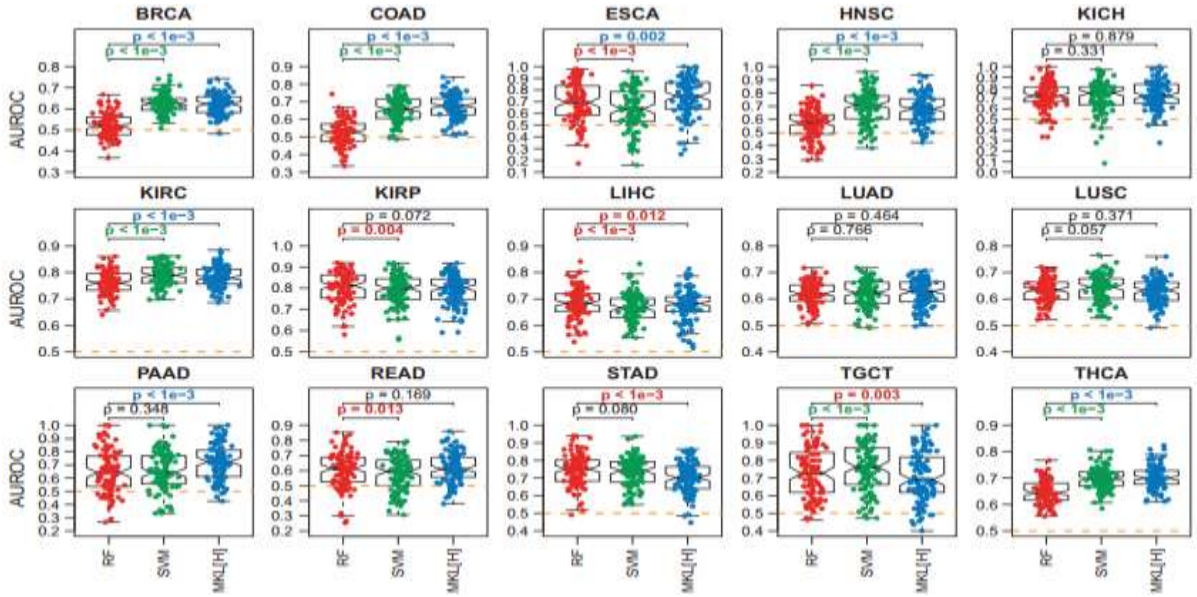
$$\forall m, \quad \eta_m^{(t+1)} = \frac{\eta_m^{(t)} \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i^{(t)}\alpha_j^{(t)} y_i y_j k_m(x_i, x_j)}}{\sum_{o=1}^{p} \eta_o^{(t)} \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i^{(t)}\alpha_j^{(t)} y_i y_j k_o(x_i, x_j)}} \qquad (۳۳\ ۱)$$

از دو مجموعه داده‌ای که یکی شامل ۱۵ مجموعه داده و دیگری شامل ۱۸ مجموعه داده RNA-seq از ۲۰ نوع بیماری سرطان موجود در پایگاه داده TCGA بود، استفاده کردند. در مجموعه داده ۱۵ تایی، تومورهای اولیه را با حاشیه نویسی مرحله I به عنوان مرحله اولیه (یعنی سرطان های موضعی) و بقیه تومورها را با حاشیه نویسی مرحله II، III یا IV به عنوان سرطان های مرحله انتهایی (به عنوان مثال گسترش منطقه ای) در نظر گرفتند و در مجموعه داده ۱۸ تایی، تومورهای حاشیه نویسی شده با مرحله I یا II را مرحله اولیه و تومورهای حاشیه دار با مرحله III یا IV را مرحله انتهایی در نظر گرفتند. همچنین مجموعه های ژنی هالمارک[۶۶] را از پایگاه داده‌های امضای مولکولی استخراج کردند که در آن هر مجموعه ژن یک حالت یا فرآیند بیولوژیکی خاص را منتقل می‌کند و بیان منسجم در سرطان‌ها را نشان می دهد و شامل ۵۰ مجموعه ژن است.

در آزمایشات برای MKL از هسته گاسین (Gaussian) استفاده کردند و تمام طبقه‌بندی های انجام شده دوکلاسه بود. سه الگوریتم یادگیری ماشین یعنی RF، SVM و MKL در مجموعه ژن‌های هالمارک (MKL[H]) را مقایسه کردند. RF و SVM از پروفایل بیان ژن تومورها برای پیش بینی مراحل پاتولوژیک آن‌ها استفاده می‌کنند (شکل ۳-۲۱ قسمت a). با این حال، علاوه بر پروفایل های بیان ژن، MKL همچنین از یک پایگاه داده مسیر / ژن استفاده می‌کند و با دور انداختن برخی از آنها در طبقه‌بندی نهایی اطلاعات اضافی در مورد تفاوت بین سرطان‌های مرحله اولیه و انتهایی را در قالب مجموعه ژن استخراج می‌کند (شکل ۳-۲۱ قسمت b). در مقایسه انجام شده بین سه الگوریتم جنگل تصادفی، ماشین بردار پشتیبان و یادگیری چند هسته‌ای با دو مجموعه داده‌ای از انواع بیماری سرطان، یادگیری چند هسته‌ای بهترین عملکرد را به دست آورد. در مجموعه داده‌ای اول، عملکرد RF و SVM مقایسه شد که SVM در ۶ تا از ۱۵ مجموعه داده (به عنوان مثال BRCA، COAD، HNSC، KIRC، TGCT و THCA) به طور قابل توجهی نتایج بهتری به دست می آورد، در حالی که RF در چهار مورد به طور قابل توجهی بهتر بود (یعنی ESCA، KIRP، LIHC و READ) و همچنین در مقایسه بین MKL و RF، یادگیری چند هسته‌ای در ۷ مجموعه داده عملکرد بهتری بدست آورد و جنگل تصادفی هم در ۳ مجموعه داده عملکرد بهتری بدست آورد (شکل ۳-۲۲ را ببینید). ترکیبی اصولی اطلاعات مجموعه ژنی به صورت توابع هسته عملکرد پیش بینی را افزایش می‌دهد حتی اگر MKL[H] از قسمت کوچکی از پروفایل های بیان ژن استفاده کند.RF و SVM از ۱۹۸۱۴ ویژگی بیان ژن استفاده کردند، در حالی که MKL[H] فقط از ۴۳۵۷ (یعنی کمتر از یک چهارم) ویژگی بیان ژن برای ژن های موجود در مجموعه های ژنی Hallmark استفاده کرد. MKL[H] عملکرد پیش

[۶۶] Hallmark

بینی را با ۱۰,۰۰٪ در BRCA، ۱۴,۸۹٪ در COAD، ۴,۳۰٪ در ESCA، ۱۱,۴۳٪ در HNSC، ۷,۵۹٪ در PAAD و ۵,۵۰٪ در THCA بهبود داد، در حالی که بیشترین افت عملکرد ۴,۳۵٪ در STAD بود. در مجموعه داده‌ای دوم هم ترتیب عملکرد مدل‌ها حفظ شد($RF < SVM < MKL$).

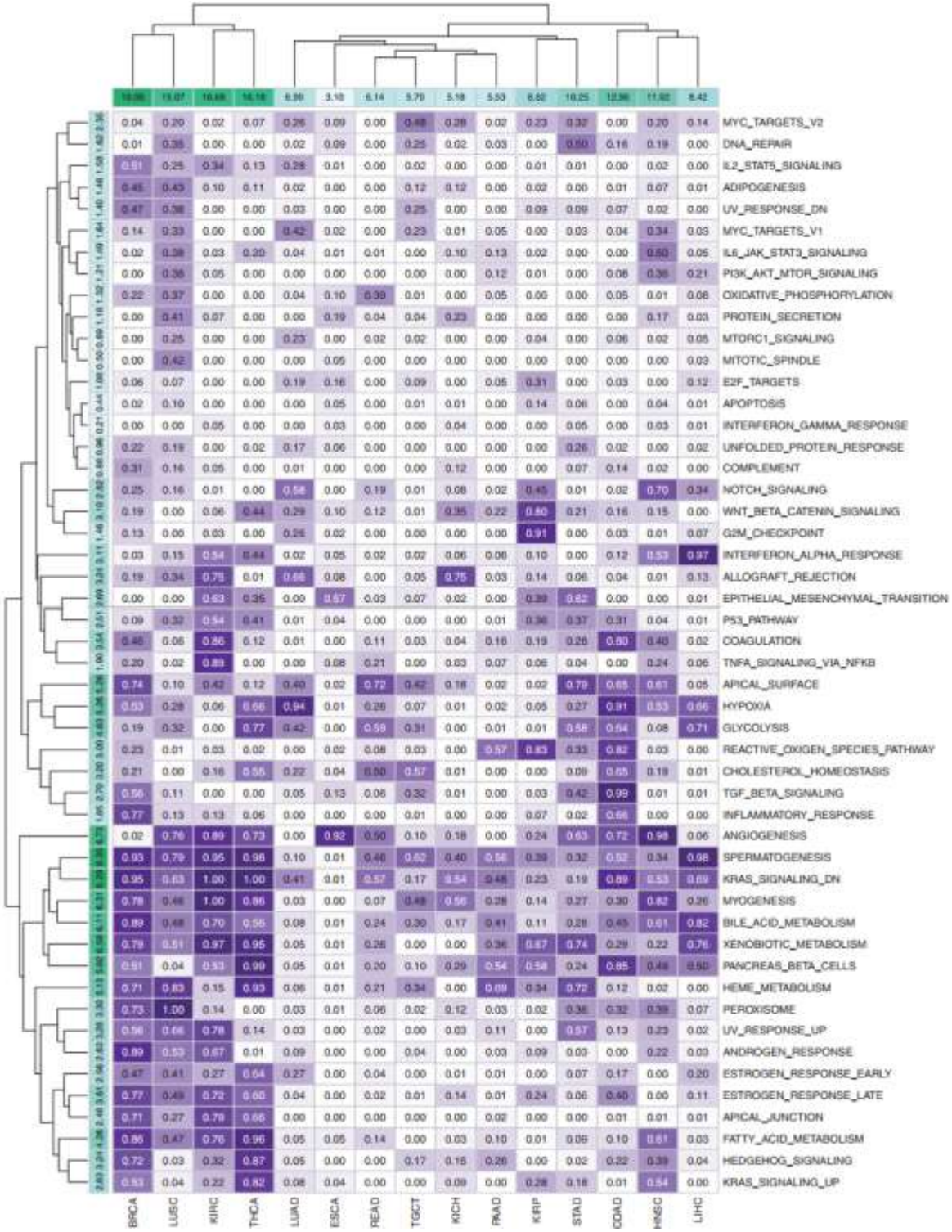


شکل ۳-۲۲ کارایی پیش‌بینی RF، SVM و MKL در ۱۵ گروه سرطان. برای نتیجه p-value، رنگ قرمز: RF بهتر می‌باشد. رنگ سبز: SVM بهتر می‌باشد. رنگ آبی: MKL بهتر می‌باشد. رنگ مشکی: تفاوتی ندارد [36].

برای نشان دادن ارتباط بیولوژیکی الگوریتم MKL[H]، توانایی آن را در شناسایی مجموعه های ژنی مربوطه بر اساس وزن هسته استنباط شده در حین آموزش تحلیل کردند. برای هر جفت مجموعه داده و مجموعه ژن، تعداد تکرارهایی را که در آن وزن هسته مربوطه غیر صفر بود را شمردند (به عنوان مثال تعداد تکرارهایی که $\eta_m \neq 0$، برقرار بود). شکل ۳-۲۳ فرکانس انتخاب ۵۰ مجموعه ژن در مجموعه Hallmark برای ۱۵ مجموعه داده در اولین مجموعه آزمایش را نشان می دهد.

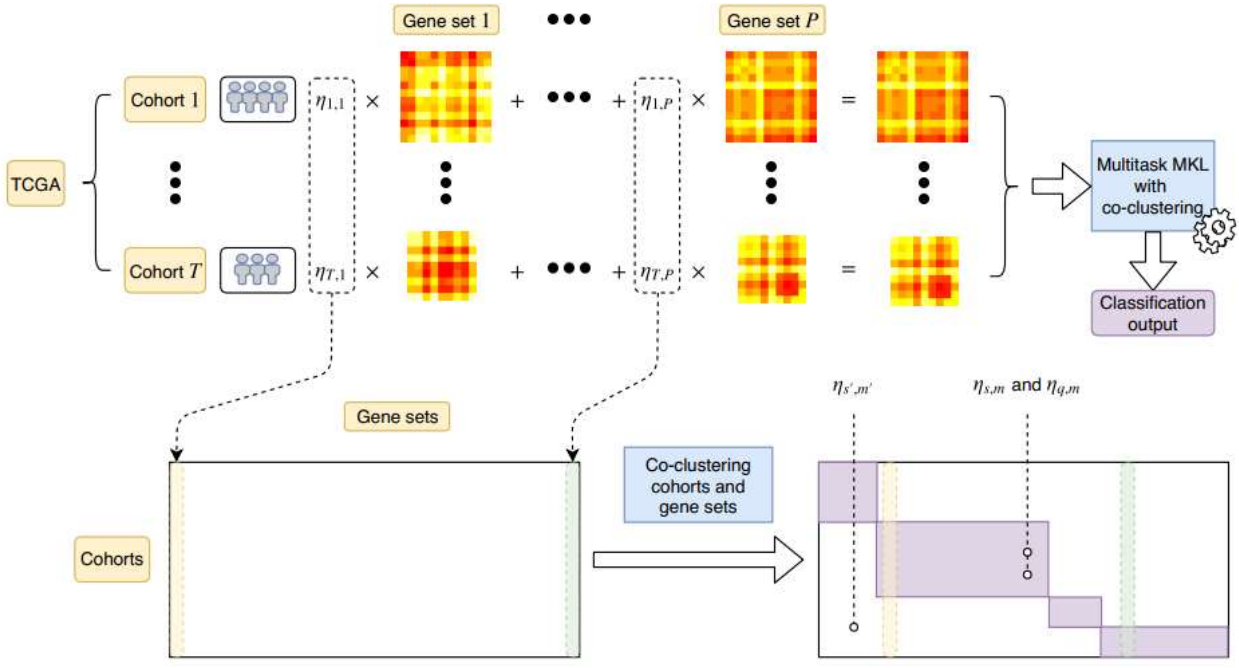
با نگاهی به ستون فراوانی‌های انتخابی می‌بینید که تشخیص برخی از انواع سرطان در مراحل اولیه و انتهایی از یکدیگر بسیار دشوارتر است. به عنوان مثال، در مجموعه داده های BRCA، KIRC، LUSC و THCA، MKL[H] به طور متوسط بیش از ۱۵ مجموعه از ۵۰ مجموعه ژن را استفاده می‌کند. با این حال در برخی از انواع بیماری‌ها مانند ESCA و PAAD، MKL[H] از مجموعه ژنی بسیار کمی استفاده می‌کند (به طور متوسط کمتر از ۶ از ۵۰ ژن تنظیم شده) و حتی عملکرد پیش بینی را در مقایسه با الگوریتم‌های RF و SVM به طور قابل توجهی بهبود می‌بخشد.

وقتی به مجموع فراوانی‌های ردیف انتخابی نگاه می‌کنید، می‌بینید که برخی از مجموعه‌های ژن به شدت در مجموعه های مختلف داده انتخاب شده‌اند. به عنوان مثال، از مجموعه های ژنی ANGIOGENESIS، KRAS_SIGNALING_DN، MYOGENESIS و SPERMATOGENESIS به طور متوسط در بیش از ۶ تا از ۱۵ مجموعه داده استفاده شد که گزارش شد مربوط به تشکیل سرطان در مراحل اولیه است. به طور مشابه چهار مجموعه ژن مربوط به متابولیسم یعنی BILE_ACID_METABOLISM، HEME_METABOLISM، PANCREAS_BETA_CELLS و XENOBI OTIC_METABOLISM به طور متوسط در بیش از ۵ مجموعه داده از ۱۵ مورد استفاده شد. برای بیماری‌های مرتبط با بافت که به ارتباط مستقیم با متابولیسم معروف هستند از جمله KIRC (کلیه)، LIHC (کبد)، PAAD (لوزالمعده) و THCA (غده تیروئید)، MKL[H] این مجموعه ژن‌های مرتبط با متابولیسم را بسیار بالا انتخاب می‌کند. فرکانس‌ها مجموعه‌های ژنی که نقش بسیار مهمی در سلول‌های اپیتلیال دارند یعنی APICAL_SURFACE و HYPOXIA به طور متوسط برای بیش از ۵ گروه از ۱۵ مجموعه داده انتخاب شدند. این دو مجموعه ژن با فراوانی بسیار بالا در BRCA (پستان)، COAD (روده بزرگ)، LUAD (ریه) و STAD (معده) که بافت‌های آنها حاوی سلول‌های اپیتلیال زیادی است، انتخاب شدند.



شکل ۳-۲۳ فراوانی انتخاب ۵۰ مجموعه ژن در مجموعه Hallmark برای ۱۵ مجموعه داده در اولین مجموعه آزمایشات. ردیف‌ها و ستون‌ها با استفاده از خوشه بندی سلسله مراتبی با فاصله اقلیدسی و توابع پیوند کامل خوشه بندی می‌شوند. مجموع فراوانی‌های ستون منتخب برای شناسایی مجموعه داده هایی که به طور متوسط تعداد بیشتری از مجموعه ژن ها را استفاده می‌کنند، گزارش شده است [36].

Rahimi و Gonen در سال ۲۰۲۰ برای پیش‌بینی سرطان‌های مرحله اولیه و مرحله انتهایی یک مدل ارائه دادند [37]. شکل ۳-۲۴ الگوریتم پیشنهادی این مقاله را نشان می دهد.



شکل ۳-۲۴ بررسی اجمالی الگوریتم یادگیری چند هسته‌ای چند وظیفه‌ای [37]

در این مطالعه از داده های ژنومی برای تمایز سرطان‌های مرحله اولیه و مرحله انتهایی استفاده کردند. از مجموعه مسیرها/ژن‌ها بهمراه داده‌های ژنومی در مدل‌های پیش‌بینی خود استفاده کردند. برای هدف مطالعه خود فقط به مجموعه ژن‌های مربوطه برای هر مجموعه مسیر/ژن نیاز داشتند. به عبارت دیگر، فعل و انفعالات بین ژن‌ها در مجموعه مسیر/ژن در اینجا مورد توجه نبود. از مجموعه های ژن هالمارک استفاده کردند. در این مطالعه سعی بر این بود که یک روش یادگیری ایجاد کنند که بتواند از اطلاعات مجموعه ژنی بهره مند شود و بتواند همه گروه‌ها را به طور همزمان مدل سازی کند. طبقه‌بندی بیماران در گروه‌های مرحله اولیه و انتهایی به عنوان یک مسئله طبقه‌بندی باینری در نظر گرفته شد.

از هر مجموعه مسیر/ژن مربوط به زیرمجموعه ای از ژن‌ها (به عنوان مثال ویژگی‌ها) برای ساخت یک عملکرد هسته استفاده شده است. سپس هسته‌ها به الگوریتم یادگیری ماشین وارد می‌شوند. با استفاده از الگوریتم تکراری با قوانین بسته به روزرسانی فرم بسته شده برای Group Lasso MKL، هسته‌های دارای بیشترین قدرت پیش‌بینی را شناسایی کردند.

انواع مختلف سرطان، علیرغم داشتن مکانیسم‌های متمایز بیولوژیکی، در مکانیسم هایشان شباهت‌هایی دارند. در این مطالعه هر گروه سرطانی به عنوان یک وظیفه مشخص در نظر گرفته شد. برای بهره برداری از این اطلاعات

از یک فرمول یادگیری چند وظیفه‌ای استفاده کردند که در آن وظایف مختلف به طور همزمان یاد گرفته می‌شوند و به این ترتیب وظایف با داده های محدود امکان بهره مندی از سایر وظایف را دارند. در یادگیری چند وظیفه‌ای میزان تشابه بین وظایف به عنوان ورودی داده می‌شود یا از طریق فرایند یادگیری استنباط می‌شود. در اینجا هدف شناسایی شباهت‌ها بین گروه‌های سرطان(یعنی وظایف) از نظر مکانیسم‌های اساسی آن‌ها است. از این رو چارچوب مناسب برای یادگیری نزدیکی بین وظایف، خوشه بندی مشترک بود. خوشه بندی مشترک به اختصاص دو مجموعه ناهمگن از موارد به تعدادی خوشه گفته می‌شود. در خوشه بندی مشترک، یک خوشه شامل گروه‌ها و مسیرها است. با توجه به سهم هر مسیر برای پیش بینی برای هر گروه که با $\eta_{s,m}$ مشخص می‌شود (s وظایف را نمایه می کند و m هسته ها را نمایه می کند)، اشتراک یادگیری چند هسته‌ای چند وظیفه‌ای (MTMKL) و مسئله خوشه بندی مشترک به صورت زیر تعریف می‌شود.

$$\underset{\eta_s \in \Delta, (\theta,\phi) \in C}{\text{Minimize}} \ \sum_{s=1}^{T} J_s(\eta_s) + \rho(\eta, \theta, \phi), \qquad (۳۴\ ۱)$$

بطوریکه Δ، $\ell_1 - norm$ وزن‌های هسته است که به صورت $\Delta = \left\{\gamma \in \mathbb{R}_+^P : \sum_{m=1}^{P} \gamma_m = 1\right\}$ تعریف می‌شود. C مجموعه ای از راه حل‌های عملی است که با دو مجموعه متغیر باینری $\chi_{k,s}$ و $\psi_{k,m}$ مشخص می‌شود و نشان می هد که گروه s (مسیر m) به خوشه k اختصاص داده شده است. $\theta_{s,m} = \sum_{k=1}^{K} \chi_{k,s} \times \psi_{k,m}$ تعیین می‌کند که گروه s و مسیر m به یک خوشه تعلق داشته باشند، در حالی که $\phi_{s,q} = \sum_{k=1}^{K} \chi_{k,s} \times \psi_{k,q}$ تعیین می‌کند که گروه‌های s و q به یک خوشه اختصاص داده شوند. $J_s(\eta_s)$ تابع هدف را برای طبقه‌بندی بهینه در گروه‌های مختلف برای یک η_s داده شده نشان می‌دهد و با حل مسئله SVM دوگانه بدست می‌آید. $\rho(\eta, \theta, \phi)$ مجازاتی را که توسط خوشه مشترک ایجاد شده است ذخیره می‌کند و به صورت زیر تعریف می شود.

$$\lambda_1 \sum_{s=1}^{T} \sum_{m=1}^{P} (1 - \theta_{s,m}) \eta_{s,m} + \lambda_2 \sum_{s=1}^{T} \sum_{m=1}^{T} \phi_{s,q} \left\| \eta_s - \eta_q \right\|^2 \qquad (۳۵\ ۱)$$

از دو عبارت جریمه در $\rho(\eta, \theta, \phi)$، اگر گروه و مجموعه ژن‌ها به یک خوشه تعلق نداشته باشند اولین مورد جریمه، استفاده از یک مجموعه ژن برای یک گروه است در حالی که دومین مورد جریمه استفاده از وزن‌های مختلف هسته برای گروه‌هایی است که در یک خوشه قرار دارند. بزرگی λ_1 نشان می‌دهد که خوشه‌بندی گروه‌ها و مجموعه ژن‌ها به چه شدت اعمال می‌شود. به طور دقیق‌تر، مقدار بسیار زیاد λ_1 باعث می‌شود که $\eta_{s,m}$ وقتی $\theta_{s,m} = 0$ برابر 0 شود. به همین ترتیب λ_2 کنترل می‌کند که چگونه خوشه بندی دقیق گروه‌ها اعمال شود به این معنی که مقدار زیاد λ_2 گروه‌های s و q را برای گرفتن وزن برابر هسته هنگام $\phi_{s,q} = 1$ مجبور می‌کند.

الگوریتم MTMKL را با یک مرحله خوشه بندی به تقسیم‌های اعتبارسنجی مختلف مجموعه داده اعمال می‌کنند و نتایج جمع بندی مشترک را برای ساخت ماتریس‌های تشابه جمع می‌کنند سپس این ماتریس‌ها را به عنوان ورودی به الگوریتم MTMKL برای ارزیابی عملکرد پیش‌بینی می‌دهند.

برای یک خوشه مشترک $(\hat{\chi}, \hat{\psi})$ از گروه‌ها و مجموعه ژن‌ها و ماتریس‌های تشابه مرتبط با $\hat{\theta}$ و $\hat{\phi}$، معادله (۱-۳۶) به شکل زیر تغییر می‌کند.

$$\underset{\eta_s \in \Delta}{\text{minimize}} \ \sum_{s=1}^{T} J_s(\eta_s) + \rho(\eta, \hat{\theta}, \hat{\phi}), \qquad (۳۶\ ۱)$$

توجه داشته باشید که به دلیل شرایط مجازات در تابع هدف (۵-۱۸)، فرمول فرم بسته دیگر نمی‌تواند برای به روزرسانی مقادیر η_s استفاده شود. از این رو یک الگوریتم صفحه برش برای حل مسئله MTMKL در (۱-۳۶) پیشنهاد می‌دهند که به شکل معادله (۱-۳۷) است.

$$\begin{aligned} &\underset{\eta_s \in \Delta, \Gamma \in \mathbb{R}_+^T}{\text{minimize}} \sum_{s=1}^{T} \Gamma_s + \rho(\eta, \hat{\theta}, \hat{\phi}), \\ &s.t. \quad \Gamma_s \geq \sum_{i=1}^{N_s} \alpha_s^i - \sum_{m=1}^{P} \eta_{s,m} K_{s,m}(\alpha_s) \qquad \forall \alpha_s \in A_s. \end{aligned} \qquad (۳۷\ ۱)$$

بطوریکه A_s نشان‌دهنده مجموعه‌ای از راه حل‌های عملی مسئله دوگانه SVM است و $K_{s,m}(\alpha_s) = 0.5 \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} \alpha_s^i \alpha_s^j y_s^i y_s^j K_{s,m}(x_s^i, x_s^j)$. به برنامه درجه دوم در (۱-۳۷) به عنوان مسئله اصلی (MP) اشاره می‌کنند. با شروع از زیرمجموعه‌های اولیه A_s، این مجموعه‌ها را با جابجایی بین MP و مشکلات SVM دوتایی تکرار می‌کنند تا زمانی که راه حل بهینه به دست آید. الگوریتم صفحه برش پیشنهادی در الگوریتم ۳-۱ که در آن UB و LB به ترتیب کران‌های بالا و پایین در بهینه (۱-۳۶) هستند به تفصیل شرح داده شده است. ζ و t_{max} پارامترهایی هستند که میزان تحمل همگرایی الگوریتم را تعیین می‌کنند. SVMهای دوگانه را می‌توان به طور مستقل برای هر گروه حل کرد که این الگوریتم را بسیار قابل تنظیم می‌کند.

الگوریتم ۳-۱ شبه کد الگوریتم صفحه برش [37]

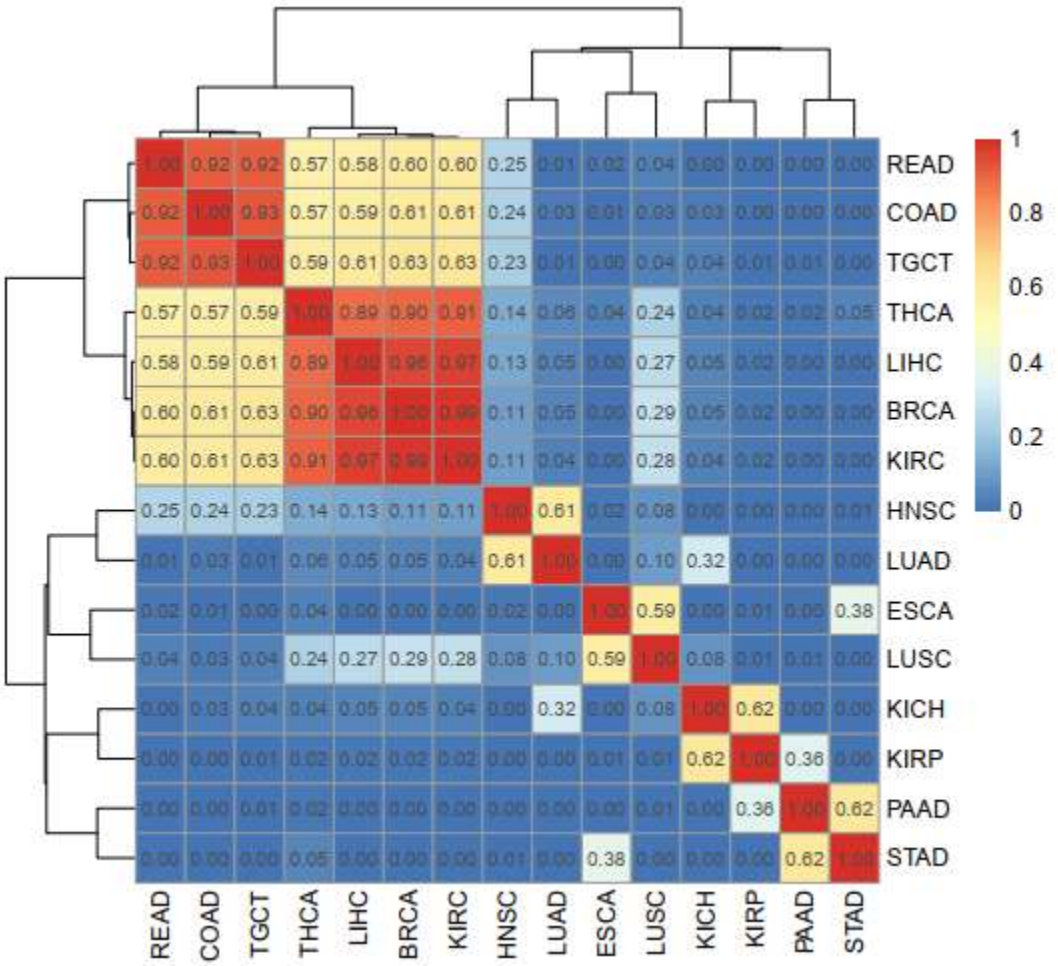
$UB \leftarrow +\infty, LB \leftarrow -\infty, \text{Gap} \leftarrow +\infty, t \leftarrow 1, \mathcal{A}_s \leftarrow \emptyset$
while Gap $> \zeta$ and $t < t_{max}$ **do**
  SOLVE MP (6) and obtain $\boldsymbol{\eta_s}$ and z^*_{MP}
  for $s \in \{1 \dots T\}$ **do**
    SOLVE dual SVM (1) for s, and obtain $\boldsymbol{\alpha_s}$ and $J_s(\boldsymbol{\eta_s})$
    $\mathcal{A}_s \leftarrow \mathcal{A}_s \cup \{\boldsymbol{\alpha_s}\}$
  end for
  $UB \leftarrow \min\{UB, \sum_s J_s(\boldsymbol{\eta_s}) + \rho(\boldsymbol{\eta}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})\}, LB \leftarrow z^*_{MP}$
  Gap $\leftarrow \frac{UB-LB}{LB}, t \leftarrow t+1$
end while

برای حل مسئله خوشه بندی مشترک در زمان مناسب از الگوریتم ابتکاری استفاده می‌کنند. در هر مرحله تعدادی از راه‌حل‌های خوشه‌بندی مشترک ناهمگن از پیش تعیین شده را انتخاب می‌کنند و وزن هسته بهینه برای هر یک از این خوشه‌های مشترک را پیدا می‌کنند. الگوریتم ۳-۲ شرح این فرآیند را ارائه می دهد.

الگوریتم ۳-۲ شبه کد الگوریتم MTMKL با گام خوشه‌بندی مشترک [37]

Initialize $\hat{\boldsymbol{\eta}}_s$ with an arbitrary set of $\boldsymbol{\eta}_s$ vectors
Derive a co-clustering $(\hat{\boldsymbol{\chi}}, \hat{\boldsymbol{\psi}})$ based on $\hat{\boldsymbol{\eta}}_s$
Let $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ be the similarity matrices associated with $(\hat{\boldsymbol{\chi}}, \hat{\boldsymbol{\psi}})$
$OPT \leftarrow (\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\chi}}, \hat{\boldsymbol{\psi}})$
$z_{OPT} \leftarrow \sum_s J_s(\hat{\boldsymbol{\eta}}_s) + \rho(\boldsymbol{\eta}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$
while (stopping criterion is not met) **do**
    Derive B best heterogeneous co-clustering solutions using the diversification algorithm based on $\hat{\boldsymbol{\eta}}_s$
    for each co-clustering solution $(\boldsymbol{\chi}, \boldsymbol{\psi})_b$ **do**
        Let $(\boldsymbol{\theta}_b, \boldsymbol{\phi}_b)$ be the similarity matrices associated with $(\boldsymbol{\chi}, \boldsymbol{\psi})_b$
        SOLVE Algorithm 1 and obtain $(\boldsymbol{\eta}_s)_b$ vectors for $(\boldsymbol{\chi}, \boldsymbol{\psi})_b$
        $z_b \leftarrow \sum_s J_s((\boldsymbol{\eta}_s)_b) + \rho((\boldsymbol{\eta})_b, \boldsymbol{\theta}_b, \boldsymbol{\phi}_b)$
    end for
    $\hat{b} = \arg\min_b \{z_b\}$
    $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\chi}}, \hat{\boldsymbol{\psi}}) \leftarrow (\boldsymbol{\eta}, \boldsymbol{\chi}, \boldsymbol{\psi})_{\hat{b}}$
    if $z_{\hat{b}} < z_{OPT}$ **then**
        $OPT \leftarrow (\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\chi}}, \hat{\boldsymbol{\psi}})$
        $z_{OPT} \leftarrow z_{\hat{b}}$
    end if
end while

بسیاری از مطالعات در ادبیات یادگیری چند وظیفه‌ای از اندازه گیری مشخصی برای تشابه بین وظایف استفاده می‌کنند که به چنین معیاری از شباهت، ماتریس شباهت گفته می‌شود. هم خوانی گروه‌ها و مجموعه‌های ژنی می‌تواند مبنایی برای استخراج چنین ماتریس‌های تشابه بین گروه‌ها (ϕ) و گروه‌ها با مجموعه‌های ژنی (θ) باشد. برای این منظور، خوشه بندی مشترک را چندین بار بر روی نمونه داده‌ها انجام می‌دهند و نتایج را به صورت ماتریس های فراوانی $\bar{\theta}$ و $\bar{\phi}$ جمع می‌کنند ، یعنی ورودی های $\bar{\theta}$ و $\bar{\phi}$ نشان می‌دهد که چند وقت یکبار دو گروه (یا یک گروه و یک مجموعه ژن) به همان خوشه اختصاص داده شده‌اند. شکل ۳-۲۵ چنین ماتریس شباهت گروه-گروه را نشان می‌دهد.

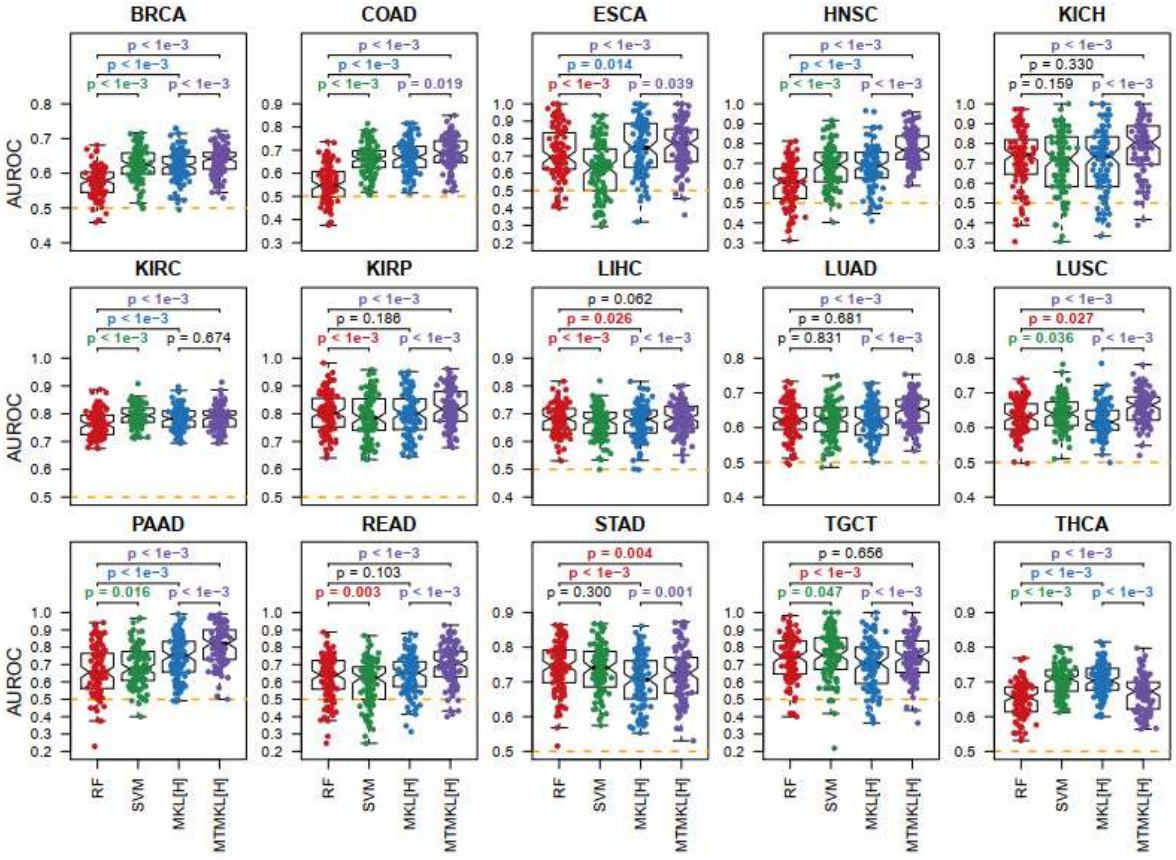


شکل ۳-۲۵ ماتریس تشابه گروه-گروه با حل ۱۰۰ تکرار مسئله MTMKL با یک مرحله خوشه بندی در ۱۵ گروه TCGA [37]

برای محک زدن از سه روش دیگر استفاده کردند که شامل RF، SVM و group Lasso MKL بود. به طور تصادفی ۲۰٪ از داده ها را در هر گروه برای آزمایش در هر تکرار نگه داشتند. برای انتخاب بهترین مجموعه پارامترها برای هر روش با تقسیم ۸۰٪ باقیمانده داده‌ها به چهار قسمت با اندازه تقریباً یکسان از یک استراتژی اعتبارسنجی 4-fold استفاده کردند. هر قسمت یک بار برای آزمایش استفاده می‌شود در حالی که از قسمت‌های باقیمانده برای آموزش استفاده شده است. این استراتژی اعتبارسنجی 4-fold ، ۱۰۰ بار تکرار شده است. برای مقایسه نتایج طبقه بندی الگوریتم های مختلف از سطح زیر منحنی مشخصه عملکرد سیستم (AUROC) استفاده کردند. مقادیر بزرگتر AUROC با عملکرد پیش بینی بهتر مطابقت دارند. همچنین از هسته گاسین برای آزمایشات استفاده کردند.

برای الگوریتم MTMKL، با استفاده از اعتبار سنجی متقابل 4-fold، پارامتر تنظیم قاعده C و پارامترهای جریمه λ_1 و λ_2 را انتخاب کردند، جایی که $\lambda_1 \in \{10^{-1}, 10^{0}, \dots, 10^{+4}\}$، $C \in \{10^{-4}, 10^{-3}, \dots, 10^{+5}\}$ و λ_2 به عنوان

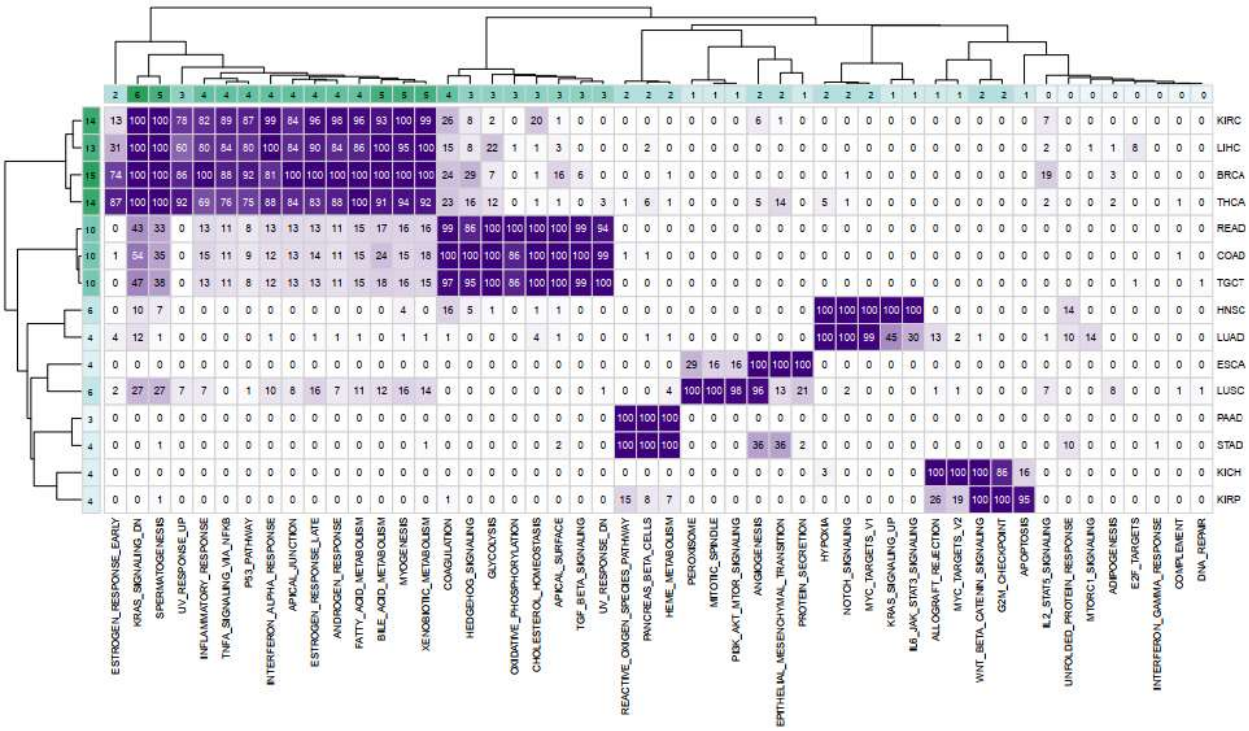
$\gamma\lambda_1$ برای مقادیر مختلف $\gamma \in \{5^{-1}, 5^0, \dots, 5^{+2}\}$ تنظیم شده است. برای خوشه‌بندی مشترک، تعداد خوشه K را از {3,4,5} و تعداد خوشه های ناهمگن B در هر تکرار را از {2,3,4,5} انتخاب می‌کنند. در الگوریتم ۳-۱ از پارامترهای $\zeta = 10^{-6}$ و $t_{max} = 400$ استفاده کردند.



شکل ۳-۲۶ عملکرد الگوریتم های RF، SVM، MKL و MTMKL روی ۱۵ مجموعه داده TCGA [37]

در مقایسه MTMKL با الگوریتم RF که به عنوان الگوریتم پایه در مقایسه استفاده می‌شود، نتایج نشان می‌دهد که MTMKL در ۱۲ تا از ۱۵ گروه از نظر آماری به طور قابل توجهی بهتر عمل می کند در حالی که RF فقط در یک گروه (به عنوان مثال STAD) قادر به پیشی گرفتن از MTMKL بود و همچنین MTMKL نسبت به SVM نیز در ۱۲ گروه عملکرد بهتری بدست آورد. الگوریتم MTMKL در تمام گروه‌ها از الگوریتم MKL بهتر عمل کرد. این نشان می‌دهد که شباهت بین گروه‌های مختلف سرطان معنی دار است و ماتریس‌های تشابهی که در این کار به دست آمد قدرت تفسیری در تشخیص سرطان‌های مرحله اولیه و مرحله انتهایی با استفاده از داده‌های ژنومی دارند. از طرف دیگر با مقایسه عملکرد پیش بینی MTMKL با MKL (شکل ۳-۲۶)، مشاهده می‌کنید که MTMKL قادر به بهبود دقت برای گروه های اندازه کوچکتر(به عنوان مثال ESCA، KICH، KIRP، PAAD، READ و TGCT) بدون به خطر انداختن صحت گروه‌های دیگر است. این انتظار می‌رفت زیرا هدف از

یادگیری چند وظیفه‌ای کمک به بهبود دقت وظایف با داده های محدود با استفاده از وظایف با داده های فراوان است. روابط بین گروه‌ها و مسیرهای مشخص شده در نتایج با بسیاری از مطالعات که تأثیر چند مسیر را بر روی انواع خاصی از سرطان بررسی می‌کنند، مطابقت دارد. روابط مشخص شده توسط الگوریتم MTMKL را نشان می‌دهد.



شکل ۳-۲۷ فراوانی انتخاب ۳۰ مجموعه ژن در مجموعه Hallmark برای ۱۵ مجموعه داده. ردیف‌ها و ستون‌ها با استفاده از خوشه بندی سلسله مراتبی با فاصله اقلیدسی و توابع پیوند کامل خوشه بندی می‌شوند [37].

در جدول زیر خلاصه‌ای از پژوهش‌های انجام شده، ذکر شده است.

جدول ۳-۱۸ خلاصه‌ای از پژوهش‌های انجام شده بر مبنای ماشین بردار پشتیبان و یادگیری چند هسته‌ای با کاربرد تشخیص نوع، مرحله و زیرگروه بیماری و تشخیص بقای بیمار

مقاله	روش	توضیحات
Bhala 2017 [24]	با استفاده از مدل مبتنی بر آستانه و ماشین بردار پشتیبان به شناسایی ژن‌ها برای پیش‌بینی مرحله بیماری پرداختند.	این نوع برنامه یادگیری ماشین که در آن حداقل تعداد ژن‌ها برای ترسیم مراحل اولیه و اواخر سرطان با استفاده از داده هایی با توان بالا استفاده می‌شود، می‌تواند بینش بهتری برای درک مکانیسم های متاستاز در سرطان های مختلف فراهم کند. این مطالعه با پیش بینی مرحله بیماران ccRCC دارای پتانسیل بالینی و همچنین پیش‌بینی است. این نوع معیارهای نمره‌گذاری/ آستانه ممکن است امضای بیان ژن پیش‌بینی کننده با تعداد محدودی از ویژگی‌ها را شناسایی کند اما تفسیر آن‌ها به دلیل داده‌های ورودی بُعدی و همبسته بسیار دشوار است [36]. کمبود عملکرد استاندارد با ارزش مانند Sn، Sp و AUC [38]. ارزش عملی بالینی این مطالعات به دلیل فقدان چندین گروه مستقل برای تأیید، هنوز محدود است [39].
Zhao 2018 [23]	انتخاب ویژگی با استفاده از رگرسیون لجستیک و سپس طبقه‌بندی سرطان روده بزرگ با استفاده از ماشین بردار پشتیبان با هسته RBF	کاهش ویژگی های زائد و انتخاب بهترین پیش‌بینی کننده‌های سرطان روده بزرگ، استفاده از عوامل سنتی(مانند سن و BMI) و عوامل ژنتیکی در کنار هم.
Du 2017 [25]	با استفاده از یادگیری چند هسته‌ای یک انتخاب ویژگی دو مرحله‌ای انجام می‌دهد. ابتدا یک مجموعه ویژگی مرتبط را انتخاب و سپس مجموعه انتخابی را فشرده می‌کند.	اثر بخشی و توانایی اجرا در انواع مختلفی از داده‌های بیان این روش‌های فقط بیان (کاملاً بر اساس مقادیر بیان ژن) قابل اعتماد نیستند زیرا ممکن است نویز در داده های بدست آمده از آزمایشات آزمایشگاهی وجود داشته باشد [40].

Speicher 2015 [27]	ادغام انواع داده مختلف با استفاده از یادگیری چند هسته‌ای برای پیش بینی زیرگروه‌های سرطان و همچنین استفاده از خوشه بندی برای تعیین میزان تاثیر داده های مختلف در پیش بینی	انعطاف پذیدی بالا در انتخاب روش کاهش ابعاد و انتخاب نوع داده. انتخاب چندین ماتریس هسته برای هر نوع داده بر اساس توابع مختلف هسته اگرچه الگوریتم پیشنهادی عملکرد خوبی در پیش بینی سرطان ریه به دست آورد، اما هنوز هم محدودیت هایی وجود دارد. به عنوان مثال، اندازه نمونه TCGA به اندازه کافی بزرگ نیست که تجزیه و تحلیل آینده را محدود می کند. هنگامی که اطلاعات بیشتری از بیماران مبتلا با SCC ریه در دسترس باشد، انتظار می‌رود که روش پیشنهادی عملکرد بهتری داشته باشد. علاوه بر این، انتظار می رود مدل های یادگیری عمیق برای استخراج ویژگی، انتخاب و تلفیق نیز استفاده شود. علاوه بر این، مدل پیشنهادی نمی‌تواند داده های از دست رفته را کنترل کند، که به کارگیری مستقیم آن در سناریوی واقعی را دشوار می کند [27].
Spicher 2018 [28]	استفاده از خوشه بندی و یادگیری چند هسته‌ای برای شناسایی زیر گروه‌های بیماری	امکان استفاده از چند نوع داده مختلف، تفسیر بیشتر خوشه‌های بیمار، کاهش نویز در هر ماتریس هسته به دلیل افزایش همگنی ویژگی‌ها با شناسایی خوشه‌های بیمار
Tao 2019 [29]	ادغام داده‌های متمایز با استفاده از یادگیری چند هسته‌ای برای شناسایی زیرگروه‌های سرطان	تلفیق داده‌های ناهمگن با استفاده از یادگیری چند هسته‌ای
Sun 2018 [32]	ادغام داده‌های ژنتیکی و تصاویر پاتولوژیک با استفاده از یادگیری چند هسته‌ای برای پیش‌بینی بقای بیماران دارای سرطان پستان	. استفاده از مجموعه داده‌های مختلف ژنتیکی و تصاویر پاتولوژیک
Zhang 2019 [34]	ادغام داده‌های ژنتیکی و تصاویر پاتولوژیک با استفاده از یادگیری چند هسته‌ای برای پیش‌بینی بقای بیماران دارای سرطان کارسینوما سلول سنگفرشی ریه	ادغام کارامد داده‌های ژنومیک و تصاویر پاتالوژیک

Wilson 2019 [35]	استفاده از داده‌های متمایز و یادگیری چند هسته‌ای برای پیش‌بینی بقای بیمار	تلفیق داده‌های ناهمگن با استفاده از یادگیری چند هسته‌ای
Rahimi 2018 [36]	استفاده از یادگیری چند هسته‌ای برای پیش‌بینی مرحله بیماری	استفاده همزمان از داده‌های بیان ژن و دانش قبلی مجموعه داده مسیر/ژن برای شناسایی مکانیسم‌های بیولوژیکی که سرطان‌های مرحله اولیه و انتهایی را از یکدیگر متمایز می‌کند و همچنین منجر به کاهش ویژگی‌های استفاده شده در الگوریتم پیشنهادی می‌شود. گروه‌های بیماری به صورت جداگانه مورد بررسی قرار گرفتند که از مدل‌های پیش‌بینی یادگیری به طور مشترک سود نمی‌برند [37].
Rahimi 2020 [37]	شناسایی شباهت بین گروه‌های سرطان با استفاده از خوشه بندی مشترک و سپس یادگیری همزمان گروه‌های سرطان با استفاده از یادگیری چند هسته‌ای و در نهایت پیش بینی گروه‌ها	یافتن شباهت بین گروه‌های مختلف سرطان و یادگیری همزمان گروه‌های مختلف سرطان برای کم رنگ کردن مشکل کمبود نمونه در برخی از گروه‌ها

فصل ۴

پیاده سازی

۴-۱ مقدمه

در این فصل ابتدا پایگاه داده TCGA را شرح می‌دهیم سپس مقاله Rahimi و Gonen ۲۰۱۸ ([36]) را پیاده سازی می‌کنیم.

۴-۲ پایگاه داده TCGA[۶۷]

سرطان یک بیماری جدی و تهدید کننده زندگی است که در همه دنیا شیوع دارد. طبق گزارش‌های Globocan برای سال ۲۰۱۸، سرطان مهمترین علت مرگ و میر در سراسر جهان است که تقریبا ۹,۶ میلیون مرگ را در بر می‌گیرد. اگر محققان تلاش خود را برای تشخیص زودهنگام سرطان انجام دهند، می‌توان این ارقام را کاهش داد. با این حال، این تلاش‌ها به مواد ژنتیکی و بالینی بی شماری از بیماران سرطانی نیاز دارد [41]. TCGA یک برنامه ژنومیک سرطانی برجسته بیش از ۲۰،۰۰۰ سرطان اولیه را به صورت مولکولی توصیف کرده و ۳۳ نمونه سرطان را شامل می شود. این تلاش مشترک بین انستیتوی ملی سرطان ایالات متحده آمریکا و موسسه ملی تحقیقات ژنوم انسانی از سال ۲۰۰۶ آغاز شد و محققان از رشته های مختلف و موسسات مختلف را گرد هم آورد. این داده ها که منجر به بهبود توانایی محققان در تشخیص، درمان و پیشگیری از سرطان شده است، برای استفاده افراد جامعه تحقیقی در دسترس عموم باقی خواهد ماند [42]. پورتال داده TCGA دیگر عملیاتی نیست و در ژوئن ۲۰۱۶ موسسه ملی سرطان برنامه GDC[۶۸] را راه اندازی کرد، یک برنامه تحقیقاتی که به اشتراک گذاری داده‌های ژنومی و بالینی بین محققان کمک می‌کند. برنامه GDC یکی از بزرگترین و جامع ترین مجموعه داده‌های ژنومیک سرطان در جهان است. GDC داده‌هایی را تولید می‌کند که در سه سطح داده طبقه بندی شده است (جدول ۴-۱ را ببینید) [43].

جدول ۴-۱ سطوح داده‌های پرتال GDC [43]

شماره سطح	نوع داده
۱	داده‌های خام[۶۹]
۲	داده‌های نرمال شده[۷۰]
۳	داده های جمع شده[۷۱]

[۶۷] The Cancer Genome Atlas
[۶۸] Genomic Data Commons
[۶۹] Raw data
[۷۰] Normalized data
[۷۱] Aggregated data

سطح ۱ نمایانگر داده های خام و غیر عادی است، سطح ۲ نشان دهنده سطح متوسط پردازش و یا عادی سازی داده ها و سطح ۳ نشان دهنده داده‌های جمع شده، عادی و یا تقسیم بندی شده است [43]. از مجموعه داده‌های این پایگاه داده در بسیاری از مقالات مرتبط با بیماری سرطان استفاده شده است.

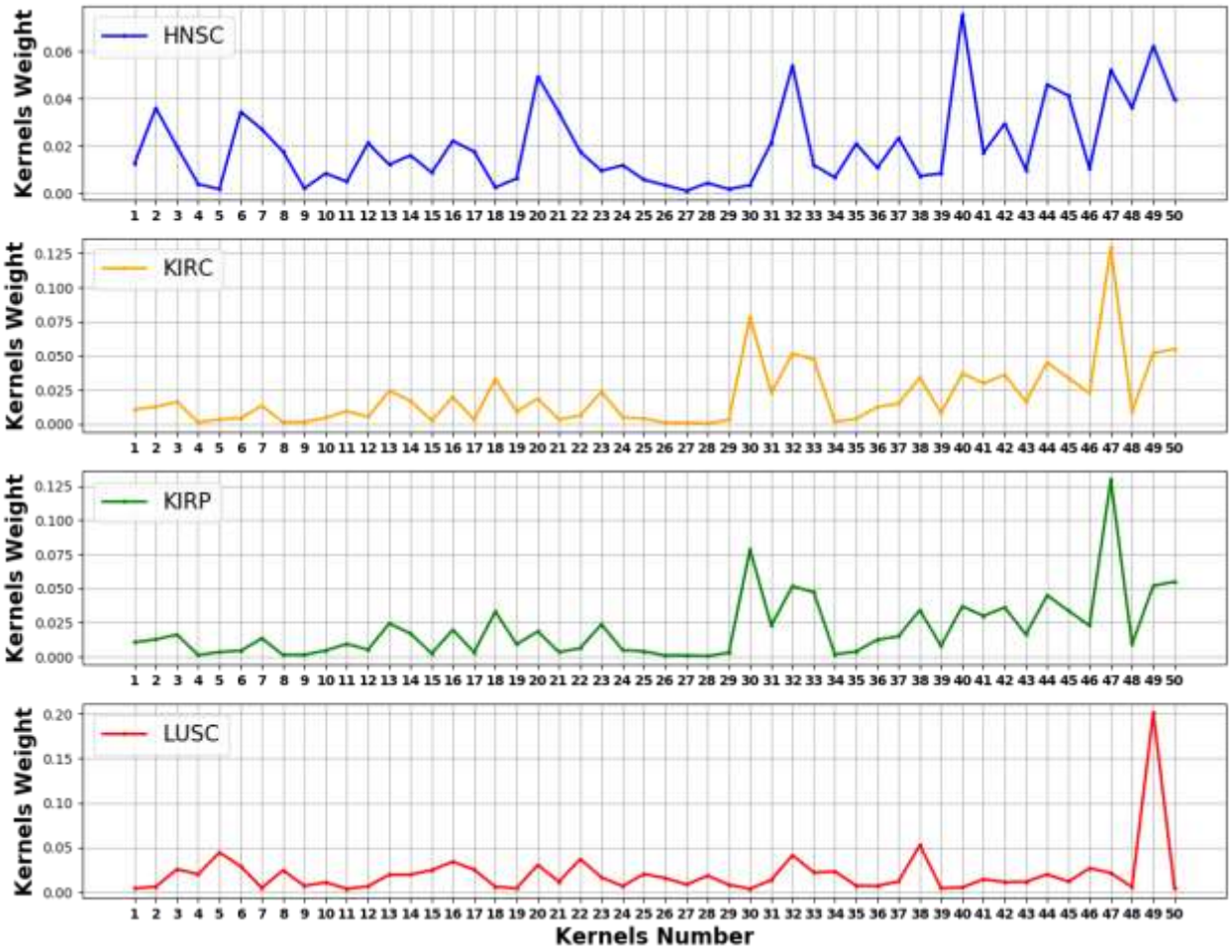
۴-۳ پیاده سازی

در این قسمت از کدهای پیاده سازی مقاله Rahimi و Gonen ۲۰۱۸ که توسط نویسنده به اشتراک گذاشته شده استفاده کردیم. از مجموعه داده‌ای شامل ۴ نوع سرطان از جمله HNSC[۷۲]، KIRC[۷۳]، KIRP[۷۴] و LUSC[۷۵] که از پایگاه داده TCGA بارگیری شده بود استفاده کردیم. همچنین از مجموعه‌های ژنی هالمارک که از پایگاه داده امضاء مولکولی استخراج شده و در آن هر مجموعه ژن یک حالت یا فرآیند بیولوژیکی خاص را منتقل می‌کند و بیان منسجم در سرطان‌ها را نشان می‌دهد، استفاده کردیم. این مجموعه شامل ۵۰ مجموعه ژن است و اندازه آن‌ها بین ۳۲ تا ۲۰۰ ژن متفاوت است. مجموعه داده شامل ۴ نوع سرطان از مقاله ([44]) دریافت شد که پیش‌پردازش شده نیز بود. از الگوریتم‌های ماشین بردار پشتیبان، یادگیری چند هسته‌ای بدون وزندهی و یادگیری چند هسته‌ای با وزندهی (Group Lasso MKL روش مقاله [36]) استفاده کردیم. AUC پیش‌بینی روش‌ها در جدول ۴-۲ ذکر شده است. در این پیاده سازی از هسته گاسین استفاده کردیم. برای هر دو الگوریتم یادگیری چند هسته‌ای با توجه به ۵۰ مجموعه ژن موجود در هالمارک، ۵۰ هسته در نظر گرفتند. برای الگوریتم یادگیری چند هسته‌ای بدون وزندهی، وزن همه هسته ها برابر بود ولی برای روش مقاله مدنظر هر هسته با توجه به نوع بیماری وزنی متفاوت داشت. وزندهی به هسته‌ها در شکل ۴-۱ نمایش داده شده است.

جدول ۴-۲ سطح زیر منحنی ROC (AUC) حاصل از پیش‌بینی الگوریتم های ماشین بردار پشتیبان، یادگیری چند هسته‌ای بدون وزندهی و یادگیری چند هسته‌ای با وزندهی (Group Lasso MKL). نتایج ذکر شده میانگین اجرای ۵ تکرار الگوریتم‌ها است.

	SVM	Unweightrd_MKL	Group Lasso MKL
HNSC	0.643	0.73	0.726
KIRC	0.729	0.783	0.788
KIRP	0.868	0.896	0.894
LUSC	0.476	0.445	0.445

[۷۲] Head and Neck squamous cell carcinoma
[۷۳] Kidney renal clear cell carcinoma
[۷۴] Kidney renal papillary cell carcinoma
[۷۵] Lung squamous cell carcinoma



شکل ۴-۱ وزن هسته‌ها با توجه به مجموعه‌های ژنی هالمارک در پیاده سازی الگوریتم یادگیری چند هسته‌ای با وزندهی (Group Lasso MKL). هر کدام از ۴ نمودار مربوط به وزن هسته‌های یکی از ۴ نوع بیماری می‌باشد. وزن هر هسته i-اُم نشان دهنده میزان اهمیت i-اُمین مجموعه داده هالمارک برای همان نوع بیماری می‌باشد.

فصل ۵

نتیجه گیری

بیوانفورماتیک دانشی است که به جنبه‌های ریاضی و محاسباتی زیست شناسی برای فهم و پردازش داده‌های زیستی می پردازد. گسترش روزافزون و حجم عظیم داده‌های ژنومی و نیاز به ذخیره، بازیابی و تحلیل مناسب این داده‌ها موجب پیدایش علم بیوانفورماتیک گردید. انجام الگوریتم‌های بیوانفورماتیک به صورت دست نویس و غیر خودکار بسیار دشوار است، به این منظور برای به دست آوردن دانش از داده‌های زیستی از ابزارها و روش‌های یادگیری ماشین استفاده می‌شود. یکی از روش‌های یادگیری ماشین که کاربرد بسیاری در بیوانفورماتیک دارد یادگیری چند هسته‌ای می‌باشد. روش‌های هسته‌ای طی دهه گذشته شاهد افزایش محبوبیت زیادی در بیوانفورماتیک بوده‌اند. در این سمینار به بررسی روش‌های مبتنی بر یادگیری چند هسته‌ای برای پیش‌بینی بیماری و بقای بیمار و همچنین پیش‌بینی مرحله و زیر گروه‌های بیماری پرداخته‌ایم. مطالب علمی موجود در این زمینه مورد بررسی قرار گرفت و مشاهده شد که استفاده از ماشین بردار پشتیبان و یادگیری چند هسته‌ای برای طبقه‌بندی اطلاعات بیماری با داده‌های ژنی مناسب بوده و نتایج خوبی حاصل شده است.

طبق مطالعات انجام شده نکته قابل توجه در روش یادگیری چند هسته‌ای استفاده از هسته‌های مختلف می‌باشد که امکان استفاده همزمان از داده‌های مختلف، انتخاب ویژگی، استفاده از هسته‌های مختلف با پارامترهای متفاوت برای مجموعه داده را فراهم می‌کند. یکی از دلایل عمده محبوبیت روش‌های هسته در بیوانفورماتیک قدرت آن‌ها در ادغام داده است. دومین مزیت روش‌های هسته این است که می‌توان آن‌ها را به راحتی روی داده‌های ساخت یافته (مثال گراف ها، مجموعه ها، سری های زمانی و رشته ها) اعمال کرد.

در بسیاری از مطالعات انجام شده از روش یادگیری چند هسته‌ای استفاده شده است و الگوریتم های مختلفی بر این اساس ارائه شده است که اکثراً تفاوت‌هایی در زمینه وزن دهی به هسته‌ها، نوع و تعداد هسته‌های استفاده شده دارند.

مراجع

[1] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 35, p. 35, 2017.

[2] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafe, A. P'erez and V. Robles, "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 86-112, 2006.

[3] J. Wang, M. Zaki, H. Toivonen and D. Shasha, "Introduction to Data Mining in Bioinformatics," in *Data Mining in Bioinformatics*, 2005, pp. 3-8.

[4] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal and Y. Zhou, "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?," *Briefings in Bioinformatics*, vol. 19, no. 3, pp. 482-494, 2018.

[۵] م. جان نثار, م. معظم جزی و س. سیدی, آشنایی با بیوانفورماتیک و کاربردهای آن, تهران: انتشارات دانش بنیان فناور, ۱۳۹۶, p. 9.

[6] S. Hochreiter, "Institute of Bioinformatics, Johannes Kepler University Linz," [Online]. Available: https://www.bioinf.jku.at/teaching/current/ws_sapvl/BioInf_I_Notes.pdf. [Accessed 04 06 2017].

[7] D. V.Volgin, "Chapter 17 - Gene Expression: Analysis and Quantitation," in *Animal Biotechnology*, 2014, pp. 307-325.

[8] P. M.Das and R. Singal, "DNA Methylation and Cancer," *clinical oncology*, vol. 22, pp. 4632-4642, 2004.

[9] M. B.Lanktree, C. T.Johansen, T. R.Joy and R. A.Hegele, "Chapter 6 - A Translational View of the Genetics of Lipodystrophy and Ectopic Fat Deposition," *Progress in Molecular Biology and Translational Science*, vol. 94, pp. 159-196, 2010.

[10] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," 7 Jun 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[11] Shen Yin and Jiapeng Yin, "Tuning kernel parameters for SVM based on expected square," *Information Sciences*, Vols. 370-371, pp. 92-102, 2016.

[12] A. Ben-Hur, C. ong, S. Sonnenburg, B. Scho ikopf and Ratach, "Support Vector Machines and Kernels for Computational Biology," *PLoS Comput Biol*, vol. 4, no. 10, 2008.

[13] K.-P. Wu and S.-D. Wang, "Choosing the kernel parameters for support vector machines by the inter-cluster," *Pattern Recognition*, pp. 710-717, 2008.

[14] R. Amami, D. Ben Ayed and N. Ellouze, "Practical Selection of SVM Supervised Parameters with Different Feature Representations for Vowel Recognition," *JDCTA*, vol. 7, 2013.

[15] L. Chen, "Support Vector Machine — Simply Explained," 7 Jan 2019. [Online]. Available: https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496.

[16] K. Borgwardt, "Kernel Methods in Bioinformatics," *Springer Handbooks of Computational Statistics*, pp. 317-334, 2011.

[17] M. Gonen and E. Alpaydın, "Multiple Kernel Learning Algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211-2268, 2011.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[19] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Computers in Biology and Medicine*, vol. 121, 2020.

[20] A. Paulson Chazhoor, "ROC curve in machine learning," 8 Jul 2019. [Online]. Available: https://towardsdatascience.com/roc-curve-in-machine-learning-fea29b14d133.

[21] A. Gaonkar, "Confused about Confusion Matrix?," 28 Aug 2020. [Online]. Available: https://medium.com/analytics-vidhya/confused-about-confusion-matrix-2ce7c52345dd.

[22] F. Dorey, "In Brief: The P Value: What Is It and What Does It Tell You?," *Clin Orthop Relat Res*, vol. 468, no. 8, pp. 2297-2298, 2010.

[23] D. Zhao, H. Liu, Y. Zheng, Y. He, D. Lu and C. Lyu, "A reliable method for colorectal cancer prediction based on feature selection and support vector machine," *Medical & Biological Engineering & Computing*, vol. 57, no. 4, pp. 901-912, 2019.

[24] S. Bhalla, K. Chaudhary, R. Kumar, M. Sehgal, H. Kaur, S. Sharma and G. P. S. Raghava, "Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer," *Scientific RepoRts*, vol. 7, no. 1, 2017.

[25] W. Du, Z. Cao, T. Song, Y. Li and Y. Liang, "A feature selection method based on multiple kernel learning with expression profiles of different types," *BioData Mining*, 2017.

[26] H. Kim, G. H. Golub and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187-198, 2005.

[27] N. K. Speicher and N. Pfeifer, "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery," *Bioinformatics*, vol. 31, no. 12, pp. 268-275, 2015.

[28] N. K. Speicher and N. Pfeifer, "An interpretable multiple kernel learning approach for the discovery of integrative cancer subtypes," *arXiv*, 2018.

[29] M. Tao, T. Song, W. Du, S. Han, C. Zuo, Y. Li, Y. Wang and Z. Yang, "Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data," *Genes*, vol. 10, no. 3, 2019.

[30] E. Whitley and J. Ball, "Statistics review 6: Nonparametric methods," *Crit Care*, vol. 6, no. 6, pp. 509-513, 2002.

[31] S. Glen, "Benjamini-Hochberg Procedur," StatisticsHowTo.com: Elementary Statistics for the rest of us!, [Online]. Available: https://www.statisticshowto.com/benjamini-hochberg-procedure/.

[32] D. Sun, A. Li, B. Tang and M. Wang, "Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 45-53, 2018.

[33] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani , D. Botstein and R. B. Altman, "Missing value estimation methods for DNA microarrays," *BIOINFORMATICS,* vol. 17, no. 6, pp. 520-525, 2001.

[34] A. Zhang, A. Li, J. He and M. Wang, "LSCDFS-MKL: A multiple kernel based method for lung squamous cell carcinomas disease-free survival prediction with pathological and genomic data," *Journal of Biomedical Informatics,* vol. 94, 2019.

[35] C. M.Wilson, K. Li, X. Yu, P.-F. Kuan and X. Wang, "Multiple-kernel learning for genomic data mining and prediction," *BMC Bioinformatics,* vol. 20, no. 1, 2019.

[36] A. Rahimi and M. Gönen, "Discriminating early- and late-stage cancers using multiple kernel learning on gene sets," *Bioinformatics,* vol. 34, no. 13, pp. 412-421, 2018.

[37] A. Rahimi and M. Gönen, "A multitask multiple kernel learning formulation for discriminating early- and late-stage cancers," *Bioinformatics,* vol. 36, no. 12, p. 3766–3772, 2020.

[38] S. Bhalla, H. Kaur, A. Dhall and G. P.S.Raghava, "Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients," *Scientific Reports,* vol. 9, no. 1, 2019.

[39] K. Yang, Y. Xiao, T. Xu, W. Yu, Y. Ruan, P. Luo and F. Cheng, "Integrative analysis reveals CRHBP inhibits renal cell carcinoma progression by regulating inflammation and apoptosis," *cancer Gene Thrapy,* vol. 27, pp. 607-618, 2020.

[40] H. FangOng, N. Mustapha, H. Hamdan, R. Rosli and A. Mustapha, "Informative top-k class associative rule for cancer biomarker discovery on microarray data," *Expert Systems with Applications,* vol. 146, 2020.

[41] Deepali , N. Goel and P. Khandnor, "TCGA: A multi-genomics material repository for cancer research," *Materials Today: Proceedings,* vol. 28, pp. 1492-1495, 2020.

[42] NATIONAL CANCER INSTITUTE, "The Cancer Genome Atlas Program," [Online]. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

[43] M. Settino and M. Cannataro, "Survey of main tools for querying and analyzing TCGA Data," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, 2018.

[44] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Computers in Biology and Medicine,* vol. 121, no. 16, 2020.