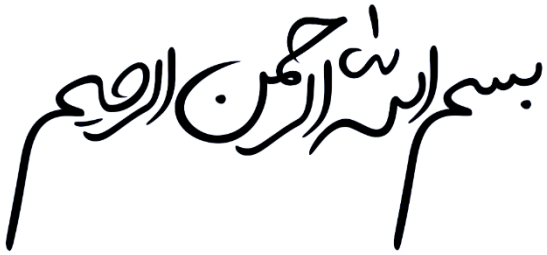

بسم الله الرحمن الرحيم

دانشگاه تربیت مدرس

دانشکده علوم زیستی

پایان‌نامه کارشناسی ارشد

رشته بیوانفورماتیک

پیش‌بینی پپتیدهای ضدسرطانی با استفاده از یادگیری ماشین

نگارنده

زینب محمدتبار

استاد راهنما

دکتر پرویز عبدالمالکی

بهمن ۱۴۰۰

چکیده

اگرچه تلاش‌های بسیار زیادی در جهت توسعه درمان‌های جدید سرطان صورت گرفته‌است اما هنوز هم سرطان یکی ازعلل عمده مرگ در جهان است. شیمی درمانی با وجود عوارض جانبی شدید بر روی سلول‌های طبیعی، هنوز هم روش اصلی مورد استفاده در درمان سرطان در مراحل پیشرفته یا متاستاتیک است. بنابراین، توسعه داروهای ضدسرطانی جدید با سمیت کم برای سلول‌های طبیعی، یک مسیر جدید برای درمان سرطان فراهم می‌کند. در حال حاضر یکی از روش‌های موثر نسبت به شیمی‌درمانی استفاده از پپتیدهای ضدسرطانی است. اما شناسایی این نوع پپتیدها به روش‌های تجربی، پرهزینه و زمانبر است. بنابراین در سال‌های اخیر استفاده از روش‌های محاسباتی جهت شناسایی پپتیدهای ضدسرطانی، مورد توجه پژوهشگران بسیاری قرارگرفته‌است. به‌همین منظور در این پژوهش نیز سعی شده‌است تا توسط یادگیری ماشین مدلی برای پیش‌بینی پپتیدهای ضدسرطانی از پپتیدهای غیرضدسرطانی ایجاد شود.

در این پژوهش با جمع‌آوری داده‌های مربوط به توالی پپتیدهای ضدسرطانی و غیرضدسرطانی، استخراج ویژگی از توالی این پپتیدها توسط کتابخانه پایتون iFeature، ساخت و آموزش دو طبقه‌بند جنگل‌تصادفی و ماشین بردارپشتیبان توسط کتابخانه پایتون scikit-learn بر روی ویژگی‌های استخراج شده، به نتایج با صحت بالای ۸۲ درصد توسط طبقه‌بند جنگل‌تصادفی بر روی داده‌های تست مستقل دست یافته‌ایم. همچنین با بررسی نتایج مدل‌های آموزش دیده برروی داده‌های تست‌مستقل، می‌توان اعلام کرد که ترکیب دو ویژگی QSOrder و APseduoAAC توسط هر دو طبقه‌بند جنگل‌تصادفی و ماشین بردارپشتیبان بالاترین عملکرد برای تشخیص پپتیدهای ضدسرطانی از بین پپتیدهای غیرضدسرطانی را با ۸۱ درصد صحت دارند.

کلید واژه: پیش‌بینی پپتیدهای ضدسرطانی، یادگیری ماشین، جنگل تصادفی، ماشین بردارپشتیبان، سرطان

فهرست مطالب

فهرست شکل‌ها و نمودارها

فهرست جدول‌ها

۱ فصل اول: مقدمه

۱-۱ مقدمه‌ای بر پپتیدهای ضدسرطانی[1]

سرطان یکی از علل عمده مرگ‌ومیر میلیون‌ها انسان در جهان است[۱, ۲]. برآورد آژانس بین‌المللی تحقیقات سرطان (IARC)[2] نشان می‌دهد که در سال ۲۰۱۸، ۱۸.۱ میلیون مورد جدید سرطان و ۹.۶ میلیون مرگ ناشی از سرطان در سراسر جهان رخ داده‌است. همچنین، آمار در سراسر جهان نشان می‌دهد که سرطان ریه، پستان و روده بزرگ بیش ازهمه انواع سرطان‌ها بوده‌است[۳].

[1] Anticancer Peptides (ACP)

[2] International Agency for Research on Cancer

یکی از روش‌های معمول در درمان سرطان، پرتودرمانی است. اما در روش پرتودرمانی علاوه بر از بین رفتن سلول‌های سرطانی، سلول‌های طبیعی نیز دچار آسیب می‌شوند. مشاهده شده‌است، که سلولهای سرطانی در برابر داروهای شیمی‌درمانی که جهت مهار تکثیر DNA به کار می‌روند به صورت مقاوم و خاموش به تکثیر خود ادامه می‌دهند[۴]. بنابراین، توسعه داروهای ضدسرطانی جدید با سمیت کم برای سلول‌های طبیعی، یک مسیر جدید برای درمان سرطان فراهم می‌کند. پپتیدها به دلیل عواملی نظیر اندازه کوچک، سنتز راحت، فعالیت و ویژگی بالا و تنوع بیولوژیکی توجه دانشمندان را به خود جلب کرده‌است. از جمله پپتیدهای درمانی که در سال‌های اخیر به منظور درمان سرطان مورد توجه قرارگرفته‌اند، می‌توان پپتیدهای کاتیونی ضدسرطانی و پپتیدهای نفوذپذیر سلولی را نام برد.

تحقیقات نشان می‌دهد که پپتیدهای ضدسرطانی، در مقایسه با روش‌های شیمی‌درمانی و پرتودرمانی، عملکردی مناسب برای شناسایی و مهار سلول‌های سرطانی بدون ایجاد مقاومت دارویی دارند[۵].

بنابراین در حال حاضر یکی از روش‌های موثردر درمان سرطان که توجه دانشمندان را به خود جلب کرده-است، استفاده از پپتیدهای ضدسرطانی است. در ادامه به بررسی ویژگی‌های زیستی پپتیدهای ضدسرطانی می‌پردازیم.

۱-۲ ویژگی‌های زیستی پپتیدهای ضدسرطانی

همانطور که گفته شد یکی از راهکارهای نوین برای درمان سرطان، استفاده از پپتیدهای ضدسرطانی است که با توجه به انتخابی بودن نسبت به سلول‌های سرطانی، عوارض کمتر و اثربخشی بالا مورد توجه بسیاری

از دانشمندان قرار گرفته‌است[۶]. یکی از انواع پپتیدهای مورد استفاده، پپتید ضدمیکروبی[۳] است که بخشی از پاسخ ایمنی ذاتی در برابر میکروب‌ها در بسیاری از گونه‌ها است. این پپتیدها دارای وزن مولکولی کم (۴۰-۱۰ اسیدآمینه) و ساختار آمفی‌پاتیک و کاتیونی هستند که به آن‌ها اجازه می‌دهد تا غشاهای منفی و سلول‌های سرطانی (همانند باکتری‌ها) را هدف قرار دهند. غشاء بیرونی سلول‌های سرطانی نسبت به سلول-های طبیعی دارای فسفولیپیدهایی با بار منفی بیشتر نظیر فسفاتیدیل سرین و گلیکوپروتئین‌ها و گلیکوز آمینوگلیکان های منفی هستند[۷]. پپتیدهای ضدسرطانی با توجه به ویژگی‌های کاتیونی و آمفی‌پاتیک خود با تعاملات الکترواستاتیک به سلول‌های سرطانی متصل شده و از این رو این پپتیدها از طریق نکروز یا آپوپتوز سمیت سلول‌های سرطانی را محدود می‌کنند[۸-۱۰].

پپتیدهای ضدسرطانی در درمان دیابت، سرطان و بیماری‌های قلبی عروقی به کار برده می‌شوند[۱۱]. اکثر پپتیدهای ضدسرطانی به صورت محلول و بدون ساختار هستند و تنها پپتید ضدمیکروبی انسانی به نام 37LL با توجه به حضور یک یا چند پیوند دی‌سولفیدی با صفحات β-sheet به شکل گلومرولی دیده شده‌است. از نظر تنوع ساختاری چند صد پپتید ضدمیکروبی مورد مطاله قرارگرفته‌است. این پپتیدها دارای ساختار آلفاهلیکسی غنی از سیستئین(نظیر Cecropins) یا دارای ساختار β-sheet(نظیر Defensins) هستند، همچنین حضور اسیدآمینه آرژنین، پرولین، هیستدین و تریپتوفان در این پپتیدها(نظیر Indolicidin) معمول است[۱۲].

[3] Antimicrobial Peptides (AMP)

Table 1
Summary of select naturally occurring cationic antimicrobial peptides with anticancer activities

Peptide	Source	Primary amino acid sequence[a]	Class	Net[b]	Anticancer activity
BMAP-28	*Bos taurus*	GGL**R**SLG**RK**IL**R**AW**KK**YGPIIVPII**R**I	α-Helix	+7	Membranolytic
HNP-1 (β-defensin)	*Homo sapiens*	AC_1YC_2**R**IPAC_3IAGE**RR**YGTC_2IYQG**R**LWAFC_3C_1	β-Sheet	+3	Membranolytic Antiangiogenic?
Lactoferricin B	*Bos taurus*	F**K**C_1**RR**WQW**R**M**KK**LGAPSITC_1V**RR**AF	β-Sheet	+8	Membranolytic Apoptosis inducer Antiangiogenic
LL-37	*Homo sapiens*	LLGDFF**RK**S**K**E**K**IG**K**EF**KR**IVQ**R**I**K**DFL**R**NLVP**R**TES	α-Helix	+6	Membranolytic
Magainin 2	*Xenopus laevis*	GIG**K**FLHSA**KK**FG**K**AFVGEIMNS	α-Helix	+3	Membranolytic Apoptosis inducer?
Melittin	*Apis mellifera*	GIGAVL**K**VLTTGLPALISWI**KRKR**QQ	α-Helix	+6	Membranolytic PLA_2[c] activator PLD[d] activator
Tachyplesin I	*Tachypleus tridentatus*	**K**WC_1F**R**VC_2Y**R**GIC_2Y**RR**C_1**R**	β-Sheet	+6	Binds hyaluronan and activates complement (C1q) Antiangiogenic? Induces cancer cell differentiation

[a] Amino acid sequences are given in one-letter code. Subscripts indicate pairings of Cys residues that form disulfide bonds. Boldface indicates cationic amino acid residues.
[b] At neutral pH.
[c] Phospholipase A_2.
[d] Phospholipase D.

شکل ۱–۱: تعدادی پپتید ضدمیکروبی طبیعی با فعالیت ضدسرطانی[۶]

در دهه‌های اخیر استفاده از این پپتیدها به عنوان عوامل ضدسرطان به عنوان یک روش درمانی نوین در نظر گرفته‌شده‌است [۱۳-۱۵].

۱-۳ طبقه‌بندی پپتیدهای ضدسرطانی

رزونانس مغناطیسی هسته‌ای NMR[۴] به عنوان یک روش مفید برای مطالعه جزئیات ساختاری بسیاری از پپتیدهای ضدمیکروبی شناخته شده، پدید آمده‌است. تجزیه و تحلیل ساختار سه بعدی این پپتیدها منجر به درک بهتری از عملکرد این پپتیدها می‌گردد[۱۶]. براساس ساختارهای شناخته شده توسط NMR پپتیدهای ضدسرطانی به پنج گروه طبقه‌بندی می‌شوند.

[4] NMR spectroscopy

۱-۳-۱ پپتیدهای ضدسرطانی با ساختار آلفا-هلیکس (*α*-Helical ACPs)

برای اولین بار در پپتید Cecropins بسیاری از ویژگی‌ها شناسایی شد و مطالعات اولیه با NMR نشان داد که پپتید Cecropin-A دارای ساختارهلیکس است. در ساختار هلیکس این پپتید ۱۵٪ الکل Hexafluoro Isopropyl شرکت دارد. در نتیجه پپتیدهای ضدمیکروبی دارای ساختار هلیکس، پپتیدهایی آمفی‌پاتیک با سطوح آبگریز و حاوی بار خالص مثبت هستند. پپتیدهای ضدمیکروبی با ساختار هلیکس بیشتر از سایر ساختارها مشاهده می‌شوند. پپتید ضدمیکروبی به نام Magainins گروه دیگری از پپتیدها با ساختارآلفا-هلیکس است. این پپتیدها از پوست نوعی قورباغه آفریقایی جدا می‌شوند. مطالعات NMR نشان داده‌است، که Cecropins نیز مانند Magainins دارای ساختارآمفی پاتیک-هلیکس در ۲۵٪ از Trifluoroethanol است[۱۷].

۱-۳-۲ پپتیدهای ضدسرطانی غنی از سیستئین

پپتیدهای نوتروفیل انسان HNP-1-2-3 اولین پپتیدهای غنی از سیستئین جدا شده از گرانول انسان بودند[۱۸]. پپتیدهای غنی از سیستئین با ۳۰ اسیدآمینه در طیف گسترده‌ای از موجودات زنده وجود دارد. این پپتیدها به صورت یک موتیف حفاظت شده حاوی شش سیستئین با سه پیوند دی‌سولفید درون مولکولی است. اکثرا موقعیت پل دی‌سولفیدی بین C1-C4، C2 C3-C5-C6 است. مطالعات کریستالوگرافی HNP-3 توسط XRAY، در ترکیب با سانتریفوژ تعادل رسوب، نشان می‌دهد که این پپتید به صورت دایمر وجود دارد[۱۹]. در مطالعات NMR بر روی ساختار Defensin مشخص گردید، که این پپتید حاوی سه رشته آنتی‌پارالل است. پپتید Drosomycin جدا شده از مگس سرکه شامل چهار پیوندهای دی‌سولفید است و ازسه رشته آنتی‌پارالل ساخته شده‌است. در این پپتید یک ساختار هلیکس بین دو رشته اول وجود دارد[۲۰].

۱-۳-۳ پپتیدهای ضدسرطانی با ساختار بتا-شیت (β-Sheet ACPs)

تعدادی از پپتیدهای ضدسرطانی با طول ۲۰ اسیدآمینه دارای یک ساختار سنجاق سری هستند. این پپتیدها دارای یک یا دو پیوند دی‌سولفیدی می‌باشند. پپتیدهای خرچنگ نعل اسبی، Tachyplesin وPolyphemusin II که به صورت موتیف کوتاه سنجاق سری هستند، این پپتید کوتاه توسط دو پیوند دی‌سولفیدی پایدار می گردد [۲۰, ۲۱]. مطالعات NMR همراه با ساختارهای 3D نشان می‌دهد که پپتید Tachyplesin شباهت بسیار زیادی با پپتیدهای جدا شده از خوک دارد. مولکول پپتیدی صفحات آنتی پارالل به یک ساختار Turn متصل شده‌اند و شامل دو پل دی‌سولفید هستند [۲۱].

۱-۳-۴ پپتیدهای ضدسرطانی غنی از اسیدآمینه‌های منظم

تعدادی از پپتیدهای ضدسرطانی دارای تعداد زیادی از یک اسیدآمینه خاص هستند. این پپتیدها دارای ساختارهای مختلف آلفا-هلیکس و بتا-شیت هستند. پپتید Histatin، غنی از اسیدآمینه هیستیدین از بزاق انسان جدا شده و این پپتید در برابر کاندید آلبیکانس فعال است[۲۲]. در حالی که پپتیدهای Cathelicidins، غنی از پرولین و دارای ساختار نامنظم هستند. پپتیدهای Tritrpticin سرشار از تریپتوفان و پپتیدهای Bactenecin Bac-5 and Bac-7 مانند پپتیدهای Cathelicidins غنی از اسیدآمینه پرولین هستند؛ در حالی که پپتید PR-39، غنی از اسیدآمینه آرژنین است[۲۳-۲۶].

۱-۳-۵ پپتیدهای ضدسرطانی با اسیدآمینه‌های تغییریافته یا اصلاح‌شده

تعدادی از پپتیدهای ضدسرطانی دارای اسیدآمینه‌های اصلاح‌شده هستند. بهترین نمونه از این پپتیدها، پپتیدهای تولید شده توسط باکتری‌ها است. از این پپتیدها می‌توان به نیسین[۵]، که توسط باکتری لاکتوکوکوس‌لاکتیس[۶] تولید می‌شود، اشاره کرد. این پپتید باکتریایی از اسیدآمینه‌های نادر مانند(3-Methyllationine, Dyhydroalanine and Dehydrobutyrine) تشکیل شده‌است و این پپتیدها برعلیه باکتری‌های گرم مثبت فعال هستند[۲۷]. از دیگر پپتیدهای ضدمیکروبی که دارای اسیدآمینه اصلاح‌شده هستند، می‌توان به پپتید Leucocin A، که از Leuconostoc gelidum جدا شده‌است، اشاره کرد. این پپتید ساختار کانفورماسیونی آمفی‌پاتیک از خود نشان می‌دهد، که این ساختارها دارای نقش مهمی در تعامل با غشا هست[۲۸].

۱-۴ روش‌های تجربی شناسایی پپتیدهای ضدسرطانی

برای شناسایی پروتئین‌ها یا پپتیدها به روش تجربی اول از همه نیازمند استخراج نمونه آزمایشگاهی و یا آزمایش‌هایی روی موجود زنده است. هم‌چنین برای کشف عملکرد موردنظر که در این پژوهش خاصیت ضدسرطانی پپتیدها مدنظر است نیازمند روش‌های آزمایشگاهی خاص خود است. تمام مراحل از یافت نمونه تا بررسی عملکرد پپتید نیازمند متخصص و تجهیزات زیستی است. این پر واضح است که انجام روش‌های آزمایشگاهی برای شناسایی عملکرد یک پپتید نیازمند زمان و هزینه مالی و انسانی قابل توجهی است. به‌همین دلیل همیشه امید است که با مطالعه بر روی روش‌های محاسباتی بتوان محققان زیستی را تا حدی

[5] Nissen

[6] Lactococcus Lactis

نجات داد، تا محققان زیستی برای انجام آزمایش‌های خود، از بین کاندیدهای میلیونی برای مطالعه و آزمایش به کاندیدهای صدتایی یا حتی هزارتایی برسند. این پدیده باعث میشود تا محققان زیستی در زمان و هزینه‌های مالی و مواد صرفه‌جویی کنند و این زمان را صرف تحقیقات با ارزش‌تری کنند. شناسایی پپتیدهای ضدسرطانی نیز از این قاعده مستثنی نمی‌شود. به‌همین دلیل در این پایان‌نامه نیز سعی می‌شود تا با انجام یک روش محاسباتی به پیش‌بینی‌ای از پپتیدهای ضدسرطانی برسیم.

۱-۵ روش‌های محاسباتی پیش‌بینی پپتیدهای ضدسرطانی

همانطور که در قسمت قبل اشاره شد، روش‌های تجربی نیازمند صرف انرژی انسانی و هزینه‌های مالی زیاد است، در نتیجه نیاز به رویکردهای محاسباتی درکنار روش‌های تجربی وجود دارد.

یکی از روش‌های محاسباتی برای پیش‌بینی پپتیدهای ضدسرطانی، استفاده از داده‌های بدست آمده توسط محققان زیستی برای یادگیری‌ماشین است. در ادامه به توضیح مختصری از الگوریتم‌های یادگیری ماشین به کاررفته در این پایان‌نامه می‌پردازیم.

۱-۵-۱ روش‌های مبتنی بر یادگیری ماشین

یادگیری ماشین به مجموعه وسیعی از روش‌های محاسباتی برای درک داده‌ها اشاره دارد که می‌تواند الگوهای پیچیده در یک مجموعه را پیدا و بر اساس آن‌ها تصمیم‌گیری کند. این روش‌ها را می‌توان به دو دسته یادگیری با نظارت[۷] و یادگیری بدون نظارت[۸] تقسیم کرد. یادگیری ماشین با نظارت، یک مدل آماری

[7] Supervised

[8] Unsupervised

برای پیش‌بینی برچسب کلاس[۹] بر اساس یک یا چند توصیف‌کننده[۱۰] (ویژگی[۱۱]) ایجاد می‌کند. در این روش با هدف پیش‌بینی دقیق کلاس‌ها برای مشاهدات آینده، مدلی[۱۲] تنظیم می‌شود که برچسب کلاس‌ها را به ویژگی‌ها مرتبط کند.

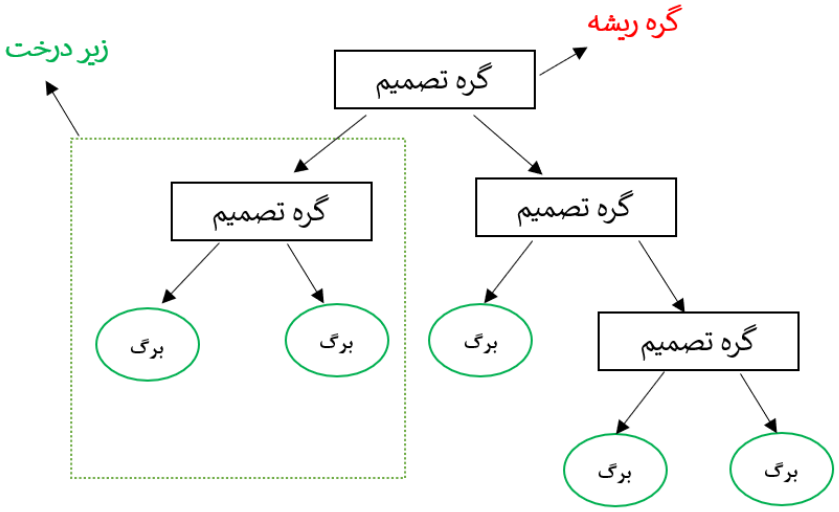
در یادگیری ماشین بدون نظارت، به توصیف‌کننده‌ها کلاسی نسبت داده نشده‌است. در این روش با پیدا کردن ساختار و روابط بین توصیف‌کننده‌ها، داده‌ها کلاستر می‌شوند.[۱۳]

در مسائلی که داده‌ها برچسب کلاس دارند؛ از روش‌های با نظارت که به آن‌ها طبقه‌بندی[۱۴] می‌گویند، استفاده می‌شود. در ادامه به اختصار طبقه‌بندهای استفاده شده در این پایان‌نامه توضیح داده می‌شود.

۱-۵-۱-۱ طبقه‌بند درخت تصمیم[۱۵]

درخت نوع خاصی از گراف است و درخت تصمیم یک روش برای پیش‌بینی بر اساس این ساختار است. به دلیل تفسیرپذیر بودن درخت‌تصمیم، این روش از پرکاربردترین مدل‌ها در مسائل یادگیری ماشین کلاسیک است. این ساختار از گره ریشه[۱۶] شروع می‌شود و به برگ‌ها[۱۷] (گره پایانی) ختم می‌شود. برگ است که مشخص می‌کند یک نمونه در چه کلاسی قرار خواهدگرفت. گره‌های داخلی با استفاده از قوانینی و با توجه به ویژگی‌ها، نمونه‌ها را به زیر مجموعه‌های مختلف تقسیم می‌کند، تا در نهایت مشخص شود هر نمونه به کدام کلاس تعلق دارد[۲۹].

[9] Class Label
[10] Descriptor
[11] Feature
[12] Predictors
[13] Clustering
[14] Classification
[15] Decision Tree Classifier
[16] Root Node
[17] Leaves



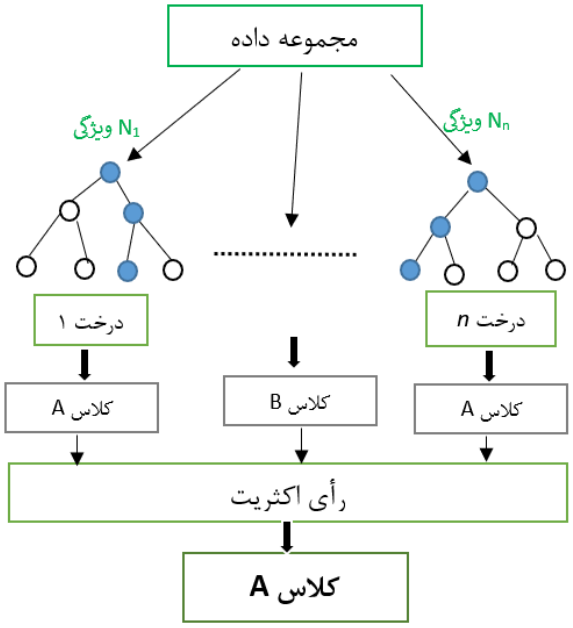
شکل ۱–۲: ساختار کلی یک درخت تصمیم

۱-۵-۱-۲ طبقه‌بند جنگل تصادفی[18]

طبقه‌بند جنگل تصادفی از تعداد زیادی درخت‌تصمیم استفاده می‌کند. به همین دلیل الگوریتم جنگل-تصادفی در دسته الگوریتم‌های یادگیری گروهی[19] قرار می‌گیرد. در این روش هر درخت‌تصمیم به صورت تصادفی تعدادی از ویژگی‌ها را انتخاب می‌کند و در مورد هر نمونه پیش‌بینی انجام می‌دهد. برای تعیین برچسب یک نمونه بین تمام برچسب‌هایی که هر درخت به آن نمونه نسبت داده است؛ رأی اکثریت گرفته می‌شود[۳۰].

[18] Random Forest Classifier

[19] Ensemble Learning



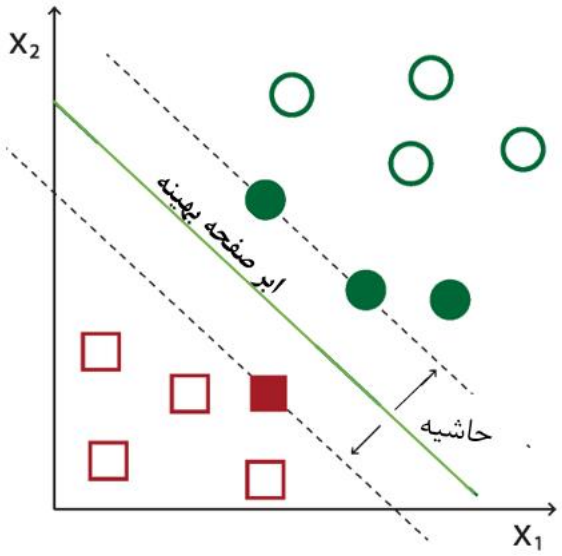
شکل ۱–۳: ساختار کلی یک طبقه‌بند جنگل‌تصادفی

۱-۵-۱-۳ طبقه‌بند ماشین بردارپشتیبان[۲۰]

ماشین بردارپشتیبان یک الگوریتم یادگیری ماشین است که سعی دارد با ایجاد یک ابرصفحه داده‌ها را به نسبت تعداد برچسب‌ها تقسیم کند. این طبقه‌بند در ایجاد ابرصفحه مناسب به دنبال بیشینه کردن فاصله نمونه‌ها از ابرصفحه است. به این فاصله حاشیه[۲۱] گفته می‌شود.

[20] Support Vector Machine

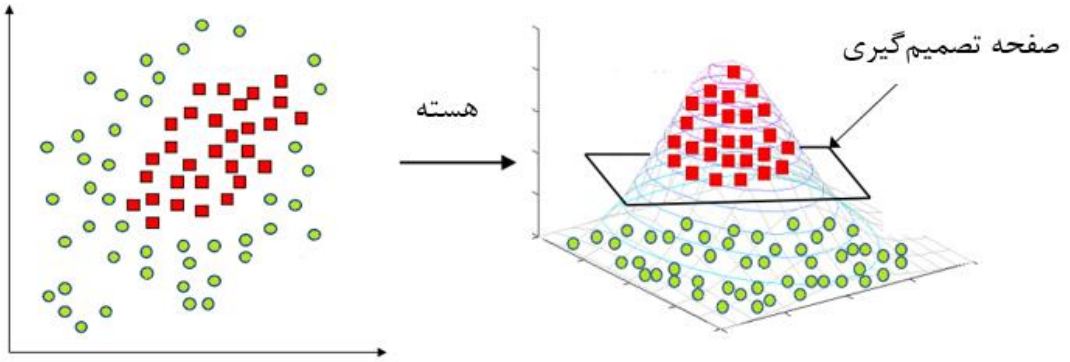
[21] Margin



شکل ۱–۴: ساختار کلی یک ماشین بردارپشتیبان دوکلاس خطی

از آنجایی که همیشه و به راحتی امکان یافتن چنین ابرصفحه‌ای نیست؛ با تابعی به نام هسته[۲۲] نمونه‌ها به فضایی جدید منتقل می‌شوند و سپس ابرصفحه مناسب ایجاد می‌شود. همچنین این الگوریتم اجازه می‌دهد تعداد کمی از نمونه‌ها در سمت نادرست ابرصفحه قرار بگیرند[۳۱].

[22] Kernel



شکل ۱–۵: مثالی از فضای نمونه پس از اعمال تابع هسته

۱-۶ پیشینه تحقیق

پایگاه داده‌های متعددی برای ذخیره داده پپتیدهای ضدمیکروبی با عملکرد ضدسرطانی که توسط محققان زیستی گزارش شده‌اند بوجودآمده‌است. همچنین مقالات متعددی در حوزه پیش‌بینی پپتیدهای ضدسرطانی به روش‌های محاسباتی در حال انتشار هستند که این نشان‌دهنده جذابیت و اهمیت موضوع تشخیص پپتیدهای ضدسرطانی از پپتیدهای غیرضدسرطانی است.

همچنین الگوریتم‌های یادگیری ماشین کلاسیک و مدرن (شبکه‌های عصبی عمیق[۲۳]) متعددی تا به امروز ارائه شده‌است. یکی از روش‌های پیش‌بینی ACPs[۲۴] استخراج ویژگی از توالی این پپتیدها است. مطالعات گسترده‌ای در حوزه استخراج ویژگی از توالی پپتیدها به صورت محاساباتی صورت گرفته‌است، که شامل استخراج ویژگی از توالی ساختار اول، ساختار دوم و سوم آن‌ها می‌شود. عمده تحقیقات برای پیش‌بینی

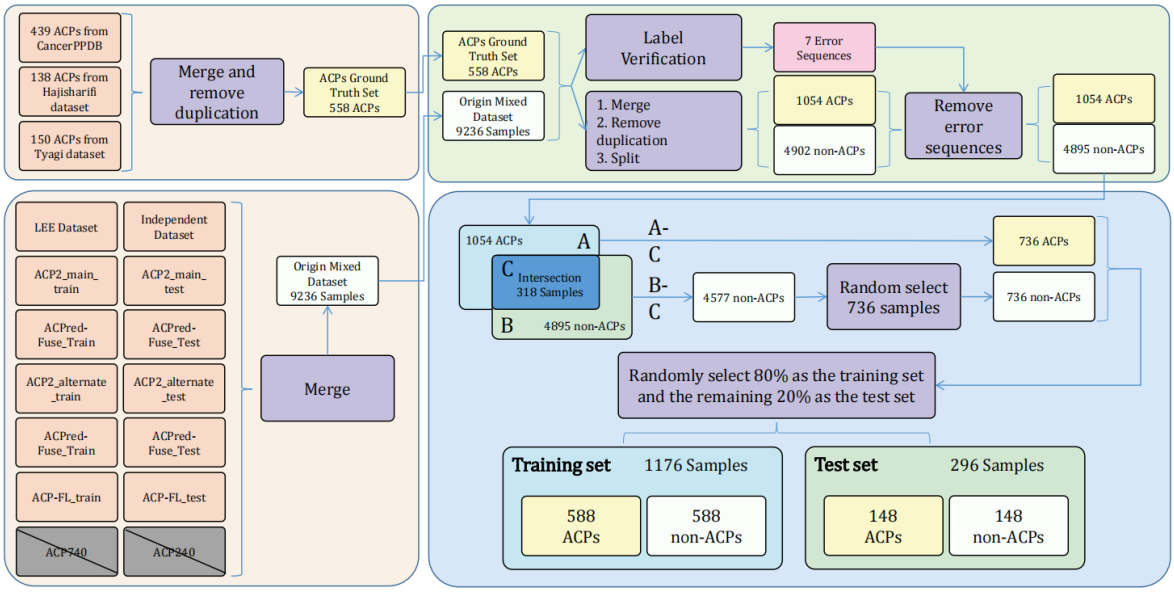
[23] Deep Neural Networks

[24] Anticancer Peptides

ACPها توسط استخراج ویژگی از ساختار اولیه این پپتیدها انجام شده‌است. بعلاوه ابزارهای متعددی برای استخراج این ویژگی‌ها در حال توسعه هستند و منتشرشده‌اند.

با مطالعه و جستجو بین پایگاه داده‌ها[۲۵] و مقالات منتشرشده در حوزه پیش‌بینی ACPs، مطالعه‌ای در سال ۲۰۲۱ تجمیعی از تمام داده‌های موجود در حوزه پپتیدهای ضدسرطانی را در مقاله خود انجام داده‌است، داده‌های این مطالعه مورد استفاده این پایان‌نامه قرار گرفته است[۳۲].

مراحل تجمیع داده بین پایگاه داده‌های پپتیدهای ضدمیکروبی و داده‌های استفاده شده در مقالات پیشین حوزه ACPs درشکل ۱-۶ مشاهده می‌شود[۳۲].



شکل ۱–۶: مراحل ادغام ساخت مجموعه‌داده از پپتیدهای ضدسرطانی و غیرضدسرطانی[۳۲]

[25] Databases

برای استخراج ویژگی از ساختار اولیه، ثانویه و یا ساختار سوم پروتئین و پپتید، ابزارهای متعددی منتشرشده و در حال توسعه هستند. با بررسی ابزارهای متفاوت برای استخراج ویژگی‌ها از ساختاراولیه توالی پروتئین و پپتیدها، پکیج پایتون iFeature که در سال ۲۰۱۸ منتشر شده است[۳۳]، پوشش دهنده نیازهای این پایان‌نامه است.

همچنین با بررسی مقالات موجود در حوزه ACP prediction به روش‌های محاسباتی و همچنین بررسی الگوریتم‌های کلاسیک یادگیری ماشین، بررسی شده‌است که RF[۲۶] و SVM[۲۷] کاندیدای خوبی برای استفاده بر روی داده‌های ACP، non-ACP[۲۸] هستند[۳۴-۳۹].

۱-۷ هدف از انجام طرح

اگرچه AMPs[۲۹] در چندین دهه قبل شناخته شده‌اند، اما تنها در دهه اخیر است که تعداد مقالات مربوط به فعالیت‌های ضدسرطانی آن‌ها افزایش یافته و از آن‌ها به عنوان پپتیدهای ضدسرطانی یاد میکنند[۶-۸, ۱۴, ۴۰]. به همین علت بر این باوریم که در سال‌های آینده، این پپتیدها به علت ویژگی‌های منحصربه‌فردشان در راستای تأثیرگذاری روی سلولهای سرطانی، پیشرفت مهمی در درمان بیماری سرطان که از بزرگترین نگرانی‌های جامعه بشری در جهان است، رقم خواهند زد. استراتژی دیگری که مورد توجه قرار گرفته‌است، استفاده ترکیبی از پپتیدها با داروهای مرسوم شیمی‌درمانی است که هزینه‌های درمان را کاهش می‌دهد و باعث به حداقل رساندن مشکل مقاومت به سرطان و جلوگیری از عود مجدد آن میشود. پیشرفت‌هایی در جهت تولید این پپتیدها در مقیاس بزرگ در جهان صورت گرفته‌است تا این روش درمانی، برای

[26] Random Forest
[27] Support Vector Machine
[28] Non Ancticancer Peptide
[29] Antimicrobal Peptides

بیماران ارزان‌تر و قابل دسترس‌تر باشد. هرچند محدودیت‌هایی ازجمله شباهت احتمالی این پپتیدها با آنتی‌ژنهای خودی یا تحریک سیستم ایمنی علیه این پپتیدها میتواند وجود داشته باشد. در نهایت با توجه به مطالب گفته‌شده میتوان پیش‌بینی کرد این پپتیدها مسیری رو به پیشرفت در جهت بهینه‌سازی روند درمان بیماری سرطان را طی کرده و میتوانند یک روش درمانی نوین و با عوارض کم را ارائه دهند. همچنین با توجه به هزینه مالی و انسانی زیاد برای شناسایی این پپتیدها به روش آزمایشگاهی، همیشه نیاز به روش‌هایی محاسباتی در کنار روش‌های آزمایشگاهی وجود دارد. به همین دلیل در این پایان‌نامه نیز سعی در بررسی پیش‌بینی پپتیدهای ضدسرطانی از بین پپتیدهای غیرضدسرطانی توسط یادگیری ماشین شده‌است.

۲ فصل دوم: مواد و روش‌ها

۲-۱ مجموعه داده[30]

مجموعه داده مورد استفاده برای پیش‌بینی پپتیدهای ضدسرطانی که در این پایان‌نامه استفاده شده و در فصل قبل نیز به آن‌ها اشاره شد به دو مجموعه تقسیم می‌شوند[۳۲].

۱- مجموعه داده مثبت[31]، شامل پپتیدهای ضدمیکروبی با عملکرد ضدسرطانی، که توسط محققان زیستی به صورت آزمایشگاهی تایید شده‌اند.

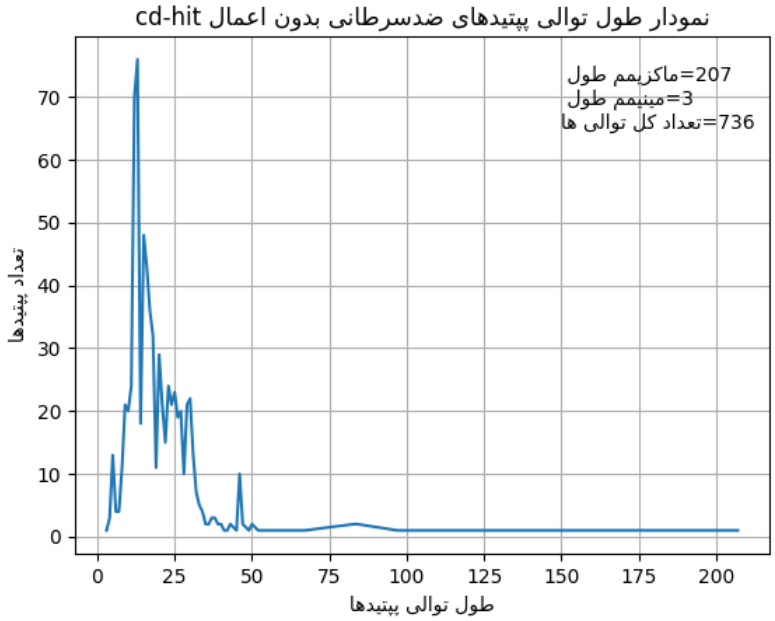
[30] Datasets

[31] Positive Dataset

۲- مجموعه داده منفی[۳۲]، شامل پپتیدهای ضدمیکروبی که عملکرد ضدسرطانی برای آن‌ها گزارش نشده‌است.

۲-۱-۱ مجموعه داده مثبت

مجموعه داده مثبت شامل پپتیدهای ضدمیکروبی با عملکرد ضدسرطانی، که توسط محققان زیستی به صورت آزمایشگاهی تایید شده‌اند، شامل ۷۳۶ پپتید ضدسرطانی(ACP) است[۳۲].

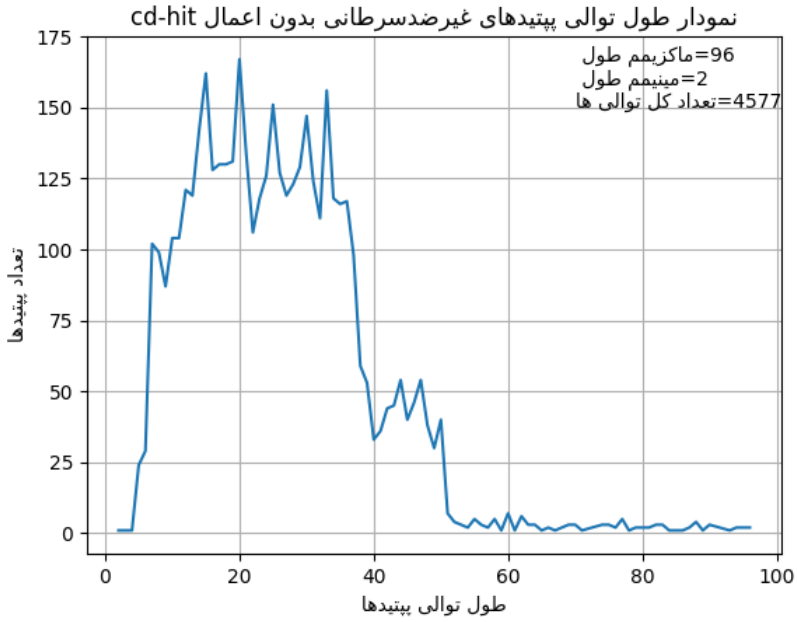


شکل ۲–۱: نمایی کلی از مجموعه داده مثبت(پپتیدهای ضدسرطانی)

[32] Negative Dataset

۲-۱-۲ مجموعه داده منفی

مجموعه داده منفی شامل پپتیدهای ضدمیکروبی که توسط محققان زیستی، عملکرد ضدسرطانی برای آن‌ها گزارش نشده‌است، شامل ۴۵۷۷ پپتید غیرضدسرطانی(non-ACP) است[۳۲].



شکل ۲-۲: نمایی کلی از مجموعه داده منفی(پپتیدهای غیرضدسرطانی)

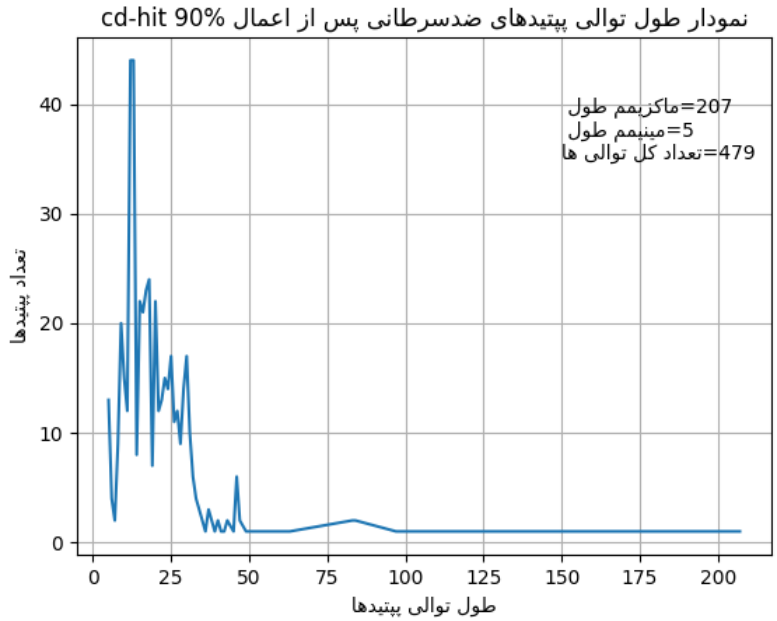
۲-۲ پیش‌پردازش مجموعه داده

۲-۲-۱ شناسایی و حذف توالی پپتیدهای مشابه توسط ابزار cd-hit

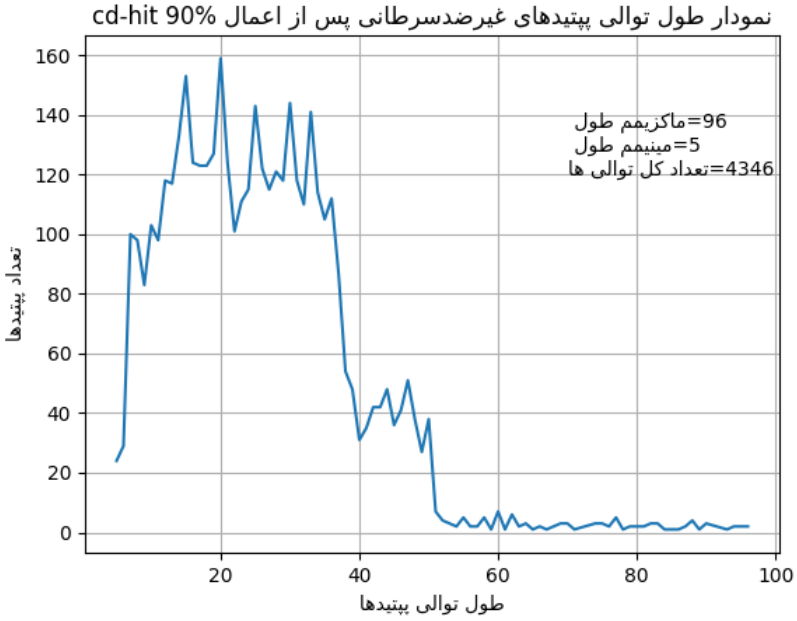
اجتماع مجموعه داده خام مثبت و مجموعه داده خام منفی، مجموعا تعداد ۵۳۱۳ پپتید است. با استفاده از ابزار cd-hit[۴۱]، ۲ بار بصورت کاملا مستقل اقدام به حذف توالی‌های مشابه شد. یکبار بر روی داده‌های

خام cd-hit با ۹۰ درصد شباهت[۳۳] و حذف پپتیدهای با طول کمتر از ۵ اسیدآمینه اجرا شد. بار دیگر مجددا بر روی داده‌های خام cd-hit با ۸۰ درصد شباهت و حذف پپتیدهای با طول کمتر از ۵ اسیدآمینه اجرا شد. نتایج بدست آمده به تفکیک داده‌های مثبت و منفی در شکل‌های زیر مشاهده می‌شود. در نهایت با اجرا cd-hit90 مجموع داده به ۴۸۲۵ داده کاهش یافت و با اجرا cd-hit80 مجموع داده به ۴۵۳۴ داده کاهش یافت. بدلیل اینکه با اعمال cd-hit80 مقدار قابل توجهی از داده‌ها کم شد، داده‌های مربوط به cd-hit90 ملاک انجام ادامه پژوهش قرار گرفتند.

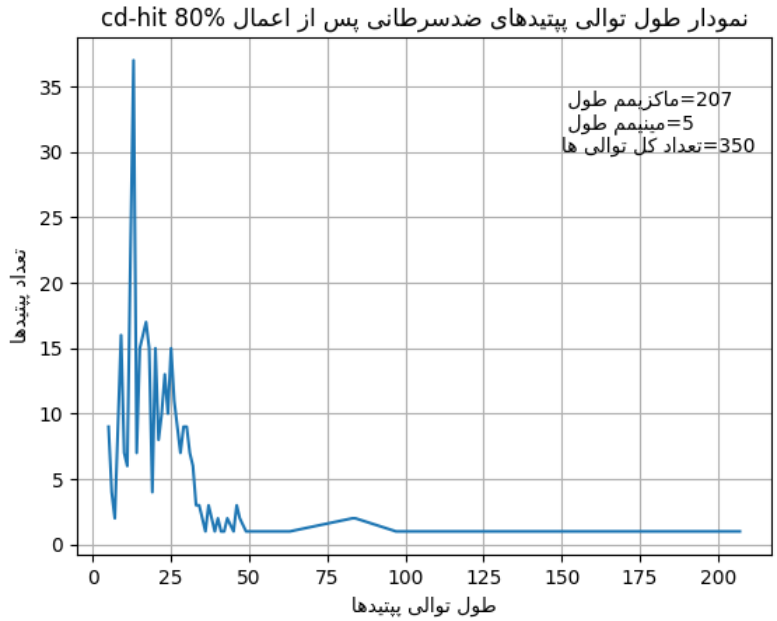
[33] Similarity



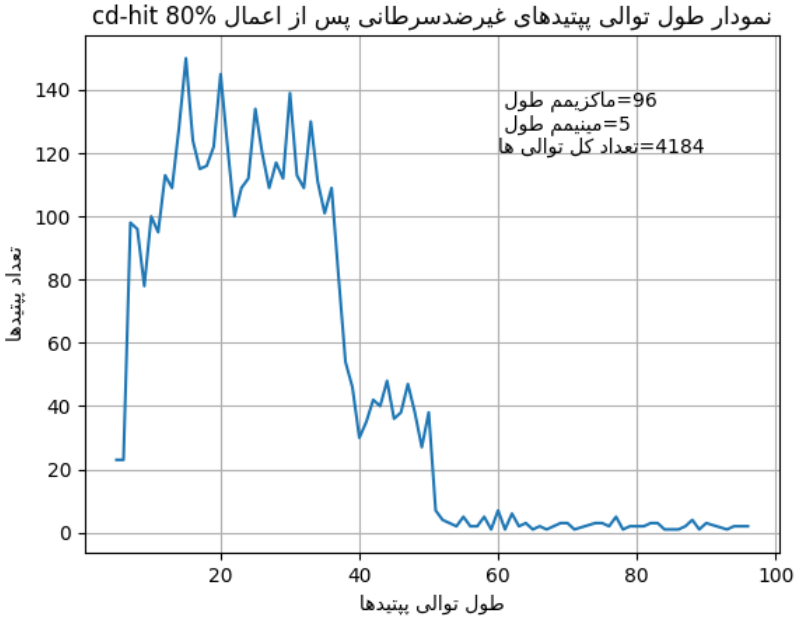
شکل ۲–۳: نمایی کلی از مجموعه داده مثبت پس از اعمال cd-hit90



شکل ۲–۴: نمایی کلی از مجموعه داده منفی پس از اعمال cd-hit90



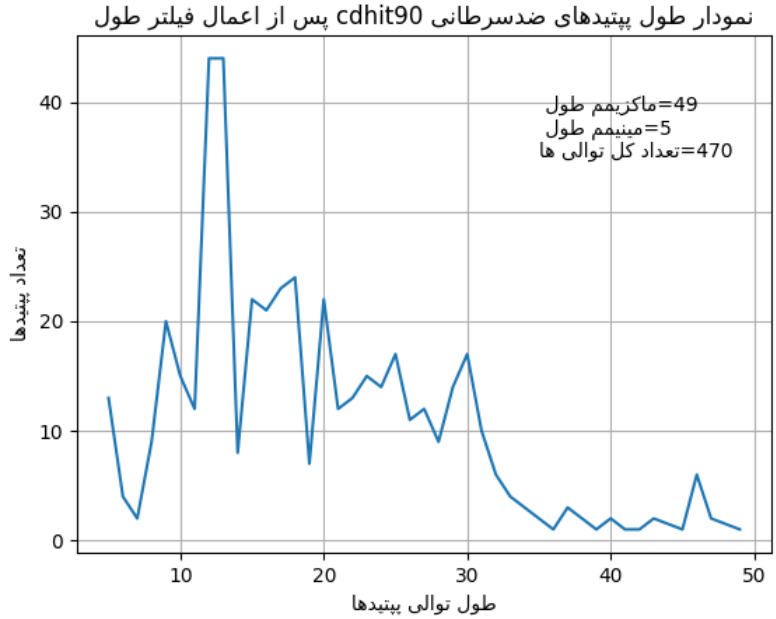
شکل ۲–۵: نمایی کلی از مجموعه داده مثبت پس از اعمال cd-hit80



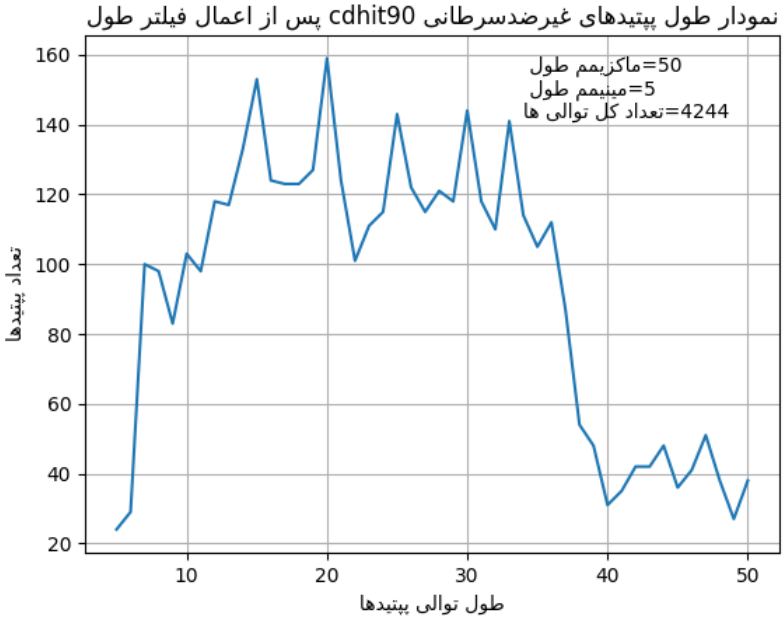
شکل ۲–۶: نمایی کلی از مجموعه داده منفی پس از اعمال cd-hit80

۲-۲-۲ فیلتر طول توالی پپتیدها

با توجه به اینکه نُرم طول پپتیدهای ضدمیکروبی در طبیعت بین ۵ تا حداکثر۵۰ اسیدآمینه دارند[۶]، به همین دلیل بر روی داده‌های cdhit-90 فیلتر با طول حداکثر۵۰ اعمال شد، که مجموعه داده به ۴۷۱۴ داده کاهش یافت.



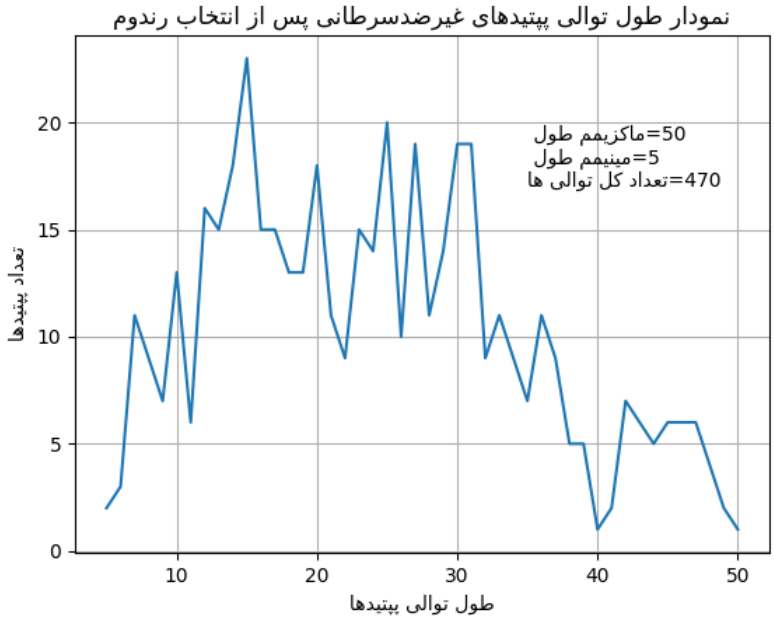
شکل ۲–۷: نمایی کلی از مجموعه داده مثبت cd-hit90 پس از اعمال فیلتر طول توالی



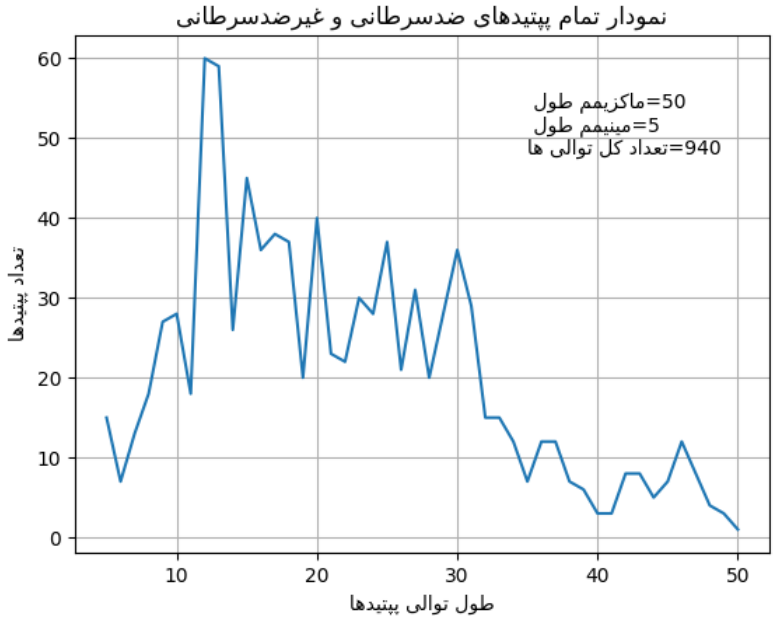
شکل ۲–۸: نمایی کلی از مجموعه داده منفی cd-hit90 پس از اعمال فیلتر طول توالی

۳-۲-۲ بالانس کردن داده‌های مثبت و منفی

پس از اعمال پیش‌پردازش بر روی داده‌های خام، تعداد داده‌های cd-hit90 به تعداد ۴۷۰ داده مثبت (ACP) و ۴۲۴۴ داده منفی (non-ACP) کاهش یافت. با توجه به تعداد بیشتر مجموعه داده منفی و همچنین عدم گزارش وسیع پپتیدهای ضدسرطانی (ACP) به صورت آزمایشگاهی، یکی از چالش‌ها انتخاب مجموعه داده منفی است. برای انتخاب داده‌های منفی به تعداد داده‌های مثبت روش‌های مختلفی وجود دارد که رایج‌ترین آن، انتخاب به صورت تصادفی به تعداد داده‌های مثبت است. با انتخاب ۴۷۰ داده منفی از بین ۴۲۴۴ داده منفی، در نهایت مجموعه‌ای ۹۴۰تایی با تعدادی مساوی از پپتیدهای ضدسرطانی و غیرضدسرطانی بدست آمد. نمایی کلی از مجموعه داده اصلی برای استفاده در ادامه پژوهش در شکل ۲-۱۰ مشاهده می‌شود.



شکل ۲–۹: نمایی کلی از مجموعه داده منفی cd-hit90 پس از انتخاب رندوم



شکل ۲–۱۰: نمایی کلی از مجموعه داده مثبت و منفی پس از اعمال پیش‌پردازش‌ها

۲-۳ مجموعه داده آموزش[۳۴] و تست[۳۵]

برای آموزش و ارزیابی مدل‌های یادگیری ماشین، مجموعه داده ۹۴۰تایی از پپتیدهای بدست آمده در مرحله قبل به نسبت ۸۰-۲۰ طوری که ۸۰ درصد برای مجموعه داده آموزش و ۲۰ درصد برای مجموعه داده تست مستقل در نظر گرفته شد.

۲-۳-۱ مجموعه داده آموزش

برای یادگیری مدل‌های یادگیری ماشین و همچنین تنظیم پارامترهای[۳۶] مدل‌ها، ۸۰ درصد از کل مجموعه داده مثبت یعنی ۳۷۶ داده مثبت و همچنین ۸۰ درصد از مجموعه داده منفی یعنی ۳۷۶ داده منفی به صورت کاملا رندوم و بدون تکرار انتخاب شدند، که در مجموع برای هر بار آموزش و تنظیم پارامترها، ۷۵۲ داده کنار گذاشته شد.

۲-۳-۲ مجموعه داده تست مستقل[۳۷]

برای ارزیابی عملکرد مدل‌های آموزش دیده ۲۰ درصد از ۹۴۰ پپتید موجود کنار گذاشته شد، طوری که این ۲۰ درصد در هیچ مرحله ای از آموزش و انتخاب پارامترهای تنظیمی استفاده نشدند. بعبارتی ۱۸۸ داده طوری که تعداد ۹۴ داده مثبت بصورت کاملا رندوم و بدون تکرار از مجموعه داده‌های مثبت و همچنین تعداد ۹۴ داده منفی بصورت کاملا رندوم و بدون تکرار از مجموعه داده‌های منفی انتخاب شدند.

---

[34] Training Dataset
[35] Test Set
[36] Hyper Parameter Tuning
[37] Independent Test Set

۲-۴ استخراج ویژگی[38]

با توجه به مطالعات پیشین در حوزه پیش‌بینی پپتیدهای ضدسرطانی، تصمیم به استخراج ویژگی از ساختار اولیه پپتیدها شد، با بررسی ابزارهای موجود برای استخراج ویژگی از توالی پپتیدها، کتابخانه پایتون iFeature نیازمندی‌های این پایان‌نامه را پوشش می‌دهد[۳۳]، در ادامه ۱۰ ویژگی استخراج شده مبتنی بر ساختار اول پپتیدها که در این پایان‌نامه استفاده شده است، به اختصار توضیح داده می‌شود[۳۴, ۳۵, ۳۷, ۳۸, ۴۲, ۴۳].

۲-۴-۱ ویژگی Pseudo-Amino Acid Composition (PseudoAAC)

فرض کنید $H_1^o(i)$، $H_2^o(i)$ و $M^o(i)$ برای $i = 1,2,\cdots,20$ به ترتیب مقادیر آب‌گریزی، آب‌دوستی و وزن زنجیره جانبی برای ۲۰ اسید آمینه معمول باشند. مقدار آب‌گریزی برای هر اسید آمینه از رابطه زیر بدست می‌آید.

$$H_1(i) = \frac{H_1^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^o(i)}{\sqrt{\frac{\sum_{i=1}^{20}[H_1^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^o(i)]^2}{20}}}$$

رابطه (۲-۱)

$H_2(i)$ و $M(i)$ نیز به ترتیب برای ویژگی‌های آب‌دوستی و وزن زنجیره جانبی برای هر اسیدآمینه مانند رابطه(۲-۱) به‌دست می‌آیند. سپس تابع correlationای به ازای هر جفت اسیدآمینه R_i و R_j مقدار میانگین این ۳ ویژگی از رابطه (۲-۲) محاسبه می‌شود.

[38] Feature Extraction

$$\Theta(R_i, R_j) = \frac{1}{3}\{[H_1(i) - H_1(j)]^2 + [H_2(i) - H_2(j)]^2 + [M(i) - M(j)]^2\}$$ رابطه (۱-۲)

سپس مجموعه عوامل مرتبط با ترتیب توالی[39] به صورت زیر تعریف می‌شود. N طول توالی و $\lambda < \mathrm{N}$ است. در اینجا λ برابر با ۴ در نظر گرفته شده‌است.

$$\theta_1 = \frac{1}{N-1}\sum_{i=1}^{N-1}\Theta(R_i, R_{i+1})$$

$$\theta_2 = \frac{1}{N-2}\sum_{i=1}^{N-2}\Theta(R_i, R_{i+2})$$ رابطه (۲-۳)

$$\vdots$$

$$\theta_\lambda = \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda}\Theta(R_i, R_{i+\lambda})$$

در نهایت برداری به طول $20 + \lambda$ در اینجا به طول ۲۴ برای هر پروتئین از رابطه (۲-۴) بدست می‌آید که به آن ویژگی PseudoAAC گفته می‌شود. f_i فراوانی نرمال شده اسیدآمینه iام و $w=0.5$ در نظر گرفته شده‌است[۳۳].

[39] Sequence Order-Correlated Factors

$$x_c = \frac{f_c}{\sum_{r=1}^{20} f_r + \omega \sum_{j=1}^{\lambda} \theta_j} \qquad (1 < c < 20)$$

رابطه (۲-۴)

$$x_c = \frac{\omega \theta_{c-20}}{\sum_{r=1}^{20} f_r + \omega \sum_{j=1}^{\lambda} \theta_j} \qquad (20 < c < 20 + \lambda)$$

۲-۴-۲ ویژگی Amphiphilic Pseudo-Amino Acid Composition (APsedoAAC)

تعریف مجموعه ویژگی‌های این ویژگی مشابه ویژگی PseudoAAC است. $H_1(i)$ و $H_2(i)$ ویژگی‌های آبگریزی و آب‌دوستی از رابطه همبستگی زیر محاسبه می‌شود.

رابطه (۲-۵)

$$H_{i,j}^1 = H_1(i)H_1(j)$$
$$H_{i,j}^2 = H_2(i)H_2(j)$$

همچنین مجموعه عوامل مرتبط با ترتیب توالی به صورت زیر تعریف می‌شود.

رابطه (۲-۶)

$$\tau_1 = \frac{1}{N-1} \sum_{i=1}^{N-1} H_{i,i+1}^1$$
$$\tau_2 = \frac{1}{N-1} \sum_{i=1}^{N-1} H_{i,i+1}^2$$
$$\tau_3 = \frac{1}{N-2} \sum_{i=1}^{N-2} H_{i,i+2}^1$$
$$\tau_4 = \frac{1}{N-2} \sum_{i=1}^{N-2} H_{i,i+2}^2$$
$$\dots$$

$$\tau_{۲\lambda-۱} = \frac{۱}{N-\lambda}\sum_{i=۱}^{N-\lambda} H_{i,i+\lambda}^{۱}$$

$$\tau_{۲\lambda} = \frac{۱}{N-\lambda}\sum_{i=۱}^{N-\lambda} H_{i,i+\lambda}^{۲}$$

سپس مجموعه‌ای از توصیف‌گرها که به آن‌ها APseudoAAC گفته می‌شود به صورت زیر تعریف می‌شوند. $w=0.5$.

$$P_c = \frac{f_c}{\sum_{r=۱}^{۲۰} f_r + w\sum_{j=۱}^{۲\lambda}\tau_j}, (۱ < c < ۲۰)$$

$$P_c = \frac{\omega\tau_u}{\sum_{r=۱}^{۲۰} f_r + w\sum_{j=۱}^{۲\lambda}\tau_j}, (۲۱ < u < ۲۰ + ۲\lambda)$$

۲-۴-۳ ویژگی Composition of k-spaced Amino Acid Pairs (CKSAAP)

ویژگی CKSAAP، تعداد تکرار جفت اسیدآمینه که توسط k اسیدآمینه از هم فاصله[۴۰] دارند را محاسبه می‌کند. در این پایان نامه k = ۰،۱،۲،۳ در نظر گرفته شده‌است. به‌همین دلیل بردار ویژگی به طول ۱۶۰۰ برای CKSAAP ساخته شد.

۲-۴-۴ ویژگی Composition/Transition/Distribution (CTD)

ویژگی CTD مخفف سه کلمه ترکیب (Composition)، توزیع (Distribution) و انتقال (Transition) است؛ این ویژگی الگوهای توزیع ویژگی‌های ساختاری یا فیزیکوشیمیایی یک اسیدآمینه در یک توالی

[40] Gap

پروتئین یا پپتید را نشان می‌دهد. برای محاسبه این ویژگی از ۱۳ ویژگی فیزیکوشیمیایی استفاده شده‌است؛ که شامل ویژگی های آبگریزی، قطبیت[۴۱]، قطبش پذیری[۴۲]، بار[۴۳]، حجم نرمال شده وندروالس[۴۴]، ساختارهای ثانویه[۴۵] و قابلیت دسترسی به حلال[۴۶] است. هر یک از این ویژگی‌ها بیست اسیدآمینه را بر اساس خواص فیزیکوشیمیایی به سه گروه تقسیم می‌کنند و سپس توالی پروتئینی بر اساس این که هر اسیدآمینه در کدام گروه قرار می‌گیرند بازنویسی می‌شوند. ویژگی‌های ترکیب، انتقال و توزیع برای توالی بازنویسی شده محاسبه می‌شود و هریک به ترتیب بردارهایی به طول ۳۹، ۳۹ و ۱۹۵ تولید می‌کنند. در نهایت بردار ویژگی CTD برای یک پروتئین از کنار هم قرار گرفتن این سه بردار ساخته می‌شود که به طول ۲۷۳ است.

[41] Polarity
[42] Polarizability
[43] Charge
[44] Normalized VanderWaals Volume
[45] Secondary Structures
[46] Solvent Accessibility

۵-۴-۲ ویژگی Dipeptide Deviation from Expected Mean (DDE)

ویژگی DDE توسط ۳ پارامتر ساخته می‌شود، ۱- ترکیب دوتایی پپتیدها[۴۷] (D_c)، میانگین نظری[۴۸] (T_m) و واریانس نظری[۴۹] (T_v). ۳ پارامترهای بیان شده و DDE در روابط زیر محاسبه می‌شوند.

$D_c(r,s)$، معیار ترکیب دوتایی برای دوپپتید 'rs' :

$$D_c(r,s) = \frac{N_{rs}}{N-1}, r,s \in \{A,C,D,\dots Y\}$$

که، N_{rs} تعداد پپتیدهای دوتایی با اسیدآمینه r و s، N طول پپتید هستند.

$T_m(r,s)$، میانگین نظری:

$$T_m(r,s) = \frac{C_r}{C_N} \times \frac{C_s}{C_N}$$

که، C_r تعداد کدون[۵۰]های کدکننده اولین اسیدآمینه و C_s تعداد کدون‌های کدکننده دومین اسیدآمینه در پپتید دوتایی 'rs' است. C_N تعداد کل کدون‌های ممکن بغیر از ۳ کدون متوقف کننده[۵۱] است.

T_v (r,s)، واریانس نظری:

$$T_v(r,s) = \frac{T_m(r,s)(1-T_m(r,s))}{N-1}$$

و، در نهایت DDE(r,s) از رابطه زیر محاسبه می‌شود:

$$DDE(r,s) = \frac{D_c(r,s) - T_m(r,s)}{\sqrt{T_v(r,s)}}$$

---

[47] Dipeptide Composition
[48] Theoretical Mean
[49] Theoretical Variance
[50] Codon
[51] Stop Codons

ویژگی DDE در نهایت بردار ویژگی به طول ۴۰۰ تولید می‌کند.

۲-۴-۶ ویژگی Moran correlation (Moran)

```
H CIDH920105
D Normalized average hydrophobicity scales (Cid et al., 1992)
R PMID:1518784
A Cid, H., Bunster, M., Canales, M. and Gazitua, F.
T Hydrophobicity and structural classes in proteins
J Protein Engineering 5, 373-375 (1992)
C CIDH920103     0.973  CIDH920104     0.970  CIDH920102     0.969
  NISK860101     0.938  BASU050102     0.931  ZHOH040103     0.926
  ROBB790101     0.921  CIDH920101     0.921  MIYS850101     0.916
  BASU050103     0.914  PLIV810101     0.914  BIOV880101     0.912
  BASU050101     0.907  WERD780101     0.905  ZHOH040101     0.904
  RADA880108     0.898  FAUJ830101     0.893  MEEJ810101     0.892
  PONP930101     0.891  SWER830101     0.890  CORJ870102     0.890
  ROSM880104     0.886  BIOV880102     0.882  MANP780101     0.879
  ARGP820101     0.867  JOND750101     0.866  RADA880102     0.861
  CASG920101     0.859  GUOD860101     0.858  ROSG850102     0.858
  NOZY710101     0.857  PONP800101     0.856  NISK800101     0.854
  BLAS910101     0.852  CORJ870107     0.848  MEEJ810102     0.844
  PONP800108     0.843  ROSM880105     0.843  MEEJ800102     0.840
  TAKK010101     0.840  EISD860101     0.839  CORJ870104     0.838
  CORJ870103     0.838  SIMZ760101     0.837  PONP800102     0.831
  LIFS790101     0.828  LEVM760106     0.828  CORJ870101     0.827
  CORJ870106     0.826  CORJ870105     0.822  GOLD730101     0.820
  ZHOH040102     0.818  PONP800107     0.818  NADH010104     0.817
  PTIO830102     0.813  VENT840101     0.813  NADH010103     0.810
  PONP800103     0.807  MEIH800103     0.804  NADH010105     0.800
  WOEC730101    -0.800  KIDA850101    -0.803  PUNT030101    -0.805
  KRIW790101    -0.816  FUKS010103    -0.821  PUNT030102    -0.822
  MEIH800102    -0.826  RACS770102    -0.830  VINM940103    -0.832
  KARP850102    -0.839  CORJ870108    -0.843  FASG890101    -0.860
  PARS000101    -0.860  KARP850101    -0.866  BULH740101    -0.871
  GRAR740102    -0.884  VINM940101    -0.885  MIYS990103    -0.886
  RACS770101    -0.887  GUYH850102    -0.892  WOLS870101    -0.899
  MIYS990105    -0.901  OOBM770103    -0.904  MIYS990104    -0.908
  VINM940102    -0.910  MIYS990102    -0.915  MIYS990101    -0.916
  MEIH800101    -0.923  GUYH850103    -0.927  PARJ860101    -0.948
I    A/L     R/K     N/M     D/F     C/P     Q/S     E/T     G/W     H/Y     I/V
     0.02   -0.42   -0.77   -1.04    0.77   -1.10   -1.14   -0.80    0.26    1.81
     1.14   -0.41    1.00    1.35   -0.09   -0.97   -0.77    1.71    1.11    1.13
//
```

شکل ۲–۱۱: یک مثال مصور از ویژگی‌های فیزیکوشیمیایی در دیتابیس AAindex [۴۴]

توصیف کننده‌های[۵۲] خودهمبستگی[۵۳] براساس توزیع خواص[۵۴] اسیدآمینه‌ها در توالی تعریف می‌شوند. خواص(ویژگی‌های) اسیدآمینه که در این توصیف‌کننده‌ها استفاده می‌شود، از دیتابیس AAindex استخراج می‌شود.

[52] Descriptors
[53] Autocorrelation
[54] Properties

دیتابیس از مسیر http://www.genome.jp/dbget/aaindex.html/ در دسترس است. ۸ شاخص

'CIDH920105', 'BHAR880101', 'CHAM820101', 'CHAM820102', 'CHOC760101', 'BIGC670101', 'CHAM810101', 'DAYM780201'

از دیتابیس AAindex استفاده شده‌اند.

تمام شاخص‌های اسیدآمینه‌ها متمرکز[55] و استانداردشده[56] توسط رابطه زیر هستند.

$$P_r = \frac{P_r - \overline{P}}{\sigma}$$

که، $\bar{P}$ میانگین خواص ۲۰ اسیدآمینه و σ انحراف معیار ۲۰ اسیدآمینه است، که $\bar{P}$ وσخود توسط رابطه زیر قابل محاسبه هستند.

$$\bar{P} = \frac{\sum_{r=۱}^{۲۰} P_r}{۲۰}, \sigma = \sqrt{\frac{۱}{۲۰}\sum_{r=۱}^{۲۰}(P_r - \bar{P})^۲}$$

طول بردار ویژگی توصیف‌کننده‌های خود همبستگی $n * lag$ است، که n تعداد شاخص‌های انتخاب‌شده از پایگاه داده AAindex است. که در این پایان‌نامه $n=8$ و $lag=4$ و درنتیجه طول بردار ویژگی‌ها ۳۲ است. توصیف‌کننده خودهمبستگی Moran توسط رابطه زیر بدست می‌آید.

$$I(d) = \frac{\frac{۱}{N-d}\sum_{i=۱}^{N-d}(P_i - \bar{P}')(P_{i+d} - \bar{P}')}{\frac{۱}{N}\sum_{i=۱}^{N}(P_i - \bar{P}')^۲}, d = ۱,۲,۳ \ldots, nlag$$

[55] Centralized

[56] Standardized

که، d، lag خودهمبستگی و $nlag$ بیشترین مقدار برای lag است که در اینجا $nlag$ برابر با ۴ در نظر گرفته ایم.

P_i و P_{i+d} ویژگی‌های اسیدآمینه‌ها در موقعیت i و $i + d$ است. $\overline{P}$ میانگین ویژگی P نیز ازرابطه زیر بدست می‌آید.

$$\bar{P}' = \frac{\sum_{i=۱}^{N} P_i}{N}$$

۲-۴-۷ ویژگی Geary correlation (Geary)

برای یک پپتید یا پروتئین توصیف کننده‌های Geary از رابطه زیر محاسبه می‌شوند.

$$C(d) = \frac{\frac{۱}{۲(N-d)} \sum_{i=۱}^{N-d} (P_i - P_{i+d})^۲}{\frac{۱}{N-۱} \sum_{i=۱}^{N} (P_i - \bar{P}')^۲}, d = ۱,۲, \dots, nlag$$

که، d، P، P_i ، P_{i+d}و nlag تعاریف مشابه ویژگی Moran دارند.

۲-۴-۸ ویژگی Normalized Moreau-Broto Autocorrelation (NMBroto)

توصیف‌کننده‌های خودهمبستگی Moreau-Broto با رابطه زیر تعریف می‌شوند.

$$AC(d) = \sum_{i=۱}^{N-d} P_i \times P_{i+d}, d = ۱,۲, \dots, nlag$$

همچنین توصیف‌کننده‌های خودهمبستگی normalized Moreau-Broto با رابطه زیر تعریف می‌شوند.

$$ATS(d) = \frac{AC(d)}{N-d}, d = ۱,۲, \dots, nlag$$

۲-۴-۹ ویژگی *k*-Spaced Conjoint Triad (KSCTriad)

توصیفگر KSCTriad مبتنی بر توصیفگر CTriad است که نه تنها اعداد سه واحد اسیدآمینه پیوسته را محاسبه می‌کند، بلکه واحدهای اسیدآمینه پیوسته را نیز در نظر می‌گیرد که با هر k باقیمانده[۵۷] از هم جدا می شوند (حداکثر مقدار پیش‌فرض k در این پایان‌نامه ۱ تنظیم شده‌است). به عنوان مثال، $AxRxT$ یک سه‌گانه با $k=1$ است. بنابراین، ابعاد بردار ویژگی کدگذاری شده در KSCTriad، 343*(k+1) = 686 است.

۲-۴-۱۰ ویژگی Quasi-sequence-order (QSOrder)

برای هر نوع اسیدآمینه، یک توصیف‌گر QSOrder را می‌توان به صورت زیر تعریف کرد:

$$X_r = \frac{f_r}{\sum_{r=۱}^{۲۰} f_r + w\sum_{d=۱}^{nlag} \tau_d}, r = ۱,۲, \dots, ۲۰$$

که در آن f_r وقوع نرمال شده اسیدآمینه نوع r و w یک عامل وزنی است (w =0.1). اینها اولین ۲۰ توصیف‌گر QSOrder هستند. ۳۰ توصیف‌گر دیگر QSOrder به صورت زیر تعریف می شوند:

$$X_d = \frac{w\tau_d - ۲۰}{\sum_{r=۱}^{۲۰} f_r + w\sum_{d=۱}^{nlag} \tau_d}, d = ۲۱,۲۲, \dots, ۲۰ + nlag$$

[57] Residues

۲-۵ پیش‌پردازش داده[58]

۴ مرحله تحت عنوان پیش‌پردازش بر روی داده‌های هر ۱۰ ویژگی استخراج شده اعمال شد.

حذف ستون‌های تکراری، ازبین ستون‌هایی که مقادیر تکراری داشتند، همه بجز اولین ستون از بین تکراری-ها حذف شدند.

حذف ستون با مقادیر یکسان، ستونی که برای تمام نمونه‌ها مقادیر یکسانی داشته باشد؛ حاوی اطلاعات مفید نیست به‌همین دلیل این ستون‌ها شناسایی و حذف شدند.

حذف نمونه‌های تکراری، نمونه‌هایی(سطرهایی) که برای تمام ویژگی‌ها مقدار یکسانی داشتند، همگی بجز اولین نمونه حذف شدند.

۲-۶ نرمالسازی داده[59]

از آنجایی که ممکن است در داده‌ها ویژگی‌های متفاوت محدوده‌های مختلفی داشته باشند، در این صورت ویژگی‌هایی که شامل دامنه اعداد بزرگتری هستند نسبت به بقیه ویژگی‌ها می‌توانند تأثیر بیشتری روی نتیجه پیش‌بینی داشته باشند که این باعث سو یافتن مدل یادگیری به سمت این ویژگی‌ها می‌شود در صورتی که ممکن است این ویژگی‌ها نسبت به ویژگی‌هایی ک دامنه کوچکتری دارند اهمیت کمتری داشته باشند که این اتفاق خوبی نیست، به همین منظور تمام ویژگی‌ها باید نرمالسازی یا استانداردسازی شوند. در این مطالعه با استفاده از روش نرمال‌سازی min-max و با رابطه (۲-۷) هر ویژگی در مجموعه داده به بازه [۰،۱] نگاشت شدند.

[58] Data preprocessing

[59] Normalization

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$ رابطه (۲-۷)

هر یک از ویژگی‌های مجموعه داده تست مستقل و آموزش با رابطه (۲-۷) به اعداد جدیدی نگاشت شدند.

۲-۷ معیارهای ارزیابی

جهت سنجش کارایی الگوریتم‌های یادگیری ماشین نیاز به معیارهایی داریم و برای تعریف معیارهای ارزیابی ابتدا احتیاج به پارامترهایی داریم که بدانیم چه تعداد از نمونه‌ها به درستی پیش‌بینی شده‌اند. با توجه به این که این مطالعه در زمینه پیش‌بینی پپتیدهای ضدسرطانی است، پارامترها و معیارهای ارزیابی بر این اساس تعریف خواهند شد.

مثبت‌صحیح[۶۰]: یک پپتید ضدسرطانی به درستی یک پپتید ضدسرطانی اعلام شده‌است. تعداد این پیش-بینی‌ها را با TP نشان می‌دهند.

مثبت‌کاذب[۶۱]: یک پپتید غیرضدسرطانی به اشتباه یک پپتید ضدسرطانی اعلام شده‌است. تعداد این پیش-بینی‌ها را با FP نشان می‌دهند.

منفی‌صحیح[۶۲]: یک پپتید غیرضدسرطانی به درستی یک پپتید غیرضدسرطانی اعلام شده‌است. تعداد این پیش‌بینی‌ها را با TN نشان می‌دهند.

منفی‌کاذب[۶۳]: یک پپتید ضدسرطانی به اشتباه یک پپتید غیرضدسرطانی اعلام شده‌است. تعداد این پیش-بینی‌ها را با FN نشان می‌دهند.

---

[60] True positive
[61] False positive
[62] True negative
[63] False negative

در ادامه معیارهای ارزیابی براساس این پارامترها بیان می‌شود.

۲-۷-۱ حساسیت[۶۴]

توانایی یک الگوریتم یادگیری ماشین برای تشخیص مقادیر مثبت‌صحیح (TP) را حساسیت یا یادآوری[۶۵] گویند و از رابطه (۲-۸) بدست می‌آید. با توجه به این رابطه می‌توان دریافت که هر چه تعداد پپتیدهای ضدسرطانی که به اشتباه پپتید غیرضدسرطانی پیش‌بینی شده‌اند؛ کمتر باشد، یعنی منفی‌کاذب (FN) کمتر باشد، رابطه (۲-۸) به عدد یک نزدیک‌تر و حساسیت الگوریتم برای پیش‌بینی بالاتر است.

$$Sensivity = \frac{TP}{TP + FN}$$ رابطه (۲-۸)

۲-۷-۲ اختصاصیت[۶۶]

توانایی یک الگوریتم یادگیری ماشین برای تشخیص مقادیر منفی‌صحیح (TN) را اختصاصیت گویند و از رابطه (۲-۸) بدست می‌آید. با توجه به این رابطه می‌توان دریافت که هر چه تعداد پپتیدهای غیرضدسرطانی که به اشتباه پپتید ضدسرطانی پیش‌بینی شده‌اند؛ کمتر باشد، یعنی مثبت کاذب (FP) کمتر باشد، رابطه (۲-۹) به عدد یک نزدیک‌تر اختصاصیت الگوریتم برای پیش‌بینی بالاتر است.

$$Specificity = \frac{TN}{TN + FP}$$ رابطه (۲-۹)

[64] Sensitivity
[65] Recall
[66] Specificity

۲-۷-۳ صحت[۶۷]

یکی‌دیگر از معیارهای ارزیابی صحت است و از رابطه (۲-۸) بدست می‌آید. این معیار نشان می‌دهد چند درصد از داده‌ها به درستی طبقه‌بندی شده‌اند.

رابطه (۲-۱۰)

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

۲-۷-۴ دقت[۶۸]

دقت نشان دهنده نسبت تعداد پپتیدهای ضدسرطانی که توسط الگوریتم یادگیری ماشین به درستی مثبت گزارش شده است به تعداد تمام پپتیدهای ضدسرطانی است؛ می‌باشد و با رابطه (۲-۱۱) محاسبه می‌شود.

رابطه (۲-۱۱)

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

۲-۸ اعتبارسنجی متقابل k-لایه[۶۹]

اعتبارسنجی متقابل k-لایه، یک روش برای ارزیابی مدل یادگیری ماشین است. در این روش مجموعه داده آموزش به k بخش افراز می‌شود. هر بار یک بخش به عنوان مجموعه داده آزمون در نظرگرفته می‌شود سپس مدل روی $k - ۱$ بخش دیگر که به عنوان مجموعه آموزش انتخاب می‌شوند اجرا می‌شود. دقت مدل

[67] Accuracy
[68] Precision
[69] *K*-fold Cross Validation

روی بخشی که به عنوان مجموعه آزمون در نظر گرفته شده بود؛ محاسبه می‌شود. این روند k بار انجام می‌شود به‌طوری که تمام داده‌ها حداقل یکبار هم در مجموعه آزمون و هم در مجموعه آموزش شرکت کرده باشند و در نهایت میانگین دقت محاسبه شده در k مرحله به عنوان دقت مدل گزارش می‌شود. برای ارزیابی مدل‌های یادگیری در این پایان نامه مقدار k برابر ۱۰ در نظر گرفته شده‌است[۴۵].

۲-۹ انتخاب ویژگی[۷۰]

انتخاب ویژگی، انتخاب یک زیرمجموعه از کل فضای مجموعه ویژگی‌ها است به طوری که ویژگی‌های انتخاب شده بیشترین ارتباط را با برچسب کلاس‌ها داشته باشند. انتخاب ویژگی‌های مناسب باعث می‌شود مدل‌های یادگیری ماشین کارآیی بهتری داشته و ارتباط بین ویژگی‌ها و کلاس‌ها قابل درک‌تر باشد. به طور کلی روش‌های انتخاب ویژگی را می‌توان به دو گروه بسته‌بندی[۷۱] و فیلتر[۷۲] تقسیم کرد.

۲-۹-۱ روش بسته‌بندی

در این روش، مدلی با زیرمجموعه‌های مختلف از ویژگی‌ها توسط حذف یا اضافه کردن آن‌ها ساخته و آموزش داده می‌شود. در نهایت زیرمجموعه‌ای از ویژگی‌ها که بالاترین عملکرد را برای مدل داشتند، انتخاب می‌شوند. از روش‌های انتخاب ویژگی به روش بسته‌بندی می‌توان انتخاب متوالی رو به جلو[۷۳] و انتخاب متوالی رو به

[70] Feature selection
[71] Wrapper
[72] Filter
[73] Sequential Forward Selection

عقب[۷۴] را نام برد. در واقع این روش یک الگوریتم جستجو است که ویژگی‌ها به عنوان ورودی الگوریتم هستند و کارآیی مدل به عنوان خروجی؛ که باید بهینه شود.

در این پایان‌نامه از روش انتخاب متوالی رو به جلو توسط کتابخانه پایتون mlxtend استفاده شده‌است[۴۶].

انتخاب متوالی رو به جلو ۳ بار توسط پارامترهای زیر اجرا و ۳ دسته ویژگی توسط الگوریتم تولید شد.

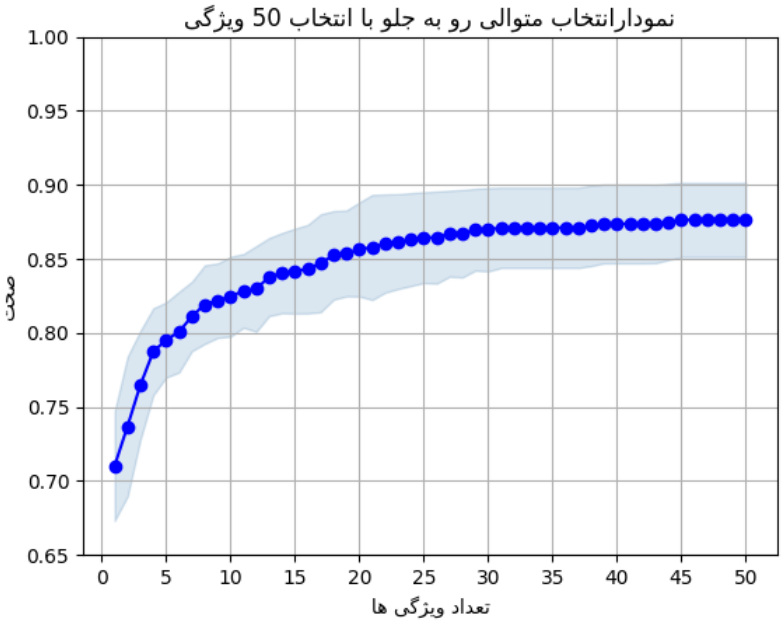
یکی از پارامترهای روش بسته‌بندی انتخاب یک مدل یادگیری ماشین برای انتخاب ویژگی‌ها است. در این پایان‌نامه از طبقه‌بند Logestic Regression برای این امر استفاده شد. معیار ارزیابی طبقه‌بند نیز صحت اعتبارسنجی ۱۰ لایه اعمال شد.

پارامتر بعدی در الگوریتم SFS[۷۵]، برای تعداد ویژگی‌های انتخابی است، که دراین‌جا ۳ مقدار ۵۰، ۱۰۰ و ۲۰۰ در نظر گرفته شد.

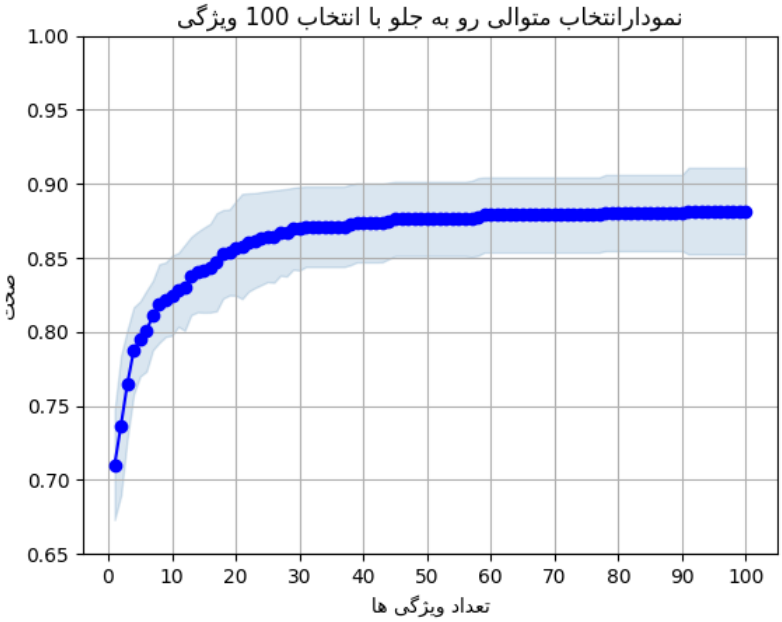
مراحل ۳ دسته ویژگی تولید شده که شامل ۵۰ ویژگی مهم، ۱۰۰ ویژگی مهم و ۲۰۰ ویژگی مهم هستند در تصاویر زیر مشاهده می‌شود. از این پس به ۵۰، ۱۰۰ و ۲۰۰ ویژگی انتخاب شده توسط الگوریتم SFS، با نام مجموعه داده SFS50، SFS100 و SFS200 اشاره خواهدشد.

[74] Sequential Backward Selection

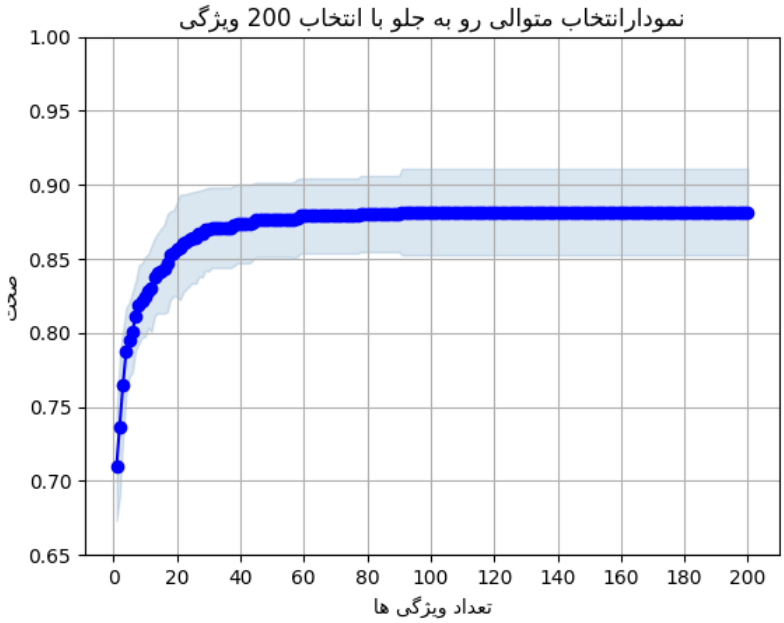
[۷۵] Sequential Forward Selection



شکل ۲–۱۲: نمودار انتخاب متوالی رو به جلو تا ۵۰ ویژگی



شکل ۲–۱۳: نمودار انتخاب متوالی رو به جلو تا ۱۰۰ ویژگی



شکل ۲–۱۴: نمودار انتخاب متوالی رو به جلو تا ۲۰۰ ویژگی

در ادامه بر روی مجموعه داده SFS50، SFS100 و SFS200 ، مدل‌های یادگیری ماشین SVM و RF اجرا شد. نتایج این مدل‌ها در فصل ۳ آمده‌است.

۲-۱۰ کاهش ابعاد[۷۶] مجموعه داده

تعداد زیاد ویژگی‌ها باعث تحمیل هزینه محاسباتی بالا به طبقه‌بندها می‌شود و به مشقت تعداد ابعاد[۷۷] معروف است. برای کاهش ابعاد می‌توان از روش تحلیل مؤلفه‌های اصلی[۷۸] استفاده کرد.

[76] Dimension Reduction
[77] Curse of Dimensionality
[78] Principal Component Analysis

به صورت ریاضی تحلیل مؤلفه‌های اصلی (PCA) یک تبدیل خطی متعامد است که داده‌ها را به دستگاه مختصات جدید می‌برد. به صورتی‌که اولین بزرگترین واریانس داده بر روی اولین محور مختصات، دومین بزرگترین واریانس بر روی دومین محور مختصات و ... قرار می‌گیرد[۴۷].

در این پایان‌نامه، این روش برای تمام ویژگی‌هایی که تا به حال بحث شده‌است با انتخاب ۱۰ مؤلفه اصلی اول اعمال شد، در ادامه بر روی این ۱۰ مؤلفه (۱۰ ویژگی) مدل‌های یادگیری ماشین SVM و RF اجرا شد. نتایج این مدل‌ها در فصل ۳ آمده‌است.

۲-۱۱ مدل‌های یادگیری ماشین اجرا شده روی مجموعه داده‌ها

همانطور که در بخش ۲-۲ بیان شد، بر روی مجموعه داده cdhit90 پیش‌پردازش اولیه انجام شد؛ سپس در بخش استخراج ویژگی (بخش ۲-۴) به ۱۰ ویژگی اشاره شد.

این ۱۰ ویژگی و همچنین ترکیبی از این ۱۰ ویژگی که در ادامه اشاره خواهد شد، از مجموعه داده پیش-پردازش شده بخش ۲-۲، استخراج شدند. سپس ۱۰مؤلفه اول هر یک از این مجموعه داده‌های بدست‌آمده توسط الگوریتم PCA از کتابخانه iFeature استخراج شد.

در نهایت بر روی هر کدام از مجموعه داده‌های بدست آمده توسط PCA و همچنین بر روی ۳ مجموعه داده SFS50، SFS100 و SFS200 (رجوع به بخش ۲-۹-۱) ۲ مدل یادگیری ماشین SVM و RF با پارامترهای تنظیمی اشاره شده در قسمت بعدی ساخته شدند. ارزیابی مدل‌ها و یافتن بهترین پارامتر تنظیمی از بین پارامترهای تعریف شده توسط اعتبارسنجی متقابل ۱۰ لایه انجام شد. در تمام مراحل از

تعریف پارامترهای تنظیمی، تنظیم پارامترهای تنظیمی، آموزش و ارزیابی مدل‌ها؛ از کتابخانه scikit-learn پایتون استفاده شد.

۱-۱۱-۲ طبقه‌بندی جنگل تصادفی

برای ساخت هر جنگل‌تصادفی ۴۳۲۰، ۴۳۲۰ = ۱۰ * ۲ * ۱۲ * ۳ * ۳ * ۲ پارامترتنظیمی، توسط قطعه کد زیر ایجاد شد.

الگوریتم انتخاب تصادفی پارامترها، ازبین ۴۳۲۰ پارامترتنظیمی، یک جنگل تصادفی بصورت کاملا تصادفی میسازد. برای یافتن بهترین پارامترتنظیمی، ۱۰۰ جنگل‌تصادفیِ تصادفی[۷۹]، توسط پارامترهای تصادفی ایجاد شده، ساخته شد.[۴۸, ۴۹]

قطعه کد زیر، فضای پارامترهای تنظیمی برای ساخت ۱۰۰ جنگل‌تصادفیِ تصادفی را نشان می‌دهد.

```
# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start=100, stop=1000, num=10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num=11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# On each iteration, the algorithm will choose a difference combination of the features.
# Altogether, there are 2 * 12 * 2 * 3 * 3 * 10 = 4320 settings!
```

[79] Random Random Forest

۲-۱۱-۲ طبقه‌بندی ماشین بردار پیشتیبان

برای ساخت هر ماشین‌بردارپشتیبان ۵۲، ۵۲ = ۱۳*۲*۲ پارامترتنظیمی، توسط قطعه کد زیر ایجاد شد.

```
kernel = ['rbf', 'linear']
gamma = ['scale', 'auto']
c_range = [x for x in np.logspace(-9, 3, 13)]
```

kernel، وظیفه اصلی هسته[۸۰] این است که مجموعه داده ورودی داده‌شده را به فرم مورد نیاز تبدیل کند. انواع مختلفی از توابع مانند تابع پایه خطی[۸۱]، چند جمله‌ای[۸۲] و شعاعی (RBF)[۸۳] وجود دارد. تابع پایه چند جمله ای و RBF برای ابر صفحه غیرخطی مفید هستند. هسته‌های چند جمله‌ای وRBF خط جداسازی را در بعد بالاتر محاسبه می‌کنند. در اینجا ۲ پارامتر هسته خطی و هسته شعاعی برای پارامترهای هسته در نظر گرفته شد.

c_range، پارامتر هزینه[۸۴] یا مجازات[۸۵]، نشان دهنده طبقه‌بندی اشتباه[۸۶] یا عبارت خطا[۸۷] است. پارامتر c به بهینه‌سازی SVM می‌گوید که چقدر خطا قابل‌تحمل است. به این ترتیب می‌تواند trade-off بین مرز تصمیم[۸۸] و اصطلاح طبقه‌بندی اشتباه را کنترل کند. مقدار کوچکتر c یک ابرصفحه با حاشیه کوچک و مقدار بزرگتر c یک ابرصفحه با حاشیه بزرگتر ایجاد می‌کند[۵۰، ۵۱].

gamma، مقدار کمتر پارامتر گاما با مجموعه داده آموزشی متناسب است، در حالی که مقدار بالاتر گاما دقیقاً با مجموعه داده آموزشی مطابقت دارد که باعث برازش بیش‌ازحد[۸۹] می‌شود. به عبارت دیگر، می‌توان

[80] Kernel
[81] Linear Basis Function
[82] Polynomial Basis Function
[83] Radial Basis Function (rbf)
[84] Cost
[85] Penalty
[86] Missclassification
[87] Error Term
[88] Decision Boundary
[89] Over-Fitting

گفت مقدار کم گاما فقط نقاط نزدیک را در محاسبه خط جداسازی در نظر می‌گیرد، در حالی که مقدار بالای گاما تمام نقاط داده را در محاسبه خط جداسازی در نظر می‌گیرد.

برای انتخاب بهترین پارامترتنظیمی، ۵۲ مدل ماشین بردارپشتیبان، به تعداد پارامترهای تنظیمیِ تعریف شده، ساخته‌شد.

۳ فصل سوم: نتایج و بحث

۳-۱ نتایج طبقه‌بندها

برای کاهش ابعاد داده‌های ۱۰ ویژگی استخراج شده و پیش‌پردازش شده، ۲ روش در نظر گرفته‌شد.

۱- استفاده از PCA و استخراج ۱۰ مؤلفه اول از داده‌هاآن‌ها

۲- استفاده از الگوریتم انتخاب ویژگی SFS، که نتیجه آن ۳ دسته مجموعه داده SFS50، SFS100 و SFS200 شد.

سپس هر مجموعه داده بدست آمده به این ۲ روش، به نسبت ۸۰-۲۰ تقسیم شدند، طوریکه ۸۰ درصد داده‌ها برای آموزش و تنظیم پارامترهای تنظیمی و ۲۰ درصد داده‌ها برای تست مستقل در

نظر گرفته شدند. سپس ۲ طبقه‌بند جنگل‌تصادفی و ماشین بردارپشتیبان بر روی ۸۰ درصد داده‌های آموزش توسط روش اعتباری‌سنجی متقابل ۱۰ لایه، با پارامترهای تنظیمی تعریف شده در بخش ۲-۱۱-۱ و ۲-۱۱-۲ تنظیم پارامتر[۹۰] شدند.

مدلی که بالاترین عملکرد(صحت) را داشته به معنی این است که پارامترهای تنظیمی آن مدل، بر روی این مجموعه داده بهترین پارامترها هستند، این مدل‌ها به عنوان بهترین مدل انتخاب شدند.

برای بار دوم، اینبار بهترین مدل‌ها با تمام ۸۰ درصد داده آموزش دیدند، سپس توسط مجموعه تست مستقل عملکرد مدل‌ها ارزیابی شد.

نتایج بهترین مدل‌ها توسط اعتبارسنجی متقابل ۱۰ لایه در مرحله تنظیم پارامترهای تنظیمی و نتایج مدل‌ها توسط تست مستقل در ادامه گزارش می‌شوند.

۲-۳ نتایج بدست آمده بر روی داده‌های ۱۰ مؤلفه اول PCA

در ادامه نتایج ۲ طبقه‌بند جنگل‌تصادفی و ماشین بردارپشتیبان بر روی ۳ سری مجموعه داده گزارش می‌شود.

سری ۱- داده‌های ۱۰ مؤلفه اول PCA که از تک‌ویژگی‌ها استخراج شد.

در این سری، بر روی ۱۰ ویژگی استخراج شده، به تفکیک ویژگی‌ها PCA و سپس RF و SVM اجرا شد.

سری ۲- داده‌های ۱۰ مؤلفه اول PCA که از ترکیب دوتایی ویژگی‌ها استخراج شد.

[90] Parameter Tuning

در این سری، ویژگی‌هایی که در سری ۱، در تست مستقل، صحت آن‌ها بالاتر از ۰.۷۳ گزارش شده، انتخاب شدند؛ که شامل ۷ دسته ویژگی شد. سپس تمام ترکیب‌های ۲تاییِ این ۷ دسته ویژگی ساخته شد و توسط PCA، ۱۰مؤلفه اول از آن‌ها استخراج شد. سپس بر روی این ۱۰مؤلفه همانند روش قبل، ۲ مدل یادگیری ماشین RF و SVM اجرا شد.

سری ۳- داده‌های ۱۰ مؤلفه اول PCA که از ترکیب **تمام** ویژگی‌ها استخراج شد.

در این سری نیز، همانند سری ۲، ویژگی‌هایی که در سری ۱، در تست مستقل، صحت آن‌ها بالاتر از ۰.۷۳ گزارش شده بود، انتخاب شدند؛ که شامل ۷ دسته ویژگی شد. سپس تمام این ۷ ویژگی‌ها باهم ترکیب شدند و توسط PCA، ۱۰مؤلفه اول از آن‌ها استخراج شد. سپس بر روی این ۱۰مؤلفه همانند روش قبل، ۲ مدل یادگیری ماشین RF و SVM اجرا شد.

در ادامه نتایج بدست آمده از ۲ طبقه‌بند جنگل تصادفی و ماشین بردارپشتیبان بر روی این ۳ سری داده گزارش می‌شود.

۳-۲-۱ نتایج مربوط به طبقه‌بند جنگل‌تصادفی

۳-۲-۱-۱ نتایج مربوط به مجموعه داده‌های سری ۱

نتایج بهترین مدل جنگل‌تصادفی بر روی داده‌های سری۱، داده‌های ۱۰ مؤلفه اول PCA که از تک‌ویژگی‌ها استخراج شد، در مرحله تنظیم پارامترهای تنظیمی و تست مستقل در جدول زیر گزارش می‌شود.

جدول ۳–۱: نتایج بهترین مدل جنگل‌تصادفی در هر تک‌ویژگی

ویژگی	صحت بهترین مدل در تنظیم پارامتر تنظیمی	معیار roc_auc بهترین مدل در تنظیم پارامتر تنظیمی	صحت (تست مستقل)	دقت (تست مستقل)	حساسیت (تست مستقل)	معیار f1 (تست مستقل)
APseudoAAC	۷۹.۰	۸۷.۰	۷۹.۰	۷۹.۰	۷۹.۰	۷۹.۰
CKSAAP	۸۰.۰	۸۹.۰	۷۹.۰	۷۸.۰	۷۸.۰	۷۸.۰
CTD	۷۶.۰	۸۴.۰	۷۴.۰	۷۴.۰	۷۴.۰	۷۴.۰
DDE	۷۷.۰	۸۵.۰	۷۴.۰	۷۴.۰	۷۴.۰	۷۴.۰
Geary	۷۱.۰	۷۷.۰	۷۱.۰	۷۱.۰	۷۲.۰	۷۱.۰
KSCTriad	۷۱.۰	۸۴.۰	۷۷.۰	۷۶.۰	۷۶.۰	۷۶.۰
Moran	۷۰.۰	۷۵.۰	۶۶.۰	۶۷.۰	۶۷.۰	۶۶.۰
NMBroto	۷۱.۰	۷۸.۰	۶۶.۰	۶۶.۰	۶۶.۰	۶۶.۰
PseudoAAC	۷۹.۰	۸۷.۰	۷۹.۰	۷۸.۰	۷۸.۰	۷۸.۰
QSOrder	۷۹.۰	۸۶.۰	۸۱.۰	۸۱.۰	۸۱.۰	۸۱.۰

۳-۲-۱-۲ نتایج مربوط به مجموعه داده‌های سری ۲

با توجه به جدول ۳-۱، ویژگی‌هایی که صحت تست مستقل آن‌ها بالاتر از ۰.۷۳ گزارش شده، انتخاب شدند؛ که شامل ۷ ویژگی APseudoAAC، CKSAAP، CTD، DDE، KSCTriad، PseudoAAC و QSOrder شد. سپس تمام ترکیب‌های ۲تاییِ این ۷ دسته ویژگی ساخته شد و توسط PCA، ۱۰مؤلفه اول از آن‌ها استخراج شد. نتایج بهترین مدل‌های جنگل‌تصادفی بر روی این سری از داده‌ها، در مرحله تنظیم پارامترهای تنظیمی و تست مستقل در جدول زیر گزارش می‌شود.

جدول ۳–۲: نتایج بهترین مدل جنگل‌تصادفی در هر ترکیب دوتایی ویژگی‌ها

ویژگی	صحت (بهترین مدل در تنظیم پارامتر تنظیمی)	معیار roc_auc (بهترین مدل در تنظیم پارامتر تنظیمی)	صحت (تست مستقل)	دقت (تست مستقل)	حساسیت (تست مستقل)	معیار f1 (تست مستقل)
QSOrder+CKSAAP	۸۲.۰	۸۹.۰	۸۰.۰	۸۰.۰	۸۰.۰	۸۰.۰
QSOrder+CTD	۷۶.۰	۸۴.۰	۷۶.۰	۷۶.۰	۷۵.۰	۷۵.۰
QSOrder+DDE	۸۳.۰	۸۹.۰	۸۰.۰	۷۹.۰	۸۰.۰	۷۹.۰
QSOrder+KSCTriad	۷۹.۰	۸۶.۰	۷۷.۰	۷۷.۰	۷۶.۰	۷۷.۰
QSOrder+PseudoAAC	۸۱.۰	۸۸.۰	۸۱.۰	۸۱.۰	۸۱.۰	۸۱.۰
QSOrder+APseudoAAC	۸۱.۰	۸۸.۰	۸۱.۰	۸۱.۰	۸۱.۰	۸۱.۰
PseudoAAC+APseudoAAC	۷۹.۰	۸۷.۰	۷۹.۰	۷۹.۰	۷۹.۰	۷۹.۰
PseudoAAC+CKSAAP	۸۲.۰	۹۰.۰	۷۸.۰	۷۷.۰	۷۸.۰	۷۷.۰
PseudoAAC+CTD	۷۵.۰	۸۴.۰	۷۶.۰	۷۵.۰	۷۵.۰	۷۵.۰
PseudoAAC+DDE	۸۱.۰	۸۹.۰	۸۰.۰	۸۰.۰	۸۰.۰	۸۰.۰
PseudoAAC+KSCTriad	۷۸.۰	۸۶.۰	۷۶.۰	۷۵.۰	۷۵.۰	۷۵.۰
CKSAAP+APseudoAAC	۸۱.۰	۹۰.۰	۷۸.۰	۷۸.۰	۷۸.۰	۷۸.۰
CKSAAP+CTD	۷۸.۰	۸۶.۰	۷۷.۰	۷۷.۰	۷۶.۰	۷۷.۰
CKSAAP+DDE	۸۰.۰	۸۹.۰	۷۵.۰	۷۵.۰	۷۴.۰	۷۴.۰
CKSAAP+KSCTriad	۸۰.۰	۸۸.۰	۷۸.۰	۷۷.۰	۷۷.۰	۷۷.۰
CTD+APseudoAAC	۷۵.۰	۸۴.۰	۷۷.۰	۷۷.۰	۷۷.۰	۷۷.۰
CTD+DDE	۷۶.۰	۸۵.۰	۷۸.۰	۷۸.۰	۷۸.۰	۷۸.۰
CTD+KSCTriad	۷۸.۰	۸۶.۰	۸۰.۰	۸۰.۰	۷۹.۰	۸۰.۰
DDE+APseudoAAC	۸۱.۰	۸۹.۰	۸۰.۰	۸۰.۰	۸۰.۰	۸۰.۰
DDE+KSCTriad	۷۷.۰	۸۶.۰	۷۵.۰	۷۵.۰	۷۴.۰	۷۴.۰
APseudoAAC+KSCTriad	۷۹.۰	۸۶.۰	۷۷.۰	۷۶.۰	۷۶.۰	۷۶.۰

۳-۲-۱-۳ نتایج مربوط به ترکیب تمام ویژگی‌ها

با توجه به جدول ۳-۱، **تمام** ۷ ویژگی‌هایی که صحت تست مستقل آن‌ها بالاتر از ۰.۷۳ گزارش شده، باهم ترکیب شدند؛ سپس توسط PCA، ۱۰مؤلفه اول از آن‌ها استخراج شد. نتایج بهترین مدل جنگل‌تصادفی بر روی این سری از داده، در مرحله تنظیم پارامترهای تنظیمی و تست مستقل در جدول زیر گزارش می‌شود.

جدول ۳–۳: نتایج بهترین مدل جنگل‌تصادفی در ترکیب تمام ۷ ویژگی‌ها

ویژگی	صحت (بهترین مدل در تنظیم پارامتر تنظیمی)	معیار roc_auc (بهترین مدل در تنظیم پارامتر تنظیمی)	صحت (تست مستقل)	دقت (تست مستقل)	حساسیت (تست مستقل)	معیار f1 (تست مستقل)
ترکیب همه ۷ ویژگی‌ها	۸۰.۰	۸۹.۰	۸۲.۰	۸۲.۰	۸۲.۰	۸۲.۰

۳-۲-۲ نتایج مربوط به طبقه‌بند ماشین بردارپشتیبان

۳-۲-۲-۱ نتایج مربوط به مجموعه داده‌های سری ۱

نتایج بهترین مدل‌های ماشین بردارپشتیبان بر روی داده‌های سری۱، داده‌های ۱۰ مؤلفه اول PCA که از تک‌ویژگی‌ها استخراج شد، در مرحله تنظیم پارامترهای تنظیمی و تست مستقل در جدول زیر گزارش می‌شود.

جدول ۳–۴: نتایج بهترین مدل ماشین بردارپشتیبان در هر تک‌ویژگی

ویژگی	صحت (بهترین مدل در تنظیم پارامتر تنظیمی)	صحت (تست مستقل)	دقت (تست مستقل)	حساسیت (تست مستقل)	معیار f1 (تست مستقل)
APseudoAAC	۷۹.۰	۷۷.۰	۷۷.۰	۷۶.۰	۷۶.۰

CKSAAP	۷۹.۰	۷۸.۰	۷۷.۰	۷۷.۰	۷۷.۰
CTD	۷۷.۰	۷۸.۰	۷۷.۰	۷۷.۰	۷۷.۰
DDE	۷۷.۰	۷۳.۰	۷۳.۰	۷۳.۰	۷۳.۰
Geary	۶۹.۰	۶۵.۰	۶۵.۰	۶۵.۰	۶۵.۰
KSCTriad	۷۷.۰	۷۳.۰	۷۲.۰	۷۳.۰	۷۲.۰
Moran	۶۹.۰	۶۴.۰	۶۳.۰	۶۳.۰	۶۳.۰
NMBroto	۷۲.۰	۷۵.۰	۷۵.۰	۷۵.۰	۷۵.۰
PseudoAAC	۸۰.۰	۷۷.۰	۷۶.۰	۷۶.۰	۷۶.۰
QSOrder	۷۹.۰	۸۰.۰	۸۰.۰	۸۰.۰	۸۰.۰

۳-۲-۲-۲ نتایج مربوط به مجموعه داده‌های سری ۲

با توجه به جدول ۳-۴، ویژگی‌هایی که صحت تست مستقل آن‌ها بالاتر از ۰.۷۳ گزارش شده، انتخاب شدند؛ که شامل ۸ ویژگی APseudoAAC، CKSAAP، CTD، DDE، KSCTriad، NMBroto ، PseudoAAC و QSOrder شد. سپس تمام ترکیب‌های ۲تاییِ این ۸ دسته ویژگی ساخته شد و توسط PCA، ۱۰مؤلفه اول از آن‌ها استخراج شد. نتایج بهترین مدل‌های ماشین بردارپشتیبان بر روی این سری از داده‌ها، در مرحله تنظیم پارامترهای تنظیمی و تست مستقل در جدول زیر گزارش می‌شود.

جدول ۳–۵: نتایج بهترین مدل جنگل‌تصادفی در هر ترکیب دوتایی ویژگی‌ها

ویژگی	صحت (بهترین مدل در تنظیم پارامتر تنظیمی)	صحت (تست مستقل)	دقت (تست مستقل)	حساسیت (تست مستقل)	معیار f1 (تست مستقل)
QSOrder+CKSAAP	۸۲.۰	۷۸.۰	۷۸.۰	۷۸.۰	۷۸.۰
QSOrder+CTD	۷۷.۰	۷۷.۰	۷۷.۰	۷۶.۰	۷۷.۰
QSOrder+DDE	۸۱.۰	۷۹.۰	۷۹.۰	۷۹.۰	۷۹.۰
QSOrder+KSCTriad	۷۸.۰	۷۱.۰	۷۰.۰	۷۱.۰	۷۰.۰
QSOrder+PseudoAAC	۸۱.۰	۷۹.۰	۷۹.۰	۷۹.۰	۷۹.۰

۸۱.۰	۸۱.۰	۸۱.۰	۸۱.۰	۸۱.۰	**QSOrder+APseudoAAC**
۷۶.۰	۷۶.۰	۷۷.۰	۷۷.۰	۷۹.۰	**PseudoAAC+APseudoAAC**
۸۰.۰	۸۰.۰	۸۰.۰	۸۰.۰	۸۱.۰	**PseudoAAC+CKSAAP**
۷۵.۰	۷۵.۰	۷۵.۰	۷۶.۰	۷۶.۰	**PseudoAAC+CTD**
۷۳.۰	۷۴.۰	۷۳.۰	۷۳.۰	۸۰.۰	**PseudoAAC+DDE**
۷۱.۰	۷۱.۰	۷۱.۰	۷۱.۰	۷۸.۰	**PseudoAAC+KSCTriad**
۷۹.۰	۷۹.۰	۷۹.۰	۸۰.۰	۸۱.۰	**CKSAAP+APseudoAAC**
۷۴.۰	۷۴.۰	۷۵.۰	۷۵.۰	۷۸.۰	**CKSAAP+CTD**
۷۷.۰	۷۶.۰	۷۸.۰	۷۸.۰	۸۰.۰	**CKSAAP+DDE**
۷۵.۰	۷۵.۰	۷۵.۰	۷۵.۰	۷۹.۰	**CKSAAP+KSCTriad**
۷۵.۰	۷۵.۰	۷۵.۰	۷۶.۰	۷۷.۰	**CTD+APseudoAAC**
۷۷.۰	۷۷.۰	۷۷.۰	۷۸.۰	۷۶.۰	**CTD+DDE**
۷۲.۰	۷۲.۰	۷۲.۰	۷۲.۰	۷۷.۰	**CTD+KSCTriad**
۷۴.۰	۷۴.۰	۷۵.۰	۷۵.۰	۸۱.۰	**DDE+APseudoAAC**
۷۴.۰	۷۴.۰	۷۳.۰	۷۴.۰	۷۸.۰	**DDE+KSCTriad**
۷۳.۰	۷۴.۰	۷۳.۰	۷۴.۰	۷۸.۰	**APseudoAAC+KSCTriad**
۷۷.۰	۷۷.۰	۷۷.۰	۷۸.۰	۸۲.۰	**NMBroto+APseudoAAC**
۷۸.۰	۷۸.۰	۷۹.۰	۷۹.۰	۸۰.۰	**NMBroto+CKSAAP**
۷۴.۰	۷۴.۰	۷۵.۰	۷۵.۰	۷۶.۰	**NMBroto+CTD**
۷۶.۰	۷۵.۰	۷۶.۰	۷۷.۰	۸۰.۰	**NMBroto+DDE**
۷۳.۰	۷۳.۰	۷۳.۰	۷۳.۰	۷۸.۰	**NMBroto+KSCTriad**
۷۸.۰	۷۸.۰	۷۸.۰	۷۹.۰	۸۲.۰	**NMBroto+PseudoAAC**
۷۹.۰	۷۹.۰	۷۹.۰	۷۹.۰	۸۰.۰	**NMBroto+QSOrder**

۳-۲-۲-۳ نتایج مربوط به ترکیب تمام ویژگی‌ها

با توجه به جدول ۳-۴، **تمام** ۸ ویژگی‌هایی که صحت تست مستقل آن‌ها بالاتر از ۰.۷۳ گزارش شده، باهم ترکیب شدند؛ سپس توسط PCA، ۱۰مؤلفه اول از آن‌ها استخراج شد. نتایج بهترین مدل‌های ماشین‌بردار

پشتیبان بر روی این سری از داده‌ها، در مرحله تنظیم پارامترهای تنظیمی و تست مستقل در جدول زیر گزارش می‌شود.

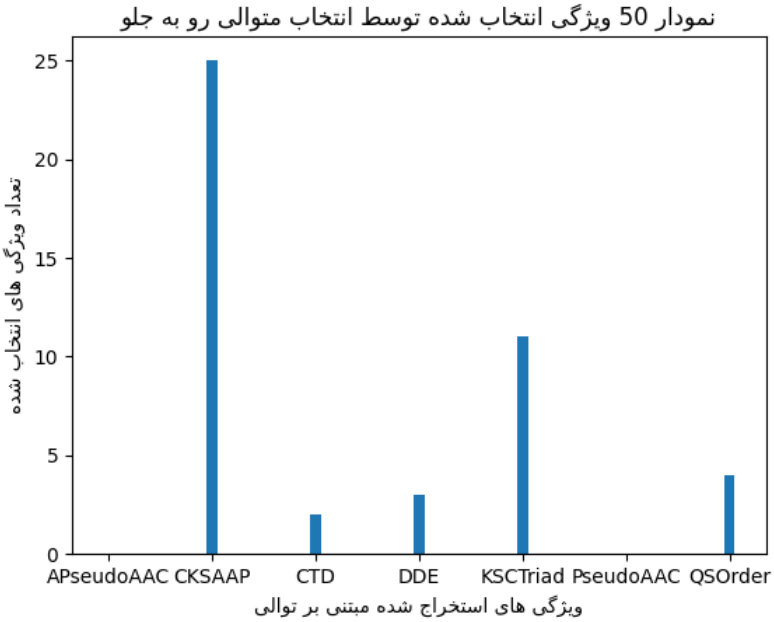
جدول ۳–۶: نتایج بهترین مدل ماشین بردارپشتیبان در ترکیب تمام ۸ ویژگی‌ها

ویژگی	صحت (بهترین مدل در تنظیم پارامتر تنظیمی)	صحت (تست مستقل)	دقت (تست مستقل)	حساسیت (تست مستقل)	معیار f1 (تست مستقل)
ترکیب همه ۸ ویژگی‌ها	۷۹.۰	۷۹.۰	۷۹.۰	۷۹.۰	۷۹.۰

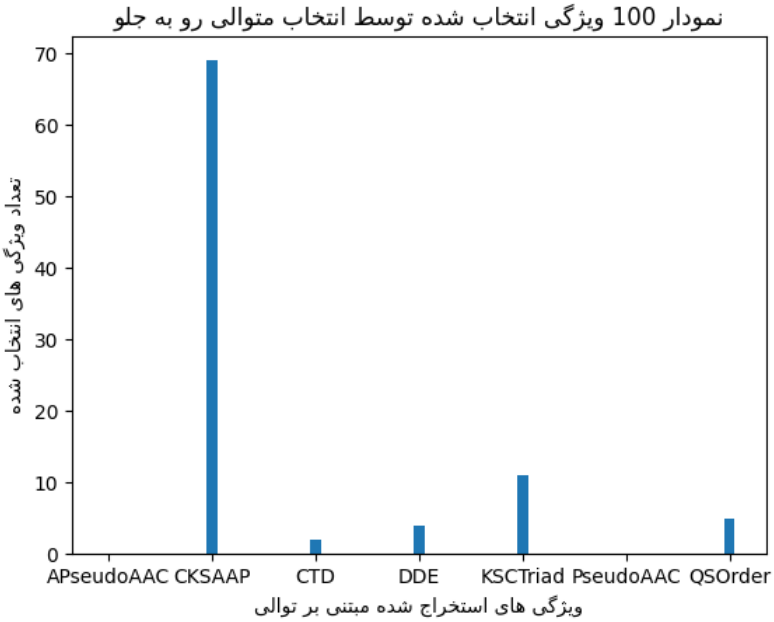
۳-۳ نتایج ویژگی‌های انتخاب شده توسط الگوریتم انتخاب متوالی رو به جلو

با کنار هم قراردادن تمام ۷ ویژگی مشخص شده در جدول ۳-۱ یک فضای ویژگی به ابعاد (۳۰۱۴، ۹۴۰) ایجاد شد. (۳۰۱۴ طول بردار ویژگی و ۹۴۰ تعداد نمونه پپتیدها)

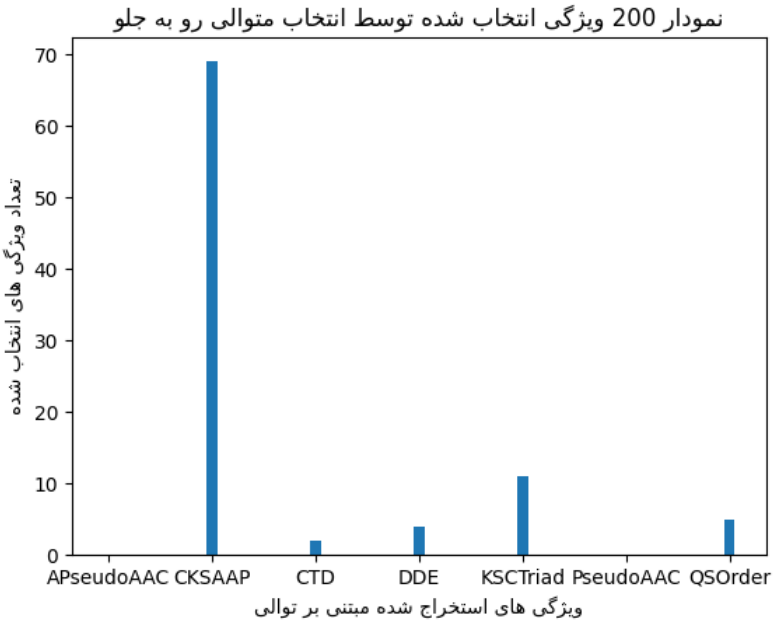
فضای ویژگی جدید به الگوریتم SFS داده شد و ۳ دسته ویژگی SFS50، SFS100 و SFS200 ساخته شد. برای درک این موضوع که چه تعداد از ۷ ویژگی در ویژگی‌های SFS50، SFS100 و SFS200 مشارکت داشته‌اند، نمودارهای زیر رسم شد. با توجه به نتایج هر ۳ دسته نمودار، بیشترین ویژگی‌های انتخاب شده، از۲ دسته ویژگی CKSAAP و KSCTriad است.



شکل ۳–۱: نمودار ۵۰ ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو



شکل ۳–۲: نمودار ۱۰۰ ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو



شکل ۳–۳: نمودار ۲۰۰ ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو

۳-۴ نتایج بدست آمده بر روی ۳ مجموعه داده SFS50، SFS100 و SFS200

در ادامه نتایج طبقه‌بند جنگل‌تصادفی بر روی ۳ مجموعه داده SFS50، SFS100 و SFS200 گزارش می‌شود.

۳-۴-۱ نتایج مربوط به طبقه‌بند جنگل‌تصادفی

نتایج بهترین مدل‌های جنگل‌تصادفی بر روی داده‌های SFS50، SFS100 و SFS200 در مرحله تنظیم پارامترهای تنظیمی و تست مستقل در جدول زیر گزارش می‌شود.

جدول ۳–۷: نتایج بهترین مدل جنگل‌تصادفی بر روی داده‌های بدست آمده از الگوریتم SFS

مجموعه داده	صحت (بهترین مدل در تنظیم پارامتر تنظیمی)	معیار roc_auc (بهترین مدل در تنظیم پارامتر تنظیمی)	صحت (تست مستقل)	دقت (تست مستقل)	حساسیت (تست مستقل)	معیار f1 (تست مستقل)
SFS50	۸۲.۰	۸۹.۰	۷۳.۰	۷۳.۰	۷۳.۰	۷۳.۰
SFS100	۸۲.۰	۸۹.۰	۷۴.۰	۷۴.۰	۷۴.۰	۷۴.۰
SFS200	۸۲.۰	۸۹.۰	۷۴.۰	۷۴.۰	۷۴.۰	۷۴.۰

٤ فصل چهارم: بحث و پیشنهادات

با توجه به نتایج طبقه‌بندها در فصل ۳، مشخص است که استفاده از روش‌های محاسباتی برای مسئله پیش‌بینی پپتیدهای ضدسرطانی روش مناسبی است، زیرا نتایج بدست آمده توسط هر ۲ طبقه‌بند جنگل-تصادفی و ماشین بردارپشتیبان با ترکیب ویژگی‌های مختلف و حتی استفاده از یک ویژگی به تنهایی همگی درصد صحت بالاتر از ۷۱ درصد را در تست مستقل گزارش کرده‌اند. همچنین با توجه دقیق‌تر به نتایج طبقه‌بند جنگل‌تصادفی مشخص می‌شود که استفاده از ویژگی‌های QSOrder، APseudoAAC، PseudoAAC و CKSAAP بصورت تکی، نتایج بالای ۷۹ تا ۸۱ درصد را در تست مستقل دارند، این در

حالی است که برای طبقه‌بند ماشین بردارپشتیبان فقط برای ویژگی QSOrder نتایج تست مستقل بین ۷۹ تا ۸۱ گزارش شد.

در ترکیب ویژگی‌ها نیز طبقه‌بند جنگل‌تصادفی به نسبت طبقه‌بند ماشین بردارپشتیبان نتایج بالاتری در تست مستقل گزارش می‌کند. در ترکیب ویژگی‌ها، ویژگی‌های QSOrder+PseudoAAC و QSOrder+APseudoAAC توسط طبقه‌بند جنگل‌تصادفی با صحت ۸۱ درصد در تست مستقل، بعنوان بیشترین صحت بین ویژگی‌های دیگر گزارش شدند. بالاترین صحت در تست مستقل توسط طبقه‌بند ماشین بردارپشتیبان نیز برای ترکیب ویژگی QSOrder+APseudoAAC گزارش شد.

این در حالی است که در مرحله آموزش و تنظیم پارامترهای تنظیمی، صحت بالای ۸۳ درصد در ترکیب ویژگی QSOrder+DDE توسط مدل جنگل‌تصادفی و صحت بالای ۸۲ درصد در ترکیب ویژگی‌های QSOrder+CKSAAP، NMBroto+APseudoAA و NMBroto+PseudoAAC گزارش شده‌است. بالاترین صحت‌های گزارش شده عموما مربوط به طبقه‌بند جنگل‌تصادفی است.

از طرفی نتایج صحت بدست آمده توسط جنگل‌تصادفی بر روی مجموعه داده‌های SFS50، SFS100 و SFS200 نسبت به صحت مجموعه داده‌های حاصل از تحلیل ۱۰ مؤلفه اول بسیار کمتر است طوری‌که با توجه به جدول ۳-۷ صحت تست مستقل به ۷۳ درصد کاهش یافته‌است. به همین دلیل به‌نظر می‌رسد که استخراج ویژگی‌های مهم توسط الگوریتم SFS برای پیش‌بینی پپتیدهای ضدسرطانی مناسب نیست.

نتایجی که از این پایان‌نامه بدست آمده، نشان می‌دهد که مسأله پیش‌بینی پپتیدهای ضدسرطانی مسأله‌ای قابل حل توسط روش‌های محاسباتی با صحت بسیار بالاتر در آینده‌ای نزدیک است. همچنین مشاهده شد

که ویژگی‌های QSOrder، APseudoAAC، PseudoAAC و CKSAAP ویژگی‌های بهتری نسبت به بقیه ویژگی‌های استخراج شده برای هر دو طبقه‌بند جنگل‌تصادفی و ماشین بردارپشتیبان هستند.

یکی از چالش‌های حال حاضر برای مسئله پیش‌بینی پپتیدهای ضدسرطانی، کمبود داده‌های آزمایشگاهی پپتیدهای ضدسرطانی است. با دسترسی به داده‌های بیشتر می‌توان به بالا رفتن کارایی مدل‌های یادگیری ماشین و در نتیجه پیش‌بینی‌های با صحت بالاتر کمک کرد. نتیجه پیش‌بینی با صحت بالاتر می‌تواند راه‌های رسیدن به درمان‌های نوین برای سرطان را هموارتر کند. همچنین از بهبودهایی که برای این مطالعه می‌توان در نظر گرفت، می‌توان به مسئله پیش‌بینی پپتیدهای ضدسرطانی برای یک نوع سرطان خاص اشاره کرد؛ که این نیازمند جمع‌آوری داده‌های بیشتر و مخصوصا جمع آوری داده‌های مثبت مربوط به آن نوع سرطان خاص است. مورد دیگری که برای بهبود این مطالعه می‌توان در نظر گرفت، استفاده از روش‌های یادگیری ماشین مدرن (برای مثال استفاده از شبکه‌های عصبی عمیق) و یا استفاده از روش‌های پردازش متن است. که این مورد نیز به دلیل ماهیت شبکه‌های عصبی عمیق، نیازمند داده‌های بسیار بیشتری است.

مراجع

١. Xi, J., et al., *Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication.* Bioinformatics, 2020. **36**(6): p. 1855-1863.

٢. Yue, Z., et al., *dbCID: a manually curated resource for exploring the driver indels in human cancer.* Briefings in bioinformatics, 2019. **20**(5): p. 1925-1933.

٣. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: a cancer journal for clinicians, 2018. **68**(6): p. 394-424.

۴. Wijdeven, R.H., et al., *Old drugs, novel ways out: Drug resistance toward cytotoxic chemotherapeutics.* Drug Resistance Updates, 2016. **28**: p. 65-81.

۵. Huang, Y., et al., *Alpha-helical cationic anticancer peptides: a promising candidate for novel anticancer drugs.* Mini reviews in medicinal chemistry, 2015. **15**(1): p. 73-81.

۶. Hoskin, D.W. and A. Ramamoorthy, *Studies on anticancer activities of antimicrobial peptides.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2008. **1778**(2): p. 357-375.

۷. Giuliani, A., G. Pirri, and S. Nicoletto, *Antimicrobial peptides: an overview of a promising class of therapeutics.* Open Life Sciences, 2007. **2**(1): p. 1-33.

۸. Gaspar, D., A.S. Veiga, and M.A. Castanho, *From antimicrobial to anticancer peptides. A review.* Frontiers in microbiology, 2013. **4**: p. 294.

۹. Ting, C.-H., et al., *The mechanisms by which pardaxin, a natural cationic antimicrobial peptide, targets the endoplasmic reticulum and induces c-FOS.* Biomaterials, 2014. **35**(11): p. 3627-3640.

۱۰. Buri, M.V., et al., *Resistance to degradation and cellular distribution are important features for the antitumor activity of gomesin.* PLoS One, 2013. **8**(11): p. e80924.

۱۱. Vijayakumar, S. and P. Lakshmi, *ACPP: A web server for prediction and design of anti-cancer peptides.* International Journal of Peptide Research and Therapeutics, 2015. **21**(1): p. 99-106.

۱۲. Reddy, K., R. Yedery, and C. Aranha, *Antimicrobial peptides: premises and promises.* International journal of antimicrobial agents, 2004. **24**(6): p. 536-547.

۱۳. Ejtehadifar, M., et al., *Anti-cancer effects of Staphylococcal Enterotoxin type B on U266 cells co-cultured with Mesenchymal Stem Cells.* Microbial pathogenesis, 2017. **113**: p. 438-444.

۱۴. Wang, Z. and G. Wang, *APD: the antimicrobial peptide database.* Nucleic acids research, 2004. **32**(suppl_1): p. D590-D592.

۱۵. Schweizer, F., *Cationic amphiphilic peptides with cancer-selective toxicity.* European journal of pharmacology, 2009. **625**(1-3): p. 190-194.

۱۶. Wüthrich, K., *NMR with proteins and nucleic acids.* Europhysics News, 1986. **17**(1): p. 11-13.

۱۷. Marion, D., M. Zasloff, and A. Bax, *A two-dimensional NMR study of the antimicrobial peptide magainin 2.* FEBS letters, 1988. **227**(1): p. 21-26.

١٨. Ganz, T., et al., *Defensins. Natural peptide antibiotics of human neutrophils.* Journal of Clinical Investigation, 1985. **76**(4): p. 1427.

١٩. Hill, C.P., et al., *Crystal structure of defensin HNP-3, an amphiphilic dimer: mechanisms of membrane permeabilization.* Science, 1991. **251**(5000): p. 1481-1485.

٢٠. Landon, C., et al., *Solution structure of drosomycin, the first inducible antifungal protein from insects.* Protein Science, 1997. **6**(9): p. 1878-1884.

٢١. Tamamura, H., et al., *A comparative study of the solution structures of tachyplesin I and a novel anti-HIV synthetic peptide, T22 ([Tyr 5, 12, Lys 7]-polyphemusin II), determined by nuclear magnetic resonance.* Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 1993. **1163**(2): p. 209-216.

٢٢. Xu, T., et al., *Anticandidal activity of major human salivary histatins.* Infection and immunity, 1991. **59**(8): p. 2549-2554.

٢٣. Selsted, M.E., et al., *Indolicidin, a novel bactericidal tridecapeptide amide from neutrophils.* Journal of Biological Chemistry, 1992. **267**(7): p. 4292-4295.

٢٤. Lawyer, C., et al., *Antimicrobial activity of a 13 amino acid tryptophan-rich peptide derived from a putative porcine precursor protein of a novel family of antibacterial peptides.* FEBS letters, 1996. **390**(1): p. 95-98.

٢٥. Gennaro, R., B. Skerlavaj, and D. Romeo, *Purification, composition, and activity of two bactenecins, antibacterial peptides of bovine neutrophils.* Infection and immunity, 1989. **57**(10): p. 3142-3146.

٢٦. AGERBERTH, B., et al., *Amino acid sequence of PR-39.* European Journal of Biochemistry, 1991. **202**(3): p. 849-854.

٢٧. De Vos, W.M., et al., *Properties of nisin Z and distribution of its gene, nisZ, in Lactococcus lactis.* Applied and environmental microbiology, 1993. **59**(1): p. 213-218.

٢٨. Fregeau Gallagher, N.L., et al., *Three-dimensional structure of leucocin A in trifluoroethanol and dodecylphosphocholine micelles: spatial location of residues critical for biological activity in type IIa bacteriocins from lactic acid bacteria.* Biochemistry, 1997. **36**(49): p. 15062-15072.

٢٩. Rokach, L. and O. Maimon, *Decision trees*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 165-192.

٣٠. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

٣١. Osuna, E.E., *Support vector machines: Training and applications*. 1998, Massachusetts Institute of Technology.

٣٢. He, W., et al., *Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides.* Bioinformatics, 2021. **37**(24): p. 4684-4693.

٣٣. Chen, Z., et al., *iFeature: a python package and web server for features extraction and selection from protein and peptide sequences.* Bioinformatics, 2018. **34**(14): p. 2499-2502.

٣٤. Hajisharifi, Z., et al., *Predicting anticancer peptides with Chou' s pseudo amino acid composition and investigating their mutagenicity via Ames test.* Journal of Theoretical Biology, 2014. **341**: p. 34-40.

٣٥. Chen, W., et al., *iACP: a sequence-based tool for identifying anticancer peptides.* Oncotarget, 2016. **7**(13): p. 16895.

٣٦. Li, F.-M. and X.-Q. Wang, *Identifying anticancer peptides by using improved hybrid compositions.* Scientific reports, 2016. **6**(1): p. 1-6.

٣٧. Manavalan, B., et al., *MLACP: machine-learning-based prediction of anticancer peptides.* Oncotarget, 2017. **8**(44): p. 77121.

٣٨. Schaduangrat, N., et al., *ACPred: a computational tool for the prediction and analysis of anticancer peptides.* Molecules, 2019. **24**(10): p. 1973.

٣٩. Boopathi, V., et al., *mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides.* International journal of molecular sciences, 2019. **20**(8): p. 1964.

٤٠. van Zoggel, H., et al., *Antitumor and angiostatic activities of the antimicrobial peptide dermaseptin B2.* 2012.

٤١. Huang, Y., et al., *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-682.

٤٢. Wei, L., et al., *ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides.* Bioinformatics, 2018. **34**(23): p. 4007-4016.

٤٣. Cao, R., et al., *DLFF-ACP: prediction of ACPs based on deep learning and multi-view features fusion.* PeerJ, 2021. **9**: p. e11906.

٤٤. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008.* Nucleic acids research, 2007. **36**(suppl_1): p. D202-D205.

٤٥. Refaeilzadeh, P., L. Tang, and H. Liu, *Cross-validation.* Encyclopedia of database systems, 2009. **5**: p. 532-538.

٤٦. Raschka, S., *MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack.* Journal of open source software, 2018. **3**(24): p. 638.

٤٧. James, G., et al., *An introduction to statistical learning*. Vol. 112. 2013: Springer.

٤٨. machinelearningmastery Authors, J.B. *How to Develop a Random Forest Ensemble in Python*. 2020 April 27]; Available from: https://machinelearningmastery.com/random-forest-ensemble-in-python/.

٤٩. Towardsdatascience Authors, W.K. *Hyperparameter Tuning the Random Forest in Python*. 2018 Jan 18]; Available from: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74.

٥٠. ScikitLearn Authors. *sklearn.svm.SVC version 1.0.2*. 2021; Available from: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

٥١. DataCamp authors, A.N. *Support Vector Machines with Scikit-learn*. 2019 December 27]; Available from: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python.

Abstract

Cancer is one of the major causes of death worldwide. To treat cancer, the use of anticancer peptides (ACPs) has attracted a lot of attention in recent years. ACPs are a unique group of small molecules that can target and kill cancer cells fast and directly. However, identifying ACPs by wet-lab experiments is time-consuming and labor-intensive. Therefore, it is significant to develop computational tools for ACPs prediction. Hence, this study tries to train machine learning models for distinguishing ACPs from non-ACPs. This study utilizes data of ACPs and non-ACPs, uses iFeature python package to extract features from peptide sequences, builds and trains Rrandom Forest and SVM models with scikit-learn python package on extracted features.

Comparing random forest models in this work, random forest with higher than 82% accuracy shows the best performance. Also comparing all models in this work, random forest and SVM models with QSOrder+APseduoAAC features show better performance (accuracy: 81%) on independent test set.

Keywords: Anticancer Peptides Prediction, Machine Learning, Random Forest, SVM, Cancer

Prediction of *Anticancer Peptides* *Using Machine Learning*

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science (M.Sc.) in Bioinformatics

Department of Biophysics
Faculty of Bioinformatics
Tarbiat Modares University

By:
Mohammadtabar Zeynab

Supervisor:
Dr. Abdolmaleki Parviz

January 2022