



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

**چكيده**

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 82 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 81 درصد صحت دارند.

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

**فهرست مطالب**

[1 فصل اول: مقدمه 1](#_Toc93745819)

[1-1 مقدمه‌ای بر پپتیدهای ضدسرطانی 1](#_Toc93745820)

[2-1 ویژگی­های زیستی پپتیدهای ضدسرطانی 2](#_Toc93745821)

[3-1 طبقه­بندی پپتیدهای ضدسرطانی 4](#_Toc93745822)

[1-3-1 پپتیدهای ضدسرطانی با ساختار آلفا-هلیکس (*α-*HelicalACPs) 5](#_Toc93745823)

[2-3-1 پپتیدهای ضدسرطانی غنی از سیستئین 5](#_Toc93745824)

[3-3-1 پپتیدهای ضدسرطانی با ساختار بتا-شیت (*β-*SheetACPs) 6](#_Toc93745825)

[4-3-1 پپتیدهای ضدسرطانی غنی از اسیدآمینه­های منظم 6](#_Toc93745826)

[5-3-1 پپتیدهای ضدسرطانی با اسیدآمینه­های تغییریافته یا اصلاح­شده 7](#_Toc93745827)

[4-1 روش‌های تجربی شناسایی پپتیدهای ضدسرطانی 7](#_Toc93745828)

[5-1 روش‌های محاسباتی پیش­بینی پپتیدهای ضدسرطانی 8](#_Toc93745829)

[1-5-1 روش‌های مبتنی بر یادگیری ماشین 8](#_Toc93745830)

[6-1 پیشینه تحقیق 13](#_Toc93745831)

[7-1 هدف از انجام طرح 15](#_Toc93745832)

[2 فصل دوم: مواد و روش‌ها 17](#_Toc93745833)

[1-2 مجموعه داده 17](#_Toc93745834)

[1-1-2 مجموعه داده مثبت 18](#_Toc93745835)

[2-1-2 مجموعه داده منفی 19](#_Toc93745836)

[2-2 پیش­پردازش مجموعه داده 19](#_Toc93745837)

[1-2-2 شناسایی و حذف توالی پپتیدهای مشابه توسط ابزار cd-hit 19](#_Toc93745838)

[2-2-2 فیلتر طول توالی پپتیدها 23](#_Toc93745839)

[3-2-2 بالانس کردن داده­های مثبت و منفی 25](#_Toc93745840)

[3-2 مجموعه داده آموزش و تست 27](#_Toc93745841)

[1-3-2 مجموعه داده آموزش 27](#_Toc93745842)

[2-3-2 مجموعه داده تست مستقل 27](#_Toc93745843)

[4-2 استخراج ویژگی 28](#_Toc93745844)

[1-4-2 ویژگی Pseudo-Amino Acid Composition (PseudoAAC) 28](#_Toc93745845)

[2-4-2 ویژگی Amphiphilic Pseudo-Amino Acid Composition (APsedoAAC) 30](#_Toc93745846)

[3-4-2 ویژگی Composition of k-spaced Amino Acid Pairs (CKSAAP) 31](#_Toc93745847)

[4-4-2 ویژگی Composition/Transition/Distribution (CTD) 31](#_Toc93745848)

[5-4-2 ویژگی Dipeptide Deviation from Expected Mean (DDE) 33](#_Toc93745849)

[6-4-2 ویژگی Moran correlation (Moran) 35](#_Toc93745850)

[7-4-2 ویژگی Geary correlation (Geary) 37](#_Toc93745851)

[8-4-2 ویژگی Normalized Moreau-Broto Autocorrelation (NMBroto) 37](#_Toc93745852)

[9-4-2 ویژگی *k*-Spaced Conjoint Triad (KSCTriad) 38](#_Toc93745853)

[10-4-2 ویژگی Quasi-sequence-order (QSOrder) 38](#_Toc93745854)

[5-2 پیش‌پردازش داده‌ 39](#_Toc93745855)

[6-2 نرمالسازی داده‌ 39](#_Toc93745856)

[7-2 ‌ معیارهای ارزیابی 40](#_Toc93745857)

[1-7-2 حساسیت 41](#_Toc93745858)

[2-7-2 اختصاصیت 41](#_Toc93745859)

[3-7-2 صحت 42](#_Toc93745860)

[4-7-2 دقت 42](#_Toc93745861)

[8-2 اعتبارسنجی متقابل -لایه 42](#_Toc93745862)

[9-2 انتخاب ویژگی 43](#_Toc93745863)

[1-9-2 روش بسته‌بندی 43](#_Toc93745864)

[10-2 کاهش ابعاد مجموعه داده 46](#_Toc93745865)

[11-2 مدل‌های یادگیری ماشین اجرا شده روی مجموعه داده‌ها 47](#_Toc93745866)

[1-11-2 طبقه‌بندی جنگل تصادفی 48](#_Toc93745867)

[2-11-2 طبقه‌بندی ماشین بردار پیشتیبان 49](#_Toc93745868)

[3 فصل سوم: نتایج و بحث 51](#_Toc93745869)

[1-3 نتایج طبقه‌بندها 51](#_Toc93745870)

[2-3 نتایج بدست آمده بر روی داده­های 10 مؤلفه اول PCA 52](#_Toc93745871)

[1-2-3 نتایج مربوط به طبقه­بند جنگل­تصادفی 53](#_Toc93745872)

[2-2-3 نتایج مربوط به طبقه‌بند ماشین بردارپشتیبان 56](#_Toc93745873)

[3-3 نتایج ویژگی­های انتخاب شده توسط الگوریتم انتخاب متوالی رو به جلو 59](#_Toc93745874)

[4-3 نتایج بدست آمده بر روی 3 مجموعه داده SFS50، SFS100 و SFS200 61](#_Toc93745875)

[1-4-3 نتایج مربوط به طبقه­بند جنگل­تصادفی 61](#_Toc93745876)

[4 فصل چهارم: بحث و پیشنهادات 63](#_Toc93745877)

**فهرست شکل­ها و نمودار­ها**

[شکل ‏1–1: تعدادی پپتید ضدمیکروبی طبیعی با فعالیت ضدسرطانی[6] 4](#_Toc93745942)

[شکل ‏1–2: ساختار کلی یک درخت تصمیم 10](#_Toc93745943)

[شکل ‏1–3: ساختار کلی یک طبقه­بند جنگل­تصادفی 11](#_Toc93745944)

[شکل ‏1–4: ساختار کلی یک ماشین بردارپشتیبان دوکلاس خطی 12](#_Toc93745945)

[شکل ‏1–5: مثالی از فضای نمونه‌ پس از اعمال تابع هسته 13](#_Toc93745946)

[شکل ‏1–6: مراحل ادغام ساخت مجموعه­داده از پپتیدهای ضدسرطانی و غیرضدسرطانی[32] 14](#_Toc93745947)

[شکل ‏2–1: نمایی کلی از مجموعه داده مثبت(پپتیدهای ضدسرطانی) 18](#_Toc93745948)

[شکل ‏2–2: نمایی کلی از مجموعه داده منفی(پپتیدهای غیرضدسرطانی) 19](#_Toc93745949)

[شکل ‏2–3: نمایی کلی از مجموعه داده مثبت پس از اعمال cd-hit90 21](#_Toc93745950)

[شکل ‏2–4: نمایی کلی از مجموعه داده منفی پس از اعمال cd-hit90 21](#_Toc93745951)

[شکل ‏2–5: نمایی کلی از مجموعه داده مثبت پس از اعمال cd-hit80 22](#_Toc93745952)

[شکل ‏2–6: نمایی کلی از مجموعه داده منفی پس از اعمال cd-hit80 22](#_Toc93745953)

[شکل ‏2–7: نمایی کلی از مجموعه داده مثبت cd-hit90 پس از اعمال فیلتر طول توالی 24](#_Toc93745954)

[شکل ‏2–8: نمایی کلی از مجموعه داده منفی cd-hit90 پس از اعمال فیلتر طول توالی 24](#_Toc93745955)

[شکل ‏2–9: نمایی کلی از مجموعه داده منفی cd-hit90 پس از انتخاب رندوم 26](#_Toc93745956)

[شکل ‏2–10: نمایی کلی از مجموعه داده مثبت و منفی پس از اعمال پیش­پردازش­ها 26](#_Toc93745957)

[شکل ‏2–11: یک مثال مصور از ویژگی­های فیزیکوشیمیایی در دیتابیس AAindex [44] 35](#_Toc93745958)

[شکل ‏2–12: نمودار انتخاب متوالی رو به جلو تا 50 ویژگی 45](#_Toc93745959)

[شکل ‏2–13: نمودار انتخاب متوالی رو به جلو تا 100 ویژگی 45](#_Toc93745960)

[شکل ‏2–14: نمودار انتخاب متوالی رو به جلو تا 200 ویژگی 46](#_Toc93745961)

[شکل ‏3–1: نمودار 50 ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو 60](#_Toc93745962)

[شکل ‏3–2: نمودار 100 ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو 60](#_Toc93745963)

[شکل ‏3–3: نمودار 200 ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو 61](#_Toc93745964)

**فهرست جدول­ها**

[جدول ‏3–1: نتایج بهترین مدل جنگل­تصادفی در هر تک­ویژگی 54](#_Toc93746077)

[جدول ‏3–2: نتایج بهترین مدل جنگل­تصادفی در هر ترکیب دوتایی ویژگی­ها 55](#_Toc93746078)

[جدول ‏3–3: نتایج بهترین مدل جنگل­تصادفی در ترکیب تمام 7 ویژگی­ها 56](#_Toc93746079)

[جدول ‏3–4: نتایج بهترین مدل ماشین بردارپشتیبان در هر تک­ویژگی 56](#_Toc93746080)

[جدول ‏3–5: نتایج بهترین مدل جنگل­تصادفی در هر ترکیب دوتایی ویژگی­ها 57](#_Toc93746081)

[جدول ‏3–6: نتایج بهترین مدل ماشین بردارپشتیبان در ترکیب تمام 8 ویژگی­ها 59](#_Toc93746082)

[جدول ‏3–7: نتایج بهترین مدل جنگل­تصادفی بر روی داده­های بدست آمده از الگوریتم SFS 62](#_Toc93746083)

# فصل اول: مقدمه

## مقدمه‌ای بر پپتید­های ضد­سرطانی[[1]](#footnote-1)

سرطان یکی از علل عمده مرگ­و­میر میلیون­ها انسان در جهان است[1, 2]. برآورد آژانس بین­المللی تحقیقات سرطان (IARC)[[2]](#footnote-2) نشان می­دهد که در سال 2018، 18.1 میلیون مورد جدید سرطان و 9.6 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[3].

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 DNAبه کار می­روند به صورت مقاوم و خاموش به تکثیر خود ادامه می­دهند[4]. 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[5].

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

## 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[6]. یکی از انواع پپتیدهای مورد استفاده، پپتید ضدمیکروبی[[3]](#footnote-3) است که بخشی از پاسخ ایمنی ذاتی در برابر میکروب­ها در بسیاری از گونه­ها است. این پپتیدها­ دارای وزن مولکولی کم (10-40 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 سلول­های طبیعی دارای فسفولیپیدهایی با بار منفی بیشتر نظیر فسفاتیدیل سرین و گلیکوپروتئین­ها و گلیکوز­آمینوگلیکان های منفی هستند[7]. 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[8-10].

پپتیدهای ضدسرطانی در درمان دیابت، سرطان و بیماری­های قلبی عروقی به کار برده می­شوند[11]. 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[12].



شکل ‏1–: تعدادی پپتید ضدمیکروبی طبیعی با فعالیت ضدسرطانی[6]

در دهه­های اخیر استفاده از این پپتید­ها به عنوان عوامل ضد­سرطان به عنوان یک روش درمانی نوین در نظر گرفته­شده­است [13-15].

## طبقه­بندی پپتیدهای ضدسرطانی

رزونانس مغناطیسی هسته­ای NMR[[4]](#footnote-4) 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[16]. براساس ساختارهای شناخته شده توسط NMR پپتیدهای ضد­سرطانی به پنج گروه طبقه­بندی می­شوند.

### پپتید­های ضد­سرطانی با ساختار آلفا-هلیکس (*α-*HelicalACPs)

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 15% 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 25% از Trifluoroethanol است[17].

### پپتید­های ضد­سرطانی غنی از سیستئین

پپتیدهای نوتروفیل انسان HNP-1-2-3 اولین پپتیدهای غنی از سیستئین جدا شده از گرانول انسان بودند[18]. پپتیدهای غنی از سیستئین با 30 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، C3-C5-C6 C2 است. مطالعات کریستالوگرافی HNP-3 توسط XRAY، در ترکیب با سانتریفوژ تعادل رسوب، نشان می­دهد که این پپتید به صورت دایمر وجود دارد[19]. 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[20].

### پپتید­های ضد­سرطانی با ساختار بتا-شیت (*β-*SheetACPs)

تعدادی از پپتیدهای ضد­سرطانی با طول20 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 [20, 21]. مطالعات NMR همراه با ساختار­های 3D نشان می­دهد که پپتید Tachyplesin شباهت بسیار زیادی با پپتیدهای جدا شده از خوک دارد. مولکول پپتیدی صفحات آنتی­پارالل به یک ساختار Turn متصل شده­اند و شامل دو پل دی­سولفید هستند [21].

### 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[22]. 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[23-26].

### 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[[5]](#footnote-5)، که توسط باکتری لاکتوکوکوس­لاکتیس[[6]](#footnote-6) 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[27]. 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[28].

## 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 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.

## 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.

### روش‌های مبتنی بر یادگیری ماشین

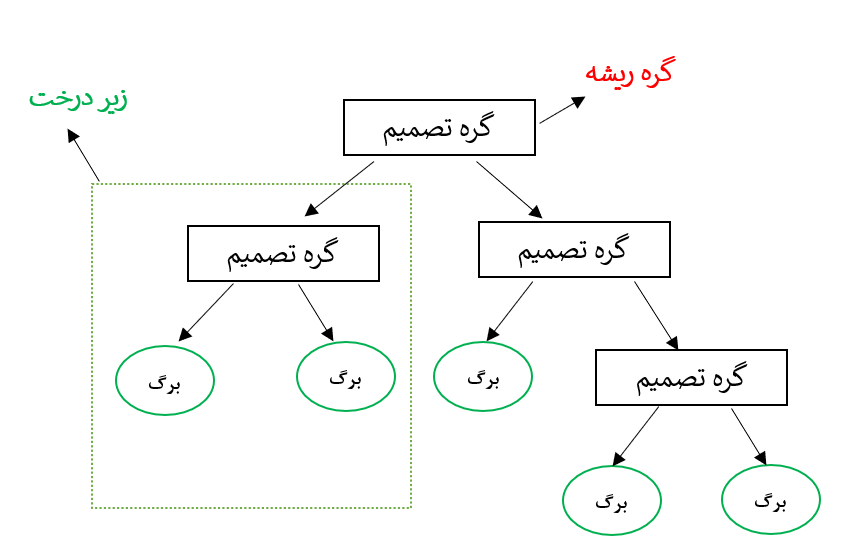
یادگیری ماشین به مجموعه وسیعی از روش‌های محاسباتی برای درک داده‌ها اشاره دارد که می‌تواند الگوهای پیچیده در یک مجموعه را پیدا و بر اساس آن‌ها تصمیم‌گیری کند. این روش‌ها را می‌توان به دو دسته یادگیری با نظارت[[7]](#footnote-7) و یادگیری بدون نظارت[[8]](#footnote-8) تقسیم کرد. یادگیری ماشین با نظارت، یک مدل آماری برای پیش‌بینی برچسب کلاس[[9]](#footnote-9) بر اساس یک یا چند توصیف­کننده[[10]](#footnote-10) (ویژگی[[11]](#footnote-11)) ایجاد می‌کند. در این روش با هدف پیش‌بینی دقیق کلاس­ها برای مشاهدات آینده، مدلی[[12]](#footnote-12) 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.[[13]](#footnote-13)

در مسائلی که داده‌ها برچسب کلاس دارند؛ از روش‌های با نظارت که به آن­ها طبقه‌بندی[[14]](#footnote-14) می­گویند، استفاده می‌شود. در ادامه به اختصار طبقه­بندهای استفاده شده در این پایان­نامه توضیح داده می‌شود.

#### طبقه­بند درخت تصمیم[[15]](#footnote-15)

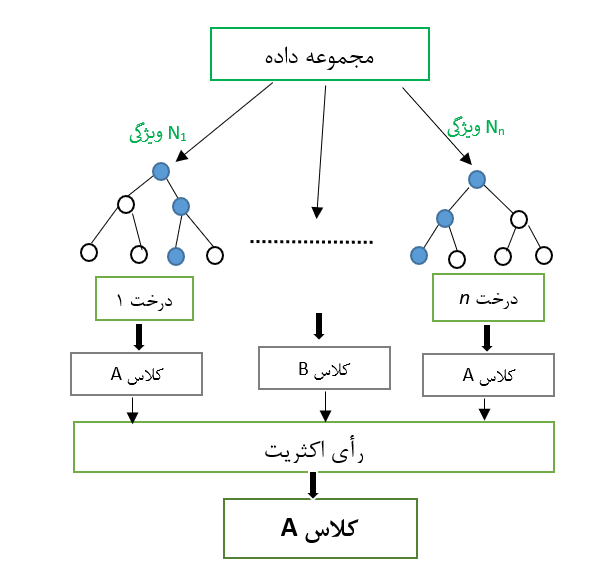
درخت نوع خاصی از گراف است و درخت تصمیم یک روش برای پیش‌بینی بر اساس این ساختار است. به دلیل تفسیر­پذیر بودن درخت­تصمیم، این روش از پرکاربردترین مدل­ها در مسائل یادگیری ماشین کلاسیک است. این ساختار از گره ریشه[[16]](#footnote-16) شروع می‌شود و به برگ­ها[[17]](#footnote-17) (گره پایانی) ختم می‌شود. برگ‌ 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[29].



شکل ‏1–: ساختار کلی یک درخت تصمیم

#### طبقه­بند جنگل تصادفی[[18]](#footnote-18)

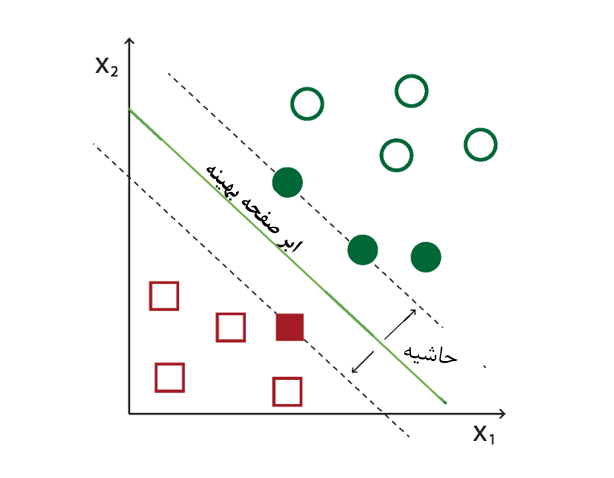
طبقه‌بند جنگل تصادفی از تعداد زیادی درخت­تصمیم استفاده می‌کند. به همین دلیل الگوریتم جنگل­تصادفی در دسته الگوریتم­های یادگیری گروهی[[19]](#footnote-19) قرار می­گیرد. در این روش هر درخت­تصمیم به صورت تصادفی تعدادی از ویژگی‌ها را انتخاب می‌کند و در مورد هر نمونه پیش‌بینی انجام می‌دهد. برای تعیین برچسب یک نمونه بین تمام برچسب‌هایی که هر درخت به آن نمونه نسبت داده است؛ رأی اکثریت گرفته می‌شود[30].



شکل ‏1–: ساختار کلی یک طبقه­بند جنگل­تصادفی

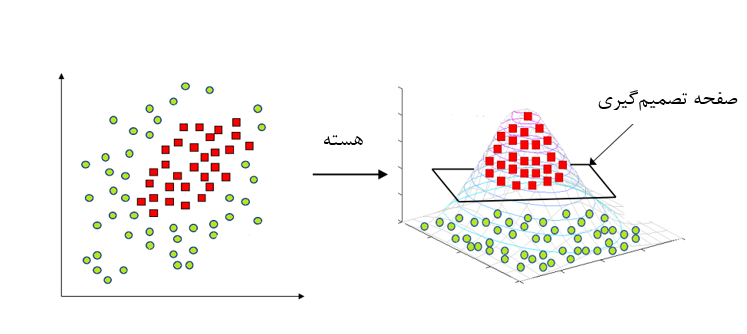
#### طبقه­بند ماشین بردار­پشتیبان[[20]](#footnote-20)

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[[21]](#footnote-21) گفته­ می­شود.



شکل ‏1–: ساختار کلی یک ماشین بردارپشتیبان دوکلاس خطی

از آنجایی که همیشه و به راحتی امکان یافتن چنین ابرصفحه‌ای نیست؛ با تابعی به نام هسته[[22]](#footnote-22) 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[31] .



شکل ‏1–: مثالی از فضای نمونه‌ پس از اعمال تابع هسته

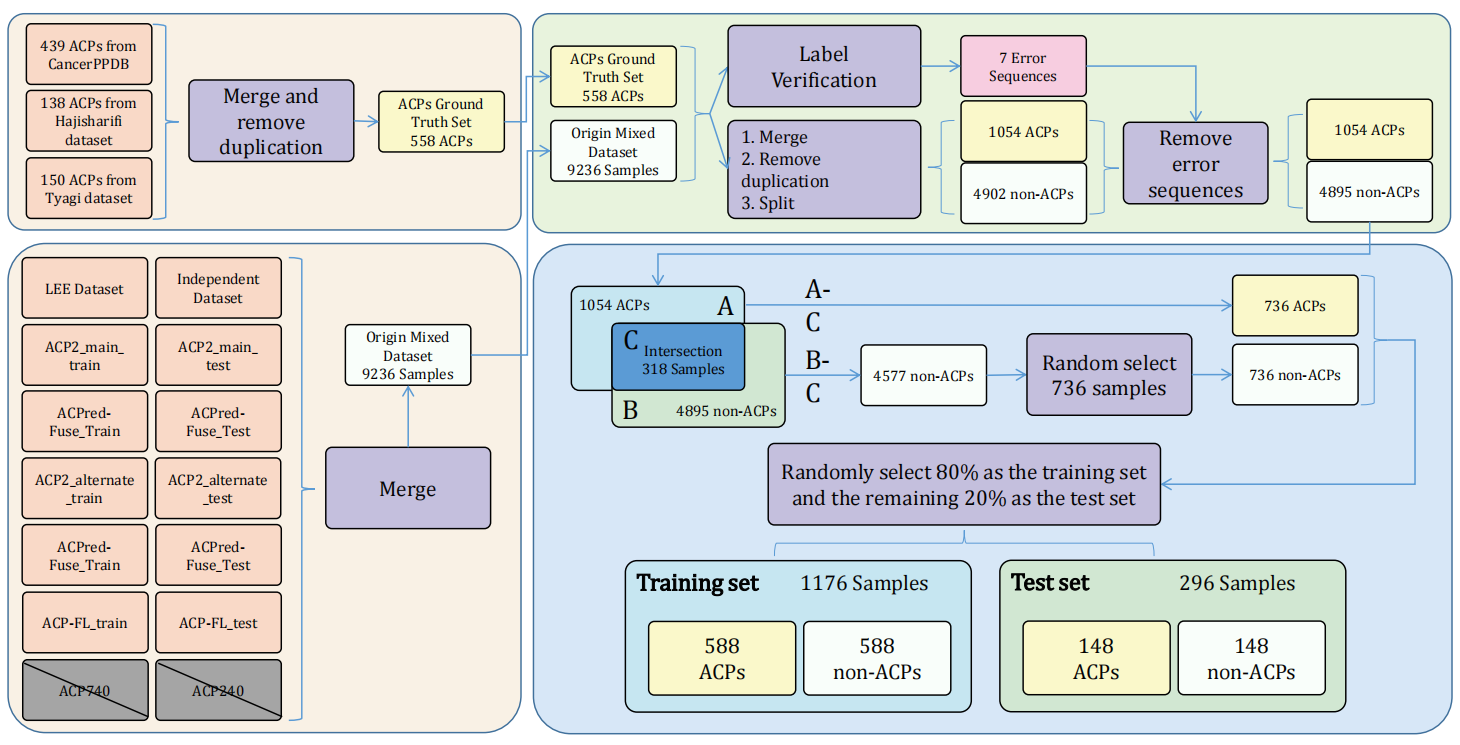
## پیشینه تحقیق

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[[23]](#footnote-23)) متعددی تا به امروز ارائه شده­است. یکی از روش­های پیش­بینی ACPs[[24]](#footnote-24) 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 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[[25]](#footnote-25) و مقالات منتشر­شده در حوزه پیش­بینی ACPs، مطالعه­ای در سال 2021 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[32].

مراحل تجمیع داده­­ بین پایگاه داده­­های پپتید­های ضدمیکروبی و داده­های استفاده شده در مقالات پیشین حوزه ACPs درشکل 1-6 مشاهده می­شود[32].

**

شکل ‏1–: مراحل ادغام ساخت مجموعه­داده­ از پپتید­های ضدسرطانی و غیر­ضد­سرطانی[32]

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 2018 منتشر شده است[33]، 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[[26]](#footnote-26) و SVM[[27]](#footnote-27) کاندیدای خوبی برای استفاده بر روی داده­های ACP، non-ACP[[28]](#footnote-28) هستند[34-39].

## هدف از انجام طرح

اگرچه AMPs[[29]](#footnote-29) 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[6-8, 14, 40]. 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 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.

# فصل دوم: مواد و روش‌ها

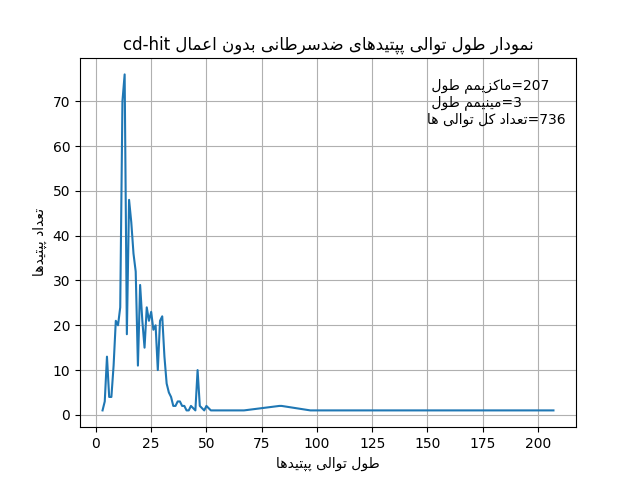
## مجموعه داده[[30]](#footnote-30)

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[32].

1. مجموعه داده مثبت[[31]](#footnote-31)، شامل پپتید­های ضد­میکروبی با عملکرد ضدسرطانی، که توسط محققان زیستی به صورت آزمایشگاهی تایید شده­اند.
2. مجموعه داده منفی[[32]](#footnote-32)، شامل پپتید­های ضدمیکروبی که عملکرد ضدسرطانی برای آن­ها گزارش نشده­است.

### مجموعه داده مثبت

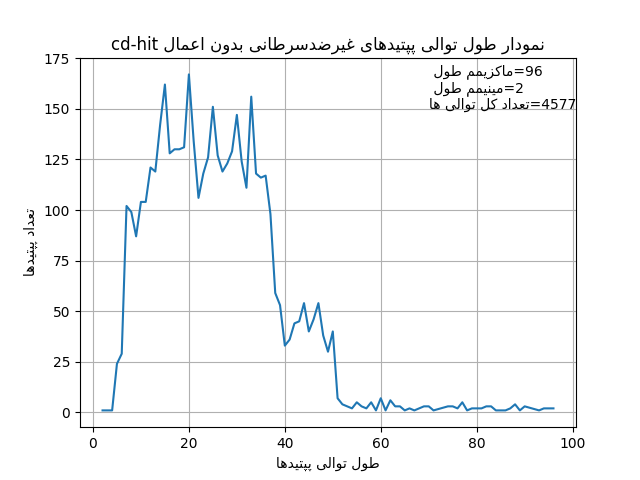
مجموعه داده مثبت شامل پپتید­های ضد­میکروبی با عملکرد ضدسرطانی، که توسط محققان زیستی به صورت آزمایشگاهی تایید شده­اند، شامل 736 پپتید ضدسرطانی(ACP) است[32].



شکل ‏2–1: نمایی کلی از مجموعه داده مثبت(پپتید­های ضدسرطانی)

### مجموعه داده منفی

مجموعه داده منفی شامل پپتید­های ضدمیکروبی که توسط محققان زیستی، عملکرد ضدسرطانی برای آن­ها گزارش نشده­است، شامل 4577 پپتید غیرضدسرطانی(non-ACP) است[32].

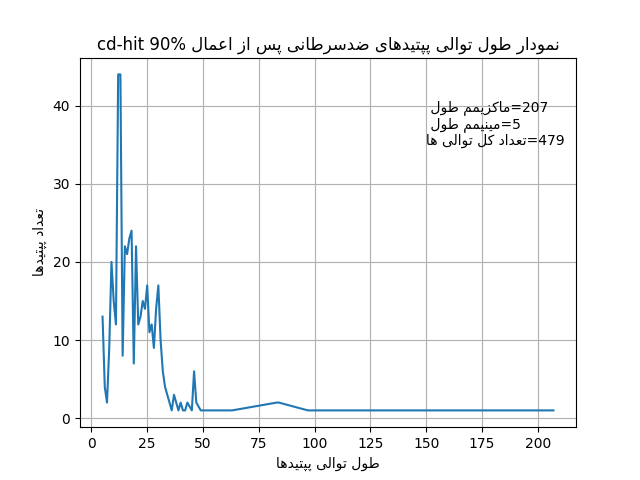


شکل ‏2–2: نمایی کلی از مجموعه داده منفی(پپتیدهای غیرضدسرطانی)

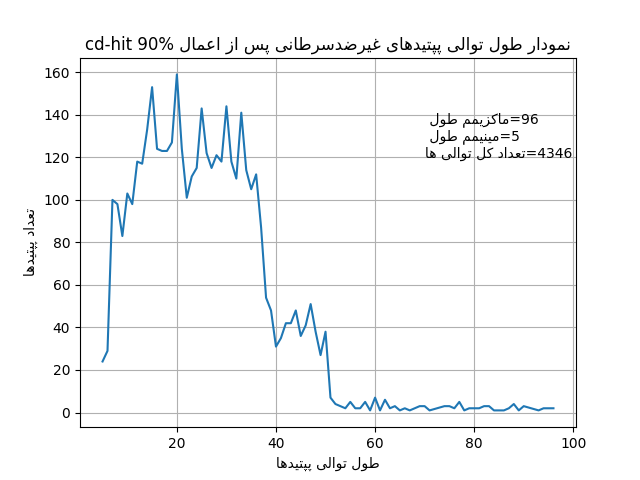
## پیش­پردازش مجموعه داده

### شناسایی و حذف توالی پپتیدهای مشابه توسط ابزار cd-hit

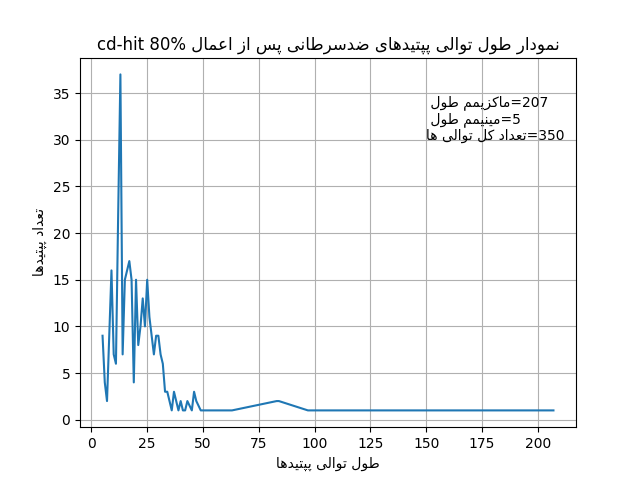
اجتماع مجموعه داده خام مثبت و مجموعه داده خام منفی، مجموعا تعداد 5313 پپتید است. با استفاده از ابزار cd­-­hit­[41]، 2 بار بصورت کاملا مستقل اقدام به حذف توالی­های مشابه شد. یکبار بر روی داده­های خام cd-hit با 90 درصد شباهت[[33]](#footnote-33) و حذف پپتید­های با طول کمتر از 5 اسید­آمینه اجرا شد. بار دیگر مجددا بر روی داده­های خام cd-hit با 80 درصد شباهت و حذف پپتید­های با طول کمتر از 5 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 4825 داده کاهش یافت و با اجراcd-hit80 مجموع داده به 4534 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.



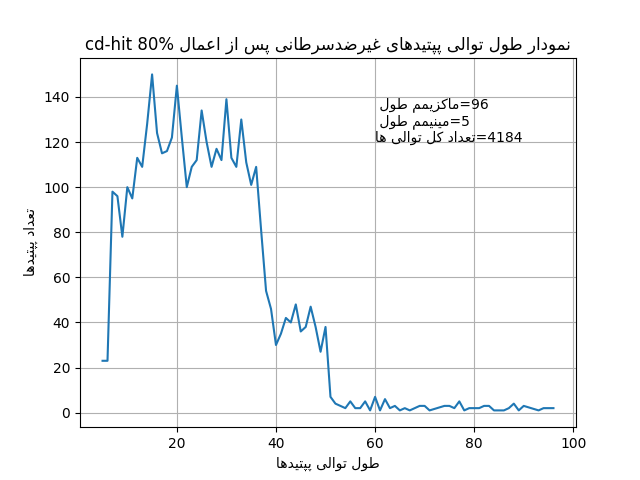
شکل ‏2–3: نمایی کلی از مجموعه داده مثبت پس از اعمال cd-hit90



شکل ‏2–4: نمایی کلی از مجموعه داده منفی پس از اعمال cd-hit90



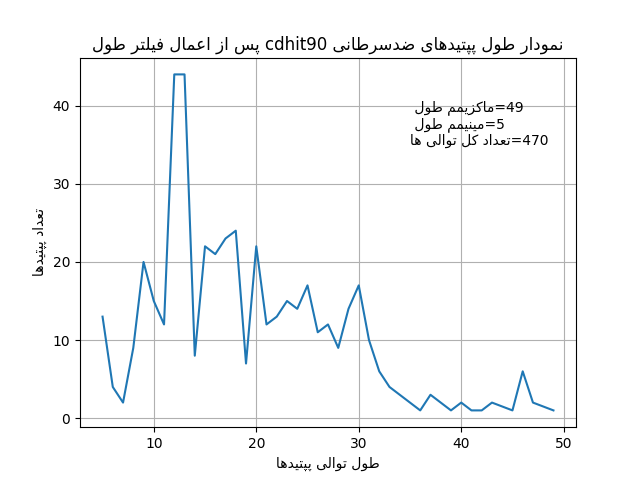
شکل ‏2–5: نمایی کلی از مجموعه داده مثبت پس از اعمال cd-hit80



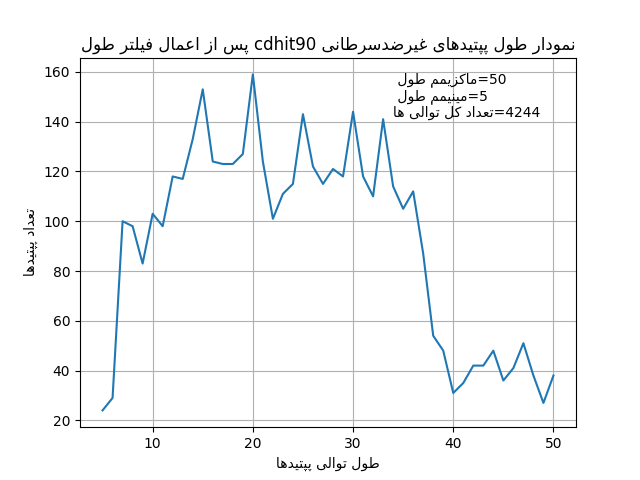
شکل ‏2–6: نمایی کلی از مجموعه داده منفی پس از اعمال cd-hit80

### فیلتر طول توالی پپتید­ها

با توجه به اینکه نُرم طول پپتید­های ضدمیکروبی در طبیعت بین 5 تا حداکثر50 اسید­آمینه دارند[6]، به همین دلیل بر روی داده­های cdhit-90 فیلتر با طول حداکثر50 اعمال شد، که مجموعه داده به 4714 داده کاهش یافت.



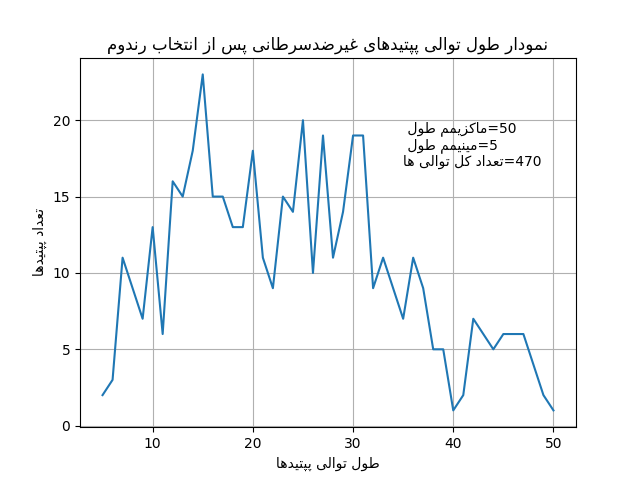
شکل ‏2–7: نمایی کلی از مجموعه داده مثبت cd-hit90 پس از اعمال فیلتر طول توالی



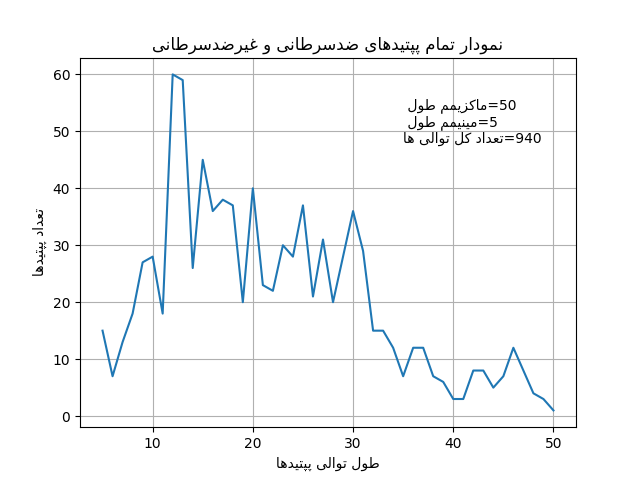
شکل ‏2–8: نمایی کلی از مجموعه داده منفی cd-hit90 پس از اعمال فیلتر طول توالی

### بالانس کردن داده­های مثبت و منفی

پس از اعمال پیش­پردازش بر روی داده­های خام، تعداد داده­های cd-hit90 به تعداد 470 داده مثبت (ACP) و 4244 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 470 داده منفی از بین 4244 داده منفی، در نهایت مجموعه­ای 940تایی با تعدادی مساوی از پپتیدهای ضدسرطانی و غیرضدسرطانی بدست آمد. نمایی کلی از مجموعه داده اصلی برای استفاده در ادامه پژوهش در شکل 2-10 مشاهده می­شود.



شکل ‏2–9: نمایی کلی از مجموعه داده منفی cd-hit90 پس از انتخاب رندوم



شکل ‏2–10: نمایی کلی از مجموعه داده مثبت و منفی پس از اعمال پیش­پردازش­ها

## مجموعه داده آموزش[[34]](#footnote-34) و تست[[35]](#footnote-35)

برای آموزش و ارزیابی مدل­های یادگیری ماشین، مجموعه داده 940تایی از پپتیدهای بدست آمده در مرحله قبل به نسبت 80-20 طوری که 80 درصد برای مجموعه داده آموزش و 20 درصد برای مجموعه داده تست مستقل در نظر گرفته شد.

### مجموعه داده آموزش

برای یادگیری مدل­های یادگیری ماشین و همچنین تنظیم پارامتر­های[[36]](#footnote-36) مدل­­ها، 80 درصد از کل مجموعه داده مثبت یعنی 376 داده مثبت و همچنین 80 درصد از مجموعه داده منفی یعنی 376 داده منفی به صورت کاملا رندوم و بدون تکرار انتخاب شدند، که در مجموع برای هر بار آموزش و تنظیم پارامترها، 752 داده کنار گذاشته شد.

### مجموعه داده تست مستقل[[37]](#footnote-37)

برای ارزیابی عملکرد مدل­های آموزش دیده 20 درصد از 940 پپتید موجود کنار گذاشته شد، طوری که این 20 درصد در هیچ مرحله ای از آموزش و انتخاب پارامترهای تنظیمی استفاده نشدند. بعبارتی 188 داده طوری که تعداد 94 داده مثبت بصورت کاملا رندوم و بدون تکرار از مجموعه داده­های مثبت و همچنین تعداد 94 داده منفی بصورت کاملا رندوم و بدون تکرار از مجموعه داده­های منفی انتخاب شدند.

## استخراج ویژگی[[38]](#footnote-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[33]، در ادامه 10 ویژگی­ استخراج شده مبتنی بر ساختار اول پپتیدها که در این پایان­نامه استفاده شده است، به اختصار توضیح داده می­شود[34, 35, 37, 38, 42, 43].

### ویژگی Pseudo-Amino Acid Composition (PseudoAAC)

فرض کنید ، و برای 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.

|  |  |
| --- | --- |
| رابطه (2-1) |  |

و نیز به ترتیب برای ویژگی­های آب­دوستی و وزن زنجیره جانبی برای هر اسید­آمینه مانند رابطه(1-2) به­دست می‌آیند. سپس تابع correlation­­ای به ازای هر جفت اسید­آمینه و مقدار میانگین این 3 ویژگی از رابطه (2-2) محاسبه می­شود.

|  |  |
| --- | --- |
| رابطه (‏2‑1) |  |

سپس مجموعه‌ عوامل مرتبط با ترتیب توالی[[39]](#footnote-39) به صورت زیر تعریف می‌شود. طول توالی و است. در اینجا λ برابر با 4 در نظر گرفته شده‌است.

|  |  |
| --- | --- |
| رابطه (2-3) |  |

در نهایت برداری به طول در اینجا به طول 24 برای هر پروتئین از رابطه (2-4) بدست می­آید که به آن ویژگی PseudoAAC گفته می­شود. فراوانی نرمال شده اسید­آمینه ام و *w=0.5* در نظر گرفته شده‌است[33].

|  |  |
| --- | --- |
| رابطه (2-4) |  |

### ویژگی Amphiphilic Pseudo-Amino Acid Composition (APsedoAAC)

تعریف مجموعه ویژگی­های این ویژگی مشابه ویژگی PseudoAAC است. و ویژگی­های آبگریزی و آب­دوستی از رابطه همبستگی زیر محاسبه می­شود.

رابطه (2-5)

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

رابطه (2-6)

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

### ویژگی Composition of k-spaced Amino Acid Pairs (CKSAAP)

ویژگی CKSAAP، تعداد تکرار جفت اسید­آمینه که توسط *k اسید­آمینه از هم فاصله[[40]](#footnote-40) دارند را محاسبه می­کند. در این پایان نامه 0،1،2،3 k =* در نظر گرفته شده­است. به­همین دلیل بردار ویژگی به طول 1600 برای CKSAAP ساخته شد.

### 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 پروتئین یا پپتید را نشان می‌دهد. برای محاسبه این ویژگی از 13 ویژگی فیزیکوشیمیایی استفاده شده‌است؛ که شامل ویژگی های آبگریزی، قطبیت[[41]](#footnote-41)، قطبش پذیری[[42]](#footnote-42)، بار[[43]](#footnote-43)، حجم نرمال شده وندروالس[[44]](#footnote-44)، ساختارهای ثانویه[[45]](#footnote-45) و قابلیت دسترسی به حلال[[46]](#footnote-46) 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 39، 39 و 195 تولید می‌کنند. در نهایت بردار ویژگی CTD برای یک پروتئین از کنار هم قرار گرفتن این سه بردار ساخته می‌شود که به طول 273 است.

### ویژگی Dipeptide Deviation from Expected Mean (DDE)

ویژگی DDE توسط 3 پارامتر ساخته می­شود، 1- ترکیب دوتایی پپتید­ها [[47]](#footnote-47)(*Dc*)، میانگین نظری[[48]](#footnote-48) (*Tm*) و واریانس نظری[[49]](#footnote-49) (*Tv*). 3 پارامترهای بیان شده و DDE در روابط زیر محاسبه می­شوند.

*Dc(r,s)، معیار ترکیب دوتایی برای دوپپتید* ‘*rs’ :*

که، *Nrs*  تعداد پپتیدهای دوتایی با اسیدآمینه *r و s، N طول پپتید هستند.*

*Tm(r,s)، میانگین نظری:*

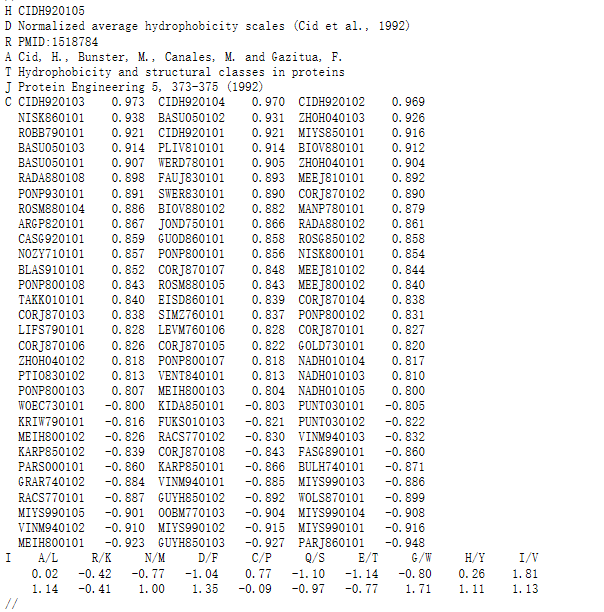
که، *Cr* تعداد کدون[[50]](#footnote-50)­های کد­کننده اولین اسید­آمینه و *Cs* تعداد کدون­های کدکننده دومین اسید­آمینه در پپتید دوتایی ‘*rs*’ است. *CN* تعداد کل کدون­های ممکن بغیر از 3 کدون متوقف کننده[[51]](#footnote-51) است.

*Tv* (*r,s*)، واریانس نظری:

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

ویژگی DDE در نهایت بردار ویژگی به طول 400 تولید می­کند.

### ویژگی Moran correlation (Moran)



شکل ‏2–11: یک مثال مصور از ویژگی­های فیزیکوشیمیایی در دیتابیس AAindex [44]

توصیف کننده­های[[52]](#footnote-52) خودهمبستگی[[53]](#footnote-53) براساس توزیع خواص[[54]](#footnote-54) اسید­آمینه­ها در توالی تعریف می­شوند. خواص(ویژگی­های) اسید­آمینه که در این توصیف­کننده­ها استفاده می­شود، از دیتابیس AAindex استخراج می­شود.

دیتابیس­ از مسیر <http://www.genome.jp/dbget/aaindex.html/> در دسترس است. 8 شاخص

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

از دیتابیس AAindex استفاده شده­اند.

تمام شاخص­های اسید­آمینه­ها متمرکز[[55]](#footnote-55) و استاندارد­شده[[56]](#footnote-56) توسط رابطه زیر هستند.

که، میانگین خواص 20 اسید­آمینه و σ انحراف معیار 20 اسید­آمینه است، که وσخود توسط رابطه زیر قابل محاسبه هستند.

طول بردار ویژگی توصیف­کننده­های خود همبستگی *n \* lag* است، که *n* تعداد شاخص­های انتخاب­شده از پایگاه دادهAAindex است. که در این پایان­نامه *n=8* و *lag=4* و درنتیجه طول بردار ویژگی­ها 32 است. توصیف­کننده خودهمبستگی Moran توسط رابطه زیر بدست می­آید.

که، *d*، *lag* خود­همبستگی و *nlag* بیشترین مقدار برای *lag* است که در اینجا *nlag* برابر با 4 در نظر گرفته ایم.

*Pi* و *Pi+d* ویژگی­های اسید­آمینه­ها در موقعیت *i* و *i + d* است. میانگین ویژگی *P* نیز ازرابطه زیر بدست می­آید.

### ویژگی Geary correlation (Geary)

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

که، d، P، Pi ، Pi+d و nlag تعاریف مشابه ویژگیMoran دارند.

### ویژگی Normalized Moreau-Broto Autocorrelation (NMBroto)

توصیف­کننده­های خود­همبستگی Moreau-Broto با رابطه زیر تعریف می­شوند.

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

### 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[[57]](#footnote-57) از هم جدا می شوند ( حداکثر مقدار پیش­فرض *k* در این پایان­نامه 1 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.

### ویژگی Quasi-sequence-order (QSOrder)

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

که در آن *fr* وقوع نرمال شده اسید­آمینه نوع *r* و *w* یک عامل وزنی است (*w* =0.1). اینها اولین 20 توصیف­گر QSOrder هستند. 30 توصیف­گر دیگر QSOrder به صورت زیر تعریف می شوند:

## پیش‌پردازش داده‌[[58]](#footnote-58)

4 مرحله تحت عنوان پیش­پردازش بر روی داده­های هر 10 ویژگی استخراج شده اعمال شد.

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

## نرمالسازی داده‌[[59]](#footnote-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 (2-7) هر ویژگی‌ در مجموعه داده به بازه نگاشت شدند.

|  |  |
| --- | --- |
| رابطه (2-7) |  |

هر یک از ویژگی‌های مجموعه داده تست مستقل و آموزش با رابطه (2-7) به اعداد جدیدی نگاشت شدند.

## ‌ 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[[60]](#footnote-60): یک پپتید ضدسرطانی به درستی یک پپتید ضدسرطانی اعلام شده­است. تعداد این پیش­بینی­ها را با TP نشان می‌دهند.

مثبت­کاذب[[61]](#footnote-61): یک پپتید غیر­ضد­سرطانی به اشتباه یک پپتید ضدسرطانی اعلام شده­است. تعداد این پیش­بینی­ها را با FP نشان می‌دهند.

منفی­صحیح[[62]](#footnote-62): یک پپتید غیر­ضدسرطانی به درستی یک پپتید غیر­ضدسرطانی اعلام شده­است. تعداد این پیش­بینی­ها را با TN نشان می‌دهند.

منفی­کاذب[[63]](#footnote-63): یک پپتید ­ضد­سرطانی به اشتباه یک پپتید غیر­ضدسرطانی اعلام شده­است. تعداد این پیش­بینی­ها را با FN نشان می‌دهند.

در ادامه معیارهای ارزیابی بر­اساس این پارامتر­ها بیان می­شود.

### حساسیت[[64]](#footnote-64)

توانایی یک الگوریتم یادگیری ماشین برای تشخیص مقادیر مثبت­صحیح(TP) را حساسیت یا یاد­آوری[[65]](#footnote-65) گویند و از رابطه (2-8) 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 (2-8) به عدد یک نزدیک‌تر و حساسیت الگوریتم برای پیش‌بینی بالاتر است.

|  |  |
| --- | --- |
| رابطه (2-8) |  |
|  |  |

### اختصاصیت[[66]](#footnote-66)

توانایی یک الگوریتم یادگیری ماشین برای تشخیص مقادیر منفی­صحیح (TN) را اختصاصیت گویند و از رابطه (2-8) 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 (2-9) به عدد یک نزدیک‌تر اختصاصیت الگوریتم برای پیش‌بینی بالاتر است.

|  |  |
| --- | --- |
| رابطه (2-9) |  |

### صحت[[67]](#footnote-67)

یکی­دیگر از معیارهای ارزیابی صحت است و از رابطه (2-8) بدست می­آید. این معیار نشان می‌دهد چند درصد از داده‌ها به درستی طبقه‌بندی شده‌اند.

رابطه (2-10)

|  |  |
| --- | --- |
|  |  |

### دقت[[68]](#footnote-68)

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 (2-11) محاسبه می‌شود.

|  |  |
| --- | --- |
| رابطه (2-11) |  |

## اعتبارسنجی متقابل -لایه[[69]](#footnote-69)

اعتبارسنجی متقابل -لایه، یک روش برای ارزیابی مدل یادگیری ماشین است. در این روش مجموعه داده آموزش به بخش افراز می­شود. هر بار یک بخش به عنوان مجموعه داده آزمون در نظر­گرفته می­شود سپس مدل روی بخش‌ دیگر که به عنوان مجموعه آموزش انتخاب می­شوند اجرا می‌شود. دقت مدل روی بخشی که به عنوان مجموعه آزمون در نظر گرفته شده بود؛ محاسبه می‌شود. این روند 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 مرحله به عنوان دقت مدل گزارش می‌شود. برای ارزیابی مدل­های یادگیری در این پایان نامه مقدار  برابر 10 در نظر گرفته شده­است[45].

## انتخاب ویژگی[[70]](#footnote-70)

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[[71]](#footnote-71) و فیلتر[[72]](#footnote-72) تقسیم کرد.

### 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[[73]](#footnote-73) و انتخاب متوالی رو به عقب[[74]](#footnote-74) 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.

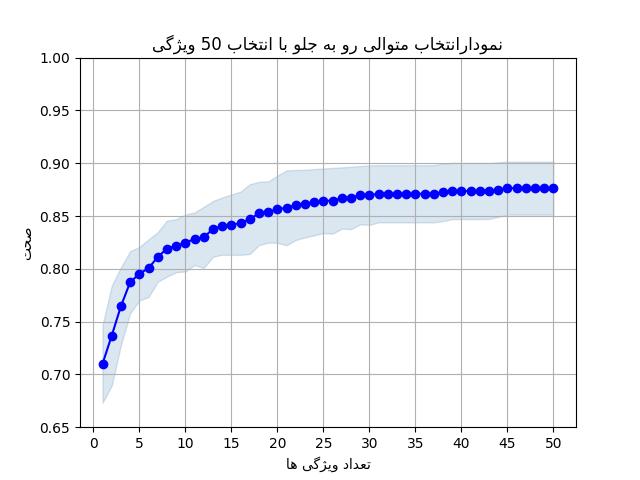
در این پایان­نامه از روش انتخاب متوالی رو به جلو توسط کتابخانه پایتون mlxtend استفاده شده­است[46].

انتخاب متوالی رو به جلو 3 بار توسط پارامتر­های زیر اجرا و 3 دسته ویژگی توسط الگوریتم تولید شد.

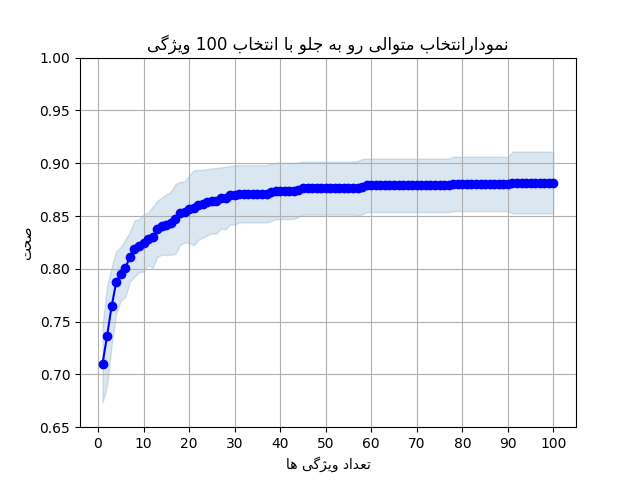
یکی از پارامتر­های روش بسته­بندی انتخاب یک مدل یادگیری ماشین برای انتخاب ویژگی­ها است. در این پایان­نامه از طبقه­بند Logestic Regression برای این امر استفاده شد. معیار ارزیابی طبقه­بند نیز صحت اعتبار­سنجی 10 لایه اعمال شد.

پارامتر بعدی در الگوریتم SFS[[75]](#footnote-75)، برای تعداد ویژگی­های انتخابی است، که دراین­جا 3 مقدار 50، 100 و 200 در نظر گرفته شد.

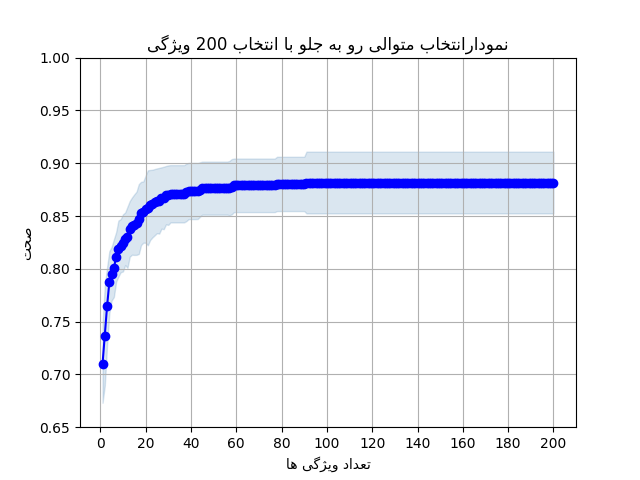
مراحل 3 دسته ویژگی تولید شده که شامل 50 ویژگی مهم، 100 ویژگی مهم و 200 ویژگی مهم هستند در تصاویر زیر مشاهده می­شود. از این پس به 50، 100 و 200 ویژگی انتخاب شده توسط الگوریتم SFS، با نام مجموعه داده SFS50، SFS100 و SFS200 اشاره خواهد­شد.



شکل ‏2–12: نمودار انتخاب متوالی رو به جلو تا 50 ویژگی



شکل ‏2–13: نمودار انتخاب متوالی رو به جلو تا 100 ویژگی



شکل ‏2–14: نمودار انتخاب متوالی رو به جلو تا 200 ویژگی

در ادامه بر روی مجموعه داده SFS50، SFS100 و SFS200 ، مدل­های یادگیری ماشین SVM و RF اجرا شد. نتایج این مدل­ها در فصل 3 آمده­است.

## کاهش ابعاد[[76]](#footnote-76) مجموعه داده

تعداد زیاد ویژگی‌ها باعث تحمیل هزینه محاسباتی بالا به طبقه‌بند­ها می‌شود و به مشقت تعداد ابعاد[[77]](#footnote-77) معروف است. برای کاهش ابعاد می‌توان از روش تحلیل مؤلفه‌های اصلی[[78]](#footnote-78) استفاده کرد.

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[47].

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 (10 ویژگی­) مدل­های یادگیری ماشین SVM و RF اجرا شد. نتایج این مدل­ها در فصل 3 آمده­است.

## مدل‌های یادگیری ماشین اجرا شده روی مجموعه داده‌ها

همانطور که در بخش 2-2 بیان شد، بر روی مجموعه داده cdhit90 پیش­پردازش اولیه انجام شد؛ سپس در بخش استخراج ویژگی (بخش 2-4) به 10 ویژگی اشاره شد.

این 10 ویژگی و همچنین ترکیبی از این 10 ویژگی که در ادامه اشاره خواهد شد، از مجموعه داده پیش­پردازش شده بخش 2-2، استخراج شدند. سپس 10مؤلفه 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 3 مجموعه داده SFS50، SFS100 و SFS200 (رجوع به بخش 2-9-1) 2 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 10 لایه انجام شد. در تمام مراحل از تعریف پارامتر­های تنظیمی، تنظیم پارامتر­های تنظیمی، آموزش و ارزیابی مدل­ها؛ از کتابخانه scikit-learn پایتون استفاده شد.

### طبقه‌بندی جنگل تصادفی

برای ساخت هر جنگل­تصادفی 4320، 4320 = 10 \* 2 \* 12 \* 3 \* 3 \* 2 پارامتر­تنظیمی، توسط قطعه کد زیر ایجاد شد.

الگوریتم انتخاب تصادفی پارامتر­ها، ازبین 4320 پارامترتنظیمی، یک جنگل تصادفی بصورت کاملا تصادفی میسازد. برای یافتن بهترین پارامتر­تنظیمی، 100 جنگل­تصادفیِ تصادفی[[79]](#footnote-79)، توسط پارامتر­های تصادفی ایجاد شده، ساخته شد.[48, 49]

قطعه کد زیر، فضای پارامتر­های تنظیمی برای ساخت 100 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!

### طبقه‌بندی ماشین بردار پیشتیبان

برای ساخت هر ماشین­بردار­پشتیبان 52، 52 = 13\*2\*2 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[[80]](#footnote-80) این است که مجموعه داده ورودی داده­شده را به فرم مورد نیاز تبدیل کند. انواع مختلفی از توابع مانند تابع پایه خطی[[81]](#footnote-81)، چند جمله­ای[[82]](#footnote-82) و شعاعی (RBF)[[83]](#footnote-83) 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 2 پارامتر هسته خطی و هسته شعاعی برای پارامتر­های هسته در نظر گرفته شد.

c\_range، پارامتر هزینه[[84]](#footnote-84) یا مجازات[[85]](#footnote-85)، نشان دهنده طبقه­بندی اشتباه[[86]](#footnote-86) یا عبارت خطا[[87]](#footnote-87) است. پارامتر c به بهینه­سازی SVM می­گوید که چقدر خطا قابل­تحمل است. به این ترتیب می­تواند trade-off بین مرز­تصمیم[[88]](#footnote-88) 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[50, 51].

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[[89]](#footnote-89) می­شود. به عبارت دیگر، می­توان 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، 52 مدل ماشین بردار­پشتیبان، به تعداد پارامتر­های تنظیمیِ تعریف شده، ساخته­شد.

# فصل سوم: نتایج و بحث

## نتایج طبقه‌بند­ها

برای کاهش ابعاد داده­های 10 ویژگی استخراج شده و پیش­پردازش شده، 2 روش در نظر گرفته­شد.

1. استفاده از PCA و استخراج 10 مؤلفه اول از داده­هاآن­ها
2. استفاده از الگوریتم انتخاب ویژگی SFS، که نتیجه آن 3 دسته مجموعه داده SFS50، SFS100 و SFS200 شد.

سپس هر مجموعه داده بدست آمده به این 2 روش، به نسبت 80-20 تقسیم شدند، طوریکه 80 درصد داده­ها برای آموزش و تنظیم پارامتر­های تنظیمی و 20 درصد داده­ها برای تست مستقل در نظر گرفته شدند. سپس 2 طبقه­بند جنگل­تصادفی و ماشین بردار­پشتیبان بر روی 80 درصد داده­های آموزش توسط روش اعتباری­سنجی متقابل 10 لایه، با پارامتر­های تنظیمی تعریف شده در بخش 2-11-1 و 2-11-2 تنظیم پارامتر[[90]](#footnote-90) 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 80 درصد داده آموزش دیدند، سپس توسط مجموعه تست مستقل عملکرد مدل­ها ارزیابی شد.

نتایج بهترین مدل­ها توسط اعتبار­سنجی متقابل 10 لایه در مرحله تنظیم پارامتر­های تنظیمی و نتایج مدل­ها توسط تست مستقل در ادامه گزارش می­شوند.

## نتایج بدست آمده بر روی داده­های 10 مؤلفه اول PCA

در ادامه نتایج 2 طبقه­بند جنگل­تصادفی و ماشین بردار­پشتیبان بر روی 3 سری مجموعه داده گزارش می­شود.

سری 1- داده­های 10 مؤلفه اول PCA که از تک­ویژگی­ها استخراج شد.

در این سری، بر روی 10 ویژگی استخراج شده، به تفکیک ویژگی­ها PCA و سپس RF و SVM اجرا شد.

سری 2- داده­های 10 مؤلفه اول PCA که از ترکیب دوتایی ویژگی­ها استخراج شد.

در این سری، ویژگی­هایی که در سری 1، در تست مستقل، صحت آن­ها بالاتر از 0.73 گزارش شده، انتخاب شدند؛ که شامل 7 دسته ویژگی شد. سپس تمام ترکیب­های 2­تاییِ این 7 دسته ویژگی ساخته شد و توسط PCA، 10­مؤلفه اول از آن­ها استخراج شد. سپس بر روی این 10­مؤلفه همانند روش قبل، 2 مدل یادگیری ماشین RF و SVM اجرا شد.

سری 3- داده­های 10 مؤلفه اول PCA که از ترکیب **تمام** ویژگی­ها استخراج شد.

در این سری نیز، همانند سری 2، ویژگی­هایی که در سری 1، در تست مستقل، صحت آن­ها بالاتر از 0.73 گزارش شده بود، انتخاب شدند؛ که شامل 7 دسته ویژگی شد. سپس تمام این 7 ویژگی­ها باهم ترکیب شدند و توسط PCA، 10­مؤلفه اول از آن­ها استخراج شد. سپس بر روی این 10­مؤلفه همانند روش قبل، 2 مدل یادگیری ماشین RF و SVM اجرا شد.

در ادامه نتایج بدست آمده از 2 طبقه­بند جنگل تصادفی و ماشین بردار­پشتیبان بر روی این 3 سری داده گزارش می­شود.

### نتایج مربوط به طبقه­بند جنگل­تصادفی

#### نتایج مربوط به مجموعه داده­های سری 1

نتایج بهترین مدل جنگل­تصادفی بر روی داده­های سری1، داده­های 10 مؤلفه اول PCA که از تک­ویژگی­ها استخراج شد، در مرحله تنظیم پارامتر­های تنظیمی و تست مستقل در جدول زیر گزارش می­شود.

جدول ‏3–1: 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** | 0.79 | 0.87 | 0.79 | 0.79 | 0.79 | 0.79 |
| **CKSAAP** | 0.80 | 0.89 | 0.79 | 0.78 | 0.78 | 0.78 |
| **CTD** | 0.76 | 0.84 | 0.74 | 0.74 | 0.74 | 0.74 |
| **DDE** | 0.77 | 0.85 | 0.74 | 0.74 | 0.74 | 0.74 |
| **Geary** | 0.71 | 0.77 | 0.71 | 0.71 | 0.72 | 0.71 |
| **KSCTriad** | 0.71 | 0.84 | 0.77 | 0.76 | 0.76 | 0.76 |
| **Moran** | 0.70 | 0.75 | 0.66 | 0.67 | 0.67 | 0.66 |
| **NMBroto** | 0.71 | 0.78 | 0.66 | 0.66 | 0.66 | 0.66 |
| **PseudoAAC** | 0.79 | 0.87 | 0.79 | 0.78 | 0.78 | 0.78 |
| **QSOrder** | 0.79 | 0.86 | 0.81 | 0.81 | 0.81 | 0.81 |

#### نتایج مربوط به مجموعه داده­های سری 2

با توجه به جدول 3-1، ویژگی­هایی که صحت تست مستقل آن­ها بالاتر از 0.73 گزارش شده، انتخاب شدند؛ که شامل 7 ویژگی APseudoAAC، CKSAAP، CTD، DDE، KSCTriad، PseudoAAC و QSOrder شد. سپس تمام ترکیب­های 2­تاییِ این 7 دسته ویژگی ساخته شد و توسط PCA، 10­مؤلفه 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 ‏3–2: 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** | 0.82 | 0.89 | 0.80 | 0.80 | 0.80 | 0.80 |
| **QSOrder+CTD** | 0.76 | 0.84 | 0.76 | 0.76 | 0.75 | 0.75 |
| **QSOrder+DDE** | 0.83 | 0.89 | 0.80 | 0.79 | 0.80 | 0.79 |
| **QSOrder+KSCTriad** | 0.79 | 0.86 | 0.77 | 0.77 | 0.76 | 0.77 |
| **QSOrder+PseudoAAC** | 0.81 | 0.88 | 0.81 | 0.81 | 0.81 | 0.81 |
| **QSOrder+APseudoAAC** | 0.81 | 0.88 | 0.81 | 0.81 | 0.81 | 0.81 |
| **PseudoAAC+APseudoAAC** | 0.79 | 0.87 | 0.79 | 0.79 | 0.79 | 0.79 |
| **PseudoAAC+CKSAAP** | 0.82 | 0.90 | 0.78 | 0.77 | 0.78 | 0.77 |
| **PseudoAAC+CTD** | 0.75 | 0.84 | 0.76 | 0.75 | 0.75 | 0.75 |
| **PseudoAAC+DDE** | 0.81 | 0.89 | 0.80 | 0.80 | 0.80 | 0.80 |
| **PseudoAAC+KSCTriad** | 0.78 | 0.86 | 0.76 | 0.75 | 0.75 | 0.75 |
| **CKSAAP+APseudoAAC** | 0.81 | 0.90 | 0.78 | 0.78 | 0.78 | 0.78 |
| **CKSAAP+CTD** | 0.78 | 0.86 | 0.77 | 0.77 | 0.76 | 0.77 |
| **CKSAAP+DDE** | 0.80 | 0.89 | 0.75 | 0.75 | 0.74 | 0.74 |
| **CKSAAP+KSCTriad** | 0.80 | 0.88 | 0.78 | 0.77 | 0.77 | 0.77 |
| **CTD+APseudoAAC** | 0.75 | 0.84 | 0.77 | 0.77 | 0.77 | 0.77 |
| **CTD+DDE** | 0.76 | 0.85 | 0.78 | 0.78 | 0.78 | 0.78 |
| **CTD+KSCTriad** | 0.78 | 0.86 | 0.80 | 0.80 | 0.79 | 0.80 |
| **DDE+APseudoAAC** | 0.81 | 0.89 | 0.80 | 0.80 | 0.80 | 0.80 |
| **DDE+KSCTriad** | 0.77 | 0.86 | 0.75 | 0.75 | 0.74 | 0.74 |
| **APseudoAAC+KSCTriad** | 0.79 | 0.86 | 0.77 | 0.76 | 0.76 | 0.76 |

#### نتایج مربوط به ترکیب تمام ویژگی­ها

با توجه به جدول 3-1، **تمام** 7 ویژگی­هایی که صحت تست مستقل آن­ها بالاتر از 0.73 گزارش شده، باهم ترکیب شدند؛ سپس توسط PCA، 10­مؤلفه 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 ‏3–3: نتایج بهترین مدل­ جنگل­تصادفی در ترکیب تمام 7 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 7 ویژگی­ها** | 0.80 | 0.89 | 0.82 | 0.82 | 0.82 | 0.82 |

### نتایج مربوط به طبقه‌بند ماشین بردار­پشتیبان

#### نتایج مربوط به مجموعه داده­های سری 1

نتایج بهترین مدل­های ماشین بردار­پشتیبان بر روی داده­های سری1، داده­های 10 مؤلفه اول PCA که از تک­ویژگی­ها استخراج شد، در مرحله تنظیم پارامتر­های تنظیمی و تست مستقل در جدول زیر گزارش می­شود.

جدول ‏3–4: 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** | 0.79 | 0.77 | 0.77 | 0.76 | 0.76 |
| **CKSAAP** | 0.79 | 0.78 | 0.77 | 0.77 | 0.77 |
| **CTD** | 0.77 | 0.78 | 0.77 | 0.77 | 0.77 |
| **DDE** | 0.77 | 0.73 | 0.73 | 0.73 | 0.73 |
| **Geary** | 0.69 | 0.65 | 0.65 | 0.65 | 0.65 |
| **KSCTriad** | 0.77 | 0.73 | 0.72 | 0.73 | 0.72 |
| **Moran** | 0.69 | 0.64 | 0.63 | 0.63 | 0.63 |
| **NMBroto** | 0.72 | 0.75 | 0.75 | 0.75 | 0.75 |
| **PseudoAAC** | 0.80 | 0.77 | 0.76 | 0.76 | 0.76 |
| **QSOrder** | 0.79 | 0.80 | 0.80 | 0.80 | 0.80 |

#### 

#### نتایج مربوط به مجموعه داده­های سری 2

با توجه به جدول 3-4، ویژگی­هایی که صحت تست مستقل آن­ها بالاتر از 0.73 گزارش شده، انتخاب شدند؛ که شامل 8 ویژگی APseudoAAC، CKSAAP، CTD، DDE، KSCTriad، NMBroto ، PseudoAAC و QSOrder شد. سپس تمام ترکیب­های 2­تاییِ این 8 دسته ویژگی ساخته شد و توسط PCA، 10­مؤلفه 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 ‏3–5: 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** | 0.82 | 0.78 | 0.78 | 0.78 | 0.78 |
| **QSOrder+CTD** | 0.77 | 0.77 | 0.77 | 0.76 | 0.77 |
| **QSOrder+DDE** | 0.81 | 0.79 | 0.79 | 0.79 | 0.79 |
| **QSOrder+KSCTriad** | 0.78 | 0.71 | 0.70 | 0.71 | 0.70 |
| **QSOrder+PseudoAAC** | 0.81 | 0.79 | 0.79 | 0.79 | 0.79 |
| **QSOrder+APseudoAAC** | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 |
| **PseudoAAC+APseudoAAC** | 0.79 | 0.77 | 0.77 | 0.76 | 0.76 |
| **PseudoAAC+CKSAAP** | 0.81 | 0.80 | 0.80 | 0.80 | 0.80 |
| **PseudoAAC+CTD** | 0.76 | 0.76 | 0.75 | 0.75 | 0.75 |
| **PseudoAAC+DDE** | 0.80 | 0.73 | 0.73 | 0.74 | 0.73 |
| **PseudoAAC+KSCTriad** | 0.78 | 0.71 | 0.71 | 0.71 | 0.71 |
| **CKSAAP+APseudoAAC** | 0.81 | 0.80 | 0.79 | 0.79 | 0.79 |
| **CKSAAP+CTD** | 0.78 | 0.75 | 0.75 | 0.74 | 0.74 |
| **CKSAAP+DDE** | 0.80 | 0.78 | 0.78 | 0.76 | 0.77 |
| **CKSAAP+KSCTriad** | 0.79 | 0.75 | 0.75 | 0.75 | 0.75 |
| **CTD+APseudoAAC** | 0.77 | 0.76 | 0.75 | 0.75 | 0.75 |
| **CTD+DDE** | 0.76 | 0.78 | 0.77 | 0.77 | 0.77 |
| **CTD+KSCTriad** | 0.77 | 0.72 | 0.72 | 0.72 | 0.72 |
| **DDE+APseudoAAC** | 0.81 | 0.75 | 0.75 | 0.74 | 0.74 |
| **DDE+KSCTriad** | 0.78 | 0.74 | 0.73 | 0.74 | 0.74 |
| **APseudoAAC+KSCTriad** | 0.78 | 0.74 | 0.73 | 0.74 | 0.73 |
| **NMBroto+APseudoAAC** | 0.82 | 0.78 | 0.77 | 0.77 | 0.77 |
| **NMBroto+CKSAAP** | 0.80 | 0.79 | 0.79 | 0.78 | 0.78 |
| **NMBroto+CTD** | 0.76 | 0.75 | 0.75 | 0.74 | 0.74 |
| **NMBroto+DDE** | 0.80 | 0.77 | 0.76 | 0.75 | 0.76 |
| **NMBroto+KSCTriad** | 0.78 | 0.73 | 0.73 | 0.73 | 0.73 |
| **NMBroto+PseudoAAC** | 0.82 | 0.79 | 0.78 | 0.78 | 0.78 |
| **NMBroto+QSOrder** | 0.80 | 0.79 | 0.79 | 0.79 | 0.79 |

#### نتایج مربوط به ترکیب تمام ویژگی­ها

با توجه به جدول 3-4، **تمام** 8 ویژگی­هایی که صحت تست مستقل آن­ها بالاتر از 0.73 گزارش شده، باهم ترکیب شدند؛ سپس توسط PCA، 10­مؤلفه 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.

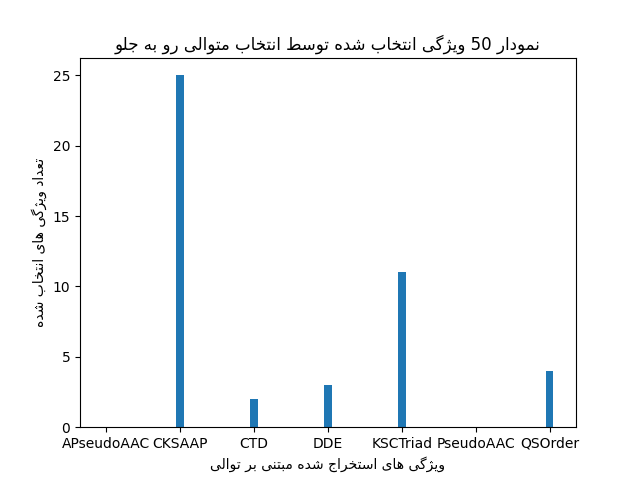
جدول ‏3–6: نتایج بهترین مدل ماشین بردار­پشتیبان در ترکیب تمام 8 ویژگی­ها

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ویژگی** | **صحت**  **(بهترین مدل در تنظیم پارامتر تنظیمی)** | **صحت**  **(تست مستقل)** | **دقت**  **(تست مستقل)** | **حساسیت**  **(تست مستقل)** | **معیار f1**  **(تست مستقل)** |
| **ترکیب همه 8 ویژگی­ها** | 0.79 | 0.79 | 0.79 | 0.79 | 0.79 |

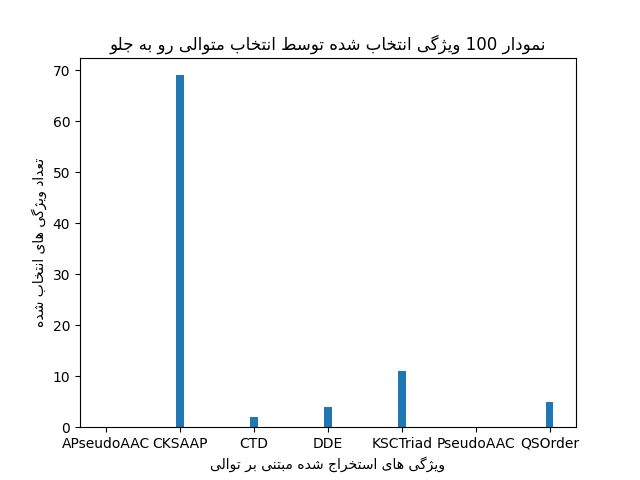
## نتایج ویژگی­های انتخاب شده توسط الگوریتم انتخاب متوالی رو به جلو

با کنار هم قرار­دادن تمام 7 ویژگی مشخص شده در جدول 3-1 یک فضای ویژگی به ابعاد (3014، 940) ایجاد شد. (3014 طول بردار ویژگی و 940 تعداد نمونه پپتیدها)

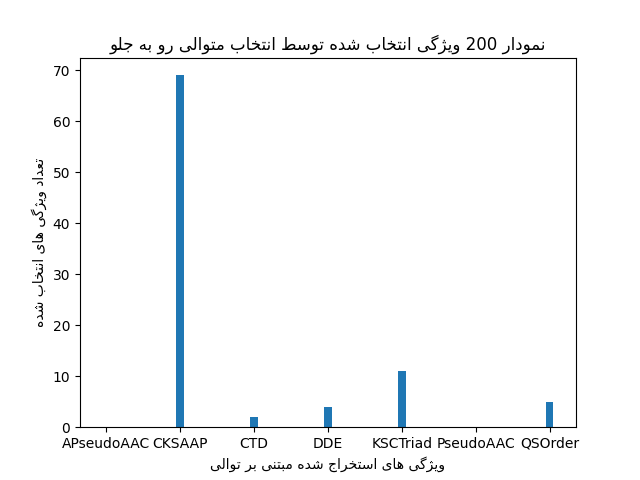
فضای ویژگی جدید به الگوریتم SFS داده شد و 3 دسته ویژگی SFS50، SFS100 و SFS200 ساخته شد. برای درک این موضوع که چه تعداد از 7 ویژگی در ویژگی­های SFS50، SFS100 و SFS200 مشارکت داشته­اند، نمودار­های زیر رسم شد. با توجه به نتایج هر 3 دسته نمودار، بیشترین ویژگی­های انتخاب شده، از2 دسته ویژگی CKSAAP­ و KSCTriad است.



شکل ‏3–1: نمودار 50 ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو



شکل ‏3–2: نمودار 100 ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو



شکل ‏3–3: نمودار 200 ویژگی انتخاب شده توسط انتخاب متوالی رو به جلو

## نتایج بدست آمده بر روی 3 مجموعه داده SFS50، SFS100 و SFS200

در ادامه نتایج طبقه­بند جنگل­تصادفی بر روی 3 مجموعه داده SFS50، SFS100 و SFS200 گزارش می­شود.

### 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 ‏3–7: 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** | 0.82 | 0.89 | 0.73 | 0.73 | 0.73 | 0.73 |
| **SFS100** | 0.82 | 0.89 | 0.74 | 0.74 | 0.74 | 0.74 |
| **SFS200** | 0.82 | 0.89 | 0.74 | 0.74 | 0.74 | 0.74 |

# فصل چهارم: بحث و پیشنهادات

با توجه به نتایج طبقه­بندها در فصل 3، مشخص است که استفاده از روش­های محاسباتی برای مسئله پیش­بینی پپتید­های ضد­سرطانی روش مناسبی است، زیرا نتایج بدست آمده توسط هر 2 طبقه­بند جنگل­تصادفی و ماشین بردار­پشتیبان با ترکیب ویژگی­های مختلف و حتی استفاده از یک ویژگی به تنهایی همگی درصد صحت بالاتر از 71 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 79 تا 81 درصد را در تست مستقل دارند، این در حالی است که برای طبقه­بند ماشین بردار­پشتیبان فقط برای ویژگی QSOrder نتایج تست مستقل بین 79 تا 81 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 81 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+APseudoAACگزارش شد.

این در حالی است که در مرحله آموزش و تنظیم پارامتر­های تنظیمی، صحت بالای 83 درصد در ترکیب ویژگی­ QSOrder+DDE توسط مدل جنگل­تصادفی و صحت بالای 82 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 10 مؤلفه اول بسیار کمتر است طوری­که با توجه به جدول 3-7 صحت تست مستقل به 73 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

1. 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.

2. 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.

3. 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.

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

5. 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.

6. 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.

7. 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.

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

9. 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.

10. 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.

11. 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.

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

13. 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.

14. Wang, Z. and G. Wang, *APD: the antimicrobial peptide database.* Nucleic acids research, 2004. **32**(suppl\_1): p. D590-D592.

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

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

17. 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.

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

19. 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.

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

21. 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.

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

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

24. 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.

25. 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.

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

27. 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.

28. 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.

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

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

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

32. 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.

33. 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.

34. 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.

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

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

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

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

39. 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.

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

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

42. 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.

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

44. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008.* Nucleic acids research, 2007. **36**(suppl\_1): p. D202-D205.

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

46. 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.

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

48. 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/>.

49. 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>.

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

51. 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

1. Anticancer Peptides (ACP) [↑](#footnote-ref-1)
2. International Agency for Research on Cancer [↑](#footnote-ref-2)
3. Antimicrobial Peptides (AMP) [↑](#footnote-ref-3)
4. NMR spectroscopy [↑](#footnote-ref-4)
5. Nissen [↑](#footnote-ref-5)
6. Lactococcus Lactis [↑](#footnote-ref-6)
7. Supervised [↑](#footnote-ref-7)
8. Unsupervised [↑](#footnote-ref-8)
9. Class Label [↑](#footnote-ref-9)
10. Descriptor [↑](#footnote-ref-10)
11. Feature [↑](#footnote-ref-11)
12. Predictors [↑](#footnote-ref-12)
13. Clustering [↑](#footnote-ref-13)
14. Classification [↑](#footnote-ref-14)
15. Decision Tree Classifier [↑](#footnote-ref-15)
16. Root Node [↑](#footnote-ref-16)
17. Leaves [↑](#footnote-ref-17)
18. Random Forest Classifier [↑](#footnote-ref-18)
19. Ensemble Learning [↑](#footnote-ref-19)
20. Support Vector Machine [↑](#footnote-ref-20)
21. Margin [↑](#footnote-ref-21)
22. Kernel [↑](#footnote-ref-22)
23. Deep Neural Networks [↑](#footnote-ref-23)
24. Anticancer Peptides [↑](#footnote-ref-24)
25. Databases [↑](#footnote-ref-25)
26. Random Forest [↑](#footnote-ref-26)
27. Support Vector Machine [↑](#footnote-ref-27)
28. Non Ancticancer Peptide [↑](#footnote-ref-28)
29. Antimicrobal Peptides [↑](#footnote-ref-29)
30. Datasets [↑](#footnote-ref-30)
31. Positive Dataset [↑](#footnote-ref-31)
32. Negative Dataset [↑](#footnote-ref-32)
33. Similarity [↑](#footnote-ref-33)
34. Training Dataset [↑](#footnote-ref-34)
35. Test Set [↑](#footnote-ref-35)
36. Hyper Parameter Tuning [↑](#footnote-ref-36)
37. Independent Test Set [↑](#footnote-ref-37)
38. Feature Extraction [↑](#footnote-ref-38)
39. Sequence Order-Correlated Factors [↑](#footnote-ref-39)
40. Gap [↑](#footnote-ref-40)
41. Polarity [↑](#footnote-ref-41)
42. Polarizability [↑](#footnote-ref-42)
43. Charge [↑](#footnote-ref-43)
44. Normalized VanderWaals Volume [↑](#footnote-ref-44)
45. Secondary Structures [↑](#footnote-ref-45)
46. Solvent Accessibility [↑](#footnote-ref-46)
47. Dipeptide Composition [↑](#footnote-ref-47)
48. Theoretical Mean [↑](#footnote-ref-48)
49. Theoretical Variance [↑](#footnote-ref-49)
50. Codon [↑](#footnote-ref-50)
51. Stop Codons [↑](#footnote-ref-51)
52. Descriptors [↑](#footnote-ref-52)
53. Autocorrelation [↑](#footnote-ref-53)
54. Properties [↑](#footnote-ref-54)
55. Centralized [↑](#footnote-ref-55)
56. Standardized [↑](#footnote-ref-56)
57. Residues [↑](#footnote-ref-57)
58. Data preprocessing [↑](#footnote-ref-58)
59. Normalization [↑](#footnote-ref-59)
60. True positive [↑](#footnote-ref-60)
61. False positive [↑](#footnote-ref-61)
62. True negative [↑](#footnote-ref-62)
63. False negative [↑](#footnote-ref-63)
64. Sensitivity [↑](#footnote-ref-64)
65. Recall [↑](#footnote-ref-65)
66. Specificity [↑](#footnote-ref-66)
67. Accuracy [↑](#footnote-ref-67)
68. Precision [↑](#footnote-ref-68)
69. *K*-fold Cross Validation [↑](#footnote-ref-69)
70. Feature selection [↑](#footnote-ref-70)
71. Wrapper [↑](#footnote-ref-71)
72. Filter [↑](#footnote-ref-72)
73. Sequential Forward Selection [↑](#footnote-ref-73)
74. Sequential Backward Selection [↑](#footnote-ref-74)
75. Sequential Forward Selection [↑](#footnote-ref-75)
76. Dimension Reduction [↑](#footnote-ref-76)
77. Curse of Dimensionality [↑](#footnote-ref-77)
78. Principal Component Analysis [↑](#footnote-ref-78)
79. Random Random Forest [↑](#footnote-ref-79)
80. Kernel [↑](#footnote-ref-80)
81. Linear Basis Function [↑](#footnote-ref-81)
82. Polynomial Basis Function [↑](#footnote-ref-82)
83. Radial Basis Function (rbf) [↑](#footnote-ref-83)
84. Cost [↑](#footnote-ref-84)
85. Penalty [↑](#footnote-ref-85)
86. Missclassification [↑](#footnote-ref-86)
87. Error Term [↑](#footnote-ref-87)
88. Decision Boundary [↑](#footnote-ref-88)
89. Over-Fitting [↑](#footnote-ref-89)
90. Parameter Tuning [↑](#footnote-ref-90)