
بسم الله الرحمن الرحيم

دانشکده : شیمی

گروه : شیمی تجزیه

پایان نامه کارشناسی ارشد

کاربرد روشهای مختلف کمومتریکس جهت پیش بینی فعالیت برخی ترکیبات دارویی

سعید نکوئی

استاد راهنما :

دکتر ناصر گودرزی

استاد مشاور:

دکتر مهدی نکوئی

شهریور ماه ۱۳۹۴

تقدیم به:

خانوادۀ عزیزم که درکمال شکیبایی همواره پشتیبان من بودند.

سپاس خدای را که هر چه دارم از اوست

به امید آنکه توفیق یابم جز خدمت به خلق

خدا نکوشم

از استاد گرامیم آقای دکتر ناصر گودرزی و استاد مشاورم

دکتر مهدی نکوئی سپاسگزارم که زحمت راهنمایی این رساله

را بر عهده گرفتند.

تعهد نامه

اینجانب سعید نکوئی دانشجوی دوره کارشناسی ارشد رشته شیمی تجزیه دانشکده شیمی دانشگاه صنعتی شاهرود نویسنده پایان نامه کاربرد روشهای مختلف کمومتریکس جهت پیش بینی فعالیت برخی ترکیبات دارویی تحت راهنمائی آقای دکتر ناصر گودرزی متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

چکیده

در بخش اول، مطالعه‌ی کمی ساختار-فعالیت (QASR) بر روی اثر بازداری (pIC_{50}) ۱۰۳ ترکیب دارویی از مشتقات هموپیپرازین، دی آمین و ۳-آمینو پیرولیدین انجام شد. برای انتخاب مهمترین توصیفگرها از دو روش رگرسیون مرحله‌ای و الگوریتم ژنتیک استفاده شد. پس از انتخاب توصیفگرها توسط این دو روش از سه روش رگرسیون خطی چندگانه (MLR)، شبکه عصبی مصنوعی (ANN) و ماشین بردار پشتیبان (SVM) جهت مدل سازی استفاده گردید. عملکرد هر مدل توسط سری تست مورد بررسی قرار گرفت. میانگین مربعات خطا (MSE) و خطای استاندارد پیش بینی (SEP) سری تست برای روش‌های SR-MLR ، SR-ANN،SR-SVM، GA-MLR، GA-ANN و GA-SVM به ترتیب برابر۰/۲۸۳، ۰/۵۳۲، ۰/۲۲۵، ۰/۴۷۵، ۰/۲۲۸، ۰/۴۷۷، ۰/۲۶۳، ۰/۵۱۳، ۰/۳۴۶، ۰/۵۸۸، ۰/۳۲۸ و ۰/۵۷۳به دست آمد.

در بخش دوم، مطالعه‌ی کمی ساختار-فعالیت (QASR) بر روی اثر بازداری (pIC_{50})۴۲ ترکیب دارویی از مشتقات آریل سولفونیل پیپرازین‌انجام شد. و برای انتخاب مهمترین توصیفگرها ازالگوریتم ژنتیک استفاده شد. برای برای مدل سازی و پیش بینی اثر بازداری (pIC_{50}) از روش‌های رگرسیون خطی چندگانه (MLR)، شبکه عصبی مصنوعی (ANN) و ماشین بردار پشتیبان (SVM) استفاد شد. عملکرد هر مدل توسط سری تست مورد بررسی قرار گرفت. که میانگین مربعات خطا (MSE) و خطای استاندارد پیش بینی (SEP) و ضریب تعیین (R^2) بسیار به یکدیگر نزدیک بوده و این بیانگر این مطلب است که هر سه مدل به کار برده شده توانایی خوبی در پیش بینی pIC_{50} ترکیبات مورد مطالعه دارند.

کلمات کلیدی: اثر بازداری، رگرسیون مرحله‌ای، الگوریتم ژنتیک، رگرسیون خطی چندگانه، شبکه عصبی مصنوعی، ماشین بردار پشتیبان

فهرست مطالب

فصل اول: مقدمه

فصل دوم: کمومتریکس

فصل سوم: مطالعه ارتباط کمی ساختار – فعالیت مشتقات هموپیپرازین، دی آمین، ۳-آمینوپیرولیدین با استفاده روش های خطی و غیر خطی

فصل چهارم: مطالعه ارتباط کمی ساختار – فعالیت مشتقات آریل سولفونیل پیپرازین با استفاده روش های خطی و غیر خطی

فهرست اشکال

فهرست جداول

فصل اول

مقدمه

۱-۱- مقدمه

یکی از مشکلاتی که جامعه بشری همیشه با آن روبرو بوده، مقابله با انواع بیماری هایی است که سلامت انسان ها را به مخاطره انداخته و همواره یکی از مهمترین دغدغه های محققان یافتن داروهای موثر، برای رفع این معضل و یا کاهش عوارض این بیماری ها بوده است. بروز انواع بیماری ها از قبیل سرطان، ایدز، بیماری های مشترک انسان و دام و مقاوم شدن ویروسها در برابر آنتی بیوتیک ها همه از جمله مواردی هستند که ذهن دانشمندان را در جهت یافتن داروهای موثر و کارآمدتر برای مقابله با این بیماری ها معطوف خود نموده اند[۱]. روندی که در گذشته منجر به کشف و توسعه داروهای جدید می شد به روش آزمون و خطا صورت می گرفت که روشی وقت گیر و هزینه بر است. مشکل دیگری که در این راه دانشمندان را آزار می دهد، عدم اطلاع آنها از فعالیت داروئی ترکیبات، قبل از انجام سنتز و بررسی تجربی آنها بوده و به همین دلیل یکی از مهم ترین اهداف شیمیدان ها و محققان دارویی پیش بینی فعالیت ترکیبات دارویی، قبل از سنتز آن ها می باشد. چرا که انجام بسیاری از آزمایش‌ها مستلزم صرف زمان و هزینه های زیادی است. از این رو نیاز به استفاده از روش های تئوری و محاسباتی که بدون انجام آزمایش بتوانند فعالیت ترکیبات دارویی را پیش بینی کنند، ضروری به نظر می رسد. ظهور علم کمومتریکس[۱] توانسته راه حلی برای رفع این مشکلات باشد[۳-۲].

۱-۲- کموکین ها

کموکین ها[۲] خانواده‌ای از سیتوکین‌های[۳] با اندازه کوچک یا پروتئینهای تولید شده توسط سلولها هستند. نام این خانواده ازخاصیت آن‌ها برای جذب سلول‌های پاسخ دهنده الهام گرفته شده است.پروتئین‌هایی که به عنوان کموکین طبقه‌بندی می‌شوند دارای ویژگی‌های ساختاری مشابهی همچون اندازه کوچک و حضور چهارعدد مولکول سیستئین در جایگاه‌های کلیدی ساختاری هستند.

۱- Chemometrics
۲- Chemokines
۳-Cytokines

برخی ازکموکین‌ها دارای نقش پیش-التهاب هستند و در هنگام التهاب، سلول‌های ایمنی را به مکـان عفونت فرا می‌خوانند[۴].

کموکین ها از طریق یـک پـروتئین غشـایی بنـام جـی پـروتئین اثـر میکننـد. جـی‌پـروتئین‌هـا گیرنده‌هایی پروتئینی هستند که یک سمت آنها خارج از سلول و سمت دیگرشان داخـل سـلول قـرار دارد (عبورکننده از عرض غشا). این سیستم نوعی پیامرسان برای مواد شیمیایی خاصی مثل هورمـون ها و پیامرسان عصبیدر سطح غشای سلول است[۵].

CCR_2b[۱] عضوی از خانواده گیرنده های بتا کموکین یا گروه CC است که عضو ثابت غشای سـلول ها هستند. عوامل التهاب زا یا بیمـاری هـای التهـابی ماننـد ورم مفاصـل، التهـاب گلـومرول، رسـوب کلسترول در جدار عروق و بیمـاری MS اغلـب بـا اسـتفاده از CCR_2b وارد سـلول هـدف مـی شـود. بازدارنده های[۲] مورد مطالعه در این تحقیق، برای دخالت در برهمکنش میان CCR_2b و ویروس هـای التهاب زا طراحی شده اند. از جمله این دسته از داروها مشـتقات همـوپیپرازین،دی آمـین و ۳-آمینـو پیرولیدین هستند که به عنوان آنتاگونیست CCR_2b عمل می نماینـد و از تحریـک ایـن گیرنـده هـا توسط ویروس ممانعت به عمل می آورند[۷-۶].

در این تحقیق اثر بازدارندگی این مشتقات مورد بررسی قرار گرفته و سعی شده است تا به کمک مطالعات QSAR مدلی با توانایی پیش بینی توان بازدارندگی ترکیبات یاد شده ارائه گردد تا بتـوان از این مدل در طراحی و ساخت ترکیبات دارویی جدید از این دسته استفاده نمود.

۱-۳-پارامترهای معرفی خواص دارویی ترکیبات

جهت بیان هر خواص دارویی ترکیبات نیـاز بـه اسـتانداردهایی اسـت کـه دارای تعریـف کـاملا مشخص و شناخته شده باشند تا به وسیله آن استانداردها بتوان پارامترهای مورد نظر را مورد مقایسه

۱-Chemotactic cytokinereceptor 2b
۲- Inhibitors

قرار داد. داروها نیز از این قاعده مستثنی نیستند و جهت مقایسه خواص دارویی آنها در برابر بیماری های مختلف از پارامترهای ویژه ای استفاده می گردد.عمل داروها در بدن انسان را فارماکودینامیک[۱] و آنچه را که بدن با دارو انجام می دهد،فارماکوکینتیک[۲] می گویند. داروهایی که وارد بدن انسان می شوند تمایل دارند تا یک گیرنده خاص یا عمل روی یک آنزیم یا پروتئین واسط را تحت تاثیر قرار دهند. در نتیجه آنها موجب می شوند که بدن انسان به طریق خاصی واکنش نشان دهد که این واکنش را پاسخ می نامند.

داروها به طور کلی به دو دسته تقسیم می شوند:

۱- آگونیست ها[۳]: مواد شیمیایی که به گیرنده یک سلول متصل شده و آن را تحریک و فعال می کنند.

۲- آنتاگونیست ها[۴]: که از تحریک گیرنده ها بوسیله آگونیست ها جلوگیری می کنند.

ابتدا گیرنده ها فعال می شوند و سپس گیرنده‌ها یا به وسیله یک پاسخ ویژه مستقیما بر روی بدن اثر می گذارند یا موجب آزادسازی هورمون ها یا داروهای درونی بدن شده و از این طریق پاسخ ویژه را ایجادمی نمایند[۸].

۱-۴-توان دارویی[۵]

در مقوله داروشناسی توان(قدرت یا دوز) دارویی میزان اثربخشی دارو بوده و مقدار دارویی است که در یک زمان مشخص استفاده می‌شود. دوز مناسب یک دارو، مقدار دارویی است که می‌تواند بهترین اثر را بدون کمترین عوارض جانبی در یک مدت زمان مشخص ایجاد کند. یک داروی بسیار قوی (مانند:مورفین،آلپرازولام و کلروپرومازین) در غلظت های خیلی کم پاسخ بزرگتری ایجاد نموده

۱- Pharmacodynamics
۲-PharmacoKinetics
۳-Agonist
۴- Antagonist
۵- Drug dose

در حالیکه داروهای با توان کمتر(مانند:ایبوپروفن و استیل سالیسیلیک اسید) در غلظت های پایین پاسخ کوچکتری تولید می کنند. پاسخ تولید شده متناسب با میل ترکیبی و درجه تاثیر دارو است.

میل ترکیبی،توانایی دارو در اتصال به گیرنده است و درجه تاثیر، ارتباط بین اشغال کننده‌ی گیرنده و توانایی آن در شروع پاسخ در سطح مولکولی، سلولی، بافت یا دستگاه می باشد. پاسخ، همان اثر یا (E) است و به میزان اتصال دارو و گیرنده ای که دارو به آن متصل شده وابسته است. با این توضیحات، قدرت دارو به میل ترکیبی و درجه تاثیر دارو وابسته می باشد. یک آگونیست، لیگاند، دارو یا هورمونی که به گیرنده متصل می شود و ایجاد پاسخ می کند را معمولا بطور مختصر با A یا D نشان می دهند.

پایین تر از یک غلظت معین از آگونیست([A])، پاسخ یا اثر (E) برای اندازگیری خیلی کوچکتر است اما در غلظت های بالاتر قابل ارزیابی می شود و با افزایش غلظت آگونیست([A])، زیاد شده تا به غلظتی برسیم که دیگر با افزایش [A]، E نتواند افزایش یابد و آن را با E_{max} نشان می دهند. E_{max} حداکثر اثر احتمالی آگونیست است. غلظتی از Aکه در آن E برابر ۵۰٪ E_{max} است را "نصف حداکثر غلظت موثر" گویند و به صورت $[A]_{50}$ یا بطور رایج با EC_{50}[۱] نشان می دهند[۸].

۱-۵-بازدارندگی

غلطت بازدارندگی ۵۰٪[۲] (IC_{50}) نماینده غلظتی از دارو است که دارای ۵۰٪ اثر بازدارندگی می باشد. به عبارت دیگرIC_{50} کمیتی است که نشان می دهد چه مقدار از یک داروی خاص یا یک ماده دیگر (بازدارنده) نیاز است تا یک فرایند ویژه بیولوژیکی را به نصف کاهش دهد. به عبارت دیگر نصف حداکثر(۵۰٪) غلظت بازدارندگی (IC) یک ماده، IC ۵۰٪ یا IC_{50} نامیده می‌شود. از این کمیت عموما به عنوان معیاری از توان داروهای آنتاگونیست در تحقیقات دارو شناسی استفاده می شود. بعضی

۱-Half maximal effective concentration
۲-Half maximal inhibitory concentration

مواقع نیز آن را به عنوان تابع لگاریتمی و به صورت تابع pIC_{50} (که برابر با $-\log IC_{50}$ است) تبدیل می نمایند، که مقادیر بزرگتر آن به صورت نمایی بیانگر فعالیت بیشتر دارو می باشند[۸].

۱-۶-تعیین مقدار IC_{50} برای یک دارو

IC_{50} یک دارو را می توان بوسیله رسم منحنی دوز بر حسب پاسخ و آزمایش اثر غلظت های مختلف از آنتاگونیست روی فعالیت آگونیست بدست آورد. مقادیر IC_{50} را می توان با تعیین غلظت مورد نیاز آنتاگونیست مورد نظر جهت بازدارندگی نیمی از پاسخ بیولوژیکی آگونیست محاسبه نمود. مقادیر IC_{50} به شدت به شرایط اندازه‌گیری وابسته هستند. بطور کلی، در غلظت بالای بازدارنده، فعالیت آگونیست کاهش بیشتری می یابد. مقدار IC_{50} با افزایش غلظت آنزیم افزایش می یابد. همچنین با توجه به نوع بازدارنده سایر عوامل نیز ممکن است روی مقدار IC_{50} اثر گذار باشند که به عنوان مثال در مورد آنزیم های وابسته به ATP، مقدار IC_{50} نیز به غلظت ATP وابسته خواهد بود به ویژه هنگامی که بازدارندگی کاملا رقابتی باشد. مقادیر IC_{50} را می توان برای مقایسه توان دو آنتاگونیست بکار برد[۸].

اندازه گیری مقدارIC_{50} معمولا به طریق آزمایشگاهی انجام می گیرد کهاین کار مستلزم صرف زمان و هزینه های زیاد می باشد. از این رو استفاده از روش های آسان، قدرتمند و کم هزینه برای تخمین و پیش بینی فعالیت ترکیبات دارویی از اهمیت بالایی برخوردار است. یکی از روش هایی که بدین منظور استفاده می شود مطالعه ارتباط کمی ساختار-فعالیت[۱](QSAR) می باشد که فعالیت ترکیبات شیمیایی را به ویژگیهای ساختاری آنها ارتباط می دهد. فرض اساسی در مطالعات QSAR این است که ساختار یک مولکول (یعنی خواص هندسی، الکترونی و چیدمان فضایی اتم ها) حاوی ویژگیهایی است که مسئول خواص فیزیکی، شیمیایی و زیستی ترکیبات شیمیایی می باشند. مدل سازی QSAR در حوزه سم شناسی متولد شد. در واقع تلاش ها برای کمی کردن روابط بین ساختار شیمیایی و پتانسیل سمیت بحرانی، قسمتی از دستاوردهای منتشر شده سم شناسی برای بیش از

۱- Quantitative structure activity relationship

۱۰۰ سال بود. در سال ۱۸۶۳ کراس[۱] در دفاع از رساله خود در دانشکده پزشکی دانشگاه استراسبرگ[۲] عنوان کرد که سمیت الکل در پستانداران با کاهش حلالیت آنها در آب افزایش پیدا می‌کند. بعد از آن در سال ۱۸۶۸ کروم‌برون[۳] و فریزر[۴] این ایده را بیان کردند که فعالیت فیزیولوژیکی یک ماده در یک سیستم زیستی معین تابعی از ساختمان شیمیایی آن است. بنابراین یک تغییر در ساختمان شیمیایی بصورت یک تغییر در فعالیت زیستی انعکاس می‌یابد. در حال حاضر علم QSAR به شاخه های مختلف شیمی، بیوشیمی، شیمی پزشکی، داروسازی و شیمی زراعی راه یافته است و همچنین به ابزاری اساسی در طراحی دارو و آفت کش تبدیل شده و نقش رو به رشدی در علوم زیست محیطی اشاره دارد. روند رو به رشد QASR در طی این سال‌ها تا حدی به توسعه گسترده و سریع روش‌ها و تکنیک‌های محاسباتی نسبت داده می‌شود که با تعیین و پالایش توصیفگرها و روش‌های متنوع مدل سازی به موفقیت بیشتر مدل‌های QSAR منجر شده است. بدیهی است نتایج این مطالعات به شفاف سازی نحوه ارتباط بین خواص مولکول و ویژگی‌های ساختمانی آنها و درک بهتر مکانیسم عمل در یک سری از ترکیبات شیمیایی کمک می‌کند. علاوه بر آن توسط مدل‌های QSAR، تخمین فعالیت زیستی (یا خاصیت، واکنش پذیری و...) یک ماده شیمیایی جدید یا تست نشده با شباهت ساختاری به ترکیباتی که فعالیت مذکور برای آنها قبلا ارزیابی شده است، امکان می‌یابد. این روش می‌تواند منجر به صرفه جویی قابل توجه‌ای در هزینه‌های تولید شده ونیاز به آزمایشات طولانی مدت را کاهش دهد[۹].

۱-۷- پیشینه کارهای انجام شده در QSAR ترکیبات دارویی

در سال ۲۰۰۹ انترس و همکارانش یک مدل ساده ارتباط کمی ساختار- فعالیت خطی را جهت مدل سازی و پیش بینی بازدارندگی گیرنده‌های CXCR3 به وسیله ۴-N-آریل-[۱و۴] دیازپان ارائه

۱-Cros
۲- Strasbourg
۳- Crum-Brown
۴- Fraser

نمودند. جهت مدل سازی از روش MLR استفاده کردند و تعداد ترکیبات مورد مطالعه شامل ۳۲ ترکیب بودند. در اینجا آنها صرفا از توصیفگرهای فیزیک و شیمیایی استفاده نمودند و مدل نهایی بدست آمده شامل ۵ توصیفگر بود. و پارامترآماری، R^2 برای سری آموزش و تست به ترتیب برابر ۰/۸۲ و ۰/۷۵ می‌باشد[۱۰].

در سال ۲۰۱۱ شهلایی و همکارانش به مطالعه ارتباط کمی ساختار-فعالیت ۲۹ ترکیب از مشتقات سیکلوبوتن دی ان جهت بازدارندگی آنتاگونیست های CCR1 با استفاده از یک شبکه عصبی مصنوعی با الگوریتم آموزشی پس انتشار پرداختند. RMSE و R^2 برای سری آموزش به ترتیب ۰/۱۸۹ و ۰/۹۰۶ و برای سری تست مقادیر ۰/۱۰۳ و ۰/۹۳۲ بدست آمد[۱۱].

در سال ۲۰۱۱ اسدالهی و همکارانش به مطالعه QSAR بر روی گیرنده‌های CXCR2 پرداختند. در این مطالعه ارتباط کمی بین ساختارهای آنتاگونیست CXCR2 و فعالیتشان به وسیله روش حداقل مربعات جزئی (PLS) مورد مطالعه قرار گرفت و از روش‌های رگرسیون مرحله‌ای و الگوریتم ژنتیک جهت انتخاب مناسب ترین توصیفگرها استفاده کردند که ۵ توصیفگر انتخاب شد. جهت مدل سازی از روش‌های MLR و PLS استفاده کردند که نتایج، برتری روش GA-PLS را نشان داد. در این روش $R^2_{training}$ و RMSEP برای بهترین مدل (GA-PLS) به ترتیب برابر ۰/۸۸ و ۰/۵۱ می‌باشد[۱۲].

در سال ۲۰۱۰ عرب جم جنگلی و همکارانش به مطالعه ساختار- فعالیت مشتقات ۵-اکسوپیرولیدن-۳-کربوکسامید پرداختند. این ترکیبات از طریق دخالت در برهمکنش میان CCR5 و HIV مانع از گسترش بیماری می‌شوند. مجموعه داده‌های مورد بررسی شامل ۱۰۴ ترکیب می‌باشد که این سری داده‌ها به طور تصادفی به سه سری آموزش، ارزیابی و تست تقسیم کردند که با استفاده از شبکه عصبی مقادیر pIC_{50} مربوط به هر یک از ترکیبات را پیش بینی کردند و مقادیر MSE برای سری تست و رد مرحلهای تک تک به ترتیب برابر ۰/۱۲۱۰ و ۰/۱۴۰۵ می‌باشد[۱۳].

در این پایان نامه هدف، ارتباط کمی ساختار فعالیت برای پیش بینی فعالیت دارویی یکسری از ترکیبات به کار گرفته شد. بر اساس تحقیقات صورت گرفته، برای ترکیبات بررسی شده در این پایان نامه قبلا مدل سازی صورت نگرفته است و این کار به عنوان تحقیقی نوین در ارتباط با این ترکیبات است. در این پایان نامه از روش های رگرسیون خطی چندگانه (MLR)، شبکه عصبی مصنوعی (ANN) و ماشین بردار پشتیبان (SVM) جهت مدل سازی استفاده گردید.

فصل دوم

کمومتریکس

۲-۱- کمومتریکس[۱]

برای درک مکانیسم فرآیندهای مختلف شیمیایی، کشف و توسعه مواد جدید، حفظ محیط زیست و زمینه های دیگر شیمی هنوز توانایی حل مسائل به طور کامل وجـــود ندارد و بـرای عملـی کـردن بعضی از مسائل نیاز به سیستم های بسیار پیچیده ای می باشد که انجام آنهـا در گـرو صـرف هزینـه های بسیار و مطالعات گسترده است. در جهت حل این مشکل، روشهای محاسباتی کمومتریکس مـی توانند مفید باشند. تجزیه و تحلیـــل آماری و ریاضی داده های شیمیایی، تحت عنـوان کمـومتریکس یاد می شود [۱۴]. به عبارتی کمومتریکس یک روش کارآمد برای خلاصه کردن اطلاعات مفید از یک سری داده مشخص و پـیش بینـی سـری دیگـر داده هاسـت. هـدف کمـومتریکس، بهبـود بخشـیدن فرآیندهای اندازه گیری و استخراج اطلاعات شیمیایی مفیدتر از داده های اندازه گیری شده فیزیکی و شیمیایی می باشد [۱۵]. کمومتریکس اولین بار توسط دانشمند سوئدی بـه نـام ولـد[۲] [۱۶] در سـال ۱۹۷۲ به کار گرفته شد و توسط کووالسکی توسعه داده شد [۱۷]. اولین مقاله با عنوان کمـومتریکس در سال ۱۹۷۵ توسط کوالسکی[۳] و اولین کتاب در زمینه‌ی کمومتریکس در سال ۱۹۷۷ توسط انجمن شیمی آمریکا منتشر شد. انتشار این مقالات روش جدیدی را در زمینه‌ی تبدیل حجم وسیع داده‌های شیمیایی به اطلاعات شیمیایی معنـی‌دار گشـود. رشـد و توسـعه‌ی روز افـزون کمـومتریکس مرهـون استفاده از رایانه در آزمایشگاه‌ها، کامپیوترهای پیشرفته و نرم‌افزارهای مناسب است. اما با ایـن وجـود نمی‌توان از این داده‌ها اطلاعات مفیدی به‌دست آورد. مگر این‌کـه ابتـدا بـا اسـتفاده از کمـومتریکس، تجزیه و تحلیل داده‌ها صورت گرفته و سپس اطلاعات از آن‌ها استخراج شود.

کاربردهای کمومتریکس در زمینه‌های مختلف شامل: داده‌پردازی، بهینه‌سازی، بهبود سیگنال بـه نویز، کالیبراسیون چندمتغیره، الگوشناسی، دسته‌بندی و آنـالیز فاکتورهاسـت. چنـدین تعریـف بـرای

۱- Chemometrics
۲- Wold
۳- Kowalski

کمومتریکس بیان شده است که غالبا در متـــن های تجزیه ای به کار می روند. یکـی از جـامع تـرین تعاریف به صورت زیر است:

کمومتریکس شاخه ای از شیمی است که از ریاضی و آمار و منطق استفاده می کند بـرای اینکـه: الف) فرآیندهای تجربی بهینه را طراحی و انتخاب کند، ب) حداکثر اطلاعات شیمیایی قابل حصول را از تحلیل اطلاعات شیمیایی فراهم کند و پ) بتوان اطلاعات بیشتری در مورد سیستم‌هـای شـیمیایی بدست آورد [۱۸]. کمومتریکس مجموعه ای از یک سری قواعـد شـیمیایی اسـت کـه بـا اسـتفاده از ریاضی، آمار و کامپیوتر در عرصه های ارزیابی و تفسیر اطلاعات، بهینه کردن و مدلسازی فرآینـدها و آزمایشات و استخراج حداکثر اطلاعات شیمیایی از داده های تجربی به ما کمک می کند. به طور کلی می توان گفت کمومتریکس، علم استفاده از کامپیوتر، ریاضی و آمار در شیمی می باشد. کمومتریکس به عنوان یک شاخه علمی جوان در دو دهه اخیر به سرعت توسعه پیدا کرده است. ایـن رشـد سـریع مدیون پیشرفت دستگاه های هوشمند و خـــودکار آزمایشگاهـــــی و همچنین امکان اسـتفاده از کامپیوترهای قدرتمند و نرم افزارهای ساده است. بنابراین کمومتریکس به عنوان یک وسـیله در همـه قسمت های شیمی و به طور زیادی در زمینه شیمی تجزیه مورد استفاده قرار گرفته است. امروزه یک شیمیدان تجزیه به صورت فزاینده ای به استفاده از روشهای آماری و ریاضی در کارهای روزمره اش به کمومتریکس نیاز پیدا می کند [۱۹].

امروزه کمومتریکس در شاخه های مختلف شیمی مورد استفاده قرار گرفته است که برخی از آنها عبارتند از: کنترل فرآیندها، تجزیه و تحلیل و شناخت الگوها، پـردازش علائـم، بهینـه کـردن شـرایط [۲۰]. یکی از زمینه های مهم کاربرد کمومتریکس در مطالعاتی است که خواص مولکول هـــا را بـه ویژگی های ساختاری آنها نسبت می دهد. از نظر شیمیدانان فعالیت ها و خواص یک ترکیب ناشـی از ویژگی های ساختاری آن است. این نوع از مطالعات به بررسـی کمـی ارتبـاط سـاختمان بـا فعالیـت، QSAR معروف می باشد.

۲-۲- ارتباط کمی ساختار- فعالیت(QSAR)

یکی از کاربردهای روش‌های کمومتریکس، یافتن ارتباط کمی ساختار- فعالیت (QSAR) و ارزیابی داده‌های تجزیه‌ای از طریق آن است. مولکول‌های مشابه با تغییر کوچکی در ساختارشان می‌توانند فعالیت های شیمیایی کاملاً متفاوتی داشته باشند. این نوع ارتباط بین ساختار مولکولی و رفتار شیمیایی در مطالعات QSAR مورد توجه می‌باشد. در واقع، QSAR سعی در پیدا کردن رابطه‌ی هماهنگ میان فعالیت شیمیایی و ویژگی‌های مولکولی، به‌منظور کاربرد این روابط برای ارزیابی فعالیت ترکیبات جدید دارد. برای انجام یک مطالعه QSAR، ابتدا باید مقادیر فعالیت شیمیایی محاسبه و تعیین شده و توصیف‌کننده‌های مناسب از طریق تجربی یا محاسبه توسط کامپیوتر تولید شوند. مرحله‌ی بعد، دست‌یابی به ارتباط ریاضی بین ساختار و فعالیت است. این ارتباط معمولا توسط روش های خطی از جمله رگرسیون خطی چندگانه[۱](MLR)، آنالیز اجزای اصلی[۲](PCA) و حداقل مربعات جزیی[۳](PLS) و روش‌های غیرخطی مانند شبکه عصبی مصنوعی[۴](ANN) و ماشین بردار پشتبان[۵] (SVM) ارتباط میان ساختار و فعالیت ترکیبات را مورد مطالعه قرار می‌دهند[۳۰-۲۱].

هر مطالعه QSAR شامل مراحل زبر می‌باشد:

۱) جمع آوری و انتخاب سری داده ها

۲) رسم و بهینه سازی ساختار ترکیبات

۳) محاسبه توصیف کننده ها

۴) تجزیه و تحلیل و ارزیابی آماری توصیف کننده ها و انتخاب موثرترین آنها

۵) ایجاد مدل های آماری

۶) تجزیه و تحلیل و ارزیابی آماری مدل ها و انتخاب مناسب ترین مدل

۱- Multiple Linear Regression
۲- Principle Component Analysis
۳- Partial Least Squares
۴- Artificial Neural Networks
۵- Support Vector Machines

۷) ارزیابی قدرت پیش بینی و اعتبار مدل بدست آمده[۳۱].

.

۲-۲-۱- جمع آوری و انتخاب سری داده ها

مرحله نخست مدل سازی جمع آوری و انتخاب یک سری ترکیباتی است که مقادیر تجربی خاصیت یا فعالیت مورد نظر آن ها از منابع قابل اعتماد در دسترس باشد. بهتر است ترکیباتی بعنوان سری داده ها انتخاب شوند که ساختارشان مشابه بوده و آن کمیتی که هدف مدلسازی است در شرایط عملی یکسان بدست آمده باشد تا نتیجه‌ی قابل اعتماد و مناسب تری بدست آید.

در قدم بعدی سری داده ها به طورتصادفی به دو قسمت سری آموزشی[۱] و سری پیش بینی[۲] تقسیم می شوند. سری آموزشی که حاوی بیشترین تعداد مولکول ها است باید نماینده ی مناسبی از کل مولکول ها باشد چون این دسته از مولکول ها برای مدل سازی استفاده می شوند. مولکول های سری پیش بینی طوری انتخاب می شوند که نماینده مولکول های موجود در سری آموزشی باشند و معمولاً ۲۰٪ از کل داده ها را به عنوان سری پیش بینی انتخاب می کنند. سری پیش بینی در هیچ یک از مراحل مدلسازی شرکت نمی کند و در واقع همان طور که از نامش پیداست به منظور ارزیابی قدرت پیش بینی و اعتبار مدل های بدست آمده مورد استفاده قرار می گیرد. شایان ذکر است که هر قدر مجموعه داده ها بزرگتر و متنوع تر باشد مدل حاصل از آن معتبرتر و قدرت پیش بینی آن نیز بالاتر خواهد بود.

۲-۲-۲- رسم و بهینه سازی ساختار ترکیبات

برای محاسبه برخی از توصیف کننده ها به اطلاعاتی از قبیل مختصات سه بعدی و حالت پایدار مولکول نیاز می باشد. روشی که در این تحقیق بکار گرفته شده، AM1 است که جزء روش های نیمه

۱- Training set
۲- Prediction set

تجربی بوده و در روش‌های نیمه تجربی که در برنامه هایی نظیر HyperChem وارد شده انـد، قبـل از محاسبه توصیف کننده ها باید ساختمان تمام مولکول های موجـود در سـری داده هـا بـا روش هـای مکانیک کوانتومی و با استفاده از رایانه بهینه شوند و پایدارترین حالت آن ها با حداقل انرژی مشخص شود[۳۲].

۲-۲-۳- محاسبه توصیفگر ها

توصیفگر ها مقادیر عددی هستند که ویژگی های ساختاری و الکترونی مولکول ها را بصورت کمی بیان می‌کنند و به پیش بینی فعالیت و خصوصیات مولکول ها در آزمایش‌های پیچیده کمک می‌کنند. به عبارت دیگر می‌توان گفت که هر توصـیفگر اطلاعـات خاصـی از مولکـول را کـه بـر کمیـت مـورد مدلسازی اثر می‌گذارد، در اختیار قرار می‌دهد. در صورتی که توصیفگرهای مناسبی انتخاب شوند می-توان مکانیسم اثر کمیت مورد مدل سازی را به طور منطقی توجیه کرد. در صورتی که توصیف‌کننده-های انتخاب شده مناسب نباشند، عمل مدل‌سازی با مشـکل مواجـه شـده و مـدل مناسـب بـه‌دسـت نخواهد آمد. بنابراین در انتخاب توصیفگرها باید دقت کرد و توصیفگرهایی می‌بایست انتخاب شوند که در ارتباط با ویژگی خاص یا فعالیت مولکول باشند.

توصیفگرهای مولکولی، به دو دسته‌ی اصلی تقسیم می‌شوند:

۱- توصیفگرهای حاصل از اندازه‌گیری‌های تجربی مانند قطبش‌پذیری، انکسار مولاری، ممان دوقطبی.

۲- توصیفگرهای مولکولی نظری که از ساختار نمادین مولکول مشتق شده و می‌توانند به دسته هـای بیشتری مطابق با انواع مختلفی از نمایش مولکولی یا براساس ابعاد توصیف مولکول تقسیم شوند.

نمایش‌مولکولی، روشی است که با آن یک مولکول از طریق فرآیند قراردادی و قواعد اختصـار بـه-صورت نمادین ارائه می‌شود. بر این اساس می توان توصیفگر ها را به طور کلی به صـورت زیـر دسـته بندی کرد:

الف) توصیفگرهای صفر بعدی

ساده‌ترین روش نمایش مولکولی، فرمول شیمیایی مربوطه است. این نمایش هیچ‌گونه اطلاعاتی از ساختار مولکول ارائه نداده و از این رو توصیفگرهای مولکولی به‌دست آمده از فرمول‌های شیمیایی، توصیفگرهای صفربعدی نامیده می‌شوند. این دسته از توصیفگرها تعداد و نوع اتم‌ها، جرم مولکولی و هر عاملی از خصوصیات اتمی را تعیین می‌کنند. (مثلاً مجموع حجم‌های اتمی واندروالس). این نوع توصیفگرها به توصیفگرهای ساختاری نیز معروفند که علاوه‌بر موارد فوق شامل توصیفگرهای مربوط به نوع پیوندها و حضور حلقه‌ها در مولکول نیز می‌شوند. تعداد کل اتم‌ها یا تعداد یک عنصر شیمیایی خاص (کربن، نیتروژن، اکسیژن، هیدروژن، فلوئور و...)، تعداد گروه‌های عاملی خاص، تعداد پیوندهای ساده، دوگانه، سه‌گانه، آروماتیک و... در مولکول، تعداد کل حلقه‌ها بر اساس تعداد اتم‌ها (حلقه های شش‌تایی، پنج‌تایی‌و...) در مولکول، وزن مولکولی و متوسط وزن اتمی، مثال‌هایی از توصیفگرهای صفر بعدی می‌باشند.

ب) توصیفگرهای یک بعدی

ارائه‌ی نمایش ساده ساختاری، می‌تواند به عنوان نمایشی یک بعدی از مولکول در نظر گرفته شود- که شامل لیستی از اجزای ساختاری از مولکول (گروه‌های عاملی یا استخلافات مورد نظر در یک مولکول) است. این دسته از توصیف‌گرها، یک بعدی نامیده می‌شوند. توصیفگرهای گروه‌های عاملی، قطعات اتم مرکزی، توصیفگرهای تجربی و خصوصیات مولکولی زیر گروه‌های این دسته می‌باشند.

ج) توصیفگرهای‌دوبعدی

نمایش دوبعدی از مولکول، مبین چگونگی اتصال اتم‌ها و ماهیت باندهای شیمیایی می‌باشد. مانند دیدگاه بر پایه‌ی‌گراف مولکولی،که یک نمایش دو بعدی را ارائه می‌کنند و معمولاً به عنوان نمایش توپولوژیکی شناخته می‌شود. توصیفگرهای مولکولی به‌دست‌آمده از الگوریتم‌های به‌کار رفته برای یک نمایش توپولوژیکی، توصیفگرهای دوبعدی نامیده می‌شوند. این دسته از توصیفگرها شامل زیرگروه-

های توپولوژیکی، شمارش‌های مسیرهای مولکولی، توصیفگرهای BCUT، ضرایب بار و توصیفگرهـای خود ارتباطی دوبعدی می‌شود.

د) توصیفگرهای سه‌بعدی

نمایش سه بعدی یک مولکول را معمولاً به‌عنوان یک شیء هندسی انعطاف‌پذیر تجسم نموده و نه تنها یک نمایش از ماهیت و اتصال اتم‌ها بلکه معیاری از صورت بندی فضایی مولکـول نیـز در اختیـار قرار می‌دهد. این نمایش‌ها، هم نمایش‌های هندسـی و توصـیفگرهـای مولکـولی مشـتق شـده از ایـن نمایش، توصیفگرهای سه‌بعدی نامیده می‌شوند. زیر گـروه‌هـای ایـن دسـته عبارتنـد از: انـدیس‌هـای آروماتیکی، خصوصیات مولکولی راندیک، توصیفگرهای هندسی که شامل مسـاحت سـایه‌ی مولکـول، اندیس کاپا، فاکتور شکل (نسبت طول به عرض مولکول)، مساحت و حجم واندروالس، محورهای اصلی چرخش و ممان اینرسی‌های اصلی مولکول می‌باشد.

۲-۲-۴-حذف توصیفگرهای نامناسب

در مراحل قبل توصیفگرها براساس تئوری و تشخیص نظری انتخاب شده و سپس محاسبه شدند. لذا این احتمال وجود دارد که این توصیف کننده ها با هم همپوشانی داشته باشند و یـا ممکـن اسـت برخی از آن ها از لحاظ توجیه کمیت مورد مدلسازی مناسب نباشند. بنابراین بایـد بـه نحـوی آنهـا را حذف کرد و تعداد توصیفگرها را قبل از شروع مدلسازی کاهش داد. بنابراین برای عملی کردن چنـین اهدافی یعنی نخست برای حذف همپوشانی، ماتریس ضرایب همبستگی [۱] تشکیل داده مـی شـود و از هر زوج توصیف کننده ای کـه دارای ضـریب همبسـتگی بـالای ۰/۹ هسـتند، توصـیفگری کـه دارای همبستگی کمتری با کمیت مورد بررسی دارد حذف می گردد. لازم به ذکر است که توصـیفگر هـایی که بیش از ۹۰٪ مقادیر یکسان یا صفر دارند نیز حذف می گردند. بـدیهی اسـت کـه هـر چـه تعـداد متغیرهای مستقل کمتر باشد دسترسی به یک مدل مطلوب آسـانتر اسـت و از پیچیـدگی محاسـبات

۱- Correlation matrix

جلوگیری به عمل می آید. که اخیرا با استفاده از برنامه ی نوشته شده در محیط نرم افزار MATLAB همبستگی توصیفگرها بررسی و توصیفگر های حذفی مشخص می گردند.

۲-۲-۵-انتخاب بهترین توصیفگرها

با وجودی که با به کار بردن روش های کاهش متغیر، تعداد توصیفگرها به میزان زیادی کاهش می یابد. اما هنوز تعداد آنها برای ساخت یک مدل مناسب زیاد است. بنابراین باید به ترتیبی، بهترین توصیفگرها را از میان توصیفگر های باقیمانده، انتخاب نمود. از جمله این روش ها می توان به روش رگرسیون مرحله ای و الگوریتم ژنتیک اشاره کرد. در ادامه به اختصار در مورد هر کدام از این روش ها توضیحاتی ارائه می گردد.

۲-۲-۵-۱-انتخاب بهترین توصیفگرها توسط رگرسیون مرحله ای(SR)

در این روش خصوصیت شیمیایی و بیولوژیکی مورد نظر به عنوان متغیر وابسته و توصیفگرها به عنوان متغیر های مستقل در نظر گرفته می شوند. در روش مرحله ای، ابتدا با توجه به ضریب تعیین، متغیر اولیه انتخاب می شود. سپس مدل دو متغیری ایجاد شده و پارامترهای آماری مدل حاصل برای سری ارزیابی مورد بررسی قرار می گیرند. در صورتی که مدل از نظر آماری بی معنا باشد، متغیر بکار گرفته شده در آن مدل حذف می گردد. سپس مدل سه متغیری ساخته می شود و دوباره پارامترهای آماری آن شامل ضریب تعیین، مقدار میانگین مربعات خطا و آماره ی F برای سری ارزیابی مورد بررسی قرار می گیرد. مدلی که ضریب تعیین و آماره F آن بیشتر و مقدار میانگین مربعات خطای آن کمتر باشد به عنوان بهترین مدل انتخاب شده و توصیفگرهای موجود در مدل به عنوان توصیفگرهای بهینه به روش رگرسیون مرحله ای در نظر گرفته می شود[۳۳].

۲-۲-۵-۲-انتخاب بهترین توصیفگر ها توسط الگوریتم ژنتیک[۱](GA)

الگوریتم ژنتیک زیر مجموعه‌ای از الگوریتم‌های تکاملی[۲] است که ازاصول انتخاب طبیعی داروین برای جستجو و بهینه سازی استفاده می‌کند و برترین روش انتخاب متغیر در مطالعات QSAR به شمار می‌رود. طبق اصول داروین گونه هایی از یک جمعیت قادر به بقا هستند که خصوصیات برتری داشته باشند و این گونه ها می‌توانند ویژگی‌های خود را به نسل‌های بعدی منتقل کنند (حفظ و بقای اصلح). با وجود اینکه فرآیند جستجو در الگوریتم ژنتیک تصادفی می‌باشد ولی این الگوریتم در زمره الگوریتم‌های تصادفی ساده قرار نمی‌گیرد، بلکه بطور کارآمدی به اکتشاف در فضای جستجو (یعنی همه راه حل‌های ممکن برای مسئله) می‌پردازد تا از یک نقطه جستجوی جدید به سمت بهترین جواب پیش رود. الگوریتم‌های ژنتیک در هر تکرار چند نقطه از فضای جستجو را بررسی می‌کنند بنابراین احتمال اینکه به یک نقطه بهینه محلی[۳] همگرا شوند، کم است. موتور آغازگر الگوریتم ژنتیک با در نظر گرفتن یک جمعیت اولیه شروع می‌شود که مجموعه‌ای است از راه حل‌های تصادفی. هر راه حل بر طبق یک الگو کد گذاری می‌شود و کروموزوم نام می‌گیرد. این کروموزوم‌ها با هدف پیدا کردن بهترین راه حل، بعد طی مراحلی در چرخه، اصلاح می‌شوند که بعد از n چرخه، نسل[۴]nام نام می‌گیرند. تابع برازش[۵] برای ارزیابی کیفیت راه حل‌های ارائه شده بکار می‌رود. جمعیت اولیه کروموزوم‌ها با استفاده از تابع برازش ارزیابی می‌شود و بر اساس آن کروموزوم‌های برتر در فرآیندی نیمه تصادفی به عنوان والد انتخاب می‌شوند (رعایت اصل بقای اصلح). سپس از عملگرهای ژنتیکی (دورگه‌سازی، باز تولید و جهش) برای ایجاد کروموزوم‌های جدید در هر نسل استفاده می‌شود. دورگه‌سازی و جهش تنوع ژنتیکی را باعث می شوند که برای فرآیند تکامل، حیاتی است. دورگه‌سازی، دو کروموزم والد را برای ایجاد کروموزوم جدید که ممکن است بهتر از والدین باشد ترکیب می‌کند. جهش، یک موقعیت

۱- Genetic Algorithm
۲- Evolutionary Algorithms
۳- Local optima
۴- Generation
۵- Fitness Function

در این کروموزوم‌ها را بطور تصادفی تغییر می‌دهد و باز تولید یعنی رونوشت از بهترین کروموزوم‌ها تا در طی فرآیند دورگه‌سازی و جهش از دست نروند. این کروکوزوم‌های جدید (فرزندان[1]) هم در جمعیت قرار می‌گیرند و نسل اول را تشکیل می‌دهند. این فرآیند تا رسیدن به شرایط خاتمه الگوریتم بارها تکرار می‌شود. شرایط خاتمه الگوریتم ژنتیک عبارتند از:

۱- رسیدن به تعداد ثابتی از نسل‌ها

۲- به پایان رساندن زمان محاسباتی اختصاص داده شده

۲- پیدا کردن نسلی با کمترین خطا

۴- عدم دستیابی به نتایج بهتر.

روند کار الگوریتم ژنتیک در مراحل زیر خلاصه می شود [۳۴ و ۹] :

۱- ایجاد یک جمعیت اولیه از توصیفگرها بصورت تصادفی

۲- ارزیابی برازش هر کروموزوم در جمعیت

۳- دسته بندی کروموزوم ها براساس برازش آنها

۴- تولید مثل بین کروموزومهای مناسب، جهش و ایجاد یک جمعیت جدید

۵- بررسی شرایط خاتمه برنامه

۶- اگر جمعیت جدید شرایط خاتمه را مهیا نکند مراحل ۲ تا ۵ دوباره تکرار می شود.

۲-۲-۶-ساختن مدل

پس از انتخاب مناسب ترین توصیگرها، با استفاده از روش‌های آماری مختلف به جستجوی مدل مناسبی پرداخته می‌شود که بتواند ارتباط بین توصیفگرهای انتخابی و پارامترهای مورد مدل‌سازی را به درستی بیان کند. برای مدل‌سازی از روش‌های گوناگون خطی و غیرخطی می‌توان استفاده کرد. در

۱- Offspring

این پایان نامه رگرسیون خطی چندگانه به عنوان یک روش خطی و از میان روش‌های غیرخطی، شبکه‌های عصبی مصنوعی و ماشین بردار پشتیبان مختصرا شرح داده می‌شوند.

۲-۲-۶-۱-رگرسیون خطی چندگانه (MLR)

رگرسیون خطی چندگانه روشی است که برای مدلسازی رابطه خطی بین یک متغیر وابسته با یک یا چند متغیر مستقل به کار برده می‌شود. متغیر وابسته گاهی مورد پیش بینی[۱] و متغیر مستقل پیش بینی کننده[۲]نامیده می‌شوند. در QASR توصیفگرها به عنوان متغیرهای مستقل در نظر گرفته شده و کمیت مورد مدلسازی در نقش متغیر وابسته ظاهر می‌شوند. رابطه کلی MLR به صورت زیر است:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_iX_i \qquad (۱-۲)$$

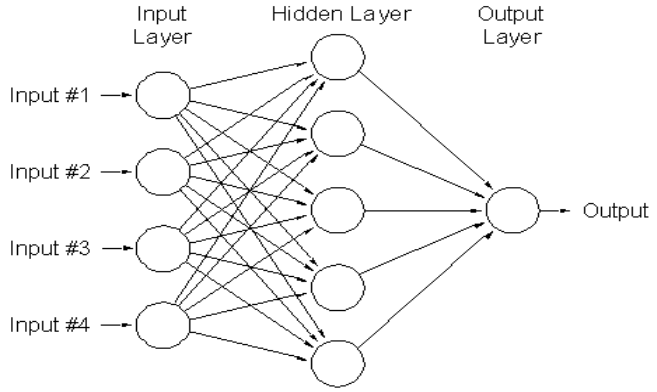
که در آن Y متغیر وابسته (فعالیت دارویی)، X_i متغیر مستقل (توصیفگر) وβ_i ضرائب رگرسیون نامیده می‌شوند. ضرایب مدل فوق به روش حداقل مربعات تخمین زده می‌شود. در این روش هر چه رابطه خطی بین متغیرهای مستقل و وابسته بیشتر باشد و برهمکنش متغیرهای مستقل با یکدیگر کمتر باشد، مدل بهتری بدست خواهد آمد.

۲-۲-۶-۲- شبکه های عصبی مصنوعی

شبکه عصبی مصنوعی یک سیستم محاسباتی است که با الهام گرفتن از ساختار و نحوه اتصال سلول های عصبی به حل مسائل می پردازد. شبکه عصبی مصنوعی از تعداد زیادی نورون (عنصر پردازش) تشکیل شده است که در سه لایه سازماندهی می‌شوند: الف) لایه ورودی که از طریق آن متغیرهای ورودی یا مستقل به شبکه داده می‌شود، ب) لایه مخفی که بین لایه ورودی و خروجی قرار دارد و ج) لایه خروجی که متغیرهای خروجی یا وابسته را ایجاد می‌کند. شکل (۲-۲) نمای ساده‌ای از

۱- Predictand
۲- Predictors

یک شبکه عصبی مصنوعی را نشان می‌دهد. در این شبکه ورودی‌ها قبل از وارد شدن به هر عنصر پردازش وزن‌دار شده و پس از جمع شدن با ترم بایاس[۱] در صورتی که جمع سیگنال‌های درونی عنصر پردازش از یک حد معین به نام حد آستانه بیشتر شد سیگنال خروجی ایجاد می‌گردد. بایاس و وزن، دو پارامتر قابل تنظیم در نورون‌ها می‌باشند. سیگنال خروجی ایجاد شده از یک تابع انتقال[۲](تابع فعالیت)[۳] عبور کرده و به عنوان ورودی برای سایر عناصر پردازش مورد استفاده قرار می‌گیرد[۳۵ و ۹].



شکل (۲-۱)- نمای ساده‌ای از یک شبکه عصبی مصنوعی[۳۵]

۲-۲-۶-۲-الف- توابع انتقال

تابع انتقال f می تواند خطی یا غیرخطی باشد. یک تابع انتقال براساس نیاز خاص حل یک مساله، مساله ای که قرار است به وسیله شبکه عصبی حل شود، انتخاب می شود. در عمل تعداد محدودی از توابع انتقال مورد استفاده قرار می گیرند.

توابع انتقال باید دو شرط اصلی داشته باشند:

۱- مقدار تابع باید در محدوده مشخصی تعریف شده باشد.

۲- مقدار تابع باید به صورت یکنواختی افزایش یابد[۳۷ و ۳۶].

سه تابع انتقال رایج در فرآیند بهینه سازی شبکه های عصبی، تابع انتقال خطی[۴]، تابع انتقال لگاریتم

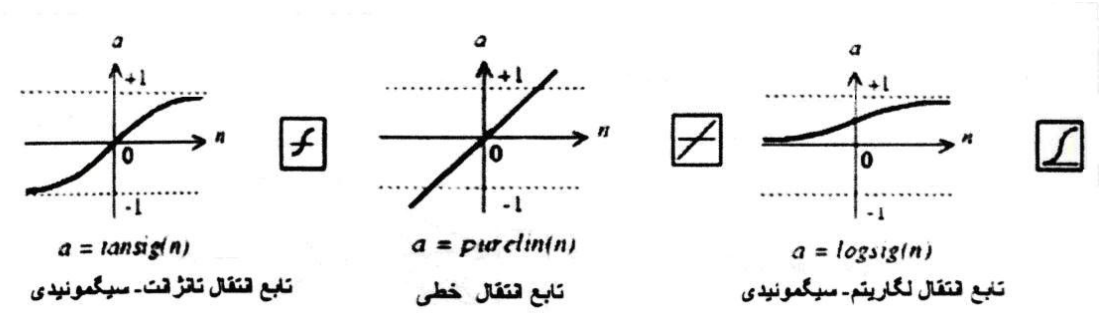
۱- Bias
۲- Transfer Function
۳- Activation Function
۴- Linear Transfer Function

سیگموید و تابع انتقال تانژانت هایپربولیک سیگموید نام دارند.

۱-تابع انتقال خطی: تابعی که به اختصار به صورت purline نمایش داده می شود یک تابع پیوسته است که معمولا برای تقریب خطی در فیلتر های خطی به کار می روند. این تابع همان مقدار ورودی را به عنوان خروجی بر می گرداند.

۲-تابع انتقال لگاریتم سیگموید: این تابع که در برنامه نرم افزاری با LOGSIG نمایش داده می شود بیشتر در شبکه های پس انتشار استفاده می شود. این تابع انتقال مقادیر ورودی را در محدوده ∞- تا محدوده∞+ در یافت کرده و خروجی بین صفر و یک تولید می نماید.

۳-تابع انتقال تانژانت سیگموید: که با TANSIG نمایش داده می شود ورودی را در محدوده ∞- تا ∞+ دریافت کرده و خروجی بین ۱ و۱- تولید می نماید.



شکل(۲-۲)- چند مورد از توابع انتقال موجود در ANN[۳۶]

۲-۲-۶-۳- ماشین بردار پشتیبان[۱](SVM)

ماشین های بردار پشتیبان یکی از روش های یادگیری با ناظر هستند. این روش از جمله روش-های نسبتا جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش های قدیمی تر مثل شبکه‌های عصبی نشان داده است. از الگوریتم آموزش دیده توسط ماشین های بردار پشتیبان می توان برای موارد زیر استفاده نمود:

۱- Support Vector Machine

۱- طبقه بندی برای سیستم های خطی و غیرخطی

۲- آنالیز رگرسیون برای سیستم های خطی و غیرخطی

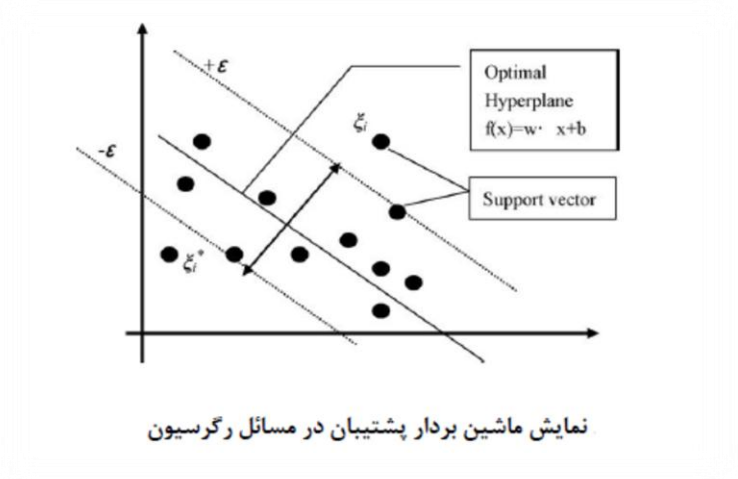
ماشین های بردار پشتیبان در مقایسه با شبکه های عصبی مصنوعی دارای دو مزیت عمده هستند که همین امر موجب توجه خاص محققین به این ماشین های فراگیر شده است:

۱- قابلیت مدل سازی توسط SVM با تعداد داده های کمتر نسبت به شبکه های عصبی

۲- تئوری قوی ماشین بردار پشتیبان قابلیت تعمیم بالایی به آن داده و از قرار گرفتن آن در بهینه محلی جلوگیری می کند[۴۰-۳۸].

۲-۲-۶-۳-الف- مبانی ماشین بردار رگرسیون

در ماشین بردار رگرسیون ایده اصلی بر مبنای نگاشت داده ها به فضایی با ابعاد بالاتر است و از طریق یک انتقال غیرخطی صورت می گیرد که در این فضای جدید رابطه خطی بین داده ها برقرار شده و می توان تابع رگرسیون خطی را بر روی داده ها اعمال کرد. در مسائل رگرسیون توسط SVM هدف پیدا کردن صفحه ای است که فاصله اش از تمامی نقاط داده ها، مینیمم باشد[۴۰].



شکل (۲-۳)-نمایش ماشین بردار پشتیبان در مسائل رگرسیون[۴۰]

فرض کنیم یک مجموعه داده بصورت زیر داشته باشیم

$$G = \left\{ (x_i, d_i)^n \right\} \qquad (۲-۲)$$

که در آن x_i مقادیر ورودی ها، d_i مقادیر خروجی و n تعداد داده ها می باشد.

ماشین بردار پشتیبان تابعی به شکل زیر بر روی داده ها در نظر می گیرد.

$$y = f(x) = \sum_{i=1}^{n} w_i \Phi(x_i) + b \qquad (۳-۲)$$

که در این تابع $\{\Phi(x_i)+b\}$ مشخص کننده ورودی هاست و $\{w_i\}'_{i=1}$ و b مقادیر ثابت می باشند. ثابت های w و b با حداقل کردن تابع زیر بدست می آیند.

$$R_{SVM}(C) = C\frac{1}{n}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2 \qquad (۴-۲)$$

$$L_\varepsilon(d,y) = \begin{cases} |d-y| \geq \varepsilon & |d-y| - \varepsilon \\ \text{بقیه موارد} & \end{cases} \qquad (۵-۲)$$

در معادله ۲-۴ R_{SVM} تابع خطای منظم شده و عبارت $C\frac{1}{n}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i)$ خطـای تجربـی اسـت کـه توسط تابع اتلاف $(L_\varepsilon(d_i, y_i))$ اندازه گیری شده و نشان می دهـد کـه در مقـادیر کمتـر از ε هـیچ خطایی وجود نخواهد داشت. $\frac{1}{2}\|w\|^2$ مشخص کننده پیچیدگی مدل می باشد. C ثابت تنظیم است و مصالحه بین خطای تجربی و پیچیدگی مدل را برقرار می کنـد. افـزایش مقـدار C در نتیجـه اهمیـت نسبی خطای تجربی است که مرتبط با افزایش عبارت تنظیم کننده است. ε انـدازه لولـه نامیـده مـی شود و بیانگر فاصله به اندازه $-\varepsilon$ و $+\varepsilon$ حول خط رگرسیون می باشد که مشـخص کننـده میـزان صحت داده های سری آموزشی است به طوری که داده هایی که داخل این لوله فرضـی قـرار بگیرنـد خطا محسوب نشده و داده های خارج از این محدوده به عنوان مقادیر خطـا محسـوب مـی شـوند. در حقیقت اگر داده ای داخل این لوله قرار بگیرد تابع اتلاف معادل صفر و اگر خارج از محدوده لوله (بـه شعاع ε اطراف تابع رگرسیون) قرار بگیرد تابع اتلاف متناسب با اختلاف بین موقعیت داده و شـعاع ε خواهد بود.

با معرفی متغیرهای سست [۱] ξ معادله ۲-۴ بصورت زیر تبدیل می شود.

(۶-۲)

$$
\begin{aligned}
&MinR\left(w,\xi_i,\xi_i^*\right)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\left(\xi_i+\xi_i^*\right)\\
&w\phi\left(x_i\right)+b-d_i\le\varepsilon+\xi_i^*\\
&d_i-w\phi\left(x_i\right)-b\le\varepsilon+\xi_i\\
&\xi_i,\xi_i^*\ge 0,\\
&i=1,...,n
\end{aligned}
$$

متغیرهای ξ_i,ξ_i^* جهت نشان دادن داده هایی است که خارج از محدوده مجاز در نظر گرفته شـده در لوله فرضی حول خط تابع قرار گرفته اند. داده های بالاتر از مقادیر $\varepsilon+$ با ξ_i و داده های کمتر از $\varepsilon-$ با ξ_i^* نشان می دهد. داده های خارج از لوله ε بردار پشتیبان [۲] نامیده می شوند، به خاطر این کـه ایـن نقاط اساس تابع رگرسیون محاسبه شده را تشکیل می دهند[۴۱].

در نهایت با معرفی دو جمله ای های لاگرانژ $\left(a_i,a_i^*\right)$ معادله ۲-۳ بصورت زیر تبدیل می شود.

(۷-۲)

$$
\begin{aligned}
&f(x,a_i,a_i^*)=\sum\left(a_i-a_i^*\right)K(x,x_i)+b\\
&a_i a_i^*=0\\
&a_i\ge 0\\
&a_i^*\ge 0\\
&i=1,.....,n
\end{aligned}
$$

و با حداقل کردن تابع ۲-۶ به فرم زیر تبدیل می شود.

۱- Slack

۲- Support Vectors

$$\phi(a_i,a_i^*)=\sum_{i=1}^{n}d_i(a_i-a_i^*)-\varepsilon\sum_{i=1}^{n}(a_i-a_i^*)-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i-a_i^*)(a_j-a_j^*)K(a_i-a_i^*) \quad (۸-۲)$$

$$\sum_{i=1}^{n}(a_i-a_i^*)=0$$

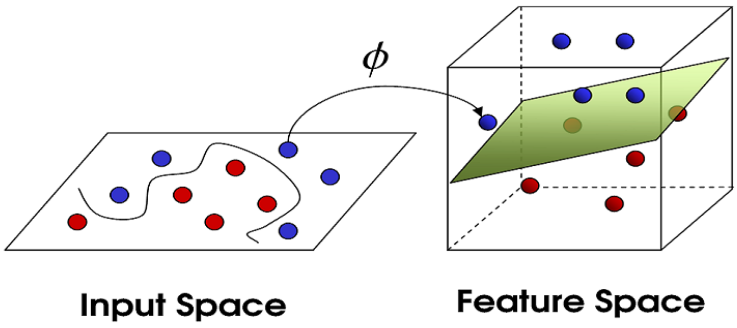
$$0\le a_i \le C, i=1,\dots,n$$

$$0\le a_i^* \le C, i=1,\dots,n$$

بر پایه شرایط کاروش – کان – تـاکر [۱] داده هـایی بـا مقـادیر $(a_i-a_i^*)$ غیرصـفر تعـداد بردارهـای پشتیبان را تشکیل می دهند. در معادله ۲-۸، K تابع کرنل است که تعریف آن به این شکل است.

$$K(x,x_i)=\phi(x).\phi(x_i) \quad (۹-۲)$$

که x و x_i بردار مقادیر ورودی را تشکیل می دهند، $\phi(x),\phi(x_i)$ مقادیر نگاشت شده این ورودی-ها در فضای جدید با ابعاد بالاتر است. ظرافت استفاده از تابع کرنـل در ایـن اسـت کـه مـا بـا فضـای ویژگی‌ها در بعد دلخواه خود سروکار داریم و بدون محاسبه $\phi(x)$ عمل رگرسیون را انجام مـی دهـیم. در ماشین بردار پشتیبان این تابع کرنل است که با نگاشت داده ها به یک فضا با ابعاد بالاتر باعث مـی شود تا یک رابطه خطی بین داده ها برقرار شده و سپس رگرسیون خطی بین آنها اعمال شود. پس در حقیقت ماشین بردار پشتیبان بر پایه الگوریتم خطی بنا نهاده شده که تابع کرنـل نقـش انتقـال داده های غیرخطی به فضای خطی را در آن اعمال می کند[۴۱].



شکل (۲-۴)- انتقال داده ها به فضایی با ابعاد بالاتر توسط تابع کرنل[۴۱]

۱- Karush – Kuhn-Tucker (kkt)

۲-۲-۶-۳-ب- انواع توابع کرنل[۴۱].

توابع کرنل زیادی می توانند مورد استقاده قرار بگیرند که استفاده از نوع تابع کرنل در هر مجموعه داده بستگی به طبیعت داده ها و نوع روابط حاکم بین آنها دارد. انواع توابع کرنل مورد استفاده:

- **تابع چند جمله ای:**

نگاشت چندجمله ای [۱] یک روش معروف جهت مدلسازی غیرخطی است.

$$K(x,x') = \langle x,x' \rangle^d \qquad (۱۰-۲)$$

$$K(x,x') = (< x,x' > +1)^d \qquad (۱۱-۲)$$

d درجه تابع چند جمله ای است و می تواند مقادیر بین ۱ تا ۶ را بپذیرد. البته ذکر این نکته هم ضروری است که تابع چند جمله ای درجه ۱ همان تابع کرنل خطی است که فقط مقدار بایاس آن متفاوت می باشد و هر چه درجه تابع چند جمله ای افزایش یابد پیچیدگی روابط بین داده ها بیشتر خواهد بود.

- **تابع خطی:**

این تابع بصورت زیر تعریف می شود.

$$K(x,y) = < x.y > +\theta \qquad (۱۲-۲)$$

که عبارت θ مقداری بایاس می باشد. هنگامی که بین داده ها ارتباط خطی برقرار باشد می توان از این تابع استفاده کرد.

۱- Polynomial

- تابع پایه شعاعی گوسین [۱] :

$$K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\delta^2}\right) \qquad (۱۳-۲)$$

پارامتر δ یک پارامتر قابل تنظیم است و بیانگر عرض تابع می باشد به طوری که هر چه عـرض تابع بیشتر شود به سمت روابط خطی سوق پیدا می کند.

- تابع پایه شعاعی نمائی [۲] :

تابع تابع پایه شعاعی نمائی یک راه حل مرحله ای برای مواقعی است که ناپیوستگی ها مورد قبول باشد و بصورت زیر تعریف می گردد:

$$K(x,x') = \exp\left(-\frac{\|x-x'\|}{2\delta^2}\right) \qquad (۱۴-۲)$$

δ ، یک پارامتر قابل تنظیم است و بیانگر عرض تابع می باشد.

- تابع پرسپترون چند لایه ای [۳] :

تابع پرسپترون چند لایه ای بصورت زیر تعریف می شود:

$$K(x,x') = \tanh(\vartheta < x.x' > + r) \qquad (۱۵-۲)$$

که پارامترهای ϑ و r ، قابل تنظیم می باشند.

- تابع سری فوریه [۴] :

این تابع در بازه $\left]-\pi/2, \pi/2\right[$ تعریف شده است و بصورت زیر معرفی می گردد:

۱- Gaussian Radial Basis Function

۲- Exponential Radial Basis Function

۳- Multi – Layer Preceptron Function

۴- Series Series Function

$$K(x,x') = \frac{Sin(N+\frac{1}{2})(x-x')}{Sin(\frac{1}{2}(x,x'))} \qquad (۱۶-۲)$$

انتخاب این تابع به عنوان تابع کرنل مورد استفاده در مدلسازی به علت ضعف آن در تنظیم پارامترهایش، رضایت بخش نمی باشد.

- **تابع Bsplines**:

Bsplines در بازه [۱ و ۱-] تعریف می شود که دارای فرم زیر می باشد:

$$K(x,x') = B_{2N+1}(x-x') \qquad (۱۷-۲)$$

۲-۲-۶-۳-ج- آموزش ماشین بردار رگرسیون

نوع تابع کرنل و همچنین پارامترهای C و ε که قبلا به آنها اشاره شد باید به دقت توسط کاربر، بهینه و انتخاب شوند چون کارایی تعمیم پذیری ماشین بردار رگرسیون تا حد زیادی بستگی به انتخاب مناسب این سه پارامتر دارد [۴۸]. همان طور که قبلا هم اشاره شد تابع اتلاف، لوله ای به شعاع ε اطراف تابع رگرسیون لحاظ کرده و در نهایت تعداد بردارهای پشتیبان برای ساختن تابع رگرسیون انتخاب می شود. مقدار خیلی زیاد ε منجر به تعداد بردار پشتیبان کمتری می شود (تعداد داده های بیشتری در داخل لوله ε قرار می گیرد) و در نتیجه منجر به تابع رگرسیونی با پیچیدگی کمتر می شود. پارامتر بعدی که باید بهینه شود C است که پارامتر ظرفیت یا تنظیم کننده نام دارد. توانمندی مدل رگرسیون به انتخاب صحیح مقدار C نیز بستگی دارد چون مقدار زیاد $(a_i - a_i^*)$ مترادف با مقدار C است. این بدان معنی است که انتخاب مقدار C متاثر از اهمیت هر یک از داده ها در سری آموزشی است. بطور مثال از یک طرف، مقدار زیاد C منجر به اختلاف زیاد بین مقدار $(a_i - a_i^*)$ در بردارهای پشتیبان خواهد شد. در این مورد بردارهای پشتیبان با مقدار بالای $(a_i - a_i^*)$ در مدل رگرسیون ایجاد شده تعیین کننده خواهند بود. از طرف دیگر، مقدار کم C می تواند منجر به

اختلاف کمی در بردارهای پشتیبان شده یا حتی مقدار $\left(a_i - a_i^*\right)$ یکسان شود. در این مورد داده های انتخاب شده به عنوان بردارهای پشتیبان در تابع رگرسیون توزیع یکسانی خواهند داشت[۲۹]. انتخاب صحیح و مناسب پارامترهای C و ε می تواند از بروز برازش بیشینه جلوگیری کند. بنابراین جهت دستیابی به یک مدل ماشین بردار پشتیبان قدرتمند با قابلیت پیش بینی خوب نیاز است که تابع کرنل مناسب به همراه پارامترهای قابل تنظیم مربوط به آن و همچنین پارامترهای C و ε به درستی بهینه شوند[۴۰].

۲-۲-۷- ارزیابی قدرت پیش بینی مدل

۲-۲-۷-۱- با استفاده از پارامترهای آماری

برای اطمینان از این که مدل بدست آمده توانایی پیش بینی نمونه‌های مختلفی از یک جمعیت را داراست، باید مدل را ارزیابی کرد. این ارزیابی با محاسبه‌ی پارامترهای آماری صورت می‌گیرد. رابطه‌ی ریاضی پارامترهای آماری استفاده شده در این پایان نامه، در ادامه توضیح خواهد داده شد.

ضریب تعیین: به عنوان یک شاخص برای بیان دقت خط رگرسیون برآورد شده، به کار می‌رود و نشان‌دهنده‌ی نسبت تغییرات متغیر وابسته توضیح داده شده توسط متغیر مستقل است. به عنوان مثال R^2 برابر با ۰/۹۲۴۳ نشان می‌دهد که ۹۲/۴۳ درصد تغییرات در متغیر وابسته می‌تواند توسط متغیر مستقل توضیح داده شود. رابطه‌ی ریاضی مربوط به ضریب تعیین به صورت زیر است:

$$R^2 = \frac{SSR}{SST} = -1 - \frac{SSE}{SST} \qquad (۱۸-۲)$$

که SSR[۱] طبق رابطه‌ی (۲-۱۹)، بیانگر مجموع مربعات انحراف مقادیر پیش بینی شده‌ی متغیر وابسته از میانگین مقادیر آن است:

$$SSR=\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad (۱۹-۲)$$

۱- Sum of Square Regression

SST[1] طبق این رابطه‌ی (۲-۲۰) نشانگر مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از میانگین مقادیر آن است:

$$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2 \qquad (۲۰-۲)$$

که در این روابط، $\hat{y}_i$ مقدار پیش بینی شده متغیر وابسته، y_i مقدار واقعی متغیر وابسته و $\bar{y}$ در هر رابطه، میانگین مقادیر متغیر وابسته است. همچنین SSE[2] نیز مبین مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از مقادیر پیش بینی شده برای آن است:

$$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad (۲۱-۲)$$

بنابراین با توجه به روابط صفحه قبل می توان نوشت:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad (۲۲-۲)$$

طبق رابطه‌ی (۲-۲۲)، اگر تمام مشاهدات بر روی خط برازش شده قرار گرفته باشند، یعنی به ازای تمام نقاط $\hat{y}_i=y_i$ باشد، مقدار R^2 برابر یک شده و هر گونه انحرافی از این حالت باعث می‌شود که مقدار R^2 از یک کوچکتر شود.

میانگین مربع خطاها[3](MSE): آماره MSE از رابطه‌ی (۲-۲۳) به دست می‌آید:

$$MSE=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n} \qquad (۲۳-۲)$$

مجموع مربع باقیمانده‌ها[4](PRESS) :

$$PRESS=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad (۲۴-۲)$$

خطای مطلق میانگین[5](MAE) :

$$MAE=\frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{n} \qquad (۲۵-۲)$$

متوسط درصد انحراف مطلق[1] (AAD):

۱- Sum of Square Total
۲- Sum of Square Error
۳- Mean Square Error
۴- Predictive Residual Sum of Square
۵- Mean Absolute Error

$$AAD=\frac{\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|}{n}\times 100 \qquad (۲۶-۲)$$

خطای استاندارد پیش بینی[۲](SEP) :

$$SEP=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \qquad (۲۷-۲)$$

۲-۲-۷-۲- با استفاده از نمودار برگشتی

در این نمودار مقادیر پیش بینی شده‌ی کمیت مورد نظر بر حسب مقادیر تجربی رسم می‌شود و به کمک ضریب تعیین (R^2) به دست آمده از نمودار، پراکندگی نقاط اطراف خط برگشت تعیین می‌شود. محدوده‌ی تغییرات بین صفر و یک است. اگر $R^2=1$ باشد، همبستگی کاملی بین دادها وجود دارد و نتایج پیش بینی شده به واقعیت نزدیک ترند. اما اگر $R^2=0$ باشد، بین داده ها هیچ گونه همبستگی وجود ندارد.

۲-۲-۷-۳- با استفاده از نمودار خطای باقیمانده

منظور از عبارت خطای باقیمانده، اختلاف بین مقادیر پیش بینی شده و مقادیر تجربی است. اگر پراکندگی مقادیر در دو طرف نمودار صفر باشد، این امر نشان‌دهنده‌ی تصادفی بودن خطاهاست. ولی اگر عمده‌ی نقاط، در این نمودار، در یک طرف صفر باشد، این بدان معناست که خطای جهت داری رخ داده است.

۲-۲-۷-۴- با استفاده از آزمون Y-تصادفی

این تکنیک برای مطالعه همبستگی‌های تصادفی مدل غیر خطی طراحی شده است. در این آزمون،

۱- Absolute Average Percent Deviation
۲- Standard Error of Prediction

مقادیر تجربی(که در اینجا بردار Y نامیده می‌شود) به صورت تصادفی در محدوده همان مقادیر، تغییر داده شده و سپس همبستگی متغیرهای مستقل با متغیرهای وابسته با استفاده از یکی از شاخص‌های آماری که معمولا R^2 است، مورد بررسی قرار می‌گیرد. اختلاف زیاد بین شاخص آماری به دست آمـده از این روش با شاخص آماری به دست آمده از مدل اصلی، نشاندهنده‌ی عدم وجود همبستگی تصادفی می‌باشد. به طور معمول این فرایند چندین بار انجام می‌شود.

۲-۲-۷-۵- با استفاده از ارزیابی متقاطع یا اعتبار سنجی تقاطعی

رایج ترین تکنیک اعتبار سنجی است که در آن در هر بار یکی یا یک گروه کوچک از داده‌ها کنار گذاشته شده و سپس برای داده‌های باقیمانده، مدلی بدست می‌آید. بعـد از آن پاسـخ بـرای داده‌هـای کنار گذاشته شده از روی این مدل پیش بینی می‌شود. این روش‌ها به ترتیب به نام‌های رد تـک تـک داده‌ها[۱] و رد گروهی از داده‌ها[۲] نامیده می شوند.

۲-۲-۸-نرم افزارهای مورد استفاده در این پژوهش

در این بخش بطور مختصر بسته نرم افزارهایی که در این پروژه مورد استفاده قرار گرفته اند شرح داده می شود.

۲-۲-۸-۱- بسته نرم افزار HyperChem

به کمک این نرم افزار ابتدا شکل سه بعدی مولکولها بطور تقریبی رسم می گردد سـپس سـاختار مولکول را می توان به روش های مختلف بهینه کرد. ایـن نـرم افـزار اطلاعـات فراوانـی نظیـر زوایـای پیوندی، زوایای پیچشی و . . . را در اختیار می گذارد. سپس این اطلاعـات بعنـوان ورودی نـرم افـزار

۱- Leave one out (LOO)
۲- Leave group out (LGO)

Dragon و سایر نرم افزارها [۴۲] مورد استفاده قرار می گیرند. نسخه های جدیدتر این نرم افزار علاوه بر شبیه سازی طیف ارتعاشی و الکترونی مولکول ها قادر به انجام بسیاری از محاسبات پیچیده الکترونی مربوط به آن نیز می‌باشند. این محاسبات منجر به تولید اطلاعاتی راجع به ساختمان هندسی مولکول، ساختار بهینه شده مولکول، گشتاور دو قطبی، انرژی سطوح تراز انرژی مولکولی، جمعیت الکترونی اوربیتالهای اتمی، بار روی اتم ها و بسیاری از اطلاعات دیگر می شود.

۲-۲-۸-۲- بسته نرم افزاری Dragon

نرم افزار Dragon توسط گروه تحقیقاتی QSAR و کمومتریکس دانشگاه میلان [۱] ارائه شد و محاسبه بیش از ۱۴۰۰ توصیف کننده شامل بیش از هجده دسته از توصیف کننده های مولکولی را امکان پذیر می کند [۴۳]. این نرم افزار با استفاده از یک فایل ورودی که می تواند فایل (*.hin) مربوط به نرم افزار HyperChem باشد، یک سری از توصیف کننده های مولکولی که محاسبه آنها مشکل و وقت گیر است را محاسبه می کند. لازم به ذکر است که این نرم افزار، ساختار هندسی بهینه مولکولها را جهت محاسبه توصیف کننده ها مورد استفاده قرار می دهد.

۲-۲-۸-۳- بسته نرم افزاری آماری SPSS[۲]

این بسته نرم افزاری [۴۴] به عنوان عمومی ترین نرم افزار آماری است که توانایی تجزیه و تحلیل کلی اطلاعات را دارا می باشد. نخستین بار در سال ۱۹۷۰ توسط جمعی از فارغ التحصیلان دانشگاه استانفورد [۳] امریکا ارائه شد و شامل یک برنامه اصلی و تعدادی زیر برنامه فرعی به زبان فرتون [۴] می باشد [۴۵]. این نرم افزار از قابلیت های فراوانی برخودار است بعنوان نمونه همان طور که در بخش های قبل ذکر شد یک دسته از توصیفگر ها که با هم همبستگی دارند باید حذف شوند برای این کار

۱- Millan
۲- Statistical Package for the Social Sciences
۳- Stanford
۴- Fortran

این نرم افزار ماتریس مربع همبستگی را که حاوی اطلاعات راجع به میزان همبستگی بین این متغیرهاست، می سازد و این هدف را عملی می سازد. مهمتر از این ویژگی، کاربرد این نرم افزار برای بدست آوردن بهترین مدل و ارزیابی مدل می باشد. لازم به ذکر است که این نرم افزار قابلیت انجام عملیات محاسباتی مختلف مثل رگرسیون چند متغیره، آنالیز تابع تشخیص، فاکتور آنالیز[۱] و آنالیز مولفه های اصلی را نیز دارد [۴۶].

۲-۲-۸-۴- بسته نرم افزاری MATLAB

MATLAB نرم افزاری با کاربرد فوق العاده برای عملیات محاسباتی است، که امکانات متعدد محاسباتی، نمایشی و برنامه نویسی را در محیطی که استفاده از آن برای کاربر آسان است، فراهم می کند. در این محیط برای مسائل و پاسخ ها از علائم و نمادهای آشنای ریاضی استفاده شده است. کاربردهای نوعی از این زبان عبارتند از:

۱) ریاضیات و محاسبات

۲) محاسبات علمی پیشرفته

۳) داده یابی

۴) مدلسازی، شبیه سازی، نمونه سازی اولیه

۵) آنالیز کردن اطلاعات، شناسایی و تجسم بخشیدن

۶) نمودارهای علمی و مهندسی

۷) کاربردهای پیشرفته، نظیر ایجاد مبدل های گرافیکی به دلخواه کاربر

MATLAB یک سیستم محاوره ای است که عنصر اساسی اطلاعات در آن رشته ای است که نیاز به تعریف ابعاد ندارد. این امر به ما اجازه می دهد بسیاری از مسائل و تکنیک های محاسباتی،

۵- Factor analysis

بخصوص آن هایی که با فرمول های برداری و ماتریسی سر و کار دارند را بسیار سریعتر از آن چه در زبان های غیر محاوره ای نظیر C و فورترن مقدور است، در کسر کوچکی از زمان حل کند. نام MATLAB ازکتابخانه عملیات ماتریسی (Matrix laboratory) گرفته شده است. در محیط های دانشگاهی، به ابزار پایه و استانداردی برای درسهای مقدماتی و پیشرفته در زمینه های ریاضی، مهندسی و علوم تبدیل شده است و در صنعت، به عنوان ابزاری برای جستجوهای تحقیقاتی با قابلیت های بالای آنالیز و توسعه، شناخته شده است. از خصوصیات برجسته MATLAB دارا بودن خانواده ای وسیع از پاسخ های دارای کاربری های خاص علمی است که با نام جعبه ابزار[۱] از آنها یاد می شود. برای بسیاری از کاربران MATLAB، جعبه ابزارها امکان یادگیری و به کاربردن تکنولوژی های خاص را فراهم کرده است. جعبه ابزارها، مجموعه ای جامع از توابع MATLAB (M فایلها) هستند که محیط MATLAB را برای کاربرد در حل کلاسهای خاصی از مسائل توسعه می دهند[۴۷].

۱- Toolbox

فصل سوم

مطالعه ارتباط کمی ساختار – فعالیت مشتقات هموپیپرازین، دی آمین، ۳-آمینوپیرولیدین با استفاده روش هایخطی و غیر خطی

۳-۱- مدل سازی فعالیت دارویی مشتقات هموپیپرازین، دی آمین و ۳-آمینو پیرولیدین به عنوان باز دارنده های CCR2

به منظور مطالعه ارتباط کمی ساختار- فعالیت از رگرسیون خطی چندگانه، شبکه عصبی مصنوعی و ماشین بردار پشیبان به عنوان روش های مدل سازی برای پیش بینی فعالیت های مشتقات هموپیپرازین، دی آمین و ۳-آمینو پیرولیدین استفاده شد. همچنین از روش رگرسیون مرحله‌ای و الگوریتم ژنتیک برای انتخاب متغیر استفاد گردید و سپس توانایی مدل های به دست آمده از هر سه روش در پیش بینی قدرت بازدارندگی این مشتقات مورد ارزیابی قرار گرفت. به طور کلی بخش تجربی این تحقیق شامل سری داده ها، رسم و بهینه سازی ساختار مولکول ها، محاسبه توصیفگرهای مولکولی مناسب، مدل سازی و ارزیابی مدل برتر است.

۳-۱-۱-انتخاب سری داده‌ها

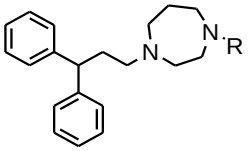
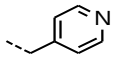
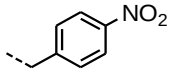
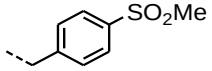
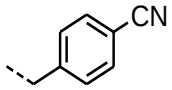
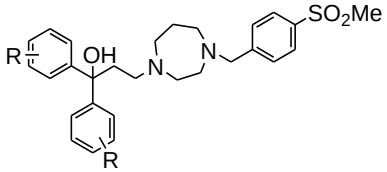
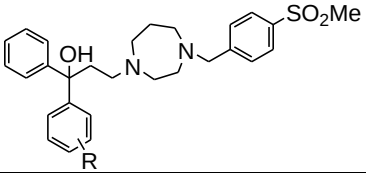
سری داده‌ها شامل فعالیت دارویی ۱۰۳ ترکیب از مشتقات هموپیپرازین، دی آمین و ۳-آمینو پیرولیدین به عنوان بازدارنده برای درمان التهاب مفاصل یا همان روماتیسم توسط روش‌های کمومتریکس مورد بررسی قرار گرفت (جدول۳-۱)[۶-۴]. قدرت بازدارندگی این ترکیبات به صورت IC_{50} گزارش شده است. IC_{50} عبارتست از مینیمم غلظتی از ترکیب دارویی که باعث ۵۰٪ اثر بازدارندگی بر روی بیماری می‌شود. این مقادیر به مقیاس لگاریتمی(pIC_{50}) تبدیل و مورد استفاده قرار گرفت.

۳-۱-۲- رسم و بهینه سازی ساختار مولکول ها

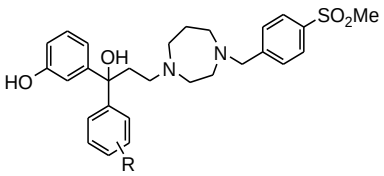
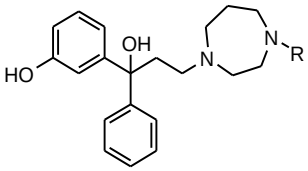
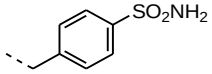
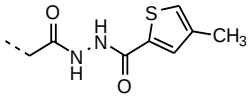
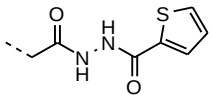
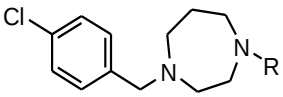
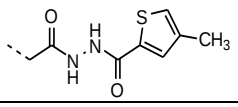
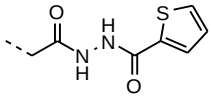
در این مرحله از مطالعه، ساختار مولکولی هر ترکیب ابتدا در نرم افزار HyperChem07 ترسیم شد. سپس با احتساب اتم های هیدروژن، ساختار سه بعدی ترکیبات با استفاد از روش های نیمه تجربی کوانتمیAM1 بهینه گردید و این بهینه سازی تا زمانی ادامه یافت که جذر میانگین مربعات گرادیان انرژی به ۰/۰۰۱ کیلوکالری بر مول برسد. با استفاده از این نرم افزار می توان اطلاعات فراوانی

نظیر زوایای پیوندی، طول پیوندها، زوایای پیچش، بار اتم هـا، انـرژی تشـکیل مولکـول و... را بدسـت آورد.

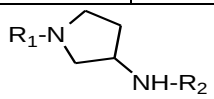
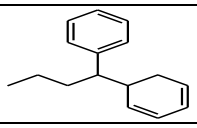
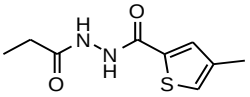
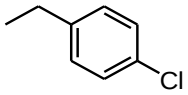
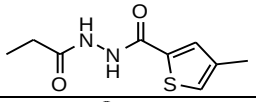
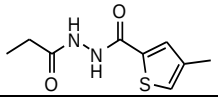
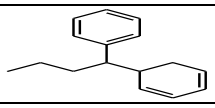
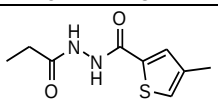
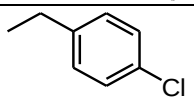
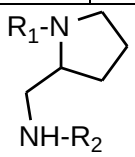
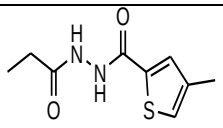
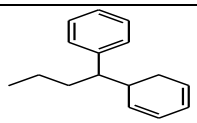
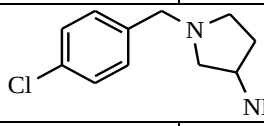
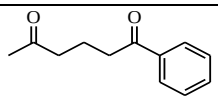
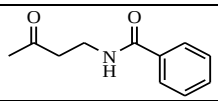
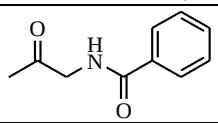
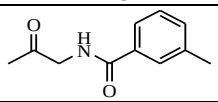
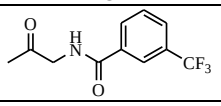
جدول(۳-۱)- ساختار ترکیبات مورد استفاده و مقدار pIC_{50} آنها

No.	R_1	R_2	pIC_{50}
			
۱		-	۴/۹۶
۲		-	۴/۷۲
۳		-	۴/۸۹
۴		-	۴/۳۷
			
۵	H	-	۵/۳۹
۶	4-NMe_2	-	۴/۳۵
۷	4-OH	-	۴/۶۸
۸	3-OH	-	۵/۸۲
۹	4-F	-	۵/۱۵
۱۰	3-F	-	۴/۴۷
۱۱	4-Cl	-	۴/۹۶
			
۱۲	3-OH	-	۶/۱۵
۱۳	3-CH_2OH	-	۵/۴۰
۱۴	3-NH_2	-	۵/۳۸
۱۵	3-NHMe	-	۵/۳۰
۱۶	3-OMe	-	۵/۱۰
۱۷	3-F	-	۵/۹۲
۱۸	3-Me	-	۴/۸۰

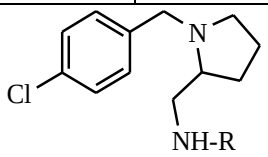
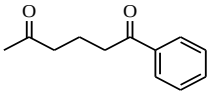
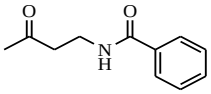
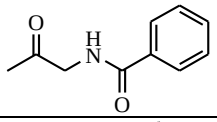
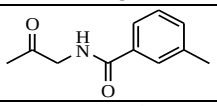
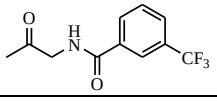
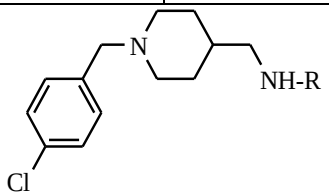
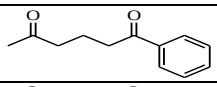
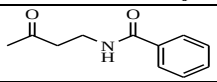
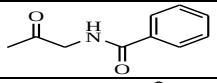
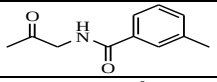
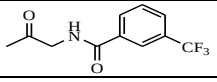
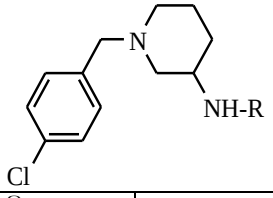
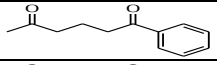
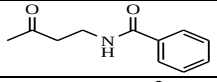
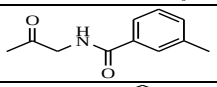
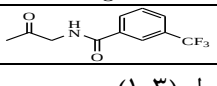
ادامه جدول (۳-۱)

No.	R_1	R_2	pIC_{50}
			
۱۹	3-F	-	۵/۶۲
۲۰	3-Cl	-	۵/۳۵
۲۱	4-F	-	۵/۸۲
۲۲	4-Cl	-	۵/۸۲
۲۳	3,5-DiF	-	۵/۰۳
			
۲۴		-	۵/۸۲
۲۵		-	۵/۱۹
۲۶		-	۴/۵۵
			
۲۷		-	۵/۱۳
۲۸		-	۴/۵۲

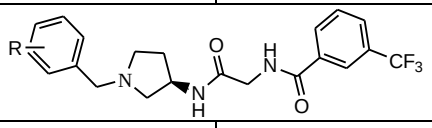
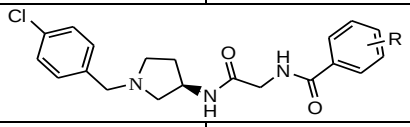
ادامه جدول (۳-۱)

No.	R_1	R_2	pIC_{50}
			
۲۹			۴/۶۴
۳۰			۴/۳۹
۳۱			۴/۷۷
۳۲			۵/۱۱
			
۳۳			۴/۵۹
			
۳۴		-	۴/۹۶
۳۵		-	۴/۶۶
۳۶		-	۴/۳۸
۳۷		-	۵/۱۶
۳۸		-	۶/۴۳

ادامه جدول (۳-۱)

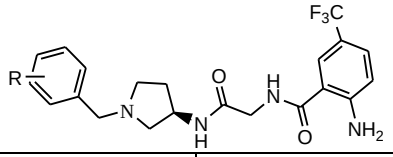
No.	R_1	R_2	pIC_{50}
			
۳۹		-	۴/۷۲
۴۰		-	۴/۴۰
۴۱		-	۴/۱۸
۴۲		-	۴/۸۵
۴۳		-	۵/۶۴
			
۴۴		-	۴/۳۰
۴۵		-	۴/۱۹
۴۶		-	۴/۹۶
۴۷		-	۵/۲۶
۴۸		-	۶/۱۵
			
۴۹		-	۴/۰۶
۵۰		-	۴/۱۳
۵۱		-	۴/۷۷
۵۲		-	۶/۱۸

ادامه جدول (۳-۱)

No.	R_1	R_2	pIC_{50}
			
۵۳	H	-	۶/۱۶
۵۴	2-Cl	-	۶/۲۰
۵۵	2-CH_3	-	۶/۰۲
۵۶	2-OCH_3	-	۵/۸۷
۵۷	3-CH_3	-	۵/۵۱
۵۸	3-OCH_3	-	۵/۵۰
۵۹	4-Cl	-	۶/۷۴
۶۰	4-CH_3	-	۶/۹۴
۶۱	4-OCH_3	-	۶/۹۴
۶۲	4-Et	-	۷/۲۳
۶۳	4-Br	-	۶/۷۸
۶۴	4-Vinyl	-	۶/۹۲
۶۵	4-CH_3S	-	۶/۶۹
۶۶	4-OH	-	۶/۶۵
۶۷	4-NHAc	-	۶/۵۲
۶۸	4-OCF_3	-	۶/۲۱
۶۹	4-F	-	۶/۰۲
۷۰	4-NO_2	-	۶/۸۱
۷۱	4-CN	-	۵/۵۸
۷۲	2,4-$(CH_3)_2$	-	۷/۲۷
۷۳	2,4-Cl_2	-	۶/۵۲
۷۴	4-OH, 3-OCH_3	-	۶/۸۲
۷۵	2-Naphthyl	-	۶/۱۲
			
۷۶	3-CH_3	-	۵/۶۲
۷۷	3-Cl	-	۵/۶۲
۷۸	4-CH_3	-	۵/۰۰
۷۹	3-F	-	۵/۳۶
۸۰	3-Br	-	۶/۱۱
۸۱	3-OCF_3	-	۶/۳۱
۸۲	3-NO_2	-	۶/۰۸
۸۳	2-NH_2, 5-NO_2	-	۶/۶۸
۸۴	2-NH_2, 5-Cl	-	۶/۱۴
۸۵	2-NH_2, 5-Br	-	۶/۱۹
۸۶	2-NH_2, 5-I	-	۶/۵۱
۸۷	2-NH_2, 5-OCF_3	-	۷/۰۶
۸۸	2-NH_2, 5-CF	-	۷/۵۹

ادامه جدول (۳-۱)

No.	R_1	R_2	pIC_{50}

			
۸۹	4-Cl	-	۷/۵۹
۹۰	4-Br	-	۶/۹۴
۹۱	4-CH_3	-	۷/۷۰
۹۲	4-Et	-	۷/۹۶
۹۳	4-Vinyl	-	۷/۷۲
۹۴	4-OCH_3	-	۷/۷۰
۹۵	4-OH	-	۷/۳۸
۹۶	4-Cl, 3-NH_2	-	۸/۳۹
۹۷	4-CH_3, 3-NH_2	-	۸/۲۱
۹۸	4-OCH_3, 3-NH_2	-	۸/۲۸
۹۹	4-OH, 3-NH_2	-	۷/۸۵
۱۰۰	4-OCH_3, 3-OH	-	۷/۲۸
۱۰۱	4-OH, 3-OCH_3	-	۷/۴۱
۱۰۲	2,4-$(CH_3)_2$	-	۸/۴۹
۱۰۳	2,4-Cl_2	-	۷/۰۲

۳-۱-۳- محاسبه توصیفگر ها

تعداد ۱۴۸۱ توصیفگر به وسیله نرم افزار Dragon محاسبه شدند. این نرم افزار برای محاسبه هیجده دسته از توصیفگرهای مولکولی مورد استفاده قرار می‌گیرد. این نرم افزار هیچگونه بهینه سازی در ساختار را انجام نداده و می تواند این تعداد توصیفگر را برای مولکول های تا ۱۵۰ اتم محاسبه نماید. به این منظور خروجی نرم افزار HyperChem برای هر ترکیب به برنامه Dragon منتقل و توصیفگرها محاسبه شدند. به این ترتیب تعداد ۱۴۸۱ توصیفگر مولکولی برای هر ترکیب محاسبه شد. نام گروه های توصیفگر قابل محاسبه توسط نرم افزارDragon بطور کامل در جدول (۳-۲) آمده است.

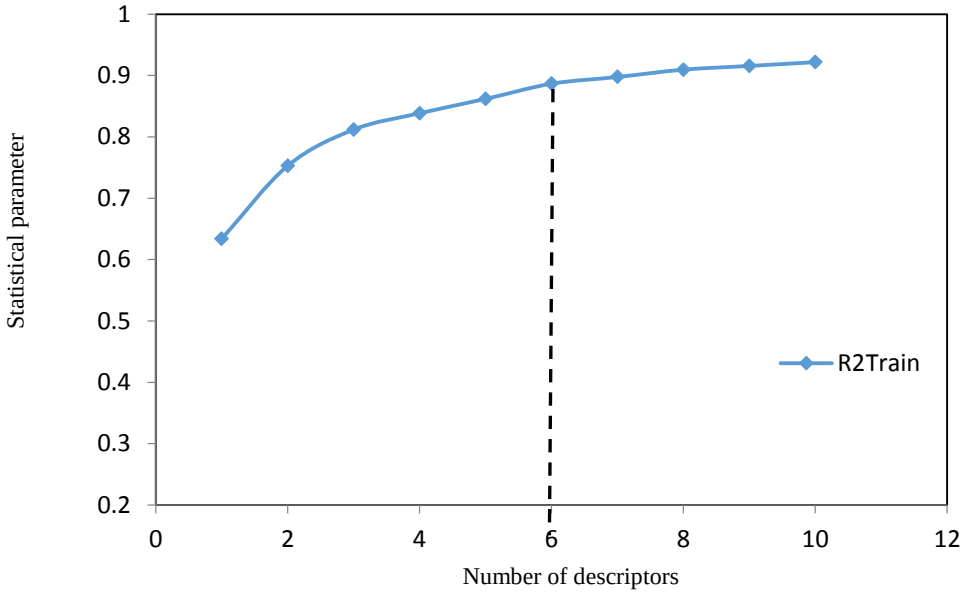
جدول(۳-۲)- فهرست گروه‌های مختلف توصیفگرهای محاسبه شده توسط نرم افزار Dragon

ردیف	نام گروه توصیفگر	ردیف	نام گروه توصیفگر
۱	توصیفگرهای زیر ساختاری[۲]	۱۰	توصیفگرهای هندسی[۱]
۲	توصیفگرهای توپولوژیکی[۴]	۱۱	توصیفگرهای RDF[۳]
۳	شمارنده های مولکولی[۶]	۱۲	توصیفگرهای سه بعدی مورس[۵]
۴	توصیفگر هایBCUT[۸]	۱۳	توصیفگرهای WHIM[۷]
۵	شاخص بار توپولوژیکی GALVEZ[۱۰]	۱۴	توصیفگرهای GETAWAY[۹]
۶	خود ارتباطی های دو بعدی[۱۲]	۱۵	گروه های عاملی[۱۱]
۷	توصیفگر های بار[۱۴]	۱۶	اجزای میان اتمی[۱۳]
۸	شاخص های آروماتیسیته[۱۶]	۱۷	توصیفگرهای تجربی[۱۵]
۹	پروفایل های مولکولی راندیک[۱۸]	۱۸	خصوصیات مولکولی[۱۷]

۱- Geometrical Descriptors
۲- Constitutional Descriptors
۳- Radial Distribution Function Descriptors
۴- Topological Descriptors
۵- 3D MORSE Descriptors
۶- Molecular Walk Counts
۷- Weighted Holistic Invariant Molecular Descriptors
۸- Burden-CAS-University of texas eigenvalues
۹- Geometry, Topology and Atom Weights Assembly
۱۰- Topological Charge Index
۱۱- Functional Groups
۱۲- 2D Autocorrelations
۱۳- Atom-Centered Fragments
۱۴- Charge Descriptors
۱۵- Empirical Descriptors
۱۶- Aromatic Indices
۱۷- Molecular Properties
۱۸- Rancid Molecular Profiles

۳-۱-۴- انتخاب توصیفگرهای مناسب

یکی از مهمترین مراحل در مطالعات QSAR انتخاب توصیفگرهای مناسب است که بتوانند فعالیت دارویی را به ساختار مولکول ارتباط دهند. با توجه به این موضوع و اینکه تعداد زیاد توصیفگرها باعث پیچیدگی محاسبات می شوند و همچنین تعداد زیادی از توصیفگرها حاوی اطلاعات یکسان و در نتیجه اثر یکسان بر روی فعالیت مهار کنندگی دارند، لذا باید کاهش تعداد متغیرها به کار گرفته شود. به همین منظور، ابتدا توصیفگرهایی که دارای مقادیر یکسانی برای مولکول ها بودند از فهرست توصیفگرها حذف شدند. سپس با بررسی همبستگی بین متغیرها، از میان هر دو توصیفگری که ضریب همبستگی بزرگتر از ۰/۹ داشتند، آنکه همبستگی کمتری با متغیر وابسته داشت، حذف و سپس رگرسیون مرحله به مرحله توسط نرم افزار SPSS جهت انتخاب توصیفگرهای مناسب برای مدلسازی مورد استفاد قرار گرفت. در روش رگرسیون مرحله ای، متغیرها یکی پس از دیگری وارد مدل شدند در این حالت، ابتدا متغیری وارد مدل می شود که بالاترین میزان همبستگی را با متغیر وابسته دارد. با ورود هر متغیر جدید، کلیه متغیرهای موجود در معادله بررسی شده و اگر هر کدام از آنها سطح معناداری خود را از دست بدهند، قبل از ورود متغیر جدید از مدل خارج می شود. به این ترتیب داده های pIC_{50} به عنوان متغیر وابسته و توصیفگرها به عنوان متغیر مستقل در نظر گرفته شده و تکنیک رگرسیون مرحله ای انجام شد. همانطور که می دانیم روش رگرسیون مرحله ایی تعداد زیادی مدل ارائه می کند. که مدل اول شامل یک توصیفگر، مدل دوم شامل دو توصیفگر و .... می باشد. با افزایش تعداد توصیفگرها بالطبع مقدار R^2 نیز افزایش می یابد. اما بدلیل پیچیدگی مدل، نمی توانیم تعداد زیادی توصیفگر را جهت مدلسازی انتخاب کنیم. بدین منظور و جهت انتخاب تعداد توصیفگرهای مناسب، نمودار پارامتر آماری (R^2_{train}) برحسب تعداد توصیفگرها رسم گردید که در شکل (۳-۱) نشان داده شده است. بر طبق این نمودار تعداد ۶ توصیفگر به عنوان توصیفگرهایی که بیشترین ارتباط را با فعالیت دارویی دارند، انتخاب شدند. این ۶ توصیفگر به همراه طبقه آنها در جدول(۳-۳) ارائه شده است.



شکل(۳-۱) نمودار پارامتر آماری (R^2_{train}) برحسب تعداد توصیفگرها

جدول (۳-۳)- توصیفگرهای انتخاب شده توسط رگرسیون خطی چندگانه مرحله به مرحله

Descriptor symbol	Type	meaning	MF[a]
X0A	Topological	Average connectivity index chi-0	۰/۷۲۱
PCWTe	Charge	Partial charge weighted topological electronic charge	۰/۱۱۱
Jhetp	Topological	Balaban-type index from polarizability weighted distance matrix	۰/۱۴۵
RDF065m	RDF	Radial distribution function-6.5/weighted by atomic masses	۰/۰۲۷
RDF145v	RDF	Radial distribution function-14.5/weighted by atomic van der Waals volumes	-۰/۰۰۳
H-052	Atom-centred fragments	H attached to C0(sp3) with 1X attached to next C	-۰/۰۰۳

[a]MeanEffect

۳-۱-۵- ارزیابی مستقل بودن توصیفگرهای انتخاب شده

به منظور ارزیابی توصیفگرهای انتخاب شده مبنی بر مستقل بودن از همدیگر در جدول (۳-۴) ماتریس همبستگی توصیفگرهای انتخاب شده آورده شده است. همانطور که در این جدول مشاهده می‌شود ضریب همبستگی بین توصیفگرهای انتخاب شده همگی کمتر از ۰/۹ می‌باشد. لذا نتایج جدول نشان می‌دهد که بین توصیفگرهای انتخاب شده همبستگی چندانی وجود نداشته و توصیفگرها تقریبا مستقل از هم هستند.

جدول (۳-۴)-ماتریس ضرایب همبستگی توصیفگرهای انتخاب شده

	X0A	PCWTe	Jhetp	RDF065m	RDF145v	H-052
X0A	۱					
PCWTe	-۰/۴۲۴	۱				
Jhetp	۰/۰۹۳	-۰/۷۳۳	۱			
RDF065m	۰/۶۴۰	-۰/۱۴۵	-۰/۰۵۶	۱		
RDF145v	-۰/۱۰۷	-۰/۲۰۱	۰/۲۸۴	-۰/۳۴۷	۱	
H-052	۰/۰۱۱	-۰/۶۲۴	۰/۸۰۱	-۰/۰۰۱	۰/۱۳۱	۱

۳-۱-۶- مدل سازی توسط رگرسیون خطی چندگانه (MLR)

پس از انتخاب مناسب ترین توصیفگرها توسط روش مرحله ای، مرحله بعدی، ایجاد مدل میان توصیفگرهای انتخاب شده و فعالیتهای ترکیبات دارویی می باشد. بین توصیف کننده ها و فعالیتهای بازدارندگی مشتقات هموپیپرازین، دی آمین و ۳-آمینو پیرولیدین برای سری آموزش با استفاده از روش MLR رابطه زیر به عنوان مدل خطی بدست آمد:

pIC_{50}= -19.2961(±4.522) + 25.306 (± 6.984) **X0A** + 0.118(±0.0164) **PCWTe** + 3.387 (±0.835) **Jhetp** + 0.036 (±0.008) **RDF065m** – 0.314 (±0.063) **RDF145v** – 0.087 (±0.030) **H-052**

سپس از معادله بدست آمده برای پیش بینی فعالیت سری تست استفاده گردید. مقادیر واقعی و پیش بینی شده فعالیت‌ها برای کلیه ترکیبات مجموعه آموزش و تست در پیوست (پ-۱) آورده شده است.

۳-۱-۷- مدل سازی به روش شبکه عصبی مصنوعی

بخاطر اینکه ممکن است مدل خطی بهترین مدل برای توصیف رفتار سری داده ها نباشد، شبکه عصبی مصنوعی به عنوان یکی از روشهای مدل‌سازی غیرخطی مورد استفاده قرار گرفت. در ابتدا کل سری داده ها به صورت تصادفی و با نسبت ۶۰:۲۰:۲۰ به سه سری آموزش، ارزیابی و تست تقسیم شدند. سپس پارامترهای موثر بر آموزش شبکه عصبی پس انتشار از قبیل تعداد ورودی ها، تعداد لایه های پنهان، تعداد گره های لایه پنهان، نوع تابع انتقال، نوع تابع آموزش، تعداد دورهای آموزش و پارامتر mu از سری ارزیابی بهینه شدند. در طی فرایند آموزش، مقادیر وزن های ارتباطی تغییر نموده و معیار ارزیابی و پارامتر مورد بررسی برای پایان دادن آموزش و بهینه شدن مقادیر وزن ها، مقدار خطای شبکه است که طبق تابع کارایی تعیین شده محاسبه می شود که در این روش، تابع کارایی، میانگین مربعات خطا (MSE) در سری ارزیابی می‌باشد.

۳-۱-۷-۱- انتخاب تعداد لایه های پنهان

متاسفانه هیچ گونه نتایج تئوری قابل دسترس یا قوانین تجربی رضایت بخش که در تعیین تعداد لایه های پنهان به ما کمک کند، وجود ندارد با این وجود برای اغلب کاربردهای شبکه عصبی در شیمی، به نظر می رسد که یک لایه مخفی کافی باشد[۳۳]. به این ترتیب شبکه عصبی تشکیل شده، یک شبکه سه لایه شامل یک لایه ورودی، یک لایه پنهان و یک لایه خروجی است که تعداد ورودی ها (توصیفگرها) با تعداد نرون های لایه ورودی برابر بوده و لایه خروجی نیز دارای یک نرون است که نشان دهنده ی pIC_{50} متناظر با هر ترکیب می باشد. ولی تعداد نرون های لایه پنهان باید بهینه شوند.

۳-۱-۷-۲- انتخاب تعداد ورودی‌های شبکه، نوع تابع آموزش، تابع انتقال و تعداد نرون‌های لایه پنهان

برای بهینه سازی تعداد ورودی های شبکه، نوع تابع آموزش، نوع تابع انتقال و تعداد گره های لایه پنهان، شبکه با ورودی های از ۲ تا ۶ توصیفگر ایجاد شد و با الگوریتم های آموزشی بایزین و لونبرگ-مارکوارت با تعداد متفاوت گره های لایه پنهان (۲ تا ۱۰) آموزش داده شد. در حالی که برای انتخاب بهترین تابع انتقال در لایه پنهان از توابع تانژانت سیگموئیدی(tansig) و لگاریتم سیگموئیدی(logsig) به عنوان تابع انتقال استفاده شد. در تمامی این شبکه ها تعداد دورهای آموزش، ثابت و برابر ۱۰۰ در نظر گرفته شد. همچنین از توابع انتقال خطی (purelin) در لایه خروجی استفاده گردید. در روند بهینه سازی پارامترهای فوق، به حداقل رساندن میانگین مربع خطاهای حاصل از بررسی سری ارزیابی به عنوان معیار انتخاب شد. نتایج بهینه سازی این پارامترها در جداول(۳-۵)، (۳-۶)، (۳-۷)، (۳-۸) گردآوری شده است.

جدول (۳-۵)- مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی لونبرگ-مارکوات و تابع انتقال لگاریتم سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)					
تعداد گره های لایه پنهان		۲	۳	۴	۵	۶
	۲	۰/۳۶۰	۰/۳۱۹	۰/۲۴۵	۰/۲۹۹	۰/۱۸۴
	۳	۰/۶۷۲	۱/۶۲۴	۰/۳۲۸	۰/۱۸۲	۴/۷۵۰
	۴	۰/۳۴۹	**۰/۱۵۳**	۰/۷۷۷	۱/۶۷۲	۰/۴۰۹
	۵	۰/۳۵۴	۰/۴۵۸	۰/۳۷۴	۰/۹۵۰	۰/۴۸۷
	۶	۱۰/۹۶۴	۰/۸۱۸	۰/۸۸۰	۰/۳۲۴	۴/۷۶۷
	۷	۰/۳۱۶	۰/۶۲۳	۱/۴۲۲	۱۴/۲۴۴	۲۰/۳۱۵
	۸	۱۲/۱۴۷	۰/۸۱۸	۷/۲۵۸	۱۳/۶۵۰	۱۶/۰۴۲
	۹	۰/۵۸۰	۶/۵۲۴	۹/۴۵۸	۲۷/۹۵۴	۱۲/۶۱۷
	۱۰	۰/۴۵۷	۱/۵۳۲	۱/۵۱۵	۱۵/۲۵۹	۲۶/۴۵۸

جدول (۳-۶)-مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی لونبرگ-مارکوات وتابع انتقال تانژانت سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)					
تعداد گره های لایه پنهان		۲	۳	۴	۵	۶
	۲	۰/۳۶۷	۰/۳۶۹	۰/۳۳۷	**۰/۱۵۱**	۰/۲۶۰
	۳	۰/۴۵۹	۰/۲۹۹	۰/۴۰۴	۰/۲۴۷	۰/۳۴۰
	۴	۰/۴۱۰	۰/۳۴۵	۰/۷۴۰	۰/۳۵۰	۵/۷۰۶
	۵	۰/۵۳۵	۰/۹۷۹	۱/۸۱۱	۲/۵۸۷	۵/۳۴۵
	۶	۰/۴۵۶	۱/۴۱۴	۲/۵۸۷	۱/۶۷۵	۱/۱۱۷
	۷	۰/۴۷۵	۳/۲۵۴	۰/۸۱۹	۲/۸۸۶	۱/۸۹۸
	۸	۰/۶۹۱	۰/۸۰۹	۰/۴۵۹	۵/۰۹۹	۱/۶۶۷
	۹	۱/۸۴۵	۳/۲۶۳	۱/۵۸۸	۶/۵۰۰	۸/۱۰۶
	۱۰	۰/۶۸۸	۲/۷۶۴	۱/۰۳۱	۴/۱۵۱	۸/۹۰۳

جدول (۳-۷)-مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی تنظیم بایزین وتابع انتقال لگاریتم سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)					
تعداد گره های لایه پنهان		۲	۳	۴	۵	۶
	۲	۰/۳۴۶	۰/۳۵۱	۰/۲۷۶	۰/۱۴۸	۰/۱۹۲
	۳	۰/۳۵۱	۰/۳۵۳	۰/۲۷۶	۰/۱۵۹	۰/۲۱۱
	۴	۰/۳۵۸	۰/۳۵۱	۰/۲۶۶	**۰/۱۴۳**	۰/۱۹۸
	۵	۰/۳۳۷	۰/۳۵۳	۰/۲۶۶	۰/۱۴۲	۰/۱۹۸
	۶	۰/۳۳۷	۰/۳۵۵	۰/۲۶۶	۰/۱۵۰	۰/۱۹۴
	۷	۰/۳۵۶	۰/۳۵۷	۰/۲۶۶	۰/۱۴۹	۰/۱۹۲
	۸	۰/۳۵۵	۰/۳۵۷	۰/۲۶۶	۰/۱۴۷	۰/۱۹۲
	۹	۰/۳۵۷	۰/۳۵۷	۰/۲۶۶	۰/۱۴۸	۰/۱۹۲
	۱۰	۰/۳۵۶	۰/۳۵۸	۰/۲۶۶	۰/۱۵۲	۰/۱۹۰

جدول (۳-۸)-مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی تنظیم بایزین وتابع انتقال

تعداد گره های لایه پنهان	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)					
		۲	۳	۴	۵	۶
	۲	۰/۳۴۷	۰/۳۵۷	۰/۲۷۷	۰/۱۶۰	۰/۱۸۷
	۳	۰/۳۵۸	۰/۳۵۶	۰/۲۶۵	۰/۱۵۳	۰/۲۰۱
	۴	۰/۳۵۸	۰/۳۵۷	۰/۲۶۵	۰/۱۴۸	۰/۱۹۷
	۵	۰/۳۵۸	۰/۳۵۷	۰/۲۶۵	۰/۱۴۹	۰/۱۹۵
	۶	۰/۳۵۷	۰/۳۵۷	۰/۲۶۵	۰/۱۵۰	۰/۱۹۵
	۷	۰/۳۵۷	۰/۳۵۷	۰/۲۶۵	۰/۱۴۸	۰/۱۹۳
	۸	۰/۳۵۶	۰/۳۵۷	۰/۲۶۵	۰/۱۴۹	۰/۱۹۳
	۹	۰/۳۵۷	۰/۳۵۷	۰/۲۶۵	**۰/۱۴۵**	۰/۱۹۳
	۱۰	۰/۳۵۶	۰/۳۵۷	۰/۲۶۵	۰/۱۵۵	۰/۱۹۳

تانژانت سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

طبق نتایج موجود در جدول (۳-۹) میانگین مربعات خطا (MSE) برای الگوریتم آموزشی تنظیم بایزین نسبت به الگوریتم آموزشی لونبرگ-مارکوارت کمتر است. همچنین تانژانت سیگموئیدی نسبت به لگاریتم سیگموئیدی خطای بیشتری دارد. در نتیجه تابع لگاریتم سیگموئیدی به عنوان تابع انتقال لایه پنهان برگزیده شده است. طبق نتایج به دست آمده از جداول، تعداد ۵ توصیفگر به عنوان تعداد بهینه ورودی ها برای شبکه و تعداد ۵ گره نیز به عنوان تعداد بهینه گره ها در لایه پنهان انتخاب شدند.

جدول (۳-۹) توابع و پارامترهای بهینه شده شبکه های بهینه (SR-ANN) بدست آمده

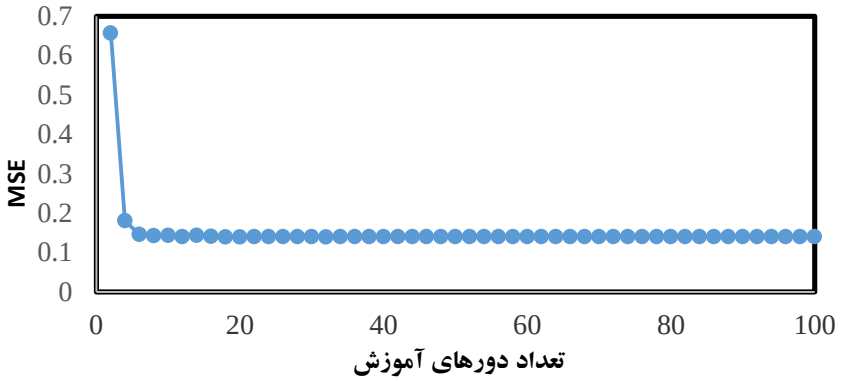
تابع انتقال	تابع آموزش	تعداد توصیف کننده	تعداد نرون لایه پنهان	تعداد دور آموزش	MSE
لگاریتم سیگموید	**تنظیم بایزین**	**۵**	**۴**	**۱۰۰**	**۰/۱۴۳**
لگاریتم سیگموید	لونبرگ-مارکوارت	۳	۴	۱۰۰	۰/۱۵۳
تانژانت سیگموید	تنظیم بایزین	۵	۹	۱۰۰	۰/۱۴۰
تانژانت سیگموید	لونبرگ-مارکوارت	۵	۲	۱۰۰	۰/۱۵۱

۳-۱-۷-۳- انتخاب تعداد دورهای آموزش

در طی آموزش شبکه، دقت تقریبی ANN برای داده های سری آموزش افزایش می یابد در حالی که توانایی پیش بینی برای ورودی های خارجی به شدت کاهش یافته که این حالت را آموزش بیش از حد می گویند و برای جلوگیری از این حالت، از تکنیک توقف زود رس استفاده می‌شود. بدین منظور با ۵ توصیفگر منتخب و قرار دادن ۵ گره در لایه پنهان، در الگوریتم آموزشی تنظیم بایزین وتابع انتقال لگاریتم سیگموئیدی تعداد چرخه آموزش تغییر داده شد و برای هر چرخه، میانگین مربع خطاهای سری ارزیابی برآورد شد (جدول(۳-۱۰)). سپس منحنی تعداد دورهای آموزش بر حسب میانگین مربع خطاها بر طبق شکل (۳-۲) ترسیم شد. مقدار بهینه تعداد دورهای آموزش جایی است که میزان خطا حداقل باشد. در نهایت مشاهده گردید تعداد دورهای آموزش ۲۰ دارای کمترین میانگین مربع خطا می باشد که به عنوان مقدار بهینه انتخاب شد.

جدول(۳-۱۰)-مقادیر میانگین مربعات خطای سری ارزیابی در تعداد دورهای آموزش مختلف

میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار
۰/۱۴۰۱	۸۲	۰/۱۴۰۲	۶۲	۰/۱۴۰۱	۴۲	۰/۱۴۰۶	۲۲	۰/۶۵۷۶	۲
۰/۱۴۰۲	۸۴	۰/۱۴۰۱	۶۴	۰/۱۴۰۲	۴۴	۰/۱۴۰۱	۲۴	۰/۱۸۱۳	۴
۰/۱۴۰۱	۸۶	۰/۱۴۰۱	۶۶	۰/۱۴۰۲	۴۶	۰/۱۴۰۰	۲۶	۰/۱۴۶۰	۶
۰/۱۴۰۱	۸۸	۰/۱۴۰۱	۶۸	۰/۱۴۰۱	۴۸	۰/۱۴۰۴	۲۸	۰/۱۴۲۸	۸
۰/۱۴۰۱	۹۰	۰/۱۴۰۲	۷۰	۰/۱۴۰۲	۵۰	۰/۱۴۰۲	۳۰	۰/۱۴۳۴	۱۰
۰/۱۴۰۱	۹۲	۰/۱۴۰۱	۷۲	۰/۱۴۰۱	۵۲	۰/۱۳۹۸	۳۲	۰/۱۴۰۳	۱۲
۰/۱۴۰۲	۹۴	۰/۱۴۰۱	۷۴	۰/۱۴۰۱	۵۴	۰/۱۴۰۲	۳۴	۰/۱۴۳۹	۱۴
۰/۱۴۰۱	۹۶	۰/۱۴۰۱	۷۶	۰/۱۴۰۲	۵۶	۰/۱۴۰۲	۳۶	۰/۱۴۰۸	۱۶
۰/۱۴۰۱	۹۸	۰/۱۴۰۱	۷۸	۰/۱۴۰۱	۵۸	۰/۱۴۰۲	۳۸	۰/۱۳۹۷	۱۸
۰/۱۴۰۱	۱۰۰	۰/۱۴۰۱	۸۰	۰/۱۴۰۱	۶۰	۰/۱۴۰۱	۴۰	**۰/۱۳۹۶**	**۲۰**



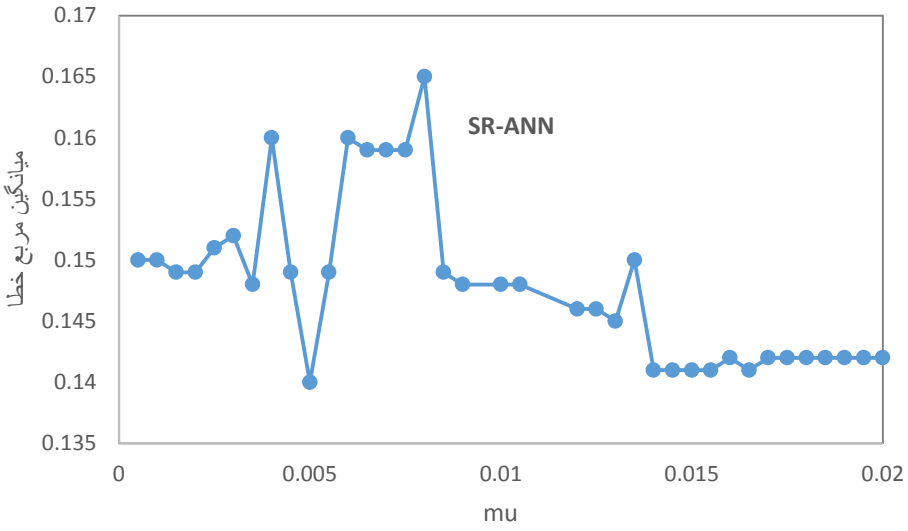
شکل (۳-۲)- منحنی تعداد دور های آموزش بر حسب میانگین مربع خطاها

۳-۱-۷-۴- بهینه کردن مقدار mu

جهت بهینه کردن مقدار mu، ساختار شبکه با ۵ متغیر ورودی، ۵ گره در لایه پنهان و الگوریتم آموزشی تنظیم بایزین و تابع انتقال لگاریتم سیگموئیدی در نظر گرفته شد. سپس مقدار mu از ۰/۰۰۰۵ تا ۰/۰ ۲ با گام های ۰/۰۰۰۵ تغییر داده شد و آنگاه برای هر مورد مقدار میانگین مربع خطای سری ارزیابی محاسبه گردید. در نهایت مقدار میانگین مربع خطا بر حسب mu رسم گردید. نقطه ای که کمترین خطا را داشت به عنوان مقدار بهینه انتخاب شد که طبق جدول (۳-۱۱) و شکل (۳-۳) مقدار بهینه ۰/۰۰۵ بدست آمد.

جدول(۳-۱۱)- مقادیر میانگین مربع خطاها برای سری ارزیابی در مقادیر مختلف mu

mu	میانگین مربع خطا	mu	میانگین مربع خطا	mu	میانگین مربع خطا
۰/۰۰۰۵	۰/۱۵۰	۰/۰۰۷۵	۰/۱۵۹	۰/۰۱۶	۰/۱۴۲
۰/۰۰۱	۰/۱۵۰	۰/۰۰۸	۰/۱۶۵	۰/۰۱۶۵	۰/۱۴۱
۰/۰۰۱۵	۰/۱۴۹	۰/۰۰۸۵	۰/۱۴۹	۰/۰۱۷	۰/۱۴۲
۰/۰۰۲	۰/۱۴۹	۰/۰۰۹	۰/۱۴۸	۰/۰۱۷۵	۰/۱۴۲
۰/۰۰۲۵	۰/۱۵۱	۰/۰۱	۰/۱۴۸	۰/۰۱۸	۰/۱۴۲
۰/۰۰۳	۰/۱۵۲	۰/۰۱۰۵	۰/۱۴۸	۰/۰۱۸۵	۰/۱۴۱
۰/۰۰۳۵	۰/۱۴۸	۰/۰۱۲	۰/۱۴۶	۰/۰۱۹	۰/۱۴۲
۰/۰۰۴	۰/۱۶۰	۰/۰۱۲۵	۰/۱۴۶	۰/۰۱۹۵	۰/۱۴۲
۰/۰۰۴۵	۰/۱۴۹	۰/۰۱۳	۰/۱۴۵	۰/۰۲	۰/۱۴۲
۰/۰۰۵	**۰/۱۴۰**	۰/۰۱۳۵	۰/۱۵۰		
۰/۰۰۵۵	۰/۱۴۹	۰/۰۱۴	۰/۱۴۱		
۰/۰۰۶	۰/۱۶۰	۰/۰۱۴۵	۰/۱۴۱		
۰/۰۰۶۵	۰/۱۵۹	۰/۰۱۵	۰/۱۴۱		
۰/۰۰۷	۰/۱۵۹	۰/۰۱۵۵	۰/۱۴۱		



شکل(۳-۳)- نمودار میانگین مربع خطا بر حسب مقادیر مختلف mu

توابع و مقادیر بهینه شده پارامترهای شبکه عصبی در جدول (۳-۱۱) ارائه شده است.

جدول (۳-۱۲) توابع و پارامترهای بهینه شده شبکه عصبی

تابع آموزش	trainbr
تابع انتقال لایه ی پنهان	logsig
تابع انتقال لایه ی خروجی	pureline
تعداد نرون لایه ی پنهان	۴
تعداد متغیرهای ورودی	۵
تعداد دورهای آموزش	۲۰
پارامترmu	۰/۰۰۵

با استفاده از مدل ANN بهینه شده مقادیر فعالیتهای بازدارندگی ترکیبات مورد نظر محاسبه گردید و نتایج به دست آمده در پیوست (پ-۱) نشان داده شده است.

۳-۱-۸- مدل سازی به روش ماشین بردار پشتیبان (SVM)

روش غیر خطی دیگری که برای مدل سازی در این پایان نامه استفاده شد، روش ماشین بردار پشتیبان است. برای انجام این روش، داده‌ها به دو دسته آموزش و تست تقسیم شدند. داده‌های سری تست همان داده‌های سری تست در رگرسیون خطی چندگانه و شبکه‌ی عصبی مصنوعی هستند. برای آموزش SVM باید پارامترهای آن که شامل تابع حساسیت وپنایک، پارامتر موازنه‌ی C و پارامتر کرنل بهینه گردد. لازم به ذکر است که در ماشین بردار پشتیبان استفاده شده در این پایان نامه از کرنل پایه شعاعی گوسین استفاده گردید. برای بهینه کردن هریک از این سه پارامتر، دو پارامتر ثابت نگه داشته می‌شود و پارامتر سوم در بازه‌ی مورد نظر تغییر داده می‌شود تا جایی که MSE کمترین مقدار شود. در این صورت مقدار آن پارامتر، به عنوان مقدار بهینه انتخاب می‌شود.

بهینه سازی سه پارامتر ماشین بردار پشتیبان با استفاده از توصیفگرهای انتخاب شده به روش SR انجام گرفت که نتایج آن در جدول (۳-۱۳) گزارش شده است. که این نتایج نشان می‌دهند که ماشین بردار پشتیبان با مقادیر ۵ برای تابع حساسیت وپنایک، ۷۰ برای پارامتر موازنه‌ی C و ۰/۱ برای پارامتر کرنل برای پیش بینی pIC_{50}، با استفاده از توصیفگرهای منتخب SR می‌باشد. مقادیر فعالیت‌های بازدارندگی ترکیبات مورد نظر در سری آموزش و تست مورد محاسبه قرار گرفته و در پیوست (پ-۱) نشان داده شده است.

جدول (۳-۱۳)- مقادیر مختلف پارامترهای ماشین بردار پشتیبان با استفاده توصیفگرهای منتخب SR و MSE آنها

پارامتر موازنه‌ی C	تابع حساسیت وپنایک	پارامتر کرنل	MSE
۵۰	۱	۰/۱	**۰/۲۶۰۱**
۵۰	۱	۰/۲	۰/۲۸۰٤
۵۰	۱	۰/۳	۰/۲۸۱۰
۵۰	۱	۰/٤	۰/۲۹۳۸
۵۰	۱	۰/۵	۰/۳۱۳۲
۵۰	۱	۰/٦	۰/۳۱۵٦
۵۰	۱	۰/۷	۰/۳۲٤۷
۵۰	۱	۰/۸	۰/۳۳۵۵
۵۰	۱	۰/۹	۰/۳۷۳۹
۵۰	۱	۱	۰/٤۳۰۹
۵۰	۱	/۱	۰/۳۰۵۳
۵۰	۲	۰/۱	۰/۲۸۷۳
۵۰	۳	۰/۱	۰/۲۷۹۳
۵۰	٤	۰/۱	۰/۲٦٦٦
۵۰	**۵**	۰/۱	**۰/۲۵٤۱**
۵۰	٦	۰/۱	۰/۲٦۸۱
۵۰	۷	۰/۱	۰/۲۷۲٦
۵۰	۸	۰/۱	۰/۲۷۵۳
۵۰	۹	۰/۱	۰/۲۷۹۱
۵۰	۱۰	۰/۱	۰/۲۸۵۳
۱۰	۵	/۱	۰/۳۰۳۲
۲۰	۵	۰/۱	۰/۲۷۸٤
۳۰	۵	۰/۱	۰/۲۷٤٦
٤۰	۵	۰/۱	۰/۲۷۳۳
۵۰	۵	۰/۱	۰/۲۷۲۱
٦۰	۵	۰/۱	۰/۲٦۷۵
۷۰	۵	۰/۱	**۰/۲٤۷۰**
۸۰	۵	۰/۱	۰/۲٦۸۳
۹۰	۵	۰/۱	۰/۲٦۹۵
۱۰۰	۵	۰/۱	۰/۲۷۱۳

۳-۱-۹- انتخاب توصیفگر به روش الگوریتم ژنتیک

در فصل دوم مفاهیم و اصول الگوریم ژنتیک مطرح گردید. همانطور که پیشتر توضیح داده شد، هدف از اجرای الگوریتم حاضر شناسایی ترکیبی از توصیفگرهاست، به گونه ای که برای پیش بینی pIC_{50} مشتقات هموپیپرازین، دی آمین و ۳-آمینو پیرولیدین بوسیله رگرسیون خطی چندگانه، شبکه عصبی مصنوعی و ماشین بردار پشتیبان مورد استفاده قرار گیرند. در فرایند اجرای این الگوریتم، ابتدا به صورت اتفاقی یک جمعیت اولیه از توصیف کننده‌ها تولید و برازندگی یا شایستگی تک تک اعضای

هر نسل یعنی توصیف کننده‌ها، محاسبه و با توجه به شایستگی‌ها، نسل های بعدی با اعمال سه عملگر جمعیت، پیوند و جهش تولید یا باز ترکیب می‌شوند. این روند جستجو برای جمعیت‌های مختلف تا حصول ملاک خاتمه ادامه می‌یابد. لازم به ذکر است که برنامه استفاده شده در الگوریتم ژنتیک یک برنامه بهینه شده می‌باشد. مقادیر بهینه شده‌ی پارامترهای جمعیت، پیوند و جهش به ترتیب ۳۰، ۰/۵ و ۰/۰۱ می باشد. با هر بار اجرای برنامه ۱۰۰ کروموزوم که بهترین پاسخ را بدهند انتخاب می شوند. تا اتمام روند الگوریتم ژنتیک n بار برنامه اجرا می‌شود. در نیمه اول اجرای برنامه بردار Yیا تابع متغیرهای وابسته همان مقادیر اصلی در نظر گرفته شده و در نیمه دوم یک بردار Yبه روش اتفاقی تولید می‌شود. حال فرض کنیم برنامه ۴۰ بار اجرا شود، در پایان ۲ بردار خواهیم داشت که اولی شامل میانگین ۲۰ اجرای اول الگوریتم بر روی مقادیر اصلی و دومی شامل میانگین ۲۰ اجرای دوم الگوریتم بر روی مقادیر تصادفی خواهد بود. به طور ساده می‌توان گفت که اجرای الگوریتم ژنتیک با بردار اصلی نشان‌دهنده‌ی توانایی الگوریتم مورد نظر برای مدل سازی اطلاعات همراه با مقادیر اشتباه است و اجرای الگوریتم برای بردار Yتصادفی نیز نشان‌دهنده‌ی توانایی الگوریتم برای مدلسازی مقادیر اشتباه می‌باشد. حال تفاوت این دو می‌تواند توانایی الگوریتم ژنتیک را در ارتباط با مقادیر حقیقی نشان دهد.

حسن عملکرد الگوریتم ژنتیک به وسیله ریشه متوسط مربع خطا (RMSE) در پیش بینی مقادیر محاسبه می گردد. لازم به ذکر است که پارامتر RMSE از رابطه‌ای (۳-۱) محاسبه میشود:

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}}{n} \qquad (۳-۱)$$

در فرمول ارائه شده n تعداد مقادیر موجود در سری مورد محاسبه است. شش توصیفگر انتخاب شده توسط الگوریتم ژنتیک برای سری داده ها در جدول(۳-۱۲) آورده شده است.

جدول (۳-۱۴)- توصیفگرهای انتخاب شده توسط الگوریتم ژنتیک

Descriptor	meaning	MF[a]
X0A	Average connectivity index chi-0	۱/۴۳۳
X5A	Average connectivity index chi-5	-۰/۴۲۰
MDDD	Mean distance degree deviation	-۰/۱۵۶
PCWTe	Partial charge weighted topological electronic descriptor	۰/۰۹۸
RDF145m	Radial distribution function – 14.5 / weighted by atomic masses	-۰/۰۰۵
RDF065p	Radial distribution function – 6.5 / weighted by atomic polarizabilities	۰/۰۵۰

[a]Mean Effect

۳-۱-۱۰- مدل رگرسیون خطی چندگانه با استفاده از توصیفگرهای انتخاب شده توسط الگوریتم ژنتیک

پس از انتخاب مناسب ترین توصیفگرها توسط الگوریتم ژنتیک، مدل سازی بین توصیفگرها و فعالیتهای بازدارندگی مشتقات هموپیپرازین، دی آمین و ۳-آمینو پیرولیدین برای سری آموزش با استفاده از روش MLR انجام و مدل خطی زیر بدست آمد:

$$\mathbf{pIC_{50}} = -16.25\ (\pm 5.382) + 44.168\ (\pm 6.261)\ \mathbf{X0A} - 112.295\ (\pm 17.837)\ \mathbf{X5A} - 0.095\ (\pm 0.015)\ \mathbf{MDDD} + 0.092\ (\pm 0.014)\ \mathbf{PCWTe} - 0.241\ (\pm 0.049)\ \mathbf{RDF145m} + 0.062\ (\pm 0.014)\ \mathbf{RDF065p}$$

سپس از معادله بدست آمده برای پیش بینی فعالیت سری تست استفاده گردید.مقادیر واقعی و پیش بینی شده فعالیت‌ها برای تمام ترکیبات در پیوست (پ-۱) آورده شده است.

۳-۱-۱۱-مدل شبکه عصبی مصنوعی با استفاده از توصیفگرهای انتخاب شده توسط الگوریتم ژنتیک

در این بخش نیز از شبکه عصبی سه لایه متشکل از یک لایه ورودی، یک لایه پنهان و یک لایه خروجی استفاده شد.

۳-۱-۱۱-۱- انتخاب تعداد گره های لایه پنهان، تعداد متغیر ورودی، نوع تابع آموزش و نوع تابع انتقال

پس از طراحی اولیه شبکه، برای این که شبکه قدرت پیش بینی مناسبی داشته باشد باید بطور مناسب آموزش داده شود و با توجه به اینکه فرایند آموزش وقتگیر است، انتخاب الگوریتم آموزشی سریعتر امری مهم است. از میان الگوریتم های آموزشی متفاوت که در جعبه ابزار متلب وجود دارد، الگوریتم لونبرگ-مارکوارت و تنظیم بایزین دارای سرعت مناسب، قابلیت پیش بینی و تعمیم پذیری بهتری نسبت به سایر روش های معمول مانند روش نزول گرادیانی می‌باشند. بر این اساس شبکه با ورودی های از ۲ تا ۶ توسط دو الگوریتم آموزشی لونبرگ-مارکوارت و تنظیم بایزین، با تعداد متفاوت گره در لایه پنهان از ۲ تا ۱۰ و همچنین توابع لگاریتمی سیگوئیدی(logsig) و تانژانت سیگموئیدی (tansig)، بعنوان توابع انتقال لایه پنهان، آموزش داده شد. در تمامی این شبکه‌ها تعداد دورهای آموزش ثابت و برابر ۱۰۰ در نظر گرفته شد. همچنین از تابع انتقال خطی (purelin) در لایه خروجی استفاده شد. معیار نیز به حداقل رساندن مربع خطا (MSE) برای سری ارزیابی در نظر گرفته شد که نتایج این محاسبات در جداول (۳-۱۵)، (۳-۱۶)، (۳-۱۷) و (۳-۱۸) آورده شده است.

جدول (۳-۱۵)-مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی لونبرگ-مارکوات و تابع انتقال لگاریتم سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)					
تعداد گره های لایه پنهان		۲	۳	۴	۵	۶
	۲	۰/۳۹۰	۰/۳۴۵	۰/۳۳۵	۰/۲۳۳	**۰/۱۲۵**
	۳	۰/۴۸۰	۰/۳۶۴	۰/۲۴۲	۰/۲۱۳	۰/۲۵۹
	۴	۰/۵۹۸	۲/۱۹۳	۱/۰۶۴	۰/۷۷۹	۰/۵۳۳
	۵	۰/۶۲۵	۰/۸۵۵	۰/۳۲۶	۱/۸۲۵۰	۳/۴۰۰
	۶	۱/۰۹۸	۱/۱۷۲	۰/۸۸۶	۱/۴۷۷	۲/۵۵۲
	۷	۰/۲۷۴	۱/۸۸۵	۸/۵۳۵	۱/۷۵۱	۰/۶۷۶
	۸	۱/۳۴۶	۲/۵۰۳	۴/۱۲۰	۷/۴۸۵	۴/۷۲۵
	۹	۱/۰۳۲	۱/۳۹۶	۲/۵۴۲	۸/۳۲۸	۳/۹۱۹
	۱۰	۰/۴۲۹	۱/۱۵۱	۱/۵۲۶	۲/۸۲۴	۱/۶۰۲

جدول (۳-۱۶)-مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی لونبرگ-مارکوات و تابع انتقال تانژانت سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)					
تعداد گره های لایه پنهان		۲	۳	۴	۵	۶
	۲	۰/۳۴۹	۰/۳۴۸	۰/۲۶۳	**۰/۱۲۴**	۰/۴۸۰
	۳	۰/۴۱۹	۰/۳۶۷	۳/۸۳۱	۰/۶۱۸	۰/۱۴۹
	۴	۰/۴۲۷	۰/۳۶۷	۰/۶۰۱	۱/۵۸۹	۰/۴۵۹
	۵	۰/۵۲۶	۵/۴۷۶	۰/۴۸۸	۲/۹۵۲	۳/۶۸۸
	۶	۲/۵۷۱	۰/۹۰۵	۷/۴۶۰	۲/۹۵۲	۰/۴۰۸
	۷	۰/۵۲۱	۸/۴۳۸	۰/۴۸۸	۶/۹۴۶	۰/۲۳۷
	۸	۳/۰۸۰	۹/۶۷۷	۷/۴۶۰	۹/۷۶۵	۱/۰۴۰
	۹	۱۰/۸۷۵	۱/۰۱۴	۴/۴۵۳	۹/۷۶۵	۸/۴۳۶
	۱۰	۰/۷۸۰	۰/۹۸۶	۱/۶۲۲	۱/۹۸۰	۱/۹۱۷

جدول (۳-۱۷)-مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی تنظیم بایزین و تابع انتقال لگاریتم سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)					
تعداد گره های لایه پنهان		۲	۳	۴	۵	۶
	۲	۰/۳۶۴	۰/۲۷۶	۰/۲۳۹	۰/۲۲۴	**۰/۱۶۳**
	۳	۰/۳۹۴	۰/۴۵۱	۰/۲۳۶	۰/۲۳۶	۰/۱۹۹
	۴	۰/۴۰۶	۰/۴۴۶	۰/۲۳۶	۰/۲۴۵	۰/۲۱۹
	۵	۰/۳۹۴	۰/۴۸۷	۰/۲۳۶	۰/۲۶۲	۰/۱۶۸
	۶	۰/۳۹۷	۰/۴۹۶	۰/۲۴۱	۰/۲۴۷	۰/۲۰۹
	۷	۰/۳۹۵	۰/۵۰۵	۰/۲۳۶	۰/۲۴۰	۰/۲۱۳
	۸	۰/۳۹۵	۰/۵۱۰	۰/۲۴۲	۰/۲۴۸	۰/۱۷۱
	۹	۰/۳۹۵	۰/۵۲۳	۰/۲۴۱	**۰/۱۳۵**	۰/۲۱۹
	۱۰	۰/۳۹۶	۰/۵۱۳	۰/۲۴۲	۰/۲۷۳	۰/۱۷۱

جدول (۳-۱۸)-مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی تنظیم بایزین و تابع انتقال تانژانت سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)					
تعداد گره های لایه پنهان		۲	۳	۴	۵	۶
	۲	۰/۳۶۳	۰/۲۷۹	۰/۲۳۸	۰/۲۴۸	**۰/۱۵۸**
	۳	۰/۳۹۵	۰/۴۳۲	۰/۲۳۸	۰/۲۵۲	۰/۲۱۰
	۴	۰/۳۹۵	۰/۴۶۰	۰/۲۳۹	۰/۲۶۲	۰/۱۸۹
	۵	۰/۳۹۵	۰/۴۷۲	۰/۲۳۹	۰/۲۶۲	۰/۱۷۱
	۶	۰/۳۹۵	۰/۴۷۷	۰/۲۴۲	۰/۲۶۱	۰/۲۱۰
	۷	۰/۳۹۶	۰/۴۸۱	۰/۲۳۸	۰/۲۶۲	۰/۲۳۸
	۸	۰/۳۹۶	۰/۴۸۵	۰/۲۴۲	۰/۲۶۲	۰/۱۷۲
	۹	۰/۳۹۶	۰/۴۹۲	۰/۲۴۱	۰/۲۵۸	۰/۱۷۴
	۱۰	۰/۳۹۶	۰/۴۸۹	۰/۲۴۲	۰/۲۶۲	۰/۱۷۵

با توجه به نتایج بدست آمده در جدول (۳-۱۹)، الگوریتم آموزشی لونبرگ-مارکوارت به عنوان تابع آموزش شبکه انتخاب گردید. همچنین تابع تانژانت سیگموئیدی دارای خطای کمتری نسبت به لگاریتم سیگموئیدی است. بنابراین تابع انتقال تانژانت سیگموئیدی به عنوان تابع انتقال لایه پنهان

استفاده شد. طبق نتایج بدست آمده ۵ توصیفگر به عنوان توصیفگرهایی که بیشترین ارتباط را با فعالیت بازدارندگی دارند، انتخاب شدند. تعداد ۲ گره نیز در لایه پنهان انتخاب شد.

جدول (۳-۱۹)- توابع و پارامترهای بهینه شده شبکه های بهینه (SR-ANN) بدست آمده

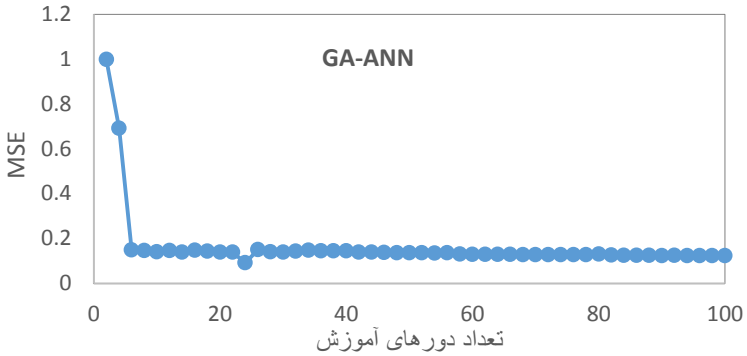
تابع انتقال	تابع آموزش	تعداد توصیف کننده	تعداد نرون لایه پنهان	تعداد دور آموزش	MSE
لگاریتم سیگموید	تنظیم بایزین	۵	۹	۱۰۰	۰/۱۳۵
لگاریتم سیگموید	لونبرگ-مارکوارت	۶	۲	۱۰۰	۰/۱۲۵
تانژانت سیگموید	تنظیم بایزین	۶	۲	۱۰۰	۰/۱۵۸
تانژانت سیگموید	**لونبرگ-مارکوارت**	**۵**	**۲**	**۱۰۰**	**۰/۱۲۴**

۳-۱-۱۱-۲- انتخاب تعداد دورهای آموزش

بدین منظور و جهت انتخاب تعداد دورهای بهینه، شبکه با الگوریتم لونبرگ-مارکوارت، ۵ توصیفگر به عنوان ورودی و ۲ گره در لایه پنهان و تابع انتقال تانژانت سیگموئیدی در تعداد دورهای متفاوت، آموزش داده شد و میانگین مربع خطاها برای سری ارزیابی ثبت گردید. جدول(۳-۲۰) و شکل (۳-۴) نشان می دهد که مقدار میانگین مربع خطا (MSE) در تعداد دورهای آموزش ۲۴ به حداقل می رسد. بنابراین ۲۴ بار تکرار به عنوان تکرارهای بهینه برای یاد گیری شبکه انتخاب شد.

جدول(۳-۲۰)- مقادیر میانگین مربع خطاها برای سری ارزیابی در تعداد دورهای آموزش مختلف

میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار
۰/۱۲۵۶	۸۲	۰/۱۲۹۶	۶۲	۰/۱۳۹۸	۴۲	۰/۱۳۹۴	۲۲	۰/۹۹۸۳	۲
۰/۱۲۵۱	۸۴	۰/۱۲۹۰	۶۴	۰/۱۳۸۶	۴۴	**۰/۰۹۱۹**	**۲۴**	۰/۶۹۰۹	۴
۰/۱۲۴۹	۸۶	۰/۱۲۸۵	۶۶	۰/۱۳۸۴	۴۶	۰/۱۵۰۱	۲۶	۰/۱۴۹۰	۶
۰/۱۲۴۵	۸۸	۰/۱۲۸۰	۶۸	۰/۱۳۶۵	۴۸	۰/۱۴۱۲	۲۸	۰/۱۴۶۸	۸
۰/۱۲۴۱	۹۰	۰/۱۲۸۱	۷۰	۰/۱۳۶۱	۵۰	۰/۱۳۹۴	۳۰	۰/۱۴۱۲	۱۰
۰/۱۲۴۲	۹۲	۰/۱۲۸۱	۷۲	۰/۱۳۵۹	۵۲	۰/۱۴۳۵	۳۲	۰/۱۴۶۸	۱۲
۰/۱۲۴۱	۹۴	۰/۱۲۸۰	۷۴	۰/۱۳۵۵	۵۴	۰/۱۴۷۱	۳۴	۰/۱۳۹۶	۱۴
۰/۱۲۴۱	۹۶	۰/۱۲۷۵	۷۶	۰/۱۳۶۵	۵۶	۰/۱۴۵۵	۳۶	۰/۱۴۷۸	۱۶
۰/۱۲۴۰	۹۸	۰/۱۲۷۱	۷۸	۰/۱۳۰۲	۵۸	۰/۱۴۵۱	۳۸	۰/۱۴۳۲	۱۸
۰/۱۲۴۰	۱۰۰	۰/۱۳۰۰	۸۰	۰/۱۲۹۶	۶۰	۰/۱۴۵۱	۴۰	۰/۱۳۸۹	۲۰



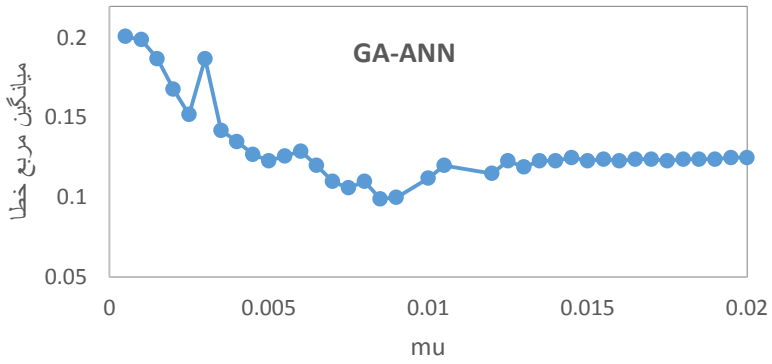
شکل (۳-۴)-منحنی تعداد چرخه های آموزش بر حسب میانگین مربع خطاها

۳-۱-۱۱-۳- بهینه کردن مقدار mu

جهت بهینه کردن مقدار mu، ساختار شبکه با ۵ متغیر ورودی، ۲ گره در لایه پنهان و الگوریتم آموزشی لونبرگ-مارکوارت و تابع انتقال تانژانت سیگموئیدی در نظر گرفته شد. سپس مقدار mu از ۰/۰۰۰۵ تا ۰/۰۲ با گام های ۰/۰۰۰۵ تغییر داده شد و آنگاه برای هر مورد مقدار میانگین مربع خطای سری ارزیابی محاسبه گردید. در نهایت مقدار میانگین مربع خطا بر حسب mu رسم گردید. نقطه ای که کمترین خطا را داشت به عنوان مقدار بهینه انتخاب شد که طبق جدول (۳-۲۱) و شکل (۳-۵) مقدار بهینه ۰/۰۰۸۵ بدست آمد.

جدول(۳-۲۱)- مقادیر میانگین مربع خطاها برای سری ارزیابی در مقادیر مختلف mu

mu	میانگین مربع خطا	mu	میانگین مربع خطا	mu	میانگین مربع خطا
۰/۰۰۰۵	۰/۲۰۱	۰/۰۰۷۵	۰/۱۰۶	۰/۰۱۶	۰/۱۲۳
۰/۰۰۱	۰/۱۹۹	۰/۰۰۸	۰/۱۱۰	۰/۰۱۶۵	۰/۱۲۴
۰/۰۰۱۵	۰/۱۸۷	**۰/۰۰۸۵**	**۰/۰۹۹**	۰/۰۱۷	۰/۱۲۴
۰/۰۰۲	۰/۱۶۸	۰/۰۰۹	۰/۱۰۰	۰/۰۱۷۵	۰/۱۲۴
۰/۰۰۲۵	۰/۱۵۲	۰/۰۱	۰/۱۱۲	۰/۰۱۸	۰/۱۲۳
۰/۰۰۳	۰/۱۸۷	۰/۰۱۰۵	۰/۱۲۰	۰/۰۱۸۵	۰/۱۲۴
۰/۰۰۳۵	۰/۱۴۲	۰/۰۱۲	۰/۱۱۵	۰/۰۱۹	۰/۱۲۴
۰/۰۰۴	۰/۱۳۵	۰/۰۱۲۵	۰/۱۲۳	۰/۰۱۹۵	۰/۱۲۵
۰/۰۰۴۵	۰/۱۲۷	۰/۰۱۳	۰/۱۱۹	۰/۰۲	۰/۱۲۵
۰/۰۰۵	۰/۱۲۳	۰/۰۱۳۵	۰/۱۲۳		
۰/۰۰۵۵	۰/۱۲۶	۰/۰۱۴	۰/۱۲۳		
۰/۰۰۶	۰/۱۲۹	۰/۰۱۴۵	۰/۱۲۵		
۰/۰۰۶۵	۰/۱۲۰	۰/۰۱۵	۰/۱۲۳		
۰/۰۰۷	۰/۱۱۰	۰/۰۱۵۵	۰/۱۲۴		



شکل(۳-۵)- نمودار میانگین مربع خطا بر حسب مقادیر مختلف mu

توابع و مقادیر بهینه شده پارامترهای مختلف شبکه عصبی در جدول (۳-۲۲) ارائه شده است.

جدول (۳-۲۲)- توابع و پارامترهای بهینه شده شبکه عصبی

تابع آموزش	trainlm
تابع انتقال لایه ی پنهان	tansig
تابع انتقال لایه ی خروجی	pureline
تعداد نرون لایه ی پنهان	۲
تعداد متغیرهای ورودی	۵
تعداد دورهای آموزش	۲۴
پارامتر mu	۰/۰۰۸۵

با استفاده از مدل ANN بهینه شده مقادیر فعالیتهای بازدارندگی ترکیبات مورد نظر در مجموعه های آموزش، ارزیابی و پیش‌بینی مورد محاسبه قرار گرفت و در پیوست (پ-۱) نشان داده شده است.

۳-۱-۱۲- مدل ماشین بردار پشتیبان با استفاده از توصیفگرهای انتخاب شده توسط الگوریتم ژنتیگ

برای آموزش SVM باید پارامترهای آن که شامل تابع حساسیت وپنایک، پارامتر موازنه‌ی C و پارامتر کرنل بهینه گردد. برای بهینه کردن هریک از این سه پارامتر، دو پارامتر ثابت نگه داشته می-

شود و پارامتر سوم در بازه‌ی مورد نظر تغییر داده می‌شود تا جایی که MSE کمترین مقدار شـود. در این صورت مقدار آن پارامتر، به عنوان مقدار بهینه انتخاب می‌شود. بهینه سازی سـه پـارامتر ماشـین بردار پشتیبان با استفاده از توصیفگرهای انتخاب شـده بـه روش GA انجـام گرفـت کـه نتـایج آن در جدول (۳-۲۳) گزارش شده است. که این نتایج نشان می‌دهند که ماشین بردار پشتیبان با مقـادیر ۷ برای تابع حساسیت وپنایک، ۸۰ برای پارامتر موازنه‌ی C و ۰/۵ برای پارامتر کرنل بـرای پـیش بینـی pIC_{50}، با استفاده از توصیفگرهای منتخب GA می‌باشد. مقادیر فعالیت‌های بازدارندگی ترکیبات مورد نظر در سری آموزش و تست مورد محاسبه و در پیوست (پ-۱) نشان داده شده است.

جدول (۳-۲۳)- مقادیر مختلف پارامترهای ماشین بردار پشتیبان با استفاده توصیفگرهای منتخب GA و MSE آنها

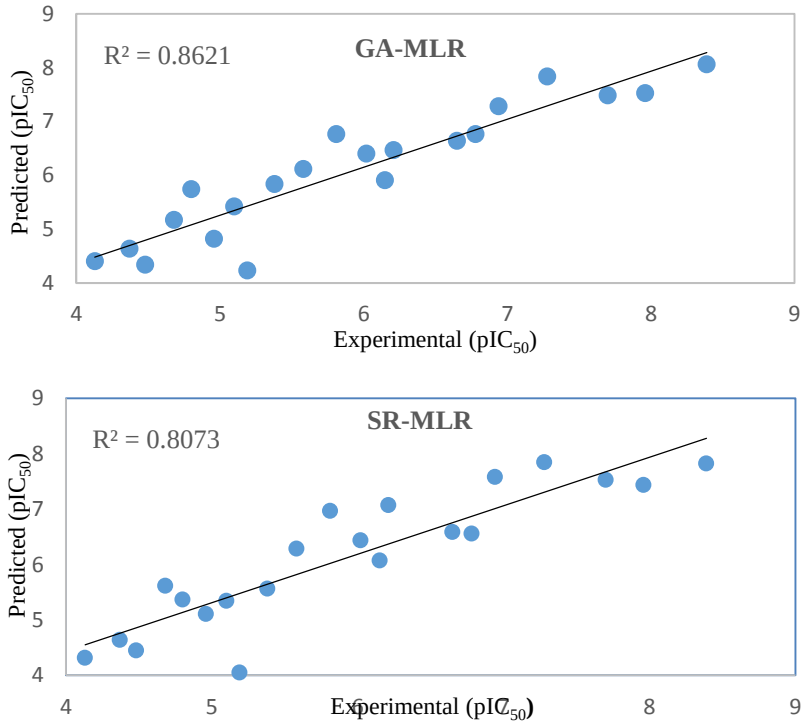
پارامتر موازنه‌ی C	تابع حساسیت وپنایک	پارامتر کرنل	MSE
۵۰	۱	۰/۱	۰/۲۴۸۸
۵۰	۱	۰/۲	۰/۲۵۱۳
۵۰	۱	۰/۳	۰/۲۴۹۴
۵۰	۱	۰/۴	۰/۲۴۸۶
۵۰	۱	**۰/۵**	**۰/۲۳۸۷**
۵۰	۱	۰/۶	۰/۲۴۸۵
۵۰	۱	۰/۷	۰/۲۸۲۷
۵۰	۱	۰/۸	۰/۳۲۱۴
۵۰	۱	۰/۹	۰/۳۷۵۶
۵۰	۱	۱	۰/۴۴۰۱
۵۰	۱	۰/۵	۰/۲۸۹۳
۵۰	۲	۰/۵	۰/۲۸۰۱
۵۰	۳	۰/۵	۰/۲۷۸۷
۵۰	۴	۰/۵	۰/۲۷۴۳
۵۰	۵	۰/۵	۰/۲۶۶۸
۵۰	۶	۰/۵	۰/۲۵۹۰
۵۰	**۷**	۰/۵	**۰/۲۵۸۲**
۵۰	۸	۰/۵	۰/۲۶۲۴
۵۰	۹	۰/۵	۰//۲۶۴۶
۵۰	۱۰	۰/۵	۰/۲۶۱۸
۱۰	۷	۰/۵	۰/۳۵۴۳
۲۰	۷	۰/۵	۰/۲۹۹۶
۳۰	۷	۰/۵	۰/۲۷۳۶
۴۰	۷	۰/۵	۰/۲۶۰۸
۵۰	۷	۰/۵	۰/۲۵۵۸
۶۰	۷	۰/۵	۰/۲۴۸۰
۷۰	۷	۰/۵	۰/۲۴۱۴
۸۰	۷	۰/۵	**۰/۲۳۸۷**
۹۰	۷	۰/۵	۰/۲۴۸۸
۱۰۰	۷	۰/۵	۰/۲۴۹۸

۳-۲-بررسی نتایج

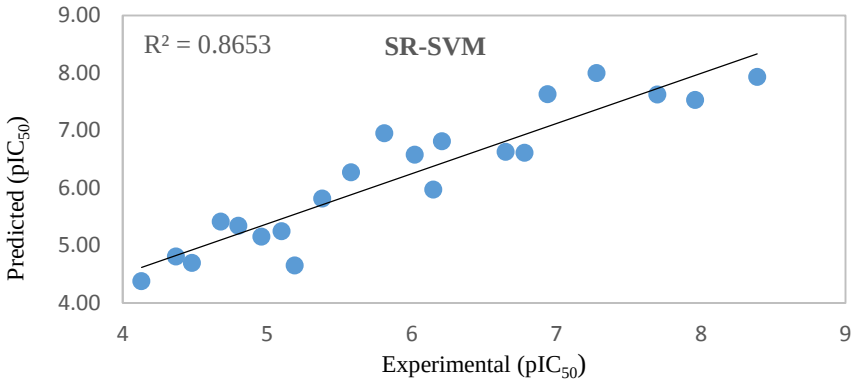
اعتبار و اهمیت معادلات پیش بینی و مدل های ساخته شده وقتی مشخص می گردد که فعالیت مولکول های که در سری تست هستند را به‌خوبی و بطور رضایت بخش و قابل قبول پیش بینی کند. در اینجا چندین روش به منظور ارزیابی توانایی مدل های ارائه شده در پیش بینی مقادیر pIC_{50} ذکر شده است.

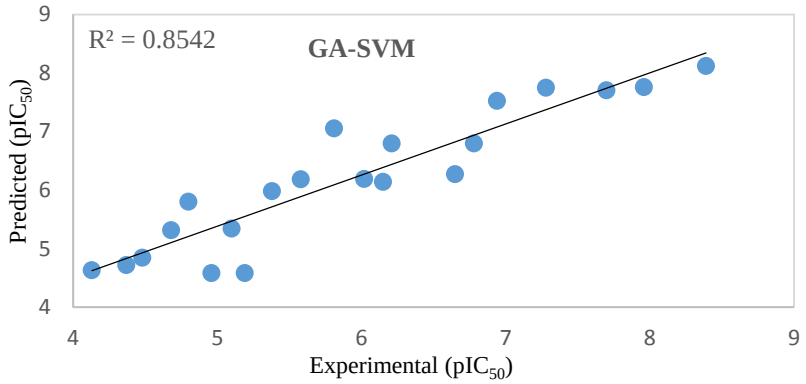
۳-۲-۱- ارزیابی مدل ها با استفاده از نمودار برگشتی

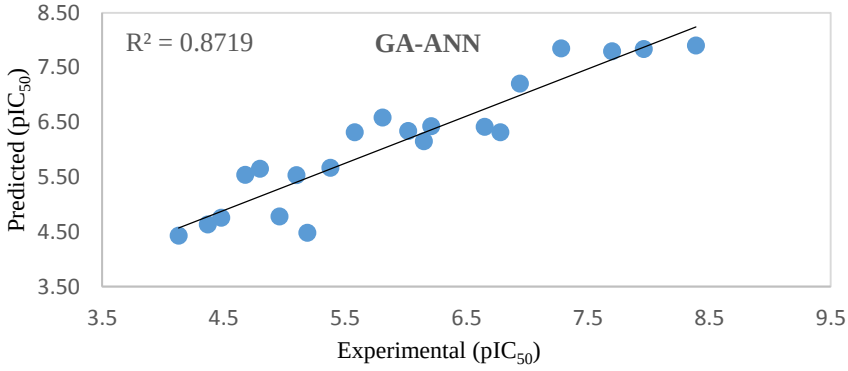
در نمودار برگشتی مقادیر پیش بینی شده برحسب مقادیر تجربی رسم می‌گردد و با توجه به مقدار ضریب تعیین (R^2) به دست آمده از نمودار، پراکندگی نقاط در اطراف خط برگشت تعیین می-شود. هر چه مقدار ضریب تعیین به یک نزدیک تر باشد، مدل ساخته شده، مدل بهتری است. نتایج حاصل از نمودارهای برگشتی مربوط به شش روش انجام شده در این کار برای سری ارزیابی و تست در شکل های (۳-۶)، (۳-۷) و (۳-۸) آورده شده است.

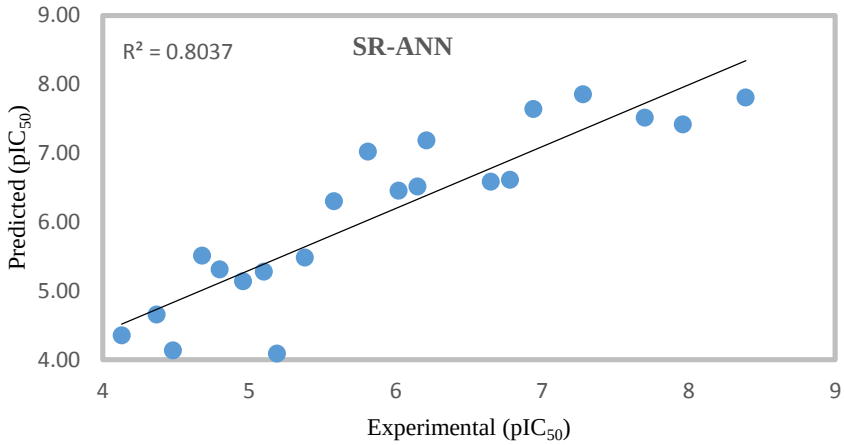


شکل (۳-۶)- نمودار مقادیر پیش بینی شده pIC_{50} بر حسب مقادیر تجربی برای سری تست

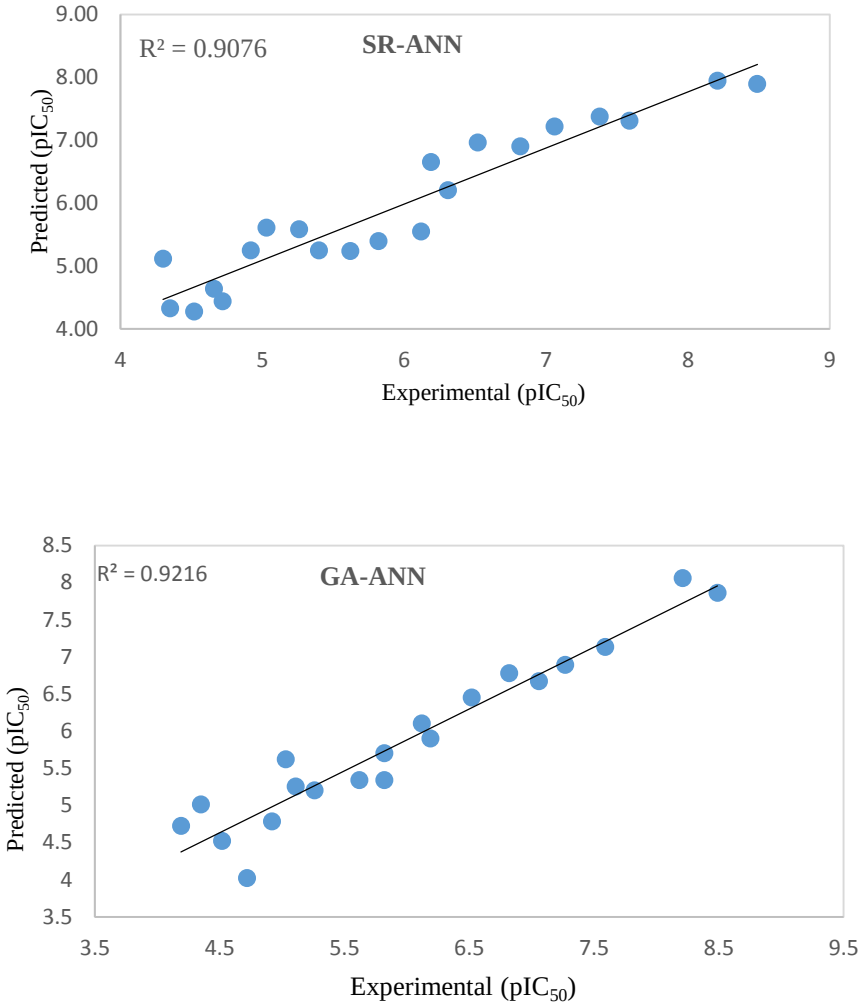








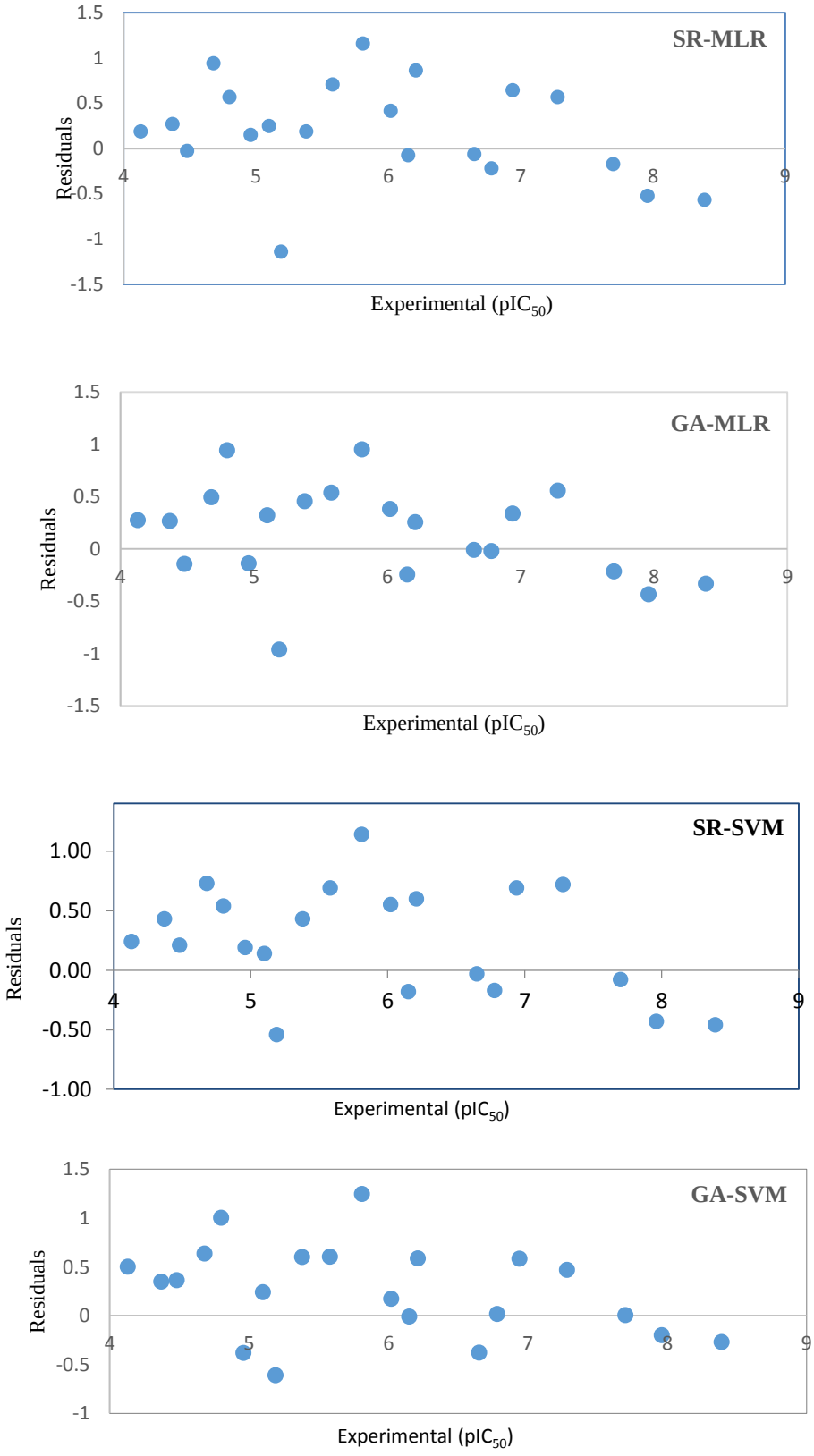
شکل (۳-۷)- نمودار مقادیر پیش بینی شده pIC_{50} بر حسب مقادیر تجربی برای سری تست



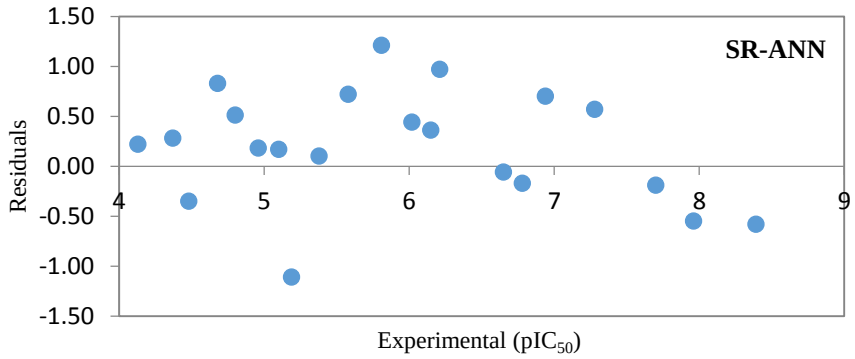
شکل (۳-۸)- نمودار مقادیر پیش بینی شده pIC_{50} بر حسب مقادیر تجربی برای سری ارزیابی

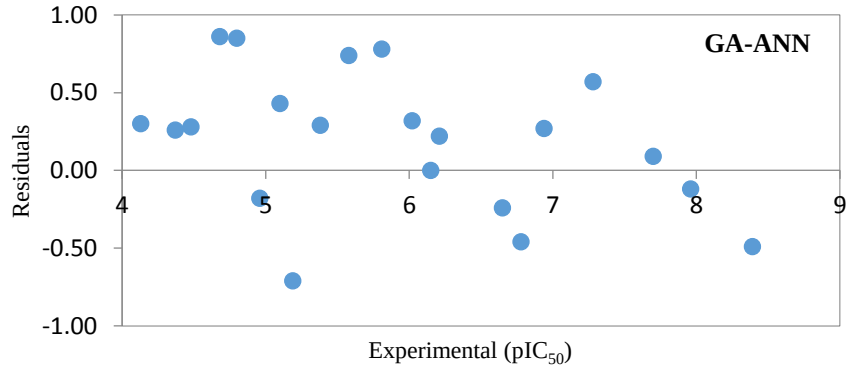
۳-۲-۲- ارزیابی مدل‌ها با استفاده از نمودار خطای باقیمانده

اختلاف مقادیر پیش بینی شده و مقادیر تجربی، خطای باقیمانده نامیده می‌شود. پراکندگی یکنواخت نقاط، حول محور افقی که بیانگر خطای باقیمانده‌ی صفر است، نشان‌دهنده‌ی آن است که خطای سیستماتیکی در مدل سازی وجود ندارد. نمودار خطای باقیمانده بر حسب مقادیر تجربی، برای مدل‌های ذکر شده در شکل‌های (۳-۹)، (۳-۱۰) و (۳-۱۱) نشان داده شده است.

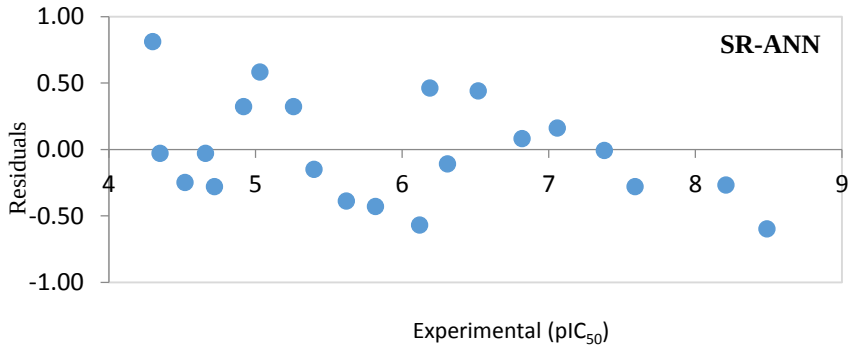


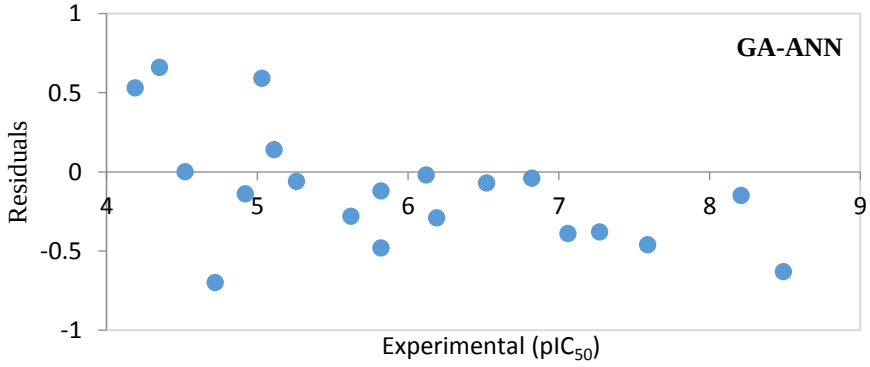
شکل (۳-۹)- نمودار مقادیر باقیمانده pIC50 بر حسب مقادیر تجربی برای سری تست





شکل (۳-۱۰)- نمودار مقادیر باقیمانده pIC50 بر حسب مقادیر تجربی برای سری تست





شکل (۳-۱۱)- نمودار مقادیر باقیمانده pIC50 بر حسب مقادیر تجربی برای سری ارزیابی

۳-۲-۳-ارزیابی مدل های برتر خطی و غیر خطی با استفاده از پارامترهای آماری

مطابق جدول (۳-۲۴) شش پارامتر آماری، جهت ارزیابی توانایی پیش بینی مدل های ساخته شده به روش های SR-MLR، SR-ANN،SR-SVM، GA-MLR، GA-ANN و GA-SVM به کار گرفته شد. نتایج حاصل از جدول (۳-۲۴) نشان می دهد که بیشترین مقدار R^2 در سری تست برای مدل GA-SVM می‌باشد. این نتایج نشان دهنده آن است که ارتباط برخی از توصیفگرها با فعالیت دارویی می‌تواند به صورت غیر خطی باشد.

جدول(۳-۲۴)- پارامترهای آماری برای مدل های انتخاب شده

		SR-MLR	SR-ANN	SR-SVM	GA-MLR	GA-ANN	GA-SVM
MSE	سری آموزش	۰/۱۳۸	۰/۱۴۲	۰/۰۳۳	۰/۱۲۸	۰/۱۴۷	۰/۱۳۸
	سری ارزیابی	-	۰/۱۴۲	-	-	۰/۱۴۷	-
	سری تست	۰/۳۲۸	۰/۳۴۶	۰/۲۶۷	۰/۲۲۸	۰/۲۲۵	۰/۲۵۷
PRESS	سری آموزش	۱۱/۳۲۱	۸/۶۹۴	۲/۷۲۶	۱۰/۵۰۹	۹/۰۱۶	۱۱/۳۲
	سری ارزیابی	-	۲/۹۸۵	-	-	۲/۲۳۹	-
	سری تست	۶/۸۹۷	۷/۲۷۰	۵/۵۹۷	۴/۸۷۸	۴/۷۴۰	۵/۴۱۳
SEP	سری آموزش	۰/۳۷۱	۰/۳۷۷	۳/۱۲	۰/۳۵۸	۰/۳۸۴	۰/۳۷۱
	سری ارزیابی	-	۰/۳۷۷	-	-	۰/۳۸۳	-
	سری تست	۰/۵۷۳	۰/۵۸۸	۰/۵۱۶	۰/۴۷۷	۰/۴۷۵	۰/۵۰۷
MAE	سری آموزش	۰/۲۹۷	۰/۳۰۵	۰/۱۳۹	۰/۲۷۹	۰/۲۹۶	۰/۳۲۵
	سری ارزیابی	-	۰/۳۱۲	-	-	۰/۳۰۶	-
	سری تست	۰/۴۶۱	۰/۴۸۹	۰/۴۳۷	۰/۳۹۴	۰/۴۰۲	۰/۴۴۸
AAD	سری آموزش	۵/۳۸۶	۵/۵۱۹	۲/۴۶۹	۵/۰۱۶	۵/۲۹۳	۵/۷۲۷
	سری ارزیابی	-	۵/۵۱۰	-	-	۵/۴۵۶	-
	سری تست	۷/۹۹۲	۸/۶۸۷	۷/۶۴۷	۷/۰۴۲	۷/۲۷۸	۸/۰۶
R^2	سری آموزش	۰/۸۸۶	۰/۸۷۱	۰/۹۷۲	۰/۸۹۴	۰/۸۶۶	۰/۹۲۰
	سری ارزیابی	-	۰/۹۰۷	-	-	۰/۹۲۱	-
	سری تست	۰/۸۰۷	۰/۸۰۳	۰/۸۶۵	۰/۸۶۲	۰/۸۷۱	۰/۸۷۷

۳-۲-۴-ارزیابی مدل ها توسط روش رد مرحله ای تک تک و گروهی

به منظور بررسی بیشتر قدرت پیش بینی مدل های خطی و غیر خطی تکنیک رد مرحله ای تک تک و گروهی مورد استفاده قرار گرفت. در روش رد مرحله ای تک تک، هر بار یکی از ترکیبات به طور تصادفی از سری داده‌ها حذف شدند و در روش رد مرحله‌ای گروهی، هر بار یک گروه از ترکیبات(۵ ترکیب) به طور تصادفی ازسری داده‌ها حذف شدند. سپس با استفاده از مدل ساخته شده توسط بقیه ترکیبات، فعالیت دارویی ترکیب یا ترکیبات حذف شده، پیش بینی شدند. این فرایند برای تمام اعضای سری داده ها تکرار شد. نتایج حاصل از رد مرحله ای و گروهی در جدول(۳-۲۵) ارائه شده است.

جدول(۳-۲۵)- پارامترهای آماری برای مدل های انتخاب شده

		SR-MLR	SR-ANN	SR-SVM	GA-MLR	GA-ANN	GA-SVM
Q^2_{LOO}	کل داده ها	۰/۸۵۷	۰/۸۶۰	۰/۸۶۱	۰/۸۶۹	۰/۸۷۷	۰/۸۷۱
Q^2_{LGO}	کل داده ها	۰/۸۳۷	۰/۸۳۹	۰/۸۴۲	۰/۸۳۲	۰/۸۴۳	۰/۸۴۹

۳-۲-۵- ارزیابی مدل های ارائه شده با استفاده از آزمون Y-تصادفی

این تکنیک ارزیابی مدل، با هدف بررسی هر گونه ارتباط تصادفی بین داده ها انجام شد. در این آزمون، متغیر وابسته بطور تصادفی بهم ریخته شد. مدل QSAR جدید با استفاد از ماتریکس متغیرهای مستقل اصلی و مقادیر تصادفی از متغیر وابسته توسعه یافت. اگر در مدل اصلی هیچ گونه ارتباط تصادفی وجود نداشته باشد، تفاوت قابل توجهی بین مقدار ضریب تعیین مدل اصلی و مدل QSAR که با پاسخ تصادفی توسعه یافته، وجود خواهد داشت. نتایج حاصل از چندین بار اجرای آزمون Y-تصادفی در جدول(۳-۲۶) نشان داده شده است. مقادیر کوچک ضریب تعیین (R^2) بیانگر عدم ارتباط شانسی در مدل توسعه یافته توسط رگرسیون خطی چندگانه، شبکه عصبی مصنوعی و ماشین بردار پشتیبان می‌باشد.

جدول (۳-۲۶)- نتایج حاصل از ۱۰ بار آزمون Y-تصادفی

تکرار	R^2_{test}					
	SR-MLR	SR-ANN	SR-SVM	GA-MLR	GA-ANN	GA-SVM
۱	۰/۰۷۸	۰/۲۱۰	۰/۱۹۸	۰/۰۹۹	۰/۱۲۵	۰/۰۰۸
۲	۰/۰۵۱	۰/۰۸۰	۰/۰۰۸	۰/۰۲۶	۰/۰۸۴	۰/۲۱۴
۳	۰/۰۸۳	۰/۱۸۰	۰/۰۱۴۸	۰/۰۶۸	۰/۰۹۶	۰/۱۴۷
۴	۰/۱۱۹	۰/۰۱۴۸	۰/۰۹۸	۰/۰۸۷	۰/۲۱۳	۰/۰۸۵
۵	۰/۰۵۰	۰/۰۱۴	۰/۱۶۲	۰/۰۸۰	۰/۱۰۵	۰/۰۹۶
۶	۰/۰۷۲	۰/۰۰۹	۰/۰۲۸	۰/۰۷۷	۰/۰۸۶	۰/۱۴۳
۷	۰/۰۸۴	۰/۰۱۷۴	۰/۰۰۹	۰/۰۳۶	۰/۰۰۸	۰/۰۵۹
۸	۰/۰۳۳	۰/۱۰۹	۰/۱۴۱	۰/۰۵۵	۰/۱۴۰	۰/۰۹۷
۹	۰/۰۱۷	۰/۰۶۱	۰/۰۸۲	۰/۰۶۰	۰/۰۸۹	۰/۰۷۷
۱۰	۰/۰۶۲	۰/۰۴۸	۰/۰۱۹	۰/۱۲۵	۰/۰۲۷	۰/۱۰۸

۳-۲-۶- بررسی ارتباط توصیفگرهای وارد شده در مدل با اثر بازدارندگی

در این بخش به طور خلاصه ارتباط بین توصیفگرهای وارد شده در مدل و اثر بازدارندگی ترکیبات، مورد بررسی قرار خواهد گرفت. با توجه به نتایج به دست آمده در مدل برتر، توصیفگرهای انتخاب شدند که هر کدام بیانگر خصوصیات متفاوتی از مولکول مورد بررسی است. توصیفگرهای انتخاب شده توسط روش رگرسیون مرحله‌ای شامل (X0A، PCWTe، Jhetp، RDF065m، RDF145v وH-052) و توصیفگرهای انتخاب شده توسط الگوریم ژنتیک شامل (X0A، X5A، MDDD، PCWTe، RDF145m و RDF065p) می‌باشند.

۳-۲-۶-۱- توصیفگرهای RDF[۱]

می‌توان گفت RDF یا تابع توزیع شعاعی مربوط به یک دسته از اتم ها، معادل توزیع احتمال یافتن یک اتم در فضای کروی به شعاع R است. رابطه (۳-۱) نحوه محاسبه تابع RDF را نشان می-دهد.

$$g(R) = \int \sum_{i}^{N-1} \sum_{j>i}^{N} A_i . A_j . e^{-B.(R-r_{ij})^2} . \qquad (۱-۳)$$

۱-Radial Distribution Function Descriptors

که f یک فاکتور مقیاس، و N تعداد اتم‌های مولکول است. همچنین r_{ij} فاصله‌ی بین دو اتم i و j و A یک ویژگی اتمی (وزن دار نشده u، جرم اتمی m، حجم واندروالس v، الکترونگاتیویته e، قطبش پذیری p) است. β یک فاکتور تسهیل کننده است که توزیع احتمال فاصله بین اتمی را مشخص می‌کند و می‌توان از آن به فاکتور دما برای تعریف جنبش اتمی تعبیر کرد. g(R) در نقاط گسسته‌ای با فواصل معین محاسبه و با مجموعه‌ای از کدهای RDF با ویژگی‌های مختلف اتمی می‌توان ساختار سه بعدی یک مولکول را به طور واضح توصیف کرد. این توصیفگرها اهمیت توزیع اتم‌ها در مولکول روی اثر بازدارندگی آن نشان می‌دهد. همچنین این توصیفگر علاوه بر فاصله اتمی، اطلاعاتی با ارزشی در مورد فاصله پیوند، انواع حلقه، سیستم‌های مسطح و غیر مسطح و نوع اتم‌ها فراهم می‌کند.

از این گروه، توصیفگرها RDF065m، RDF145v، RDF145m و RDF065p در مدل ارائه شده برای پیش بینی اثر بازدارندگی انتخاب شده‌اند. اثر متوسط توصیفگرهای RDF065m و RDF065p مثبت ذکر شده یعنی با افزایش مقدار این توصیفگرها pIC_{50} افزایش می‌یابد و اثر متوسط توصیفگرها RDF145v و RDF145m منفی ذکر شده یعنی با افزایش مقدار این توصیفگرها، pIC_{50} نیز کاهش می‌یابد.

۳-۲-۶-۲- توصیفگر ACFC[۱]

کدهای اجزای با مزیت اتمی (ACFC) یک کد با مزیت اتمی با دامنه‌ی کوتاه است که هر اتم را با نوع اتم، انواع پیوند و انواع اتم‌های همسایه توصیف می‌کند. هر مولکول کاملا با یک کد اجزا که تعداد اتم‌های غیر هیدروژن است معرفی می‌شود. این توصیفگرها اطلاعات شیمیایی زیادی را در رابطه با گروه‌های عاملی در اطراف یک اتم مرکزی و مولکول ارائه می‌دهند.

۱-Atom-Centered Fragments Code

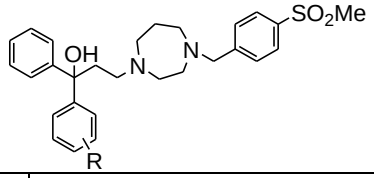
ثابت‌های اتمی آب گریز خوزه-کریپن [۱] (توصیف کننده‌های چربی دوستی): چربی دوستی میزانی از توزیع ترکیبات بین یک فاز آلی و یک فاز آبی است که معمولا با ضرایب جداسازی P به صورت توزیع غلظت یک ترکیب در فازهای آلی و آبی از تحت شرایط تعدلی زیر تعریف می‌شود:

$$P=\frac{[C]_{org}}{[C]_{aq}} \qquad (۲-۳)$$

که $[C]_{org}$ و $[C]_{aq}$، غلظت‌های حل شده در فازهای آلی و آبی هستند.

توصیفگر H-052 (H متصل به کربن با هیبریداسیون sp^3 و عدد اکسایش صفر)، یکی دیگر از توصیفگرهای از نوع اجزای با مرکزیت اتمی می‌باشد که اولین همسایه اتم کربن، اتم هیدروژن می‌باشد همچنین، این توصیفگر نمایانگر تعداد اتم‌های هیدروژن متصل به اتم کربن با هیبریداسیون sp3 می باشد. اثر متوسط این توصیفگر منفی است و نشان‌دهنده‌ی این است که با افزایش مقدار این توصیفگر، pIC_{50} کاهش می‌یابد. در جدول (۳-۲۵) در سری ترکیبات افزایش مقدار این توصیفگر به معنای افزایش تعداد هیدروژن‌های متصل به اتم کربن است که طول زنجیره بلندتر می‌شود و برهمکنش حلال و حل شونده کمتر می‌شود در نتیجه pIC_{50} کاهش می‌یابد.

جدول (۳-۲۷)- مثال هایی از مقدار توصیفگر H-052 بر pIC_{50}



نام ترکیب	مقدار H-052	pIC_{50}
3-OH	۳۴	۶/۱۵
$3\text{-}NH_2$	۳۵	۵/۳۸
3-Me	۳۶	۴/۸۰

۲-Ghose Crippen

۳-۲-۶-۳- توصیفگرهای گروه Topological

این توصیفگرها بر اساس نمایش گراف مولکول می‌باشند. در این گراف‌ها هر نقطه نشان‌دهنده‌ی یک اتم بوده و خطوط بین نقاط نیز نشان‌دهنده‌ی پیوند شیمیایی بین اتم‌ها می‌باشد. معمولا در گراف‌های مولکولی اتم هیدروژن را نشان نمی‌دهند. این توصیفگرها اطلاعاتی راجع به ساختمان، اندازه، شکل، تقارن، شاخه‌دار شدن، نحوه اتصال اتم‌ها و نوع اتم‌های موجود در یک مولکول را در اختیار ما قرار می‌دهند. محاسبه این توصیفگرها به سادگی از روی ساختمان دو بعدی مولکول‌ها امکان پذیر می‌باشد.

از این گروه توصیفگرها در مدل ارائه شده توصیفگرJhetp و MDDD انتخاب شدند. اثر متوسط برای توصیفگرJhetp مثبت است که نشان می‌دهد با افزایش آن مقدار pIC_{50} افزایش یافته و در نتیجه اثر بخشی دارو زیاد می‌شود. و اثر متوسط برای توصیفگرMDDD منفی است که نشان می‌دهد با افزایش آن مقدار pIC_{50} کاهش می‌یابد.

۳-۲-۶-۴- توصیفگرهای Connectivity Indices

یک دسته‌ی مهم از شاخص‌هایی که در دسته‌ی توصیفگرهای توپولوژیکی طبقه‌بندی می‌شود، شاخص‌های ارتباطی می‌باشد. این شاخص‌ها برای اولین بار توسط یک دانشمند کروات به نام راندیک[۱] در سال ۱۹۷۵ بنیان نهاده شد. و بعدها توسط هال [۲] و کایر [۳] تغییراتی روی آنها انجام گرفت و به نام توصیفگرهای مولکولی معروف شدند. آنها شاخصی به نام کاپا را تعریف کردند که با افزایش اندازه

۱-Randic
۲-Hall
۳-Kier

مولکول، مقدار این شاخص بزرگ تر و با افزایش شاخه‌دار شدن مولکول مقدار این شاخص کوچکتر می‌شود.

که از این گروه توصیفگرها، X0A و X5A انتخاب شدند. این شاخص‌های ارتباطی می‌توانند به صورت نسبتا رضایت بخشی بین مولکول‌های پیچیده تمایز قائل شوند. بنابراین در بررسی فعالیت بیولوژیکی و شیمیایی بسیاری از ترکیبات شرکت دارند. اثر متوسط توصیفگر X0A مثبت بوده و نشان‌دهنده‌ی این است که با افزایش مقدار این توصیفگر، pIC_{50} افزایش می‌یابد.و اثر متوسط توصیفگر X5A منفی بوده و نشان‌دهنده‌ی این است که با افزایش مقدار این توصیفگر، pIC_{50}کاهش می‌یابد.

۳-۲-۶-۵- توصیفگرهای Charge

این توصیفگرها اطلاعاتی در مورد بار الکتریکی مولکول و بار جزئی روی اتمها می‌دهند و قادرند اطلاعاتی از جمله ماکزیمم بار مثبت، ماکزیمم بار منفی، کل بارهای مثبت و منفی و ... را نشان دهند. توصیفگر PCWTe در این گروه از توصیفگرها قرار می‌گیرد. این توصیفگر اطلاعاتی راجع به میزان برهم کنش‌های الکتریکی بین آن اتم، یا سایر اتم‌های موجود در مولکول باشد. اثر متوسط توصیفگر PCWTe مثبت بوده و نشان‌دهنده‌ی این است که با افزایش مقدار این توصیفگر، pIC_{50} افزایش می‌یابد.

۳-۲-۷- بررسی میزان مشارکت توصیفگرهای منتخب شبکه عصبی

میزان مشارکت توصیفگرهای منتخب به صورت زیر تعیین شد:

۱- توصیفگر مورد نظر به همراه اوزان مربوطه‌اش از شبکه بهینه حذف گردید.

۲- مقدار متغیر وابسته با استفاده از توصیفگرهای باقیمانده برای هر یک از ترکیبات سری ارزیابی پیش بینی گردید.

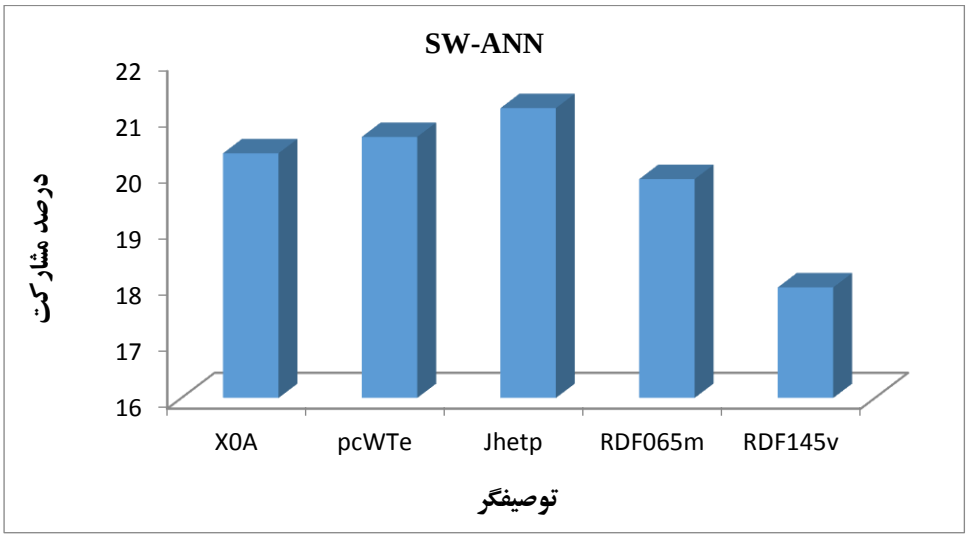
۳- میانگین خطای مطلق (MAE) حاصل از ترکیبات سری ارزیابی محاسبه شد.

۴- مراحل ۱ تا ۳ برای دیگر توصیفگرهای منتخب نیز تکرا شد.

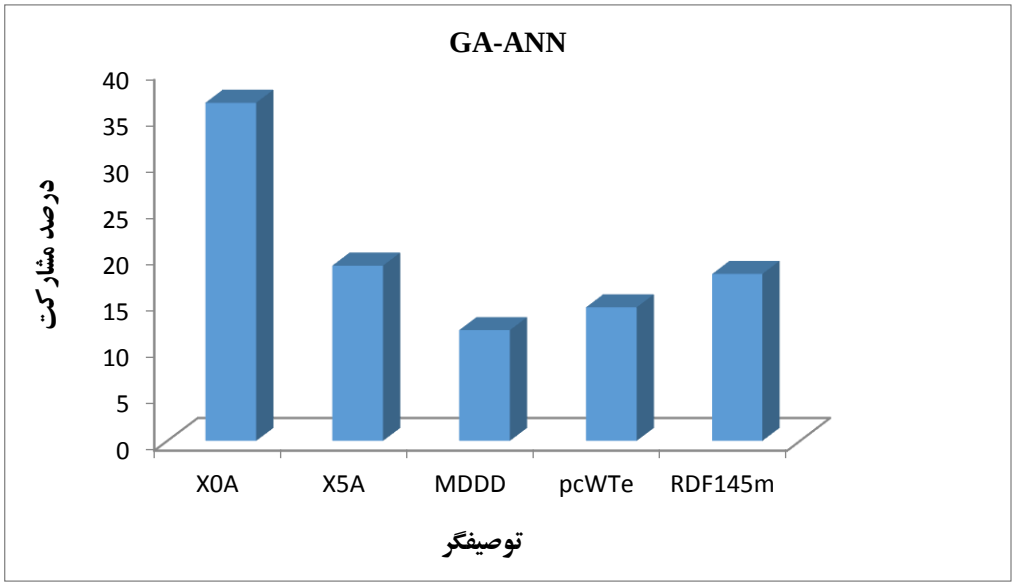
۵- سرانجام درصد مشارکت هر توصیفگر توسط رابطه‌ی (۳-۳) برآورد شد.

$$c_i = 100\frac{\Delta m_i}{\sum_{i=1}^{N}\Delta m_i} \qquad (۳-۳)$$

در این رابطه c_i درصد مشارکت توصیفگر حذف شده iام، N تعداد توصیفگرهای مدل و Δm_i میانگین خطای مطلق حاصل از سری ارزیابی در غیاب توصیفگرiام را نشان می‌دهد. که بر این اساس درصد مشارکت توصیفگرهای منتخب در ترکیبات مورد بررسی در شکل‌های(۳-۱۲) و(۳-۱۳) ارائه شده است.



شکل (۳-۱۲)- درصد مشارکت توصیگرها در مدل بهینه



شکل (۳-۱۳)- درصد مشارکت توصیگرها در مدل بهینه

بر اساس شکل‌های فوق، توصیف کننده‌ی Jhetp در مدل SW-ANN و توصیف کننده‌ی X0A در مدل GA-ANN بیشترین مشارکت را دارا می‌باشند.این توصیفگرها اطلاعاتی راجع به ساختمان، اندازه، شکل، تقارن، شاخه‌دار شدن، نحوه اتصال اتم‌ها و نوع اتم‌های موجود در یک مولکول را در اختیار ما قرار می‌دهند. و با افزایش مقدار این توصیفگرها باعث افزایش pIC_{50} می‌شود.

۳-۳- نتیجه گیری

در این فصل از روش‌های مختلف کمومتریکس برای پیش بینی فعالیت ترکیبات دارویی(IC_{50})، مشتقات هموپیپرازین، دی آمین، ۳-آمینوپیرولیدین، استفاده شده است. از آنجایی که اندازه‌گیری فعالیت دارویی بیشتر ترکیبات به صورت تجربی با صرف هزینه، زمان و پیچیدگی زیاد همراه است، دست یابی به مقادیر IC_{50}، با روش‌های تجربی مقرون به صرفه نیست. بنابراین، پیش بینی آن با استفاده از کمومتریکس از اهمیت بالایی برخوردار است.

برای انتخاب توصیفگرهای مناسب از روش‌های SR و GA استفاده شد. سپس این توصیفگرها برای مدل سازی خطی و غیر خطی مورد استفاده قرار گرفت.

نتایج جدول (۳-۲۴) نشان می‌دهد که از بین شش روش استفاده شده، روش GA-SVM با مقدار ضریب تعیین بیشتر و MSE کمتر، روش مناسبی برای پیش بینی فعالیت ترکیبات دارویی است. در نتیجه ماشین بردار پشتیبان روش کارآمدی برای پیش بینی فعالیت ترکیبات دارویی است.

۳-۴- آینده‌نگری

برای انتخاب توصیفگرهای مناسب و معتبر می‌توان از سایر روش‌ها مثل الگوریتم جستجوی ممنوع[۱]، الگوریتم اجتماع مورچگان[۲] و الگوریتم بهینه سازی جمعیت ذره‌ها[۳] استفاده نمود و نتایج را با این کار مقاسیه کرد.

برای مدل سازی می‌توان از روش‌های حداقل مربعات ماشین‌های بردار پشتیبان (LS-SVM)[۴] و سیستم استنتاج فازی-عصبی سازگار (ANFIS)[۵] استفاده کرد.

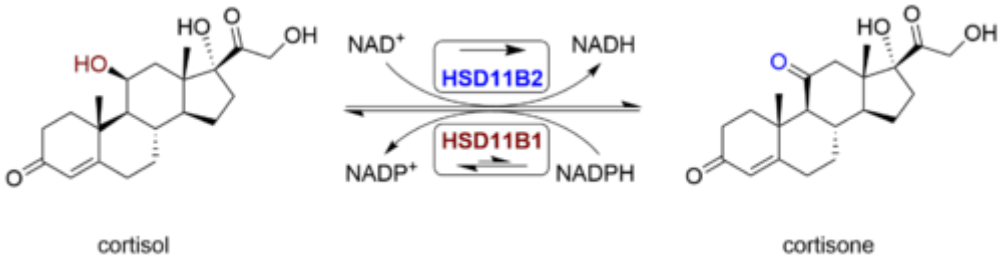
۱- Tabu Search
۲- Ant Colony Optimization
۳- Honey-bee Matikg Optimization Algorithm
۴- Least Square Support Vector Machines
۵- Adaptive neuron-fuzzy inference system

فصل چهارم

مطالعه ارتباط کمی ساختار – فعالیت مشتقات آریل سولفونیل پیپرازین با استفاده روش های خطی و غیر خطی

۴-۱- مقدمه

در بدن انسانها آنزیم 11β-HSD1[۱] عمل تبدیل گلوکوکورتیکوئید کورتیزون غیرفعال را بـه شـکل فعال آن یعنی کورتیزول برعهده دارد. تنظیم مقدار گلوکوکورتیکوئید توسط HPA[۲] صورت می‌گیرد و عمل معکوس تبدیل کورتیزول به کورتیزون توسط 11β-HSD2 صورت می گیرد.



شکل(۴-۱)- برهمکنش میان کورتیزون و کورتیزول توسط 11β-HSD1 و 11β-HSD2[۴۸].

عدم تنظیم گلوکوکورتیکوئیدها در کبد و بافتهای چرب بدن باعث ایجاد بیماری هایی مانند دیابـت و سیندرم می شود. بنابرین تنظیم زیادی میزان کورتیزول توسـط بازدارنـده هـای 11β-HSD1 ممکـن است برای معالجه دیابت ها بسیار مفید باشد. همچنین بازدارنده های 11β-HSD1 مـی تواننـد بـرای معالجه سیندرم و بیماری های قلبی عروقی بکار برده شوند. قدرت بالای تنظیم غلظت کـورتیزول در کبد و بافتهای چرب بدن بدون اثرات سیستماتیکی انتشار غلظت توسط بازدارنـده هـای 11β-HSD1 باعث ایجاد انگیزه در بررسی میزان بازدارندگی آنها گردیده است[۴۸].

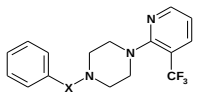
در سال ۲۰۰۲ بارف و همکارانش ۲-آمینوتیازول سولفونامید را به عنـوان اولـین بازدارنـده هـای 11β-HSD1 گزارش کردند. سپس بازدارنده های مختلفی توسط گروههای دیگر گزارش شـد[۴۹]. در این کار تعدادی از ترکیبات آریل سولفونیل پیپرازین‌ها به عنوان بازدارنـده هـای 11β-HSD1 کـه در سال ۲۰۰۸ توسط سان و همکارانش گزارش شده است، توسـط روشـهای کمـومتریکس جهـت مـدل سازی مورد بررسی قرار گرفتند.

۱- 11β-hydroxysteroid dehydrogenase type 1
۲- Hypothalamic–pituitary–adrenal axis

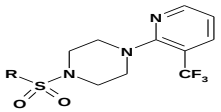
۴-۲- انتخاب سری داده‌ها

سری داده‌ها شامل فعالیت دارویی ۴۲ ترکیب از آریل سولفونیل پیپرازین توسط روش‌های کمومتریکس مورد بررسی قرار گرفت (جدول۴-۱) [۴۸]. در این کار قدرت بازدارندگی این ترکیبات به صورت IC_{50} گزارش شده است. IC_{50} عبارتست از مینیمم غلظتی از ترکیب دارویی که باعث ۵۰٪ اثر بازدارندگی بر روی بیماری می‌شود. این مقادیر به مقیاس لگاریتمی (pIC_{50}) تبدیل شده و مورد استفاده قرار گرفته است.

جدول (۴-۱)- ساختار ترکیبات مورد استفاده و مقدار pIC_{50} آنها

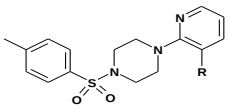
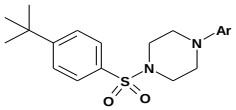


No.	X	pIC_{50}
۱	4-Me-Ph	۷/۸۰
۲	$-SO_2-$	۷/۳۵
۳	-C(O)-	۶/۵۸
۵	$-CH_2-$	۶/۲۶
۶	$-CH_2SO_2-$	۷/۷۰

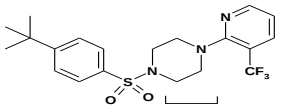
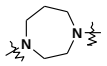
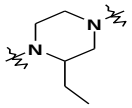
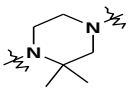
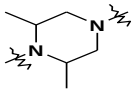
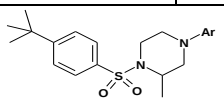


No.	R	pIC_{50}
۷	4-Et-Ph	۷/۷۴
۸	4-Pr-Ph	۷/۸۰
۹	4-*i*pr-Ph	۷/۸۲
۱۰	4-*t*Bu-Ph	۸/۴۰
۱۱	4-F-Ph	۷/۳۴

ادامه جدول (۴-۱)

۱۲	4-Cl-Ph	۷/۸۵
۱۳	4-CF_3-Ph	۷/۸۵
۱۴	4-OMe-Ph	۷/۷۲
۱۵	4-NO_2-Ph	۷/۴۳
۱۶	4-CN-Ph	۷/۳۰
۱۷	4-Ph-Ph	۷/۱۳
۱۸	2- Naphthyl	۷/۸۹
۱۹	3- Naphthyl	۷/۸۹
۲۰	3-Cl-Ph	۷/۲۶
۲۱	3-CF_3-Ph	۷/۱۹
۲۲	2,4-diCl-Ph	۸/۱۵
		
No.	R	pIC_{50}
۲٤	CH_3	٦/۸۲
۲٥	Cl	٦/۹۳
۲٦	NO_2	٦/٥٤
۲۷	CN	٦/۲۰
		
No.	Ar	pIC_{50}
۲۸	2-CF_3-Ph	۷/۸۲
۲۹	2-Me-Ph	۷/٤۹
۳۰	2-Cl-Ph	۷/۰۲
۳۱	(2-F)Ph	٦/۹۷
۳۲	2-NO_2-Ph	۷/٤۳
۳۳	4-NO_2-Ph	۸/۰۰
۳٤	4-Cl-Ph	۷/۰٦
۳٥	4-Me-Ph	٦/٤۱
۳٦	4-F-Ph	۷/۳۸
۳۷	4-OMe-Ph	٦/٦۳

ادامه جدول (۴-۱)

		
No.	Ring	pIC_{50}
۳۸		۸/۵۲
۳۹		۸/۰۵
٤۰		۷/۷۷
٤۱		۷/۲٤
٤۲		۷/۲۱
		
No.	Ar	pIC_{50}
٤۳	R-3-CF_3-2-Pyridyl	۷/۹٦
٤٥	R-4-NO_2-Ph	۷/٥٤

٤-۲-۱- رسم و بهینه سازی ساختار مولکول ها

جهت رسم و بهینه سازی ساختار مولکول ها همانند روش گفته شده در بخش ۳-۱-۲ انجام شد.

۴-۲-۲- محاسبه توصیفگر ها

جهت محاسبه توصیفگرها به وسیله نرم افزار Dragon و به صورت شرح داده شده در بخش ۳-۱-۳ انجام شد.

۴-۲-۳- انتخاب توصیفگرها به روش الگوریتم ژنتیک

لازم به ذکر است که در این بخش جهت انتخاب متغیر، فقط از روش الگوریتم ژنتیک استفاده شده است. در فصل دوم مفاهیم و اصول الگوریم ژنتیک مطرح گردید. همانطور که پیشتر توضیح داده شد، هدف از اجرای الگوریتم حاضر شناسایی ترکیبی از توصیفگرهاست، به گونه ای که برای پیش بینی pIC_{50} مشتقات آریل سولفونیل پیپرازین بوسیله روش‌های MLR، ANN و SVM مورد استفاده قرار گیرند. در فرایند اجرای این الگوریتم، ابتدا به صورت اتفاقی یک جمعیت اولیه از توصیف کننده‌ها تولید می‌شود، برازندگی یا شایستگی تک تک اعضای هر نسل یعنی توصیفگرها، محاسبه و با توجه به شایستگی‌ها، نسل های بعدی با اعمال سه عملگر جمعیت، پیوند و جهش تولید یا باز ترکیب می‌شوند. این روند جستجو برای جمعیت‌های مختلف تا حصول ملاک خاتمه ادامه می‌یابد. لازم به ذکر است که برنامه استفاده شده در الگوریتم ژنتیک یک برنامه بهینه شده می‌باشد. مقادیر بهینه شده‌ی پارامترهای جمعیت، پیوند و جهش به ترتیب ۳۰، ۰/۵ و ۰/۰۱ می باشد. پنج توصیفگر انتخاب شده توسط الگوریتم ژنتیک برای سری داده ها در جدول(۴-۲) آورده شده است.

جدول (۴-۲)- توصیفگرهای انتخاب شده توسط الگوریتم ژنتیک

نشان	نام کامل	اثر متوسط
VRA2	Average connectivity index chi-0	۲/۲۰۰
GATS5e	Average connectivity index chi-5	۲/۲۹۰
RDF040m	Mean distance degree deviation	-۰/۷۹۵
RDF035p	Partial charge weighted topological electronic descriptor	۰/۸٦٤
Te	Radial distribution function – 14.5 / weighted by atomic masses	-۳/٥٦۰

۴-۲-۴- ارزیابی مستقل بودن توصیفگرهای انتخاب شده

به منظور ارزیابی توصیفگرهای انتخاب شده مبنی بر مستقل بودن از همدیگر در جدول (۴-۳) ماتریس همبستگی توصیفگرهای انتخاب شده آورده شده است. همانطور که در این جدول مشاهده می‌شود ضریب همبستگی بین توصیفگرهای انتخاب شده همگی کمتر از ۰/۹ می‌باشد. لذا نتایج جدول نشان می دهد توصیفگرهای انتخاب شده مستقل از یکدیگر هستند.

جدول (۴-۳)- ماتریس ضرایب همبستگی توصیفگرهای انتخاب شده

	VRA2	GATS5e	RDF040m	RDF035p	Te
VRA2	۱				
GATS5e	-۰/۱۴۶	۱			
RDF040m	-۰/۲۹۵	-۰/۲۱۲	۱		
RDF035p	۰/۴۴۱	-۰/۰۳۶	-۰/۲۶۳	۱	
Te	۰/۵۹۳	-۰/۲۳۲	-۰/۲۳۳	۰/۶۷۲	۱

۴-۲-۵- مدل رگرسیون خطی چندگانه با استفاده از توصیفگرهای انتخاب شده توسط الگوریتم ژنتیک

پس از انتخاب مناسب ترین توصیف کننده ها توسط الگوریتم ژنتیک، مدل سازی بین توصیف کننده ها و فعالیتهای بازدارندگی مشتقات آریل سولفونیل پیپرازین برای سری آموزش با استفاده از روش MLR انجام و مدل خطی زیر بدست آمد:

pIC_{50}= 8.301 (±0.723) -0.255 (±0.058) VRA2 -1.807 (±0.359) GATS5e +0.047 (±0.012) RDF040m - 0.069 (±0.017) RDF035p +0.174 (±0.025) Te

سپس از معادله بدست آمده برای پیش بینی فعالیت سری تست استفاده گردید. مقادیر واقعی و پیش بینی شده فعالیت‌ها برای تمام ترکیبات در پیوست (پ-۲)آورده شده است.

۴-۲-۶- مدل شبکه عصبی مصنوعی با استفاده از توصیفگرهای انتخاب شده توسط الگوریتم ژنتیک

در این بخش نیز از شبکه عصبی سه لایه متشکل از یک لایه ورودی، یک لایه پنهان و یک لایه خروجی استفاده شد.

۴-۲-۶-۱- انتخاب تعداد گره های لایه پنهان، تعداد متغیر ورودی، نوع تابع آموزش و نوع تابع انتقال

شبکه با ورودی های از ۲ تا ۵ توسط دو الگوریتم آموزشی لونبرگ-مارکوارت و تنظیم بایزین، با تعداد متفاوت گره در لایه پنهان از ۲ تا ۱۰ و همچنین توابع لگاریتمی سیگوئیدی(logsig) و تانژانت سیگموئیدی (tansig)، بعنوان توابع انتقال لایه پنهان، آموزش داده شد. در تمامی این شبکه‌ها تعداد دورهای آموزش ثابت و برابر ۱۰۰ در نظر گرفته شد. همچنین از تابع انتقال خطی (purelin) در لایه خروجی استفاده شد. معیار نیز به حداقل رساندن مربع خطا (MSE) برای سرس ارزیابی در نظر گرفته شد که نتایج این محاسبات در جداول (۴-۴)، (۴-۵)، (۴-۶) و (۴-۷) آورده شده است.

جدول (۴-۴)- مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی تنظم بایزین و تابع انتقال تانژانت سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)				
تعداد گره های لایه پنهان		۲	۳	۴	۵
	۲	۰/۲۸۴	۰/۲۳۶	۰/۲۴۳	۰/۱۴۷
	۳	۰/۲۷۹	۰/۲۴۰	۰/۲۷۴	**۰/۱۲۹**
	۴	۰/۲۷۸	۰/۴۴۸	۰/۴۴۸	۰/۱۷۷
	۵	۰/۲۷۷	۰/۲۳۲	۰/۴۴۸	۰/۲۲۵
	۶	۰/۲۷۷	۰/۲۳۱	۰/۲۷۸	۰/۲۳۱
	۷	۰/۲۷۷	۰/۲۵۳	۰/۲۳۷	۰/۱۳۶
	۸	۰/۲۷۷	۰/۲۳۳	۰/۲۳۶	۰/۲۰۱
	۹	۰/۲۷۶	۰/۲۳۰	۰/۲۶۲	۰/۱۳۸
	۱۰	۰/۲۷۶	۰/۲۳۳	۰/۲۴۳	۰/۱۳۸

جدول (۴-۵)- مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی لونبرگ-مارکوات و تابع انتقال تانژانت سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

تعداد گره های لایه پنهان	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)				
	۲	۳	۴	۵
۲	۰/۶۶۹	۱/۰۵۱	۰/۲۵۹	۲/۶۷۲
۳	۰/۳۴۲	۵/۲۵۸	۰/۵۶۳	۰/۴۱۴
۴	۱/۳۱۴	۰/۴۹۱	۲/۱۵۰	۰/۱۴۵
۵	۰/۴۸۷	۱/۲۸۳	۲/۲۱۸	۰/۴۶۳
۶	۰/۴۳۶	۱/۶۴۸	۲/۸۶۲	۰/۴۸۱
۷	۰/۵۴۲	۳/۵۴۴	۰/۳۱۹	۰/۳۹۶
۸	۰/۲۳۱	۰/۳۴۸	۰/۷۴۴	۰/۴۷۱
۹	۳/۲۳۲	۱/۲۶۶	۱/۶۲۳	۰/۳۶۹
۱۰	**۰/۰۹۶**	۰/۹۱۳	۱/۷۱۳	۰/۴۶۳

جدول (۴-۶)- مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی لونبرگ-مارکوات و تابع انتقال لگاریتم سیگموئیدی با تعداد ورودی های مختلف و تعداد گره های متفاوت در لایه پنهان

تعداد گره های لایه پنهان	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)				
	۲	۳	۴	۵
۲	۰/۳۰۱	۰/۴۵۳	۰/۳۱۵	۴/۴۸۸
۳	۰/۳۷۸	۱/۷۲۰	۰/۳۶۷	**۰/۲۳۸**
۴	۱/۴۹۰	۳/۵۹۸	۵/۷۸۱	۰/۳۱۲
۵	۰/۹۱۶	۰/۸۰۸	۱/۴۲۷	۰/۳۵۰
۶	۳/۹۴۹	۱/۳۴۹	۲/۱۴۷	۰/۹۴۷
۷	۰/۳۹۸	۵/۳۵۷	۲/۸۹۲	۰/۴۴۹
۸	۰/۷۱۲	۰/۳۳۷	۱/۱۳۹	۰/۳۱۴
۹	۱/۷۲۶	۱/۰۴۴	۱/۷۹۸	۰/۳۹۴
۱۰	۳/۲۴۹	۲/۸۶۲	۱/۶۸۴	۰/۳۶۹

جدول (۴-۷)- مقادیر میانگین مربعات خطای سری ارزیابی برای شبکه با الگوریتم آموزشی تنظیم بایزین و تابع انتقال

تعداد گره های لایه پنهان	تعداد متغیرهای ورودی شبکه(تعداد توصیفگرها)				
	۲	۳	۴	۵	
۲	۰/۲۸۳	۰/۲۳۶	۰/۲۴۱	۰/۱۴۸	
۳	۰/۲۷۹	۰/۴۰۸	۰/۲۴۱	۰/۱۳۶	
۴	۰/۲۷۷	۰/۲۳۲	۰/۲۳۶	۰/۴۴۸	
۵	۰/۲۷۷	۰/۲۳۱	۰/۴۴۸	۰/۲۱۹	
۶	۰/۲۷۷	۰/۲۳۱	۰/۲۳۶	۰/۲۹۹	
۷	۰/۲۷۶	۰/۲۳۱	۰/۲۳۴	۰/۱۳۶	
۸	۰/۲۷۶	۰/۲۳۱	۰/۲۳۵	**۰/۱۱۶**	
۹	۰/۲۷۶	۰/۲۳۱	۰/۲۳۴	۰/۱۳۷	
۱۰	۰/۲۷۶	۰/۲۳۲	۰/۲۳۳	۰/۱۴۷	

طبق نتایج موجود در

جدول (۴-۸) میانگین مربعات خطا (MSE) برای الگوریتم آموزشی تنظیم بایزین نسبت به الگـوریتم آموزشی لونبرگ-مارکوارت بیشتر است. همچنین تانژانت سیگموئیدی نسبت به لگاریتم سـیگموئیدی خطای کمتری دارد. در نتیجه تابع تانژانت سیگموئیدی به عنوان تابع انتقال لایه پنهان برگزیده شـده است. طبق نتایج به دست آمده از جداول، تعداد ۲ توصیفگر به عنوان تعـداد بهینـه ورودی هـا بـرای شبکه و تعداد ۱۰ گره نیز به عنوان تعداد بهینه گره ها در لایه پنهان انتخاب شدند.

جدول (۴-۸) توابع و پارامترهای بهینه شده شبکه های بهینه (SR-ANN) بدست آمده

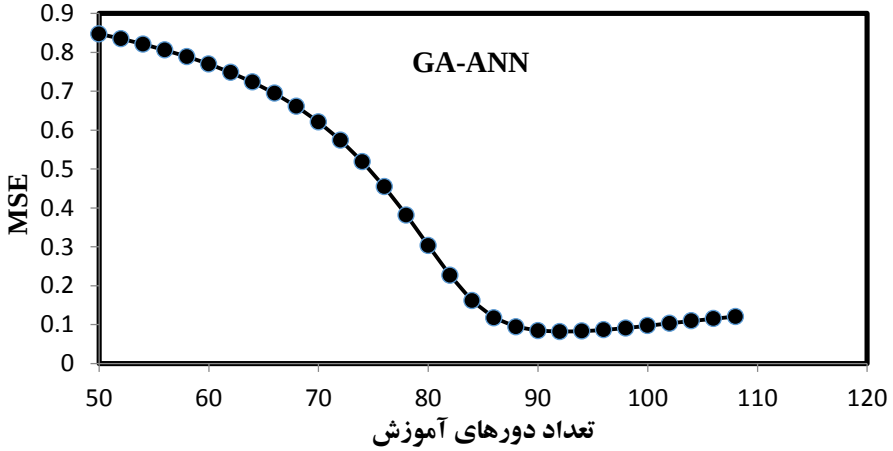
تابع انتقال	تابع آموزش	تعداد توصیف کننده	تعداد نرون لایه پنهان	تعداد دور آموزش	MSE
لگاریتم سیگموید	تنظیم بایزین	۵	۸	۱۰۰	۰/۱۱۶
لگاریتم سیگموید	لونبرگ-مارکوارت	۵	۳	۱۰۰	۰/۲۳۸
تانژانت سیگموید	تنظیم بایزین	۵	۳	۱۰۰	۰/۱۲۹
تانژانت سیگموید	**لونبرگ-مارکوارت**	**۲**	**۱۰**	**۱۰۰**	**۰/۰۹۶**

۴-۲-۶-۲- انتخاب تعداد دورهای آموزش

بدین منظور و جهت انتخاب تعداد دورهای بهینه، شبکه با الگوریتم لونبرگ-مارکوارت، ۲ توصیفگر به عنوان ورودی و ۱۰ گره در لایه پنهان و تابع انتقال تانژانت سیگموئیدی در تعداد دورهای متفاوت، آموزش داده شد و میانگین مربع خطاها برای سری ارزیابی ثبت گردید. جدول (۴-۹) و شکل (۴-۲) نشان می دهد که مقدار میانگین مربع خطا (MSE) در تعداد دورهای آموزش۹۲ به حداقل می رسد. بنابراین ۹۲ بار تکرار به عنوان تکرارهای بهینه برای یاد گیری شبکه انتخاب شد.

جدول (۴-۹)- مقادیر میانگین مربع خطاها برای سری ارزیابی در تعداد دورهای آموزش مختلف

میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار	میانگین مربع خطا	تعداد تکرار
۰/۰۹۱	۹۸	۰/۱۱۷	۸۶	۰/۵۱۸	۷۴	۰/۷۴۸	۶۲	۰/۸۴۷	۵۰
۰/۰۹۶	۱۰۰	۰/۰۹۴	۸۸	۰/۴۵۴	۷۶	۰/۷۲۳	۶۴	۰/۸۳۴	۵۲
۰/۱۰۳	۱۰۲	۰/۰۸۴	۹۰	۰/۳۸۱	۷۸	۰/۶۹۴	۶۶	۰/۸۲۱	۵۴
۰/۱۰۹	۱۰۴	**۰/۰۸۱**	**۹۲**	۰/۳۰۳	۸۰	۰/۶۶۰	۶۸	۰/۸۰۵	۵۶
۰/۱۱۵	۱۰۶	۰/۰۸۲	۹۴	۰/۲۲۶	۸۲	۰/۶۲۱	۷۰	۰/۷۸۹	۵۸
۰/۱۲۰	۱۰۸	۰/۰۸۶	۹۶	۰/۱۶۱	۸۴	۰/۵۷۴	۷۲	۰/۷۷۰	۶۰



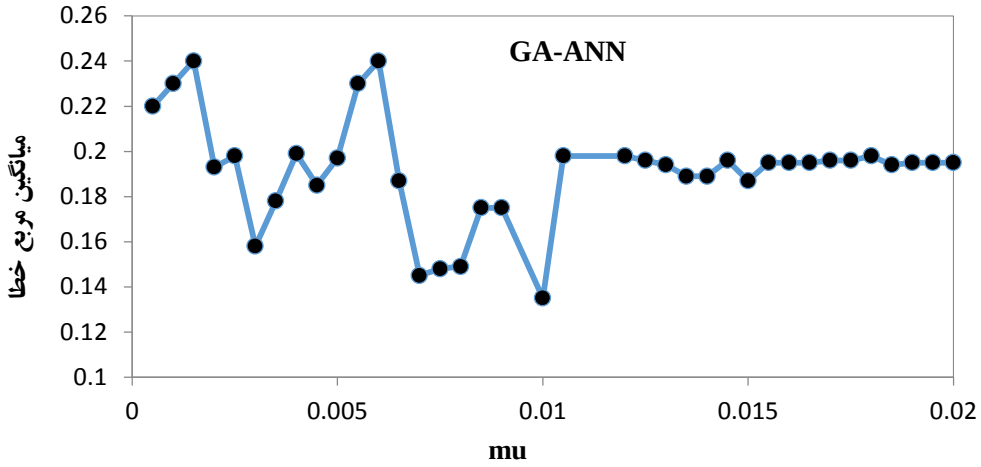
شکل (۴-۲)- منحنی تعداد چرخه های آموزش بر حسب میانگین مربع خطاها

۴-۲-۶-۳- بهینه کردن مقدار mu

جهت بهینه کردن مقدار mu، ساختار شبکه با ۲ متغیر ورودی، ۱۰ گره در لایه پنهان و الگوریتم آموزشی لونبرگ-مارکوارت و تابع انتقال تانژانت سیگموئیدی در نظر گرفته شد. سپس مقدار mu از ۰/۰۰۰۵ تا ۰/۰۲ با گام های ۰/۰۰۰۵ تغییر داده شد و آنگاه برای هر مورد مقدار میانگین مربع خطای سری ارزیابی محاسبه گردید. در نهایت مقدار میانگین مربع خطا بر حسب mu رسم گردید. نقطه ای که کمترین خطا را داشت به عنوان مقدار بهینه انتخاب شد که طبق جدول (۴-۱۰) و شکل (۴-۳) مقدار بهینه ۰/۰۱ بدست آمد.

جدول(۴-۱۰) مقادیر میانگین مربع خطاها برای سری ارزیابی در مقادیر مختلف mu

mu	میانگین مربع خطا	mu	میانگین مربع خطا	mu	میانگین مربع خطا
۰/۰۰۰۵	۰/۲۲	۰/۰۰۷۵	۰/۱۴۸	۰/۰۱۶	۰/۱۹۵
۰/۰۰۱	۰/۲۳	۰/۰۰۸	۰/۱۴۹	۰/۰۱۶۵	۰/۱۹۵
۰/۰۰۱۵	۰/۲۴	۰/۰۰۸۵	۰/۱۷۵	۰/۰۱۷	۰/۱۹۶
۰/۰۰۲	۰/۱۹۳	۰/۰۰۹	۰/۱۷۵	۰/۰۱۷۵	۰/۱۹۶
۰/۰۰۲۵	۰/۱۹۸	**۰/۰۱**	**۰/۱۳۵**	۰/۰۱۸	۰/۱۹۸
۰/۰۰۳	۰/۱۵۸	۰/۰۱۰۵	۰/۱۹۸	۰/۰۱۸۵	۰/۱۹۴
۰/۰۰۳۵	۰/۱۷۸	۰/۰۱۲	۰/۱۹۸	۰/۰۱۹	۰/۱۹۵
۰/۰۰۴	۰/۱۹۹	۰/۰۱۲۵	۰/۱۹۶	۰/۰۱۹۵	۰/۱۹۵
۰/۰۰۴۵	۰/۱۸۵	۰/۰۱۳	۰/۱۹۴	۰/۰۲	۰/۱۹۵
۰/۰۰۵	۰/۱۹۷	۰/۰۱۳۵	۰/۱۸۹		
۰/۰۰۵۵	۰/۲۳	۰/۰۱۴	۰/۱۸۹		
۰/۰۰۶	۰/۲۴	۰/۰۱۴۵	۰/۱۹۶		
۰/۰۰۶۵	۰/۱۸۷	۰/۰۱۵	۰/۱۸۷		
۰/۰۰۷	۰/۱۴۵	۰/۰۱۵۵	۰/۱۹۵		



شکل(۴-۳)- نمودار میانگین مربع خطا بر حسب مقادیر مختلف mu

توابع و مقادیر بهینه شده پارامترهای مختلف شبکه عصبی در جدول (۴-۱۱) ارائه شده است.

جدول (۴-۱۱) توابع و پارامترهای بهینه شده شبکه عصبی

تابع آموزش	trainlm
تابع انتقال لایه ی پنهان	tansig
تابع انتقال لایه ی خروجی	pureline
تعداد نرون لایه ی پنهان	۱۰
تعداد متغیرهای ورودی	۲
تعداد دورهای آموزش	۹۲
پارامتر mu	۰/۰۱

با استفاده از مدل ANN بهینه شده مقادیر فعالیتهای بازدارندگی ترکیبات مورد نظر در مجموعه های آموزش، ارزیابی و پیش‌بینی مورد محاسبه قرار گرفت و در پیوست (پ-۲) نشان داده شده است.

۴-۲-۷- مدل سازی به روش ماشین بردار پشتیبان (SVM)

در این فصل همانند فصل دوم بهینه سازی سه پارامتر ماشین بردار پشتیبان با استفاده از توصیفگرهای انتخاب شده به روش GA انجام گرفت که نتایج آن در جدول (۴-۱۲) گزارش شده

است. که این نتایج نشان می‌دهند که ماشین بردار پشتیبان با مقادیر ۲ برای تابع حساسیت وپنایک، ۲۰ برای پارامتر موازنه‌ی C و ۰/۱ برای پارامتر کرنل برای پیش بینی pIC_{50}، با استفاده از توصیفگرهای منتخب GA می‌باشد. مقادیر فعالیت‌های بازدارندگی ترکیبات مورد نظر در سری آموزش و تست مورد محاسبه و در پیوست (پ-۲) نشان داده شده است.

جدول (۴-۱۲)- مقادیر مختلف پارامترهای ماشین بردار پشتیبان با استفاده توصیگرهای منتخب GA و MSE آنها

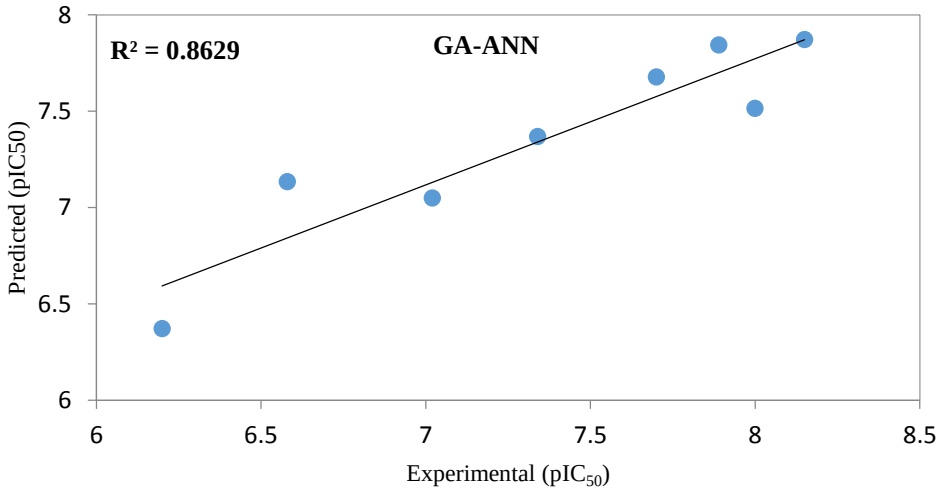
پارامتر موازنه‌ی C	تابع حساسیت وپنایک	پارامتر کرنل	MSE
۵۰	۱	۰/۱	**۰/۱۱۰۲**
۵۰	۱	۰/۲	۰/۱۲۲۹
۵۰	۱	۰/۳	۰/۱۴۰۹
۵۰	۱	۰/٤	۰/۱٦۲۲
۵۰	۱	۰/۵	۰/۱٦۹٤
۵۰	۱	۰/٦	۰/۱۸۲٦
۵۰	۱	۰/۷	۰/۲۰۵٦
۵۰	۱	۰/۸	۰/۲۵۱۲
۵۰	۱	۰/۹	۰/۳۱٤۹
۵۰	۱	۱	۰/۳۹٦۷
۵۰	۱	۰/۱	۰/٤۰۳٤
۵۰	**۲**	**۰/۱**	**۰/۱۰۵۳**
۵۰	۳	۰/۱	۰/۱۱٦۳
۵۰	٤	۰/۱	۰/۱۲٤۱
۵۰	۵	۰/۱	۰/۱۲۸۵
۵۰	٦	۰/۱	۰/۱۳۱۰
۵۰	۷	۰/۱	۰/۱۳۲۲
۵۰	۸	۰/۱	۰/۱۳۵٦
۵۰	۹	۰/۱	۰/۱٤۲۷
۵۰	۱۰	۰/۱	۰/۱٤٤۹
۱۰	۲	۰/۱	۰/۳۳۲۰
۲۰	**۲**	**۰/۱**	**۰/۳۲۸۷**
۳۰	۲	۰/۱	۰/۳۳۲۰
٤۰	۲	۰/۱	۰/۳۳۲۰
۵۰	۲	۰/۱	۰/۳۳۲۰
٦۰	۲	۰/۱	۰/۳۳۲۰
۷۰	۲	۰/۱	۰/۳۳۲۰
۸۰	۲	۰/۱	۰/۳۳۲۰
۹۰	۲	۰/۱	۰/۳۳۲۰
۱۰۰	۲	۰/۱	۰/۳۳۲۰

۴-۲-۸- بررسی نتایج

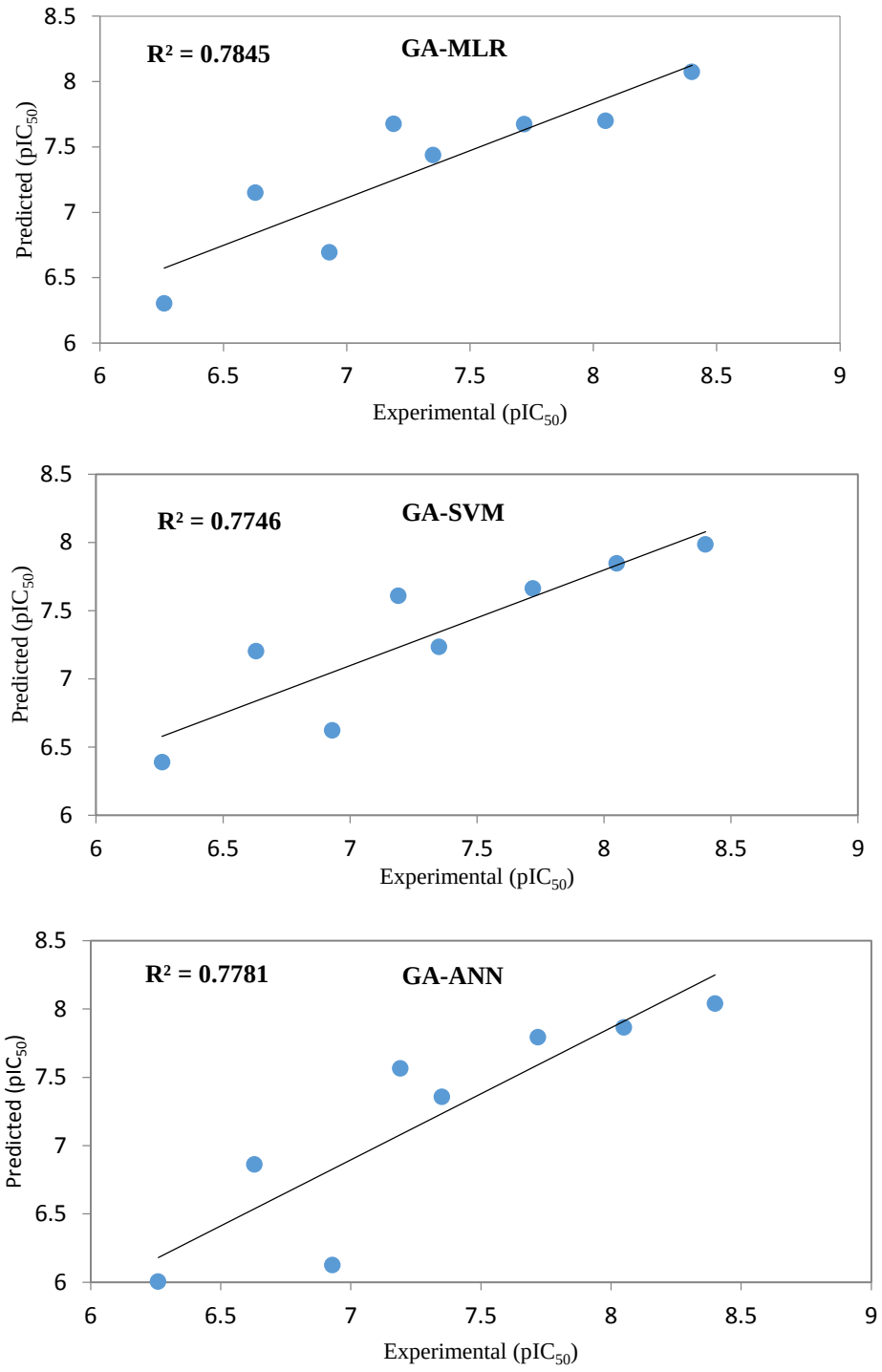
اعتبار و اهمیت معادلات پیش بینی و مدل های ساخته شده وقتی مشخص می گردد که فعالیت مولکول های که در سری تست هستند را به‌خوبی و بطور رضایت بخش و قابل قبول پیش بینی کند. در اینجا چندین روش به منظور ارزیابی توانایی مدل های ارائه شده در پیش بینی مقادیر pIC_{50} ذکر شده است.

۴-۲-۸-۱- ارزیابی مدل ها با استفاده از نمودار برگشتی

در نمودار برگشتی مقادیر پیش بینی شده برحسب مقادیر تجربی رسم می‌گردد و با توجه به مقدار ضریب تعیین (R^2) به دست آمده از نمودار، پراکندگی نقاط در اطراف خط برگشت تعیین می-شود. هر چه مقدار ضریب تعیین به یک نزدیک تر باشد، مدل ساخته شده، مدل بهتری است. نتایج حاصل از نمودارهای برگشتی مربوط به سه روش انجام شده در این کار برای سری ارزیابی و تست در شکل های (۴-۴) و (۴-۵) آورده شده است.



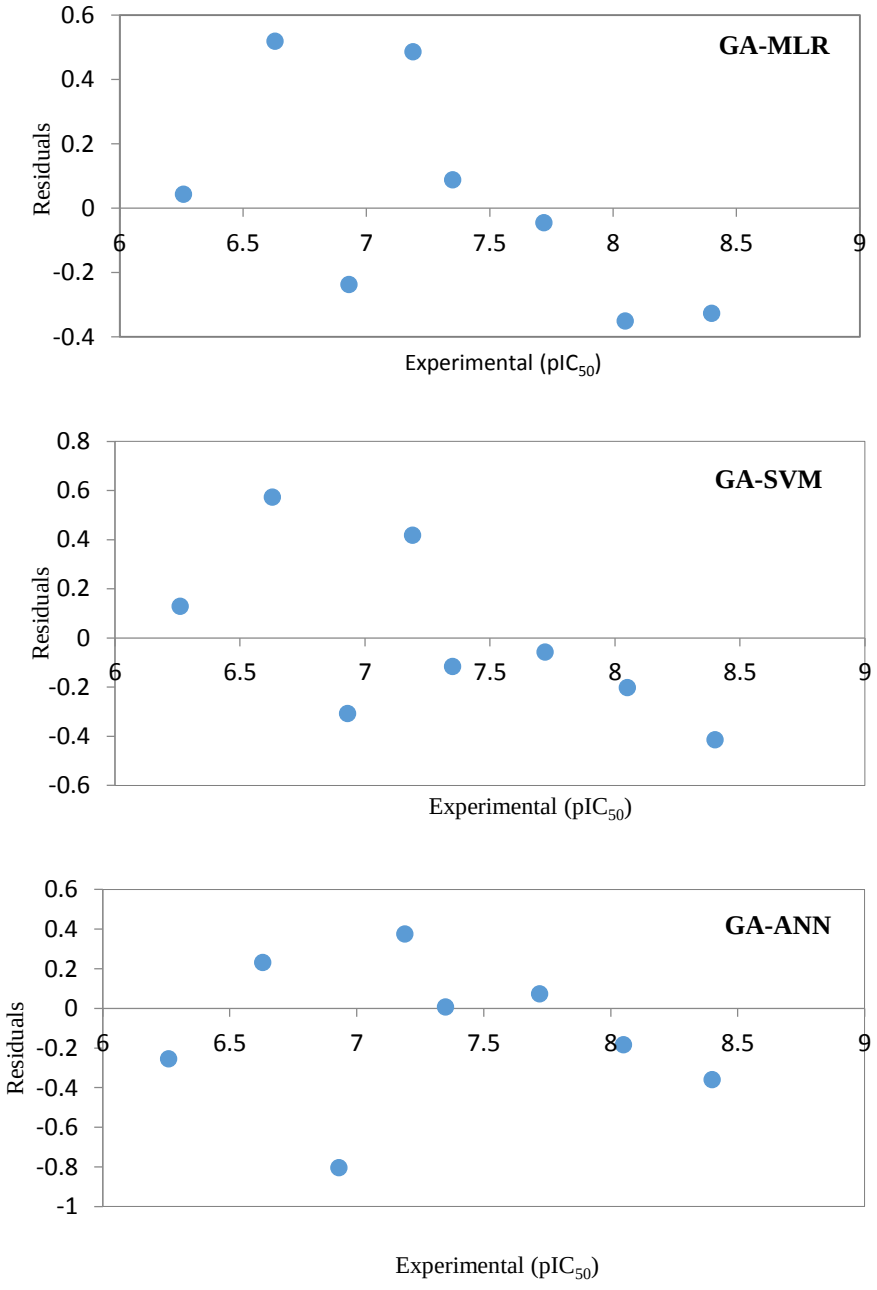
شکل (۴-۴)- نمودار مقادیر پیش بینی شده pIC_{50} بر حسب مقادیر تجربی برای سری ارزیابی



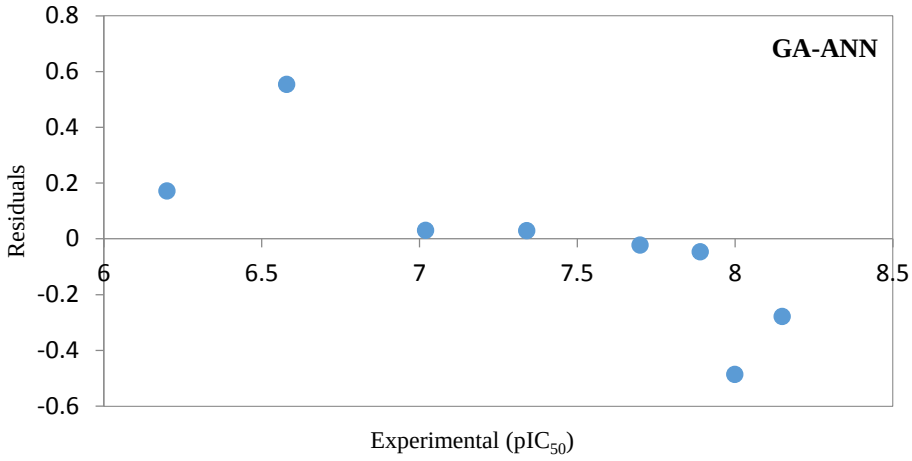
شکل (۴-۵)- نمودار مقادیر پیش بینی شده pIC_{50} بر حسب مقادیر تجربی برای سری تست

۴-۲-۸-۲- ارزیابی مدل‌ها با استفاده از نمودار خطای باقیمانده

اختلاف مقادیر پیش بینی شده و مقادیر تجربی، خطـای باقیمانـده نامیـده مـی‌شـود. پراکنـدگی یکنواخت نقاط، حول محور افقی که بیانگر خطای باقیمانده‌ی صفر است، نشان‌دهنده‌ی آن اسـت کـه خطای سیستماتیکی در مدل سازی وجود ندارد. نمودار خطای باقیمانده بر حسب مقادیر تجربی، برای مدل‌های ذکر شده در شکل‌های (۴-۶) و (۴-۷) نشان داده شده است.



شکل (۴-۶)- نمودار مقادیر باقیمانده pIC50 بر حسب مقادیر تجربی برای سری تست



شکل (۴-۷)- نمودار مقادیر باقیمانده pIC50 بر حسب مقادیر تجربی برای سری ارزیابی

۴-۲-۸-۳- ارزیابی مدل های برتر خطی و غیر خطی با استفاده از پارامترهای آماری

مطابق جدول (۴-۱۳) شش پارامتر آماری، جهت ارزیابی توانایی پیش بینیی مدل های ساخته شده به روش های GA-MLR، GA-ANN و GA-SVM به کار گرفته شد. نتایج حاصل از جدول (۴-۱۳) نشان می دهد که پارامترهای آماری مختلف برای سری تست مدل‌های مختلف از جمله MSE، SEP، MAE و R^2 بسیار به یکدیگر نزدیک بوده و این بیانگر این مطلب است که هر سه مدل به کار برده شده توانایی خوبی در پیش بینی IC_{50} ترکیبات مورد مطالعه دارند.

جدول (۴-۱۳)- پارامترهای آماری برای مدل های انتخاب شده

		GA-MLR	GA-ANN	GA-SVM
MSE	سری آموزش	۰/۰۴۶	۰/۰۰۱	۰/۰۱
	سری ارزیابی	-	۰/۰۸۱	-
	سری تست	۰/۱	۰/۱۳۴	۰/۱۱۱
Press	سری آموزش	۱/۵۸۱	۰/۰۲۷	۰/۲۶۳
	سری ارزیابی	-	۰/۶۵۴	-
	سری تست	۰/۸۰۲	۱/۰۷۹	۰/۸۹۰
SEP	سری آموزش	۰/۲۱۵	۰/۰۳۲	۰/۰۹
	سری ارزیابی	-	۰/۲۸۶	-
	سری تست	۰/۳۱۶	۰/۳۶۶	۰/۳۳۵
MAE	سری آموزش	۰/۱۷۵	۰/۰۱۸	۰/۰۸
	سری ارزیابی	-	۰/۲۰۲	-
	سری تست	۰/۲۶۱	۰/۲۸۶	۰/۲۸۶
AAD	سری آموزش	۲/۳۷۵	۰/۲۳۷	۱/۱۳
	سری ارزیابی	-	۲/۷۹۹	-
	سری تست	۳/۵۸۹	۴/۰۰۱	۰/۹۳۶
R^2	سری آموزش	۰/۸۲۷	۰/۹۹۵	۰/۹۷۱
	سری ارزیابی	-	۰/۸۶۲	-
	سری تست	۰/۷۸۴	۰/۷۷۸	۰/۷۷۴

۴-۲-۸-۴- ارزیابی مدل ها توسط روش رد مرحله ای تک تک و گروهی

به منظور بررسی بیشتر قدرت پیش بینی مدل های خطی و غیر خطی تکنیک رد مرحله ای تک تک و گروهی مورد استفاده قرار گرفت. در روش رد مرحله ای تک تک، هر بار یکی از ترکیبات به طور تصادفی از سری داده‌ها حذف شدند و در روش رد مرحله‌ای گروهی، هر بار یک گروه از ترکیبات(۵

ترکیب) به طور تصادفی ازسری داده‌ها حذف شدند. سپس با استفاده از مدل ساخته شده توسط بقیه ترکیبات، فعالیت دارویی ترکیب یا ترکیبات حذف شده، پیش بینی شدند. این فرایند برای تمام اعضای سری داده ها تکرار شد. نتایج حاصل از رد مرحله ای و گروهی در جدول(۴-۱۴) ارائه شده است.

جدول (۴-۱۴)- پارامترهای آماری برای مدل های انتخاب شده

		GA-MLR	GA-ANN	GA-SVM
Q^2_{LOO}	کل داده ها	۰/۷۴۸	۰/۷۶۸	۰/۷۵۴
Q^2_{LGO}	کل داده ها	۰/۷۳۱	۰/۷۳۶	۰/۷۲۴

۴-۲-۸-۵- ارزیابی مدل های ارائه شده با استفاده از آزمون Y-تصادفی

این تکنیک ارزیابی مدل، با هدف بررسی هر گونه ارتباط تصادفی بین داده ها انجام شد. در این آزمون، متغیر وابسته بطور تصادفی بهم ریخته شد. مدل QSAR جدید با استفاد از ماتریکس متغیرهای مستقل اصلی و مقادیر تصادفی از متغیر وابسته توسعه یافت. اگر در مدل اصلی هیچ گونه ارتباط تصادفی وجود نداشته باشد، تفاوت قابل توجهی بین مقدار ضریب تعیین مدل اصلی و مدل QSAR که با پاسخ تصادفی توسعه یافته، وجود خواهد داشت. نتایج حاصل از چندین بار اجرای آزمون Y-تصادفی در جدول (۴-۱۵) نشان داده شده است. مقادیر کوچک ضریب تعیین (R^2) بیانگر عدم ارتباط شانسی در مدل توسعه یافته توسط رگرسیون خطی چندگانه، شبکه عصبی مصنوعی و ماشین بردار پشتیبان می‌باشد.

جدول (۴-۱۵)- نتایج حاصل از ۱۰ بار آزمون Y-تصادفی

تکرار	R^2_{test}		
	GA-MLR	GA-ANN	GA-SVM
۱	۰/۰۳۱۱	۰/۰۲۴۸	۰/۰۱۲۸
۲	۰/۰۰۴۶	۰/۰۱۶۵	۰/۱۴۹۷
۳	۰/۰۷۲۶	۰/۰۰۹۶	۰/۰۰۱۲
۴	۰/۰۰۴۸	۰/۱۶۹۸	۰/۰۳۵۱
۵	۰/۰۲۱۴	۰/۰۳۶۵	۰/۰۱۲۷
۶	۰/۱۳۵۰	۰/۰۱۲۵	۰/۰۰۹۴
۷	۰/۰۰۰۷	۰/۰۹۵۴	۰/۰۱۰۵
۸	۰/۰۵۰۷	۰/۰۷۵۴	۰/۰۱۷۴
۹	۰/۱۹۷۴	۰/۱۳۲۵	۰/۱۴۵۲
۱۰	۰/۰۹۳۹	۰/۰۰۷۸	۰/۰۹۴۶

.

۴-۲-۹- بررسی ارتباط توصیفگرهای وارد شده در مدل با اثر بازدارندگی

در این بخش به طور خلاصه ارتباط بین توصیفگرهای وارد شده در مدل و اثر بازدارندگی ترکیبات، مورد بررسی قرار خواهد گرفت. با توجه به نتایج به دست آمده در مدل برتر، توصیفگرهای انتخاب شدند که هر کدام بیانگر خصوصیات متفاوتی از مولکول مورد بررسی است. توصیفگرهای انتخاب شده توسط روش توسط الگوریم ژنتیک شامل (VRA2، GATS5e، RDF040m، RDF035p و Te) می‌باشند.

۴-۲-۹-۱- توصیفگرهای RDF[۱]

RDF یا تابع توضیع شعاعی مربوط به یک دسته از اتم ها، معادل توزیع احتمال یافتن یک اتم در

۱- Radial Distribution Function Descriptors

فضای کروی به شعاع R است. رابطه (۴-۱) نحوه محاسبه تابع RDF را نشان می‌دهد.

$$g(R) = \int . \sum_{i}^{N-1} \sum_{j>i}^{N} A_i . A_j . e^{-B.(R-r_{ij})^2} \qquad (۱-٤)$$

که f یک فاکتور مقیاس، و N تعداد اتم‌های مولکول است. همچنین r_{ij} فاصله‌ی بین دو اتم i و j و A یک ویژگی اتمی (وزن دار نشده u، جرم اتمی m، حجم واندروالس v، الکترونگاتیویته e، قطبش پذیری p) است. β یک فاکتور تسهیل کننده است که توزیع احتمال فاصله بین اتمی را مشخص می‌کند و می‌توان از آن به فاکتور دما برای تعریف جنبش اتمی تعبیر کرد. g(R) در نقاط گسسته‌ای با فواصل معین محاسبه و با مجموعه‌ای از کدهای RDF با ویژگی‌های مختلف اتمی می‌توان ساختار سه بعدی یک مولکول را به طور واضح توصیف کرد. این توصیف کننده‌ها اهمیت توزیع اتم‌ها در مولکول روی اثر بازدارندگی آن نشان می‌دهد. همچنین این توصیف کننده علاوه بر فاصله اتمی، اطلاعاتی با ارزشی در مورد فاصله پیوند، انواع حلقه، سیستم‌های مسطح و غیر مسطح و نوع اتم‌ها فراهم می‌کند.

از این گروه، توصیفگرها RDF040m، RDF035p در مدل ارائه شده برای پیش بینی اثر بازدارندگی انتخاب شده‌اند. اثر متوسط توصیفگرهای RDF035p مثبت ذکر شده یعنی با افزایش مقدار این توصیفگرها pIC_{50} افزایش می‌یابد و اثر متوسط توصیفگرهای RDF040m منفی ذکر شده یعنی با افزایش مقدار این توصیفگرها pIC_{50} نیز کاهش می‌یابد.

۴-۲-۹-۲- توصیفگرهای گروه Topological

این توصیفگرها بر اساس نمایش گراف مولکول می‌باشند. در این گراف‌ها هر نقطه نشان‌دهنده‌ی یک اتم بوده و خطوط بین نقاط نیز نشان‌دهنده‌ی پیوند شیمیایی بین اتم‌ها می‌باشد. معمولا در گراف‌های مولکولی اتم هیدروژن را نشان نمی‌دهند. این توصیفگرها اطلاعاتی راجع به ساختمان، اندازه، شکل، تقارن، شاخه‌دار شدن، نحوه اتصال اتم‌ها و نوع اتم‌های موجود در یک مولکول را در اختیار ما قرار می‌دهند. محاسبه این توصیفگرها به سادگی از روی ساختمان دو بعدی مولکول‌ها

امکان پذیر می‌باشد. از این گروه توصیفگرها در مدل ارائـه شـده توصـیفگر VRA2 انتخـاب شـد. اثـر متوسط برای این توصیفگر مثبت است که نشان می‌دهد با افزایش آن مقدار pIC_{50} افزایش می‌یابد.

۴-۲-۹-۳- توصیفگرهای WHIM

این توصیفگرها که به وسیله گروه تحقیقـی تودسـچینی[۱] و همکـارانش توسـعه یافتـه‌انـد حـاوی اطلاعات سه بعدی هستند و بر اساس تجزیه اجزاء اصلی ماتریکس کوواریانس وزن دار شده حاصـل از کوردینه‌های کارتزین مولکول می‌باشند. اهمیت این توصیفگرها بخاطر تاثیر کنفورماسیون ترکیـب در طول برهمکنش با هدف می‌باشد. این توصیفگرها اطلاعات سه بعدی در مورد اندازه، شـکل، تقـارن و توزیع اتمی در مولکول‌ها در اختیـار مـا قـرار مـی‌دهنـد. در رابطـه (۴-۲) روش کلـی محاسـبه ایـن توصیفگرها را بیان می‌کند:

$$S_{jk} = \frac{\sum_{i=1}^{A} w_i (q_{ij} - \bar{q}_i)(q_{ik} - \bar{q}_k)}{\sum_{i=1}^{A} w_i} \qquad (۴-۲)$$

که S_{jk} کوواریانس وزن‌دار بین کوئوردینه‌های اتمی jام و kام، A تعداد اتم‌ها، w_i وزن iامین اتـم، q_{ij} و q_{ik} به ترتیب jامین و kامین کئوردینه‌های iامین اتم و $\bar{q}$ مقدار میانگین مربوطه را نشان می‌دهند.

توصیفگر WHIM به دو دسته مهم تقسیم مـی‌شـوند: توصـیفگرهای جهـت‌دار[۲] و توصـیفگرهای کروی[۳]. گروه توصیفگرهای جهت‌دار توصیفگرهایی در مورد اندازه، شکل و تقـارن مولکـولی در جهـت محور اصلی از مولکول ارائه می‌دهند ولی اطلاعـات توصـیفگرهای کـروی بـا دیـد کـروی از مولکـول محاسبه می‌شود. از این دسته توصیفگرها، توصیفگر Te می‌باشد. که اثر متوسط این توصـیفگر منفـی بوده واثر منفی روی pIC_{50} میگذارد.

۱- Todeschini
۲- Directiona WHIM descriptors
۳- Global WHIM descriptors

۴-۲-۹-۴- توصیفگرهای 2D autocorrelation[۱]

توصیفگرهای دو بعدی از گراف مولکولی و از طریق محاسبه اوزان اتم‌های انتهایی کل مسیرها با طول مسیر مورد نظر، به دست می‌آیند. نوعی از این توصیفگرها گروه Geary autocorrelation است که ضریب گری نام دارد و بدین صورت محاسبه می‌شود:

$$C(d) = \frac{\frac{1}{2\Delta}\sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij}\cdot(w_i - w_j)^2}{\frac{1}{A-1}\cdot\sum_{i=1}^{A}(w_i - \bar{w}_j)^2} \qquad (۴-۳)$$

که w یک ویژگی اتم، $\bar{w}$ میانگین مقدار آن ویژگی روی مولکول، A تعداد اتم‌ها و d فاصله‌ی توپولوژیکی است و $\boldsymbol{\delta}ij$ نیز که به تابع کرونکر[۲] معروف است، در حالتی که $d=d_{ij}$ باشد مقدار این تابع یک است، و در غیر این صورت صفر است و Δ نیز مجموعه‌ی $\boldsymbol{\delta}$ هاست. مقدار این توصیفگرها که از جنس فاصله است، از صفر تا بی‌نهایت متغیر است. از میان این توصیفگرها GATS5e وارد مدل برتر شد که نقش خاصیت فیزیکی و الکترونگاتیویته‌ی ساندرسون را بیان می‌کند. این توصیفگر دارای اثر متوسط مثبت بوده، بدین معنی که با افزایش مقدار این توصیفگر، pIC_{50} افزایش می‌یابد.

۴-۲-۱۰- بررسی میزان مشارکت توصیفگرهای منتخب شبکه عصبی

میزان مشارکت توصیفگرهای منتخب به صورت زیر تعیین شد:

۱- توصیفگر مورد نظر به همراه اوزان مربوطه‌اش از شبکه بهینه حذف گردید.

۲- مقدار متغیر وابسته با استفاده از توصیفگرهای باقیمانده برای هر یک از ترکیبات سری ارزیابی پیش بینی گردید.

۳- میانگین خطای مطلق (MAE) حاصل از ترکیبات سری ارزیابی محاسبه شد.

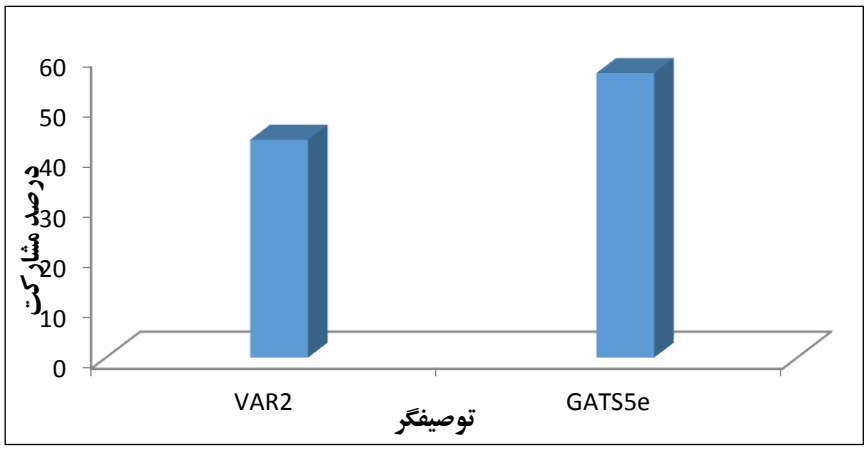
۴- مراحل ۱ تا ۳ برای دیگر توصیفگرهای منتخب نیز تکرا شد.

۱- Reciprocal distance Wiener-type index
۲- Kronecker

۵- سرانجام درصد مشارکت هر توصیفگر توسط رابطه‌ی (۴-۴) برآورد شد.

$$c_i = 100\frac{\Delta m_i}{\sum_{i=1}^{N}\Delta m_i} \qquad (۴-۴)$$

در این رابطه c_i درصد مشارکت توصیفگرحذف شده iام، N تعداد توصیفگرهای مدل و Δm_i میانگین خطای مطلق حاصل از سری ارزیابی در غیاب توصیفگر iام را نشان می‌دهد. که بر این اساس درصد مشارکت توصیفگرهای منتخب در ترکیبات مورد بررسی درشکل (۴-۸) ارائه شده است.



شکل(۴-۸)- درصد مشارکت توصیفگرها در مدل بهینه

بر اساس شکل فوق، توصیفگر GATS5 بیشترین اثر مشارکت را دارا می‌باشد. که نقش خاصیت فیزیکی و الکترونگاتیویته‌ی ساندرسون را بیان می‌کند.

۴-۳- نتیجه گیری

در این فصل از روش‌های مختلف کمومتریکس برای پیش بینی فعالیت ترکیبات دارویی(IC_{50})، مشتقات آریل سولفونیل پیپرازین، استفاده شده است. از آنجایی که اندازه‌گیری فعالیت دارویی بیشتر ترکیبات به صورت تجربی با صرف هزینه، زمان و پیچیدگی زیاد همراه است، دست یابی به مقادیر IC_{50}، با روش‌های تجربی مقرون به صرفه نیست. بنابراین، پیش بینی آن با استفاده از کمومتریکس از

اهمیت بالایی برخوردار است. برای انتخاب توصیفگرهای مناسب از روش GA استفاده شد. سپس این توصیفگرها برای مدل سازی خطی و غیر خطی مورد استفاده قرار گرفت.

نتایج جدول (۴-۱۳) نشان می‌دهد که پارامترهای آماری مختلف برای سری تست مدل‌های مختلف از جمله MSE، SEP، MAE و R^2 بسیار به یکدیگر نزدیک بوده و این بیانگر این مطلب است که هر سه مدل به کار برده شده توانایی خوبی در پیش بینی IC_{50} ترکیبات مورد مطالعه دارند.

۴-۴- آینده‌نگری

برای انتخاب توصیفگرهای مناسب و معتبر می‌توان از سایر روش‌ها مثل الگوریتم جستجوی ممنوع، الگوریتم اجتماع مورچگان و الگوریتم بهینه سازی جمعیت ذره‌ها استفاده نمود و نتایج را با این کار مقاسیه کرد.

برای مدل سازی می‌توان از روش‌های حداقل مربعات ماشین‌های بردار پشتیبان (LS-SVM) و سیستم استنتاج فازی-عصبی سازگار (ANFIS) استفاده کرد.

پیوست

جدول(پ-۱)-مقادیر پیش بینی شده pIC_{50} برای سری آموزش و ارزیابی*

شماره ترکیب	مقدار تجربی	مقادیر پیش بینی شده(pIC_{50})					
		SR-MLR	SR-SVM	SR-ANN	GA-MLR	GA-SVM	GA-ANN
۱	۴/۹۶	۴/۵۸	۴/۹۵	۴/۶۲	۴/۶۱	۴/۹۵	۴/۶۲
۲	۴/۷۲	۴/۸۰	۴/۷۳	۴/۷۷	۴/۸۸	۴/۷۳	۴/۵۵
۳	۴/۸۹	۴/۵۱	۴/۸۸	۴/۷۵	۴/۷۲	۴/۸۸	۴/۷۵
۵	۵/۳۹	۵/۱۶	۵/۳۸	۵/۱۲	۵/۴۳	۵/۴۰	۵/۴۳
*۶	۴/۳۵	۴/۵۸	۴/۳۶	۴/۳۲	۴/۹۱	۴/۳۵	۴/۰۱
۸	۵/۸۲	۵/۶۳	۵/۸۱	۵/۵۶	۵/۶۰	۵/۸۱	۵/۴۳
۹	۵/۱۵	۵/۵۷	۵/۱۶	۵/۵۵	۵/۲۴	۵/۱۳	۵/۰۴
۱۰	۴/۴۷	۴/۴۹	۴/۰۹	۴/۴۶	۵/۳۲	۴/۴۸	۴/۳۵
۱۱	۴/۹۶	۵/۶۵	۴/۹۷	۵/۶۴	۵/۴۴	۴/۹۷	۵/۳۲
۱۲	۶/۱۵	۵/۴۶	۵/۷۸	۵/۴۰	۵/۶۷	۶/۱۴	۵/۶۶
*۱۳	۵/۴	۵/۳۴	۵/۳۹	۵/۲۵	۵/۴۷	۵/۴۱	۵/۴۳
۱۵	۵/۳	۵/۴۶	۵/۳۱	۵/۳۶	۵/۴۷	۵/۳۱	۵/۴۵
*۱۷	۴/۹۲	۵/۲۹	۵/۰۶	۵/۲۴	۵/۲۵	۴/۹۳	۴/۷۸
۱۹	۵/۶۲	۵/۵۳	۵/۶۱	۵/۴۹	۵/۷۰	۵/۶۳	۵/۷۱
۲۰	۵/۳۵	۵/۵۸	۵/۳۶	۵/۵۳	۵/۲۳	۵/۳۴	۵/۳۴
۲۱	۵/۸۲	۵/۴۹	۵/۶۵	۵/۴۳	۴/۸۵	۵/۸۱	۵/۷۶
۲۲	۵/۸۲	۵/۶۷	۵/۶۹	۵/۶۴	۵/۵۹	۵/۸۱	۵/۵۱
*۲۳	۵/۰۳	۵/۶۴	۵/۱۴	۵/۶۱	۵/۱۸	۵/۰۴	۵/۶۲
*۲۴	۵/۸۲	۵/۴۵	۵/۸۱	۵/۳۹	۵/۶۸	۵/۸۱	۵/۷
۲۶	۴/۵۵	۳/۹۹	۴/۵۴	۴/۱۸	۴/۱۷	۴/۵۴	۴/۴۷
۲۷	۵/۱۳	۴/۹۴	۵/۱۲	۴/۹۴	۴/۹۵	۵/۱۳	۵/۲۴
*۲۸	۴/۵۲	۴/۰۹	۴/۵۱	۴/۲۷	۴/۱۳	۴/۵۱	۴/۵۲
۳۰	۴/۳۹	۴/۷۰	۴/۴۰	۴/۶۴	۴/۹۳	۴/۴۰	۴/۸۸
۳۱	۴/۷۷	۵/۱۷	۴/۷۸	۴/۸۸	۴/۹۹	۴/۷۷	۴/۹۱
۳۲	۵/۱۱	۵/۰۵	۵/۱۰	۴/۸۷	۵/۳۲	۵/۱۲	۵/۲۱
*۳۴	۴/۶۶	۴/۸۵	۴/۶۵	۴/۶۳	۴/۶۳	۴/۶۵	۵/۰۹
۳۵	۴/۳۸	۴/۸۷	۴/۳۹	۴/۷۶	۴/۹۶	۴/۳۹	۴/۶۷
۳۶	۵/۱۶	۵/۳۰	۵/۲۱	۵/۱۶	۵/۳۷	۵/۱۶	۵/۲۳
۳۷	۶/۴۳	۵/۸۶	۶/۴۲	۵/۸۱	۵/۹۲	۶/۴۲	۵/۸۲
*۳۸	۴/۷۲	۴/۲۰	۴/۷۱	۴/۴۴	۴/۴۲	۴/۷۱	۴/۰۲
۳۹	۴/۴	۴/۱۶	۴/۳۹	۴/۱۶	۴/۳۷	۴/۴۰	۴/۹۱
۴۰	۴/۱۸	۴/۳۱	۴/۱۹	۴/۴۶	۴/۲۳	۴/۱۸	۴/۴۴
۴۱	۴/۸۵	۴/۷۳	۴/۸۴	۴/۸۷	۴/۷۸	۴/۸۴	۴/۶۳
۴۲	۵/۶۴	۵/۶۳	۵/۶۵	۵/۹۳	۶/۰۲	۵/۶۵	۵/۸۱
*۴۳	۴/۳	۴/۷۳	۴/۳۱	۵/۱۱	۴/۴۷	۴/۳۱	۴/۴۸
۴۴	۴/۱۹	۴/۷۶	۴/۲۰	۴/۹۵	۴/۳۵	۴/۲۰	۴/۴۲

ادامه جدول(پ-۱)

شماره ترکیب	مقدار تجربی	مقادیر پیش بینی شده(pIC_{50})					
		SR-MLR	SR-SVM	SR-ANN	GA-MLR	GA-SVM	GA-ANN
۴۵	۴/۹۶	۴/۸۵	۴/۹۵	۵/۰۹	۴/۷۴	۴/۹۵	۴/۷۶
*۴۶	۵/۲۶	۵/۳۲	۵/۲۷	۵/۵۸	۵/۴۲	۵/۲۷	۵/۲
۴۸	۴/۰۶	۴/۰۳	۴/۰۷	۴/۱۸	۴/۱۸	۴/۰۷	۴/۲۴
۵۰	۴/۷۷	۴/۸۷	۴/۷۸	۴/۹۹	۴/۷۲	۴/۷۶	۴/۶۷
۵۱	۶/۱۸	۵/۶۰	۶/۱۷	۵/۸۷	۵/۸۴	۶/۱۷	۵/۴۹
۵۲	۶/۱۳	۶/۰۱	۶/۱۵	۵/۹۱	۶/۱۱	۶/۱۵	۶/۰۵
۵۳	۶/۲	۶/۲۲	۶/۲۱	۶/۲۲	۶/۲۴	۶/۲۱	۶/۲۲
۵۴	۶/۰۲	۶/۴۶	۶/۰۳	۶/۴۶	۶/۶۸	۶/۰۳	۶/۷۶
۵۵	۵/۸۷	۶/۴۸	۵/۹۱	۶/۴۶	۶/۴۸	۵/۸۸	۶/۵۵
۵۶	۵/۵۱	۶/۳۳	۵/۷۳	۶/۳۰	۶/۵۱	۵/۵۵	۶/۵۸
۵۷	۵/۵	۶/۳۵	۵/۶۷	۶/۳۲	۶/۲۹	۵/۵۱	۶/۳۰
۵۸	۶/۷۴	۶/۱۳	۶/۳۸	۶/۱۳	۶/۲۳	۶/۶۹	۶/۲۲
۵۹	۶/۹۴	۶/۷۱	۶/۹۳	۶/۷۵	۶/۷۰	۶/۹۳	۶/۷۸
۶۰	۶/۹۴	۶/۶۹	۶/۹۳	۶/۷۰	۶/۶۳	۶/۹۳	۶/۷۶
۶۱	۷/۲۳	۶/۹۸	۷/۲۲	۷/۰۱	۷/۰۲	۷/۲۲	۷/۱۵
۶۳	۶/۹۲	۶/۵۹	۶/۸۲	۶/۶۰	۶/۷۲	۶/۹۳	۶/۸۵
۶۴	۶/۶۹	۶/۶۵	۶/۶۸	۶/۶۱	۶/۵۳	۶/۶۸	۶/۶۰
*۶۶	۶/۵۲	۶/۹۶	۶/۵۳	۶/۹۶	۶/۰۵	۶/۵۱	۶/۴۵
۷۱	۷/۲۷	۶/۷۱	۷/۱۳	۶/۷۸	۶/۹۶	۷/۲۶	۷/۱۲
۷۲	۶/۵۲	۶/۴۱	۶/۵۱	۶/۵۲	۶/۴۲	۶/۵۱	۶/۴۸
*۷۳	۶/۸۲	۶/۸۳	۶/۸۳	۶/۹۳	۶/۷۸	۶/۸۳	۶/۷۸
*۷۴	۶/۱۲	۵/۹۱	۶/۱۱	۵/۵۵	۵/۷۴	۶/۱۱	۶/۱
۷۵	۵/۶۲	۵/۳۳	۵/۴۱	۵/۱۸	۵/۵۰	۵/۶۱	۵/۳۹
*۷۶	۵/۶۲	۵/۳۲	۵/۶۱	۵/۲۳	۵/۳۶	۵/۶۱	۵/۳۴
۷۷	۵	۵/۶۴	۵/۲۲	۵/۵۰	۵/۶۶	۵/۱۱	۵/۵۵
۷۸	۵/۳۶	۵/۳۱	۵/۳۷	۵/۱۹	۵/۴۵	۵/۳۷	۵/۳۷
۷۹	۶/۱۱	۵/۹۴	۶/۱۰	۵/۸۶	۵/۵۱	۶/۱۰	۵/۴۷
*۸۰	۶/۳۱	۶/۲۰	۶/۳۰	۶/۲۰	۶/۱۵	۶/۳۲	۶/۱۳
۸۱	۶/۰۸	۵/۹۸	۶/۰۷	۵/۸۶	۵/۹۳	۶/۰۷	۵/۸۳
۸۲	۶/۶۸	۶/۵۴	۶/۶۷	۶/۴۵	۶/۶۲	۶/۶۷	۶/۶۸
۸۳	۶/۱۴	۶/۰۲	۶/۱۳	۵/۹۳	۶/۰۳	۶/۱۳	۵/۹۲
*۸۴	۶/۱۹	۶/۶۹	۶/۲۰	۶/۶۵	۶/۴۷	۶/۲۰	۵/۹
۸۵	۶/۵۱	۶/۸۲	۶/۵۰	۶/۷۹	۶/۲۵	۶/۵۰	۶/۱۸
*۸۶	۷/۰۶	۷/۱۷	۷/۰۷	۷/۲۲	۶/۶۳	۷/۰۵	۶/۶۷
*۸۷	۷/۵۹	۷/۲۳	۷/۵۸	۷/۳۱	۷/۲۰	۷/۵۸	۷/۱۳
۸۸	۷/۵۹	۷/۴۷	۷/۵۸	۷/۵۳	۷/۵۳	۷/۶۰	۷/۶۷

ادامه جدول(پ-۱)

شماره ترکیب	مقدار تجربی	مقادیر پیش بینی شده(pIC_{50})					
		SR-MLR	SR-SVM	SR-ANN	GA-MLR	GA-SVM	GA-ANN
۹۰	۷/۷	۷/۲۵	۷/۶۹	۷/۲۷	۷/۵۱	۷/۶۹	۷/۶۳
۹۲	۷/۷۲	۷/۵۴	۷/۷۳	۷/۵۳	۷/۵۳	۷/۷۳	۷/۶۵
*۹۴	۷/۳۸	۷/۳۹	۷/۳۹	۷/۳۷	۷/۴۴	۷/۳۹	۷/۴۸
*۹۶	۸/۲۱	۸/۰۵	۸/۲۰	۷/۹۴	۸/۲۴	۸/۲۰	۸/۰۶
۹۷	۸/۲۸	۸/۲۴	۸/۲۷	۸/۱۲	۸/۴۳	۸/۲۷	۸/۱۹
۹۸	۷/۸۵	۷/۹۱	۷/۸۶	۷/۸۶	۸/۰۳	۷/۸۶	۷/۹۱
۱۰۰	۷/۴۱	۷/۵۷	۷/۴۲	۷/۶۰	۷/۴۰	۷/۴۰	۷/۴۰
*۱۰۱	۸/۴۹	۷/۹۱	۸/۴۰	۷/۸۹	۸/۲۲	۸/۴۸	۷/۸۶
۱۰۲	۷/۰۲	۷/۴۳	۷/۰۵	۷/۵۵	۷/۶۲	۷/۰۳	۷/۷۴
۱۰۳	۴/۵۹	۴/۷۸	۴/۶۰	۴/۹۳	۵/۱۳	۴/۶	۴/۴۶

*Validation set

جدول (پ-۱)- مقادیر پیش بینی شده pIC_{50} برای سری تست

شماره ترکیب	مقدار تجربی	مقادیر پیش بینی شده(pIC_{50})					
		SR-MLR	SR-SVM	SR-ANN	GA-MLR	GA-SVM	GA-ANN
۴	۴/۳۷	۴/۶۴	۴/۸۰	۴/۶۵	۴/۶۴	۴/۷۲	۴/۶۳
۷	۴/۶۸	۵/۶۲	۵/۴۱	۵/۵۱	۵/۱۷	۵/۳۱	۵/۵۴
۱۴	۵/۳۸	۵/۵۷	۵/۸۱	۵/۴۸	۵/۸۴	۵/۹۸	۵/۶۷
۱۶	۵/۱	۵/۳۵	۵/۲۴	۵/۲۷	۵/۴۲	۵/۳۴	۵/۵۳
۱۸	۴/۸	۵/۳۷	۵/۳۴	۵/۳۱	۵/۷۴	۵/۸۰	۵/۶۵
۲۵	۵/۱۹	۴/۰۵	۴/۶۵	۴/۰۸	۴/۲۳	۴/۵۸	۴/۴۸
۲۹	۴/۴۸	۴/۴۵	۴/۶۹	۴/۱۳	۴/۳۴	۴/۸۴	۴/۷۶
۳۳	۴/۹۶	۵/۱۱	۵/۱۵	۵/۱۴	۴/۸۲	۴/۵۸	۴/۷۸
۴۷	۶/۱۵	۶/۰۷	۵/۹۷	۶/۵۱	۵/۹۱	۶/۱۴	۶/۱۵
۴۹	۴/۱۳	۴/۳۲	۴/۳۷	۴/۳۵	۴/۴۱	۴/۶۳	۴/۴۳
۶۲	۶/۷۸	۶/۵۶	۶/۶۱	۶/۶۱	۶/۷۶	۶/۸۰	۶/۳۲
۶۵	۶/۶۵	۶/۵۹	۶/۶۲	۶/۵۹	۶/۶۴	۶/۲۷	۶/۴۱
۶۷	۶/۲۱	۷/۰۷	۶/۸۱	۷/۱۸	۶/۴۷	۶/۷۹	۶/۴۳
۶۸	۶/۰۲	۶/۴۴	۶/۵۷	۶/۴۶	۶/۴۰	۶/۱۹	۶/۳۴
۶۹	۵/۸۱	۶/۹۷	+۶/۹۵	۷/۰۲	۶/۷۶	۷/۰۵	۶/۵۹
۷۰	۵/۵۸	۶/۲۹	۶/۲۷	۶/۳۰	۶/۱۲	۶/۱۸	۶/۳۲
۸۹	۶/۹۴	۷/۵۸	۷/۶۳	۷/۶۴	۷/۲۸	۷/۵۲	۷/۲۱
۹۱	۷/۹۶	۷/۴۴	۷/۵۳	۷/۴۱	۷/۵۳	۷/۷۶	۷/۸۴
۹۳	۷/۷	۷/۵۳	۷/۶۲	۷/۵۱	۷/۴۸	۷/۷۱	۷/۷۹
۹۵	۸/۳۹	۷/۸۲	۷/۹۳	۷/۸۱	۸/۰۶	۸/۱۲	۷/۹۰
۹۹	۷/۲۸	۷/۸۵	۸/۰۰	۷/۸۵	۷/۸۴	۷/۷۵	۷/۸۵

ادامه جدول(پ-۱)

شماره ترکیب	مقدار تجربی	مقادیر درصد خطا					
		SR-MLR	SR-SVM	SR-ANN	GA-MLR	GA-SVM	GA-ANN
۱	۴/۹۶	-۷/۶۳	-۰/۲۰	۶/۸۵	-۷/۱۱	-۰/۲۰	۶/۸۵
۲	۴/۷۲	۱/۶۳	۰/۲۱	-۱/۰۶	۳/۳۷	۰/۲۱	۳/۶۰
۳	۴/۸۹	-۷/۸۷	-۰/۲۰	۲/۸۶	-۳/۴۴	-۰/۲۰	۲/۸۶
۵	۵/۳۹	-۴/۲۹	-۰/۱۹	۵	۰/۷۳	۰/۱۹	-۰/۷۴
*۶	۴/۳۵	۵/۲۹	۰/۲۳	-۰/۶۹	۱۲/۸۳	۰/۲۳	۷/۸۲
۸	۵/۸۲	-۳/۲۹	-۰/۱۷	۴/۴۷	۳/۷۸	-۰/۱۷	۶/۷۰
۹	۵/۱۵	۸/۱۸	۰/۱۹	-۷/۷۶	۱/۸۱	-۰/۱۹	۲/۱۴
۱۰	۴/۴۷	۲۲/۷۲	۱۳/۷۶	-۰/۲۲	۱۹/۰۷	۰/۲۲	۲/۶۸
۱۱	۴/۹۶	۱۳/۹۰	۰/۲۰	-۱۳/۷۰	۹/۵۹	۰/۲۰	-۷/۲۶
۱۲	۶/۱۵	-۱۱/۱۵	-۶/۰۸	۱۲/۱۹	-۷/۷۴	-۰/۱۶	۷/۹۶
*۱۳	۵/۴	-۱/۱۴	-۰/۱۹	-۱/۱۳	۱/۳۶	۰/۱۹	-۰/۵۶
۱۵	۵/۳	۲/۹۶	۰/۱۹	-۶/۵۰	۳/۱۷	۰/۱۹	-۲/۸۳
*۱۷	۴/۹۲	۷/۵۵	۲/۸۸	۲/۳۱	۶/۶۳	۰/۲۰	۲/۸۴
۱۹	۵/۶۲	-۱/۵۵	-۰/۱۸	-۳/۳۶	۱/۴۲	۰/۱۸	-۱/۶۰
۲۰	۵/۳۵	۴/۲۷	۰/۱۹	۶/۷۰	-۲/۲۲	-۰/۱۹	۰/۱۸
۲۱	۵/۸۲	-۵/۶۵	-۲/۸۶	۳/۰۹	-۱۶/۶۷	-۰/۱۷	۱/۰۳
۲۲	۵/۸۲	-۲/۵۶	-۲/۲۵	-۱۱/۵۳	-۳/۹۸	-۰/۱۷	۵/۳۲
*۲۳	۵/۰۳	-۱۲/۰۷	-۲/۲۱	۷/۳۹	۳/۰۲	۰/۲۰	-۱۱/۷۲
*۲۴	۵/۸۲	-۶/۴۲	-۰/۱۷	۸/۱۳	-۲/۴۵	-۰/۱۷	۲/۰۶
۲۶	۴/۵۵	-۱۲/۲۳	-۰/۲۲	۳/۷۰	-۸/۲۵	-۰/۲۲	۱/۷۶
۲۷	۵/۱۳	-۳/۶۷	-۰/۱۹	۵/۵۳	-۳/۴۷	۰/۱۹	-۲/۱۴
*۲۸	۴/۵۲	-۹/۴۳	۰/۲۲	-۵/۶۹	-۸/۶۸	-۰/۲۲	۰
۳۰	۴/۳۹	۷/۱۰	۰/۲۳	-۲/۳۰	۱۲/۲۲	۰/۲۳	-۱۱/۱۶
۳۱	۴/۷۷	۸/۳۸	۰/۲۱	۴/۷۰	۴/۵۶	۰/۲۱	-۲/۹۳
۳۲	۵/۱۱	-۱/۲۷	-۰/۲۰	۰/۶۴	۴/۱۴	۰/۲۰	-۱/۹۶
*۳۴	۴/۶۶	۴/۱۳	-۰/۲۱	-۸/۶۷	-۰/۵۹	-۰/۲۱	-۹/۲۲
۳۵	۴/۳۸	۱۱/۲۰	۰/۲۳	-۰/۲۲	۱۳/۱۸	۰/۲۳	-۶/۶۲
۳۶	۵/۱۶	۲/۷۰	۰/۹۴	۰	۴/۱۴	۰/۰۰	۱/۳۶
۳۷	۶/۴۳	-۸/۹۰	-۰/۱۶	-۹/۶	-۷/۹۴	-۰/۱۶	-۹/۵
*۳۸	۴/۷۲	-۱۰/۹۶	-۰/۲۱	-۵/۹۳	-۶/۳۴	-۰/۲۱	-۱۴/۸
۳۹	۴/۴	-۵/۵۰	-۰/۲۳	-۵/۴۵	-۰/۶۹	۰/۲۳	۱۱/۵۹
۴۰	۴/۱۸	۳/۰۵	۰/۲۴	۶/۶۹	۱/۲۲	۰/۲۴	۶/۲۲
۴۱	۴/۸۵	-۲/۴۷	-۰/۲۱	۰/۴۱	-۱/۳۶	۰/۲۱	-۴/۵۳
۴۲	۵/۶۴	-۰/۲۰	۰/۱۸	۵/۶۷	۶/۷۷	۰/۱۸	۳
*۴۳	۴/۳	۱۰/۰۱	۰/۲۳	۱۸/۸۳	۴/۰۳	۰/۲۳	۴/۱۸
۴۴	۴/۱۹	۱۳/۷۲	۰/۲۴	۱۸/۱۳	۳/۸۳	۰/۲۴	۵/۴۸

ادامه جدول(پ-۱)

شماره ترکیب	مقدار تجربی	مقادیر درصد خطا					
		SR-MLR	SR-SVM	SR-ANN	GA-MLR	GA-SVM	GA-ANN
۴۵	۴/۹۶	-۲/۲۲	-/۲۰	۲/۶۲	-۴/۴۹	-/۲۰	-۴
*۴۶	۵/۲۶	۱/۰۹	۰/۱۹	۱۱/۲۶	۳/۰۰	۰/۱۹	-۱/۱۴
۴۸	۴/۰۶	-۰/۷۷	۰/۲۵	۲/۹۵	۲/۹۸	۰/۲۵	۴/۴۳
۵۰	۴/۷۷	۲/۰۷	۰/۲۱	۴/۶۱	-۱/۰۴	-۰/۲۱	-۲/۰۹
۵۱	۶/۱۸	-۹/۳۹	-۰/۱۶	-۵	-۵/۵۸	-۰/۱۶	۱۱/۱۶
۵۲	۶/۱۳	-۲/۴۹	-۰/۱۶	-۳/۵۸	-۰/۷۸	-۰/۱۶	۱/۳۰
۵۳	۶/۲	۰/۳۱	۰/۱۶	۰/۳۲	۰/۶۳	۰/۱۶	-۰/۳۲
۵۴	۶/۰۲	۷/۳۸	۰/۱۷	۷/۳۱	۱۰/۹۰	۰/۱۷	-۱۲/۲۹
۵۵	۵/۸۷	۱۰/۳۱	۰/۶۳	۱۰	۱۰/۴۷	۰/۱۷	-۱۱/۵۸
۵۶	۵/۵۱	۱۴/۸۱	۳/۹۹	۱۴/۳۳	۱۸/۱۴	۰/۷۳	-۱۹/۴۱
۵۷	۵/۵	۱۵/۴۸	۳/۱۷	۱۴/۹۰	۱۴/۴۴	۰/۱۸	۷/۷۱
۵۸	۶/۷۴	-۹/۰۶	-۵/۳۱	-۹	-۷/۵۸	-۰/۷۴	۲/۳
۵۹	۶/۹۴	-۳/۲۷	-۰/۱۴	-۲/۷۳	-۳/۴۷	-۰/۱۴	۲/۳
۶۰	۶/۹۴	-۳/۶۷	-۰/۱۴	۳/۴۵	-۴/۴۴	-۰/۱۴	۲/۵۹
۶۱	۷/۲۳	-۳/۵۰	-۰/۱۴	۳/۰۴	-۲/۸۶	-۰/۱۴	۱/۱۰
۶۳	۶/۹۲	-۴/۷۳	-۱/۴۰	۴/۶۲	-۲/۹۴	۰/۱۴	۱/۰۱
۶۴	۶/۶۹	-۰/۶۳	-۰/۱۵	۱/۱۹	-۲/۴۳	-۰/۱۵	۱/۳۴
*۶۶	۶/۵۲	۶/۷۷	۰/۱۵	-۶/۷۴	-۷/۲۲	-۰/۱۵	۱/۰۷
۷۱	۷/۲۷	-۷/۶۷	-۱/۸۷	۶/۷۴	-۴/۳۲	-۰/۱۴	۲/۰۶
۷۲	۶/۵۲	-۱/۶۴	-۰/۱۵	۰	-۱/۵۸	-۰/۱۵	۰/۶۱
*۷۳	۶/۸۲	۰/۲۱	۰/۱۵	-۱/۶۱	۰/۶۵	۰/۱۵	۰/۵۸
*۷۴	۶/۱۲	-۳/۴۲	-۰/۱۶	۹/۳۱	-۶/۲۸	-۰/۱۶	۰/۳۲
۷۵	۵/۶۲	-۵/۲۴	-۳/۸۲	۷/۸۲	-۲/۱۴	-۰/۱۸	۴/۰۹
*۷۶	۵/۶۲	-۵/۴۱	-۰/۱۸	۶/۹۳	-۴/۵۷	-۰/۱۸	۴/۹۸
۷۷	۵	۱۲/۸۷	۴/۴۵	-۱۰	۱۳/۱۱	۲/۲۰	-۱۱
۷۸	۵/۳۶	-۰/۹۱	۰/۱۹	۳/۱۷	۱/۶۹	۰/۱۹	-۰/۱۸
۷۹	۶/۱۱	-۲/۷۴	-۰/۱۶	۴/۰۹	-۹/۹۰	-۰/۱۶	۱۰/۴۷
*۸۰	۶/۳۱	-۱/۷۰	-۰/۱۶	۱/۷۴	-۲/۵۴	۰/۱۶	۲/۸۵
۸۱	۶/۰۸	-۱/۶۷	۰/۱۶	۳/۶۱	-۲/۴۵	-۰/۱۶	۴/۱۱
۸۲	۶/۶۸	۲/۱۰	-۰/۱۵	۳/۴۲	-۰/۸۸	-۰/۱۵	۳/۵۸
۸۳	۶/۱۴	-۱/۹۱	-۰/۱۶	-۷/۴۳	-۱/۸۷	-۰/۱۶	۴/۶۸
*۸۴	۶/۱۹	۸/۰۷	۰/۱۶	-۴/۳۰	۴/۴۵	۰/۱۶	۵/۰۶
۸۵	۶/۵۱	۴/۷۹	-۰/۱۵	-۲/۲۶	-۴/۰۴	-۰/۱۵	۵/۵۲
*۸۶	۷/۰۶	۱/۵۱	۰/۱۴	۳/۶۸	-۶/۱۲	-۰/۱۴	۶/۰۶
*۸۷	۷/۵۹	-۴/۷۷	-۰/۱۳	۰/۷۹	-۵/۱۴	-۰/۱۳	-۱/۰۵
۸۸	۷/۵۹	-۱/۶۳	-۰/۱۳	۰/۷۹	-۰/۷۷	۰/۱۳	-۱/۰۵

ادامه جدول(پ-۱)

شماره ترکیب	مقدار تجربی	مقادیر درصد خطا					
		SR-MLR	SR-SVM	SR-ANN	GA-MLR	GA-SVM	GA-ANN
۹۰	۷/۷	-۵/۷۹	-۰/۱۳	۵/۸۹	-۲/۴۱	-۰/۱۳	۰/۹
۹۲	۷/۷۲	-۲/۳۹	۰/۱۳	۲/۴۶	-۲/۴۹	۰/۱۳	۰/۹
*۹۴	۷/۳۸	۰/۰۸	۰/۱۴	-۰/۱۳۵	۰/۷۷	۰/۱۴	-۱/۳۵
*۹۶	۸/۲۱	-۱/۹۲	-۰/۱۲	۳/۲۸	۰/۴۲	-۰/۱۲	۱/۸۲
۹۷	۸/۲۸	-۰/۴۹	-۰/۱۲	۱/۹۳	۱/۷۷	-۰/۱۲	۱/۰۸
۹۸	۷/۸۵	۰/۷۹	۰/۱۳	-۰/۱۲۷	۲/۲۳	۰/۱۳	-۰/۷۶۴
۱۰۰	۷/۴۱	۲/۱۷	۰/۱۳	-۲/۵۶	-۰/۲۰	-۰/۱۳	۰/۱۳۵
*۱۰۱	۸/۴۹	-۶/۸۰	-۱/۱۰	۷/۰۶	-۳/۲۱	-۰/۱۲	۷/۴۲
۱۰۲	۷/۰۲	۵/۸۷	۰/۴۱	-۷/۵۴	۸/۴۹	۰/۱۴	-۱۰/۲۵
۱۰۳	۴/۵۹	۴/۱۶	۰/۲۲	-۷/۴۰	۱۱/۶۶	۰/۲۲	۲/۸۳

ادامه جدول(پ-۱)

شماره ترکیب	مقدار تجربی	مقادیر درصد خطا					
		SR-MLR	SR-SVM	SR-ANN	GA-MLR	GA-SVM	GA-ANN
۴	۴/۳۷	۶/۱۶	۹/۹۴	۶/۴۵	۶/۱۳	۷/۹۸	۶/۰۰
۷	۴/۶۸	۲۰/۰۵	۱۵/۵۷	۱۷/۶۶	۱۰/۵۴	۱۳/۵۷	۱۸/۳۴
۱۴	۵/۳۸	۳/۴۶	۸/۰۰	۱/۸۴	۸/۴۹	۱۱/۱۴	۵/۳۲
۱۶	۵/۱	۴/۸۳	۲/۷۶	۳/۴۱	۶/۳۲	۴/۶۸	۸/۵۲
۱۸	۴/۸	۱۱/۸۰	۱۱/۲۱	۱۰/۶۴	۱۹/۶۲	۲۰/۸۵	۱۷/۶۸
۲۵	۵/۱۹	-۲۱/۹۷	-۱۰/۵۰	-۲۱/۳۲	-۱۸/۵۰	-۱۱/۷۵	-۱۳/۷۱
۲۹	۴/۴۸	-۰/۶۳	۴/۷۶	-۷/۷۶	-۳/۲۰	۸/۱۲	۶/۱۷
۳۳	۴/۹۶	۳/۰۳	۳/۷۷	۳/۵۹	-۲/۸۰	-۷/۷۰	-۳/۶۷
۴۷	۶/۱۵	-۱/۲۴	-۲/۹۳	۴/۹۱	-۳/۹۸	-۰/۱۷	۰/۰۸
۴۹	۴/۱۳	۴/۵۰	۵/۸۵	۵/۳۲	۶/۶۸	۱۲/۰۹	۷/۱۹
۶۲	۶/۷۸	-۳/۲۶	-۲/۵۰	-۲/۵۲	-۰/۲۸	۰/۲۷	-۶/۸۲
۶۵	۶/۶۵	-۰/۹۳	-۰/۴۰	-۰/۹۷	-۰/۱۳	-۵/۷۱	-۳/۵۵
۶۷	۶/۲۱	۱۳/۸۷	۹/۶۷	۱۵/۶۷	۴/۱۳	۹/۴۱	۳/۵۰
۶۸	۶/۰۲	۶/۹۱	۹/۲۰	۷/۲۳	۶/۳۶	۲/۸۴	۵/۲۶
۶۹	۵/۸۱	۱۹/۹۱	۱۹/۵۴	۲۰/۸۱	۱۶/۳۶	۲۱/۴۲	۱۳/۳۸
۷۰	۵/۵۸	۱۲/۳۵	۱۲/۲۹	۱۲/۸۶	۹/۶۵	۱۰/۸۳	۱۳/۱۹
۸۹	۶/۹۴	۹/۲۶	۹/۹۳	۱۰/۰۷	۴/۸۸	۸/۳۹	۳/۸۳
۹۱	۷/۹۶	-۶/۵۷	-۵/۴۱	-۶/۸۵	-۵/۴۴	-۲/۵۳	-۱/۵۷
۹۳	۷/۷	-۲/۲۵	-۱/۰۰	-۲/۴۳	-۲/۸۰	۰/۰۷	۱/۲۱
۹۵	۸/۳۹	-۶/۷۹	-۵/۵۴	-۶/۹۴	-۳/۹۶	-۳/۲۳	-۵/۸۲
۹۹	۷/۲۸	۷/۷۸	۹/۸۳	۷/۸۵	۷/۶۴	۶/۴۵	۷/۸۱

جدول (پ-۲)- مقادیر پیش بینی شده pIC_{50} برای سری آموزش و ارزیابی*

شماره ترکیب	مقدار تجربی	مقادیر پیش بینی شده(pIC_{50})			مقادیر درصد خطا		
		GA-MLR	GA-SVM	GA-ANN	GA-MLR	GA-SVM	GA-ANN
۱	۷/۸	۷/۷۶	۷/۷۹	۷/۸۳	۵/۱۳	۰/۱۳	-۰/۳۸
*۳	۶/۵۸	۶/۶۵	۶/۵۹	۷/۱۳	-۱/۰۶	-۰/۱۵	-۸/۳۵
*۶	۷/۷	۷/۵۸	۷/۶۹	۷/۶۷	۱/۵۶	۰/۱۳	۰/۳۹
۷	۷/۷۴	۷/۶۵	۷/۷۳	۷/۳۶	۱/۱۶	۰/۱۳	۴/۹۱
۸	۷/۸	۷/۵۲	۷/۷۹	۷/۸۴	۳/۵۹	۰/۱۳	-۰/۵۱
۹	۷/۸۲	۷/۶۳	۷/۸۱	۷/۸۴	۲/۴۳	۰/۱۳	-۰/۲۶
*۱۱	۷/۳۴	۷/۶۴	۷/۳۵	۷/۳۶	-۴/۰۹	-۰/۱۴	-۰/۲۷
۱۲	۷/۸۵	۷/۷۸	۷/۸۴	۷/۸۰	۰/۸۹	۰/۱۳	۰/۶۴
۱۳	۷/۸۵	۷/۹۱	۷/۸۴	۷/۸۷	-۰/۷۶	۰/۱۳	-۰/۲۵
۱۵	۷/۴۳	۷/۲۷	۷/۴۴	۷/۴۴	۲/۱۵	-۰/۱۳	-۰/۱۳
۱۶	۷/۳	۷/۱۶	۷/۳۰	۷/۲۹	۱/۹۲	۰	۰/۱۴
۱۷	۷/۱۳	۷/۵۳	۷/۱۴	۷/۱۳	-۵/۶۱	-۰/۱۴	۰
*۱۸	۷/۸۹	۷/۵۵	۷/۸۸	۷/۸۱	۴/۳۱	۰/۱۴	۱/۰۱
۱۹	۷/۸۹	۷/۷۴	۷/۸۸	۷/۸۴	۱/۹۰	۰/۱۴	۰/۶۳
۲۰	۷/۲۶	۷/۷۸	۷/۲۷	۷/۲۵	-۷/۱۶	-۰/۱۴	۰/۱۴
*۲۲	۸/۱۵	۸/۰۱	۸/۱۴	۷/۸۷	۱/۷۲	۰/۱۴	۳/۳۴
۲۴	۶/۸۲	۶/۸۸	۶/۸۳	۶/۸۱	-۰/۸۸	-۰/۱۴	۰/۱۴
۲۶	۶/۵۴	۶/۵۸	۶/۵۵	۶/۵۴	-۰/۶۱	-۰/۱۴	۰
*۲۷	۶/۲	۶/۲۳	۶/۲۱	۶/۳۷	-۰/۴۸	-۰/۱۶	-۲/۷۴
۲۸	۷/۸۲	۷/۸۲	۷/۸۳	۷/۸۱	۰	-۰/۱۳	۰/۱۳
۲۹	۷/۴۹	۷/۱۵	۷/۴۸	۷/۴۷	۴/۵۴	۰/۱۳	۰/۲۶
*۳۰	۷/۰۲	۷/۴۱	۷/۰۳	۷/۰۴	۵/۵۶	-۰/۱۳	-۰/۲۶
۳۱	۶/۹۷	۶/۸۰	۳/۹۶	۶/۹۷	۲/۴۴	۰/۱۴	۰
۳۲	۷/۴۳	۷/۴۹	۷/۴۴	۷/۴۵	-۰/۸۱	-۰/۱۴	-۰/۲۷
*۳۳	۸	۷/۷۵	۷/۹۹	۷/۵۱	۳/۱۳	۰/۱۳	۶/۱۳
۳۴	۷/۰۶	۷/۴۳	۷/۰۷	۷/۰۵	-۵/۲۴	-۰/۱۴	۰/۱۴
۳۵	۶/۴۱	۶/۶۳	۶/۴۲	۶/۴۰	-۳/۴۳	-۰/۱۴	۰/۱۴
۳۶	۷/۳۸	۷/۵۳	۷/۳۷	۷/۳۷	-۲/۰۳	۰/۱۴	۰/۱۴
۳۸	۸/۵۲	۸/۶۴	۸/۵۱	۸/۵۱	-۱/۴۱	۰/۱۴	۰/۱۴
۴۰	۷/۷۷	۷/۶۲	۷/۷۶	۷/۷۸	۱/۶۷	۰/۱۴	-۰/۱۴
۴۱	۷/۲۴	۷/۱۲	۷/۲۳	۷/۲۴	۱/۶۶	۰/۱۴	۰
۴۲	۷/۲۱	۷/۱۲	۷/۲	۷/۲۰	۱/۲۵	۰/۱۴	۰/۱۴
۴۳	۷/۹۶	۷/۷۲	۷/۹۵	۷/۸۶	۳/۰۲	۰/۱۴	۱/۲۶
۴۵	۷/۵۴	۷/۶۴	۷/۵۳	۷/۵۳	-۱/۳۳	۰/۱۴	۰/۱۴

*validation set

جدول (پ-۲)- مقادیر پیش بینی شده pIC_{50} برای سری تست

شماره ترکیب	مقدار تجربی	مقادیر پیش بینی شده(pIC_{50})			مقادیر درصد خطا		
		GA-MLR	GA-SVM	GA-ANN	GA-MLR	GA-SVM	GA-ANN
۲	۷/۳۵	۷/۴۳	۷/۲۳	۷/۳۵	-۱/۰۹	۱/۶۳	۰
۵	۶/۲۶	۶/۳۰	۶/۳۸	۶/۰۰	-۰/۶۴	-۱/۹۲	۴/۱۵
۱۰	۸/۴	۸/۰۷	۷/۹۸	۸/۰۳	۳/۹۳	۰/۲۴	۴/۴
۱۴	۷/۷۲	۷/۶۷	۷/۶۶	۷/۷۹	۰/۶۵	۰/۷۸	-۰/۹۱
۲۱	۷/۱۹	۷/۶۷	۷/۶۰	۷/۵۶	۶/۶۸	-۵/۷۰	-۵/۱۵
۲۵	۶/۹۶	۶/۶۹	۶/۶۲	۶/۱۲	۳/۸۸	۴/۸۸	۰/۱۲
۳۷	۶/۶۳	۷/۱۴	۷/۲۰	۶/۸۶	-۷/۶۹	-۸/۶	-۳/۴۷
۳۹	۸/۰۵	۷/۶۹	۷/۸۴	۷/۸۶	۴/۴۷	۲/۶۱	۲/۳۶

منابع

[1] G. Melagraki, A. Afantitis, K. Makridima, H. Sarimveis and O. Igglessi-Markopoulou. (2006) "Prediction of toxicity using a novel RBF neural network training methodology" **Journal of molecular modeling**, 12, 3, pp 297-305.

[2] B. Hemmateenejad, M.A. Safarpour, R. Miri and F. Taghavi. (2004) "Application of ab initio theory to QSAR study of 1, 4-dihydropyridine-based calcium channel blockers using GA-MLR and PC-GA-ANN procedures" **Journal of computational chemistry**, 25, 12, pp 1495-1503.

[3] O. Deeb and B. Hemmateenejad. (2007) "ANN-QSAR Model of Drug-binding to Human Serum Albumin" **Chemical biology & drug design**, 70, 1, pp 19-29.

[4] M. Imai, T. Shiota, K.-i. Kataoka, C.M. Tarby, W.J. Moree, T. Tsutsumi, M. Sudo, M.M. Ramirez-Weinhouse, D. Comer and C.-M. Sun. (2004) "Small molecule inhibitors of the CCR2b receptor. Part 1: Discovery and optimization of homopiperazine derivatives" **Bioorganic & medicinal chemistry letters**, 14, 21, pp 5407-5411.

[5] W.J. Moree, K.-i. Kataoka, M.M. Ramirez-Weinhouse, T. Shiota, M. Imai, M. Sudo, T. Tsutsumi, N. Endo, Y. Muroga and T. Hada. (2004) "Small molecule antagonists of the CCR2b receptor. Part 2: discovery process and initial structure–activity relationships of diamine derivatives" **Bioorganic & medicinal chemistry letters**, 14, 21, pp 5413-5416.

[6] W.J. Moree, K.-i. Kataoka, M.M. Ramirez-Weinhouse, T. Shiota, M. Imai, T. Tsutsumi, M. Sudo, N. Endo, Y. Muroga and T. Hada. (2008) "Potent antagonists of the CCR2b receptor. Part 3: SAR of the (R)-3-aminopyrrolidine series" **Bioorganic & medicinal chemistry letters**, 18, 6, pp 1869-1873.

[۷] L. Saghaie, M. Shahlaei, A. Fassihi, A. Madadkar-Sobhani, M.B. Gholivand and A. Pourhossein. (2011) "QSAR Analysis for Some Diaryl-substituted Pyrazoles as CCR2 Inhibitors by GA-Stepwise MLR" **Chemical biology & drug design**, 77, 1, pp 75-85.

[8] اشرفی م، (۱۳۸۹)، پایان نامه کارشناسی ارشد:"مطالعه‌ی کمی ساختار-فعالیت برخی از ترکیبات تیوکربامات ها به عنوان گروهی جدید از بازدارنده های غیر نوکلئوزیدی HIV "، دانشکده شیمی، دانشگاه شاهرود.

[9] بهرام م، خیامیان ت، عبدالهی ح، فاطمی م، قاسمی ج، کمپانی زارع م، ناصری ع، همتی نـژاد ب، (۱۳۹۴) **"مقدمهای بر کمومتریکس"** چاپ اول، انتشـارات پژوهشـی نـو آوران شـریف، تهـران، ص ۱۷۲.

[۱۰] A. Afantitis, G. Melagraki, H. Sarimveis, O. Igglessi-Markopoulou and G. Kollias. (2009) "A novel QSAR model for predicting the inhibition of CXCR3 receptor by 4-N-aryl-[1, 4] diazepaneureas" **European journal of medicinal chemistry**, 44, 2, pp 877-884.

[۱۱] M. Shahlaei, A. Fassihi, L. Saghaie, E. Arkan and A. Pourhossein. (2011) "A QSAR study of some cyclobutenediones as CCR1 antagonists by artificial neural networks based on principal component analysis" **Daru: journal of Faculty of Pharmacy, Tehran University of Medical Sciences**, 19, 5, pp 376.

[۱۲] T. Asadollahi, S. Dadfarnia, A.M.H. Shabani, J.B. Ghasemi and M. Sarkhosh. (2011) "QSAR models for CXCR2 receptor antagonists based on the genetic algorithm for data preprocessing prior to application of the PLS linear regression method and design of the new compounds using in silico virtual screening" **Molecules**, 16, 3, pp 1928-1955.

[13] محمدرضائی م، (۱۳۸۹)، پایان نامه کارشناسی ارشد:"مطالعه‌ی کمی ساختار-فعالیـت برخـی از ترکیبات سولفونانیلید به عنوان گروهی جدید از داروهای ضد سرطان و ضد HIV برخی از ترکیبات"، دانشکده شیمی، دانشگاه شاهرود.

[۱۴] W. AcreeJr and R. Meyer, Absorption and Luminescence Probes. In Encyclopedia of Analytical Chemistry: Theory and Instrumentation, in, John Wiley & Sons, Ltd. Chichester, UK, 2000

[۱۵] S. Wold. (1995) "Chemometrics; what do we mean with it, and what do we want from it?" **Chemometrics and Intelligent Laboratory Systems**, 30, 1, pp 109-115.

[۱۶] J.R. McGill and B.R. Kowalski.(1978) "Classification of mass spectra via pattern recognition" **Journal of Chemical Information and Computer Sciences**, 18, 1, pp 52-55.

[۱۷] D.L. Massart, B. Vandeginste, S. Deming, Y. Michotte and L. Kaufman. (1988) "Chemometrics: a textbook".

[۱۸] P.C. Jurs. (1996) "**Computer software applications in chemistry**" John Wiley & Sons.

[۱۹] D.L. Massart, B. Vandeginste, S. Deming, Y. Michotte and L. Kaufman. (1988) "Chemometrics: a textbook" .

[۲۰] J.C. Miller and J.N. Miller. (1988) "Statistics for analytical chemistry".

[۲۱] B. Hemmateenejad, M.A. Safarpour, R. Miri and F. Taghavi. (2004) "Application of ab initio theory to QSAR study of 1, 4-dihydropyridine-based calcium channel blockers using GA-MLR and PC-GA-ANN procedures" **Journal of computational chemistry**, 25, 12, pp 1495-1503.

[۲۲] O. Deeb and B. Hemmateenejad. (2007) "ANN-QSAR Model of Drug-binding to Human Serum Albumin" **Chemical biology & drug design**, 70, 1, pp 19-29.

[۲۳] H. González-Díaz, I. Bonet, C. Terán, E. De Clercq, R. Bello, M.M. García, L. Santana and E. Uriarte. (2007) "ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds" **European journal of medicinal chemistry**, 42, 5, pp 580-585.

[۲۴] M. Shahlaei, A. Fassihi and L. Saghaie. (2010) "Application of PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: a comparative study" **European journal of medicinal chemistry**, 45, 4, pp 1572-1582.

[۲۵] M. Goodarzi, M.P. Freitas and R. Jensen. (2009) "Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3, 5-dimethylbenzyl) uracil derivatives using MLR, PLS and SVM regressions" **Chemometrics and intelligent laboratory systems**, 98, 2, pp 123-129.

[۲۶] S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali and P. Norouzi. (2008) "Exploring QSARs for Antiviral Activity of 4-Alkylamino-6-(2-hydroxyethyl)-2-methylthiopyrimidines by Support Vector Machine" **Chemical biology & drug design**, 72, 3, pp 205-216.

[۲۷] E. Pourbasheer, S. Riahi, M.R. Ganjali and P. Norouzi. (2011) "QSAR study of C allosteric binding site of HCV NS5B polymerase inhibitors by support vector machine" **Molecular diversity**, 15, 3, pp 645-653.

[۲۸] M.H. Fatemi and S. Gharaghani. (2007) "A novel QSAR model for prediction of apoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine" **Bioorganic & medicinal chemistry**, 15, 24, pp 7746-7754.

[۲۹] M. Fernandez, J. Caballero, L. Fernandez and A. Sarai. (2011) "Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM)" **Molecular diversity**, 15, 1, pp 269-289.

[۳۰] Z. Cheng, Y. Zhang and W. Fu. (2010) "QSAR study of carboxylic acid derivatives as HIV-1 Integrase inhibitors" **European journal of medicinal chemistry**, 45, 9, pp 3970-3980.

[31] قرقانی، س، (۱۳۸۷)، پایان نامه کارشناسی ارشد: "مدلسازی و پیش بینی فعالیت ضد سرطانی ۴-آریل-۴- H-کرومن ها با استفاده از اصول QSAR و مدلسازی و پیش بینی ضریب برگزیدگی الکترود انتخابگر آنیون"، دانشگاه مازندران.

[32] لواین ای. ان ، اسلامپور غ، پارسافر غ، مقاری ع، نجفی ب، (۱۳۸۷) **"شیمی کوانتومی"** جلد سوم، چاپ اول، انتشارات فاطمی.

[33] طاهرزاده م، (۱۳۹۳)، پایان نامه کارشناسی ارشد:"مطالعه‌ی کمی ساختار-خاصیت ضریب فعالیت در رقت بی نهایت ترکیبات آلی و آب در محیط مایع یونی ۱-بوتیل ۱متیل پیرولیدینیوم تریسیانومتانید"، دانشکده شیمی، دانشگاه شاهرود.

[34] D.E. Golberg. (1989) "Genetic algorithms in search, optimization, and machine learning" **Addionwesley**, 1989.

[35] I. Aleksander and H. Morton. (1990) "**An introduction to neural computing**" 3, Chapman & Hall London.

[36] L.C. Jain and V.R. Vemuri.(1998) "**Industrial applications of neural networks**" 3, CRC press.

[37] J.H. Holland. (1975) "Adaptation in natural and artificial system: an introduction with application to biology, control and artificial intelligence" **Ann Arbor, University of Michigan Press**.

[38] C. Cortes and V. Vapnik.(1995) "Support-vector networks" **Machine learning**, 20, 3, pp 273-297.

[39] L. Pettus and R. Wurz. (2008) "For recent reviews on the design of small molecule inhibitors for p38 MAP kinase, see" **Curr. Top. Med. Chem**, 8, pp 1452.

[40] L. Wang. (2005) "**Support Vector Machines: theory and applications**" 177, Springer Science & Business Media.

[41] M. Martínez-Ramón and C. Christodoulou. (2005) "Support vector machines for antenna array processing and electromagnetics" **Synthesis Lectures on Computational Electromagnetics**, 1, 1, pp 1-120.

[42] R. HyperChem. (2002) "7.0 for windows, Hypercube" **Inc.: Gainesvill**

[43] http:/www.disat.unimib.it%252Fchm%252F

[44] SPSS for windows, The statistical package for IBMPC, SPSS Inc.,http://www.spss.com

[45] M.J. Norušis. (1992) "**SPSS/PC+ professional statistics version 5.0**" SPSS, Inc.

[46] http://www.condor-bcm.tmc.edu./Statistics/Spss.htm

[47] MATLAB 7.8, the Math Work, Inc., Natick, MA, USA.

[48] D. Sun, Z. Wang, Y. Di, J.C. Jaen, M. Labelle, J. Ma, S. Miao, A. Sudom, L. Tang and C.S. Tomooka. (2008) "Discovery and initial SAR of arylsulfonylpiperazine inhibitors of 11β-hydroxysteroid dehydrogenase type 1 (11β-HSD1)" **Bioorganic & medicinal chemistry letters**, 18, 12, pp 3513-3516.

[49] T. Barf and M. Williams. (2002) "Recent progress in 11-beta-hydroxysteroid dehydrogenase type 1 (11-beta-HSD1) inhibitor development" **Drugs Future**, 3

Abstract

In the first section, quantitative structure-activity relationship (QSAR) study was conducted on the inhibition effect (pIC_{50}) of 103 drug compounds of homopiperazine, diamine, 3-aminopyrrolidine derivatives.Two variable selection methods of stepwise regression (SR) and genetic algorithm (GA) were applied to select the important descriptors. The three methods of multiple linear regression (MLR), artificial neural network (ANN) and support vector machines (SVM) were used for construction of methods with selected descriptors for prediction of pIC_{50} of these compounds. The performance of each model was investigated by the test set. The mean square error (MSE) and standard error of prediction (SEP) for the test sets of SR-MLR, SR-ANN, SR-SVM, GA-MLR, GA-ANN and GA-SVM were 0.328 , 0.573 and 0.346 , 0.588 and 0.276 , 0.513 and 0.228 , 0.477 and 0.225 , 0.475 and 0.257, 0.532 respectively.

In the second section, quantitative structure-activity relationship (QSAR) study was conducted on the inhibition effect (pIC_{50}) of 42 drug compounds of Arylsulfonylpiperazine derivatives. The best calculated descriptors were selected by genetic algorithm. The multiple linear regression (MLR), artificial neural network (ANN) and support vector machines (SVM) were applied construction of models to predict of the inhibition effect (pIC_{50}) of these compounds. The performance of each model was investigated by test set. The mean squared error (MSE) and standard error of prediction (SEP) and the coefficient of determination (R^2) very close to each other, and this reflects to this fact, that all three models have good ability to predict the pIC_{50} of the studied compounds.

Keywords: Inhibition effect, Stepwise Regression (SR), Genetic Algorithm (GA), Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Support Vector Machines (SVM)

Shahrood University of Technology

Faculty of Chemistry

Application of different chemometrics methods to predict the activity of some drug compounds

SaeedNekoei

Supervisor:

Dr. N. Goudarzi

Advisor:

Dr. M. Nekoei

September2015