مجله شیمی کوانتومی و اسپکتروسکوپی
۱۳:٤۲ (۱٤۰۲) ٦۲-۸۰
Journal of Quantum Chemistry and Spectroscopy (JQCS)

# پیش بینی فعالیت ضدسرطانی برخی مشتقات N-آریل۲و۲-دی کلرواستامید و آریل۲و۲-دی کلرواستات با استفاده از روشهای خطی و غیرخطی مطالعه ارتباط کمی ساختار-فعالیت(QSAR)

**مهدی نکوئی*، مجید محمدحسینی، طیبه جواهری**

گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

تاریخ ثبت اولیه:۱۴۰۲/۰۴/۰۳، تاریخ دریافت نسخه اصلاح شده:۱۴۰۲/۰۶/۲۵ ، تاریخ پذیرش قطعی:۱۴۰۲/۰۷/۱۲

## چکیده

مطالعه ارتباط کمی-ساختار فعالیت(QSAR) جهت پیش‌بینی فعالیت دارویی($IC_{50}$) برخی از مشتقات N-آریل۲و۲-دی کلرواستامید و آریل۲و۲-دی کلرواستات با استفاده از روشهای رگرسیون خطی چندگانه (MLR) و شبکه های عصبی مصنوعی (ANN) به عنوان روشهای خطی و غیرخطی جهت درمان سرطان انجام شد. در ابتدا ساختار ترکیبات مورد نظر توسط نرم افزار هایپرکم رسم و بهینه گردید. ترکیبات رسم شده جهت محاسبه توصیف‌کننده‌های مولکولی به نرم افزار دراگون وارد و تعداد ۱۴۸۱ توصیف کننده برای هر مولکول محاسبه شد. جهت انتخاب مناسب ترین توصیف کننده‌ها از روش رگرسیون مرحله‌ایی استفاده گردید. پس از انتخاب مناسب ترین توصیف کننده‌ها، از دو روش MLR و ANN به عنوان روش‌های خطی و غیر خطی مطالعه ارتباط کمی-ساختار فعالیت جهت مدلسازی و پیش‌بینی فعالیت دارویی ترکیبات، استفاده شد. عملکرد هر مدل توسط چندین پارامتر آماری مورد ارزیابی قرار گرفت. نتایج بدست آمده نشان از برتری روش ANN نسبت به MLR دارد.

***واژه های کلیدی : ارتباط کمی ساختار- فعالیت، ، مشتقات N-آریل۲و۲-دی کلرواستامید - آریل۲و۲-دی کلرواستات، رگرسیون خطی چند گانه، شبکه های عصبی مصنوعی***

## ۱. مقدمه

با توجه به گسترش انواع بیماری ها، یکی از موضوعات مورد توجه و با اهمیت، طراحی، سنتز و تولید دارو می باشد. روندی که در گذشته منجر به کشف و توسعه داروهای جدید می‌شد به روش آزمون و خطا صورت می‌گرفت که روشی وقت گیر و هزینه‌بر بوده است. مشکل دیگری که پیش‌روی محققان است، 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 شیمیدان‌ها و محققان دارویی پیش‌بینی فعالیت ترکیبات دارویی، قبل از

***عهده دار مکاتبات:** مهدی نکوئی
**نشانی:** گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران
**تلفن:**۰۲۳۳۲۳۹۴۲۸۹ **پست الکترونیک:**E-mail:m_nekoei1356@yahoo.com

سنتز آن‌ها می‌باشد. چرا که انجام بسیاری از آزمایشات مستلزم صرف زمان و هزینه‌های زیادی است. از این رو نیاز به استفاده از روش‌های تئوری و محاسباتی که بدون انجام آزمایش بتوانند ویژگی و یا فعالیت ترکیبات دارویی را پیش‌بینی کنند ضروری به نظر می‌رسد. ظهور علم کمومتریکس توانسته راه حلی برای رفع این مشکلات باشد[۴-۱]. یکی از مهم‌ترین زمینه‌های کاربرد روش‌های کمومتریکس، مطالعه ارتباط بین خواص مولکول‌ها با ویژگی‌های ساختاری آن‌هاست. این نوع مطالعات که با عنوان ارتباط کمی ساختار- فعالیت [۱](QSAR) معروف شده‌اند، به بررسی نحوه ارتباط بین خواص مختلف مولکول‌ها با مشخصات ساختاری و ذاتی آن‌ها می‌پردازند[۱۰-۵]. بررسی ساختار شیمیایی و فعالیت ترکیبات، پیش بینی فعالیت ترکیبات جدید را بر اساس اطلاعات مرتبط به ساختار شیمیایی آن‌ها امکان‌پذیر می‌سازد. $IC_{50}$ غلظتی از یک دارو است که منجر به ۵۰٪ اثر مهاری در نمونه مورد بررسی می‌گردد. روشهای مختلفی برای اندازه گیری این پارامتر وجود دارد. اما از آنجاییکه این اندازه‌گیری‌ها، وقت گیر و هزینه بر می باشند، استفاده از روش‌هایی برای تخمین این پارامتر($IC_{50}$) ضروری به نظر می‌رسد. برای پیش بینی فعالیت دارویی ترکیبات شیمیایی، روش ارتباط کمی ساختار - فعالیت (QSAR) روش مطمئن و مناسبی می باشد. QSAR ارتباط ریاضی بین ساختار و فعالیت دسته‌ایی از ترکیبات دارویی را توصیف می‌کند. در سالیان اخیر، روند رو به افزایش مقالات منتشره در منابع علمی بر اساسQSAR، دلالت بر جایگاه منحصر به فرد این دیدگاه در شیمی نظری و تعمیق بینش دانشمندان در فهم و توجیه آن دارد[۱۴-۱۱]. از جمله روشهایی که جهت مدل سازی و پیش‌بینی فعالیت ترکیبات داروئی مورد استفاده قرار می‌گیرند، می‌توان به روش‌های خطی از جمله رگرسیون خطی چندگانه(MLR)، کمترین مربعات جزیی و روش‌های غیرخطی مانند شبکه‌های عصبی مصنوعی (ANN)، ماشین بردار پشتیبان، روش های فازی عصبی و... اشاره نمود[۱۹-۱۵].

هدف اصلی این تحقیق ارائه‌ی مدل‌های مناسب جهت پیش‌بینی فعالیت داروئی برخی از مشتقات N-آریل۲و۲-دی کلرواستامید و آریل۲و۲-دی کلرواستات در برابر بیماری سرطان با استفاده از روش های SW- MLR و SW-ANN می‌باشد.

## ۲. روش‌های محاسباتی

### ۲-۱. انتخاب سری داده‌ها

سری داده‌ها شامل فعالیت دارویی۳۷ ترکیب از مشتقات N-آریل۲و۲-دی کلرواستامید و آریل۲و۲-دی کلرواستات به عنوان بازدارنده برای درمان سرطان مورد بررسی قرار گرفت[۲۰]. 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. در این کار این ترکیبات به صورت تصادفی به دو گروه سری آموزش و سری تست تقسیم شدند، سری آموزش شامل ۳۰ مولکول(٪۸۰) و سری تست شامل ۷ مولکول(٪۲۰) می‌باشد. مقادیر $pIC_{50}$ به عنوان متغیر وابسته و توصیف کننده‌ها به عنوان متغیر مستقل انتخاب شدند. سری آموزش جهت ایجاد یک مدل مناسب و سری تست جهت

[1] Quantitative structure activity relationship

ارزیابی مدل مورد استفاده قرار گرفت. لازم به ذکر است جهت مقایسه، سری پیش بینی (تست) در هر دو روش دارای ترکیبات یکسان می باشد. (در ANN از سری ارزیابی برای بررسی و جلوگیری از برازش اضافی استفاده گردید).

**۲-۲. رسم و بهینه سازی ساختار مولکول‌ها**

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 استفاده از روش نیمه تجربی کوانتمیAM1 بهینه گردید. این بهینه سازی تا زمانی ادامه یافت که جذر میانگین مربعات گرادیان انرژی به ۰/۰۰۱ کیلوکالری بر مول رسید. با استفاده از این نرم افزار می توان اطلاعات فراوانی نظیر زوایای پیوندی، طول پیوندها، زوایای پیچش، بار اتم ها، انرژی تشکیل مولکول و... را بدست آورد. برخی دیگر از قابلیت‌های این نرم افزار عبارتند از: توانایی نمایش ساختار مولکولی با قابلیت کنترل آن (از جمله انتخاب، چرخش، تبدیل و تغییر اندازه ساختار مولکولی)، دارای ابزارها و متدهای محاسباتی مختلف (از جمله تعیین تراز انرژی) و امکان تعریف نوع اتم، جرم اتمی و سایر ویژگی ها و ....

**۲-۳. محاسبه توصیف کننده‌های مولکولی**

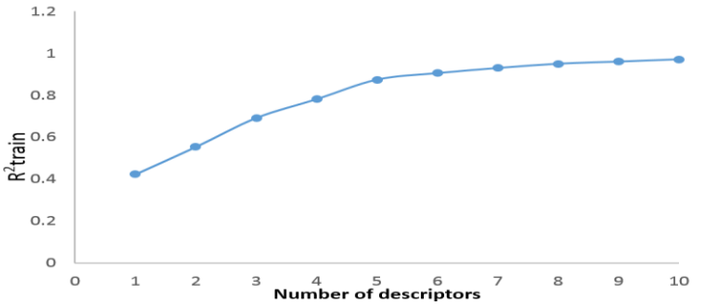
توصیف کننده‌ها مقادیر عددی هستند که ویژگیهای مختلف مولکول را نشان می‌دهند. توصیف‌کننده‌های مولکولی نتیجه نهایی یک استدلال و روش ریاضی است که اطلاعات شیمیایی را به رمز تبدیل می کند و آنها را به صورت یک نماد نشان می‌دهد که ارائه دهنده یک ویژگی مولکول به صورت یک عدد مفید می‌باشد. هر یک از این توصیف کننده‌ها، اطلاعات خاصی از مولکول را در اختیار می گذارند. توصیف‌کننده‌های مولکولی مختلفی برای اهداف گوناگون به کار برده شده‌اند. اختلاف این توصیف‌کننده‌ها در پیچیدگی اطلاعات رمزگزاری شده و زمان مورد نیاز برای محاسبه می‌باشد. اولین نوع توصیف‌کننده ها، توصیف‌کننده‌های توپولوژی می‌باشند. این توصیف‌کننده‌ها از روی گراف‌های مولکولی بدست می‌آیند و جزء ساده‌ترین نوع توصیف‌کننده‌ها می‌باشند و به ساختار فضایی مولکول ارتباطی نداشته و تنها به نوع اتم، نوع پیوندها و نحوه ارتباط اتم‌ها به یکدیگر وابسته است. از جمله این توصیف کننده ها می‌توان به تعداد اتم‌ها، شاخص‌های ارتباطی مولکولی و وزن مولکولی و ... اشاره کرد. دومین نوع توصیف کننده‌ها، توصیف کننده‌های هندسی است. این توصیف‌کننده‌ها با ساختار سه بعدی مولکول‌ها در ارتباط می‌باشند. برای محاسبه این توصیف‌کننده‌ها ابتدا می‌بایست ساختار فضایی مولکول‌ها بهینه شود. برخی از این توصیف کننده‌ها عبارتند از: حجم مولکولی، مساحت سطح و مساحت سطح در دسترس حلال. توصیف‌کننده‌های شیمی کوانتومی از جمله توصیف‌کننده‌های دیگری هستند که با استفاده از بهینه‌سازی نیمه تجربی ساختار مولکول‌ها در نرم‌افزارهای مختلف بدست می‌آیند. از جمله این توصیف کننده‌ها می‌توان به انرژی بالاترین تراز اشغال شده، انرژی پایین ترین تراز اشغال نشده، بار و الکترونگاتیویته اتم‌ها و ... اشاره کرد. توصیف‌کننده‌های دیگر، توصیف کننده های فیزیکوشیمیایی هستند که این توصیف‌کننده‌ها بیانگر بعضی از خواص فیزیکوشیمیایی مولکول‌ها می‌باشند که به ساختار مولکول وابستگی شدیدی نشان می‌دهند. از قبیل: ضریب تقسیم آب-اکتانول، ویسکوزیته، میزان

حلالیت ترکیبات در آب، شکست مولکولی، نقطه ذوب و نقطه جوش. توصیف‌کننده‌های ارتباطی مولکولی نیز از جمله توصیف کننده‌های مهم دیگری هستند که اطلاعاتی از جمله اندازه و ساختار مولکول، مرتبه شاخه‌دار شدن و نحوه ارتباط اتم‌ها در مولکول را بیان می‌کنند[۲۱].

برای محاسبه توصیف کننده ها، بعد از رسم ساختارهای مولکولی به کمک نرم افزار Hyper Chem و بهینه سازی ساختار آنها، این ساختارها به نرم افزارDragon وارد شده و توصیف کننده‌های مولکولی به تعداد ۱۴۸۱ مورد به وسیله این نرم افزار محاسبه شدند.

### ۲-٤. انتخاب مناسب‌ترین توصیف‌کننده‌ها

جهت انتخاب مناسب ترین توصیف کننده ها از روش رگرسیون مرحله ایی[1] استفاده شد. در روش رگرسیون مرحله ای، متغیرها یکی پس از دیگری وارد مدل شدند در این حالت، ابتدا متغیری وارد مدل می‌شود 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 بدهد، قبل از ورود متغیر جدید از مدل خارج می‌شود. به این ترتیب داده‌های $pIC_{50}$ به عنوان متغیر وابسته و توصیف‌کننده‌ها به عنوان متغیر مستقل در نظر گرفته شده و تکنیک رگرسیون مرحله‌ای انجام شد. همانطور که می‌دانیم روش رگرسیون مرحله‌ایی تعداد زیادی مدل ارائه می‌کند. که مدل اول شامل یک توصیف‌کننده، مدل دوم شامل دو توصیف‌کننده و ... می‌باشد. با افزایش تعداد توصیف‌کننده‌ها بالطبع مقدار $R^2$ افزایش و RMSE[2]( خطای جذر میانگین مربعات)کاهش می‌یابد. اما بدلیل پیچیدگی مدل، نمی‌توانیم تعداد زیادی توصیف کننده را جهت مدل‌سازی انتخاب کنیم. بدین منظور و جهت انتخاب تعداد توصیف کننده‌های مناسب، نمودار پارامترهای مختلف آماری از جمله $R^2$ برحسب تعداد توصیف کننده‌ها رسم گردید که در شکل۱ نشان داده شده است. بر طبق این شکل، تعداد ۵ توصیف کننده به عنوان توصیف کننده‌هایی که بیشترین ارتباط را با فعالیت دارویی($pIC_{50}$) دارند، انتخاب شدند. زیرا بعد از ۵ توصیف‌کننده، مقدار $R^2$ تغییر محسوسی نمی‌کند.این ۵ توصیف کننده به همراه مفهوم و نوع آنها در جدول۱ ارائه شده است.




**شکل۱. نمودار پارامتر آماری ($R^2$train) برحسب تعداد توصیف کننده‌ها**

[1] Stepwise
[2]Root-mean-square-error

**جدول۱. توصیف‌کننده‌های انتخاب شده توسط رگرسیون خطی چندگانه مرحله به مرحله**

| نشانه توصیف کننده | مفهوم توصیف کننده | نوع توصیف کننده |
|---|---|---|
| MWC09 | Moleculer walk count of order 09 | Molecularwalk counts |
| RDF020e | Radial Distribtion Function-2.0 | RDF descriptors |
| HGM | Geometric mean on the leverage magnitude | GETAWAY descriptors |
| HATS3v | Leverge weighted auto correlation of lag 3/weighted by atomic van der waals volumes | GETAWAY descriptors. |
| HATS5e | Leverge weighted auto correlation of lag 5/weighted by atomic sanderson electronegativities | GETAWAY descriptors. |

### ۲-۵. ارزیابی توصیف کننده ها

به منظور ارزیابی توصیف‌کننده‌های انتخاب شده مبنی بر مستقل بودن از همدیگر در جدول۲ ماتریس همبستگی توصیف‌کننده-های انتخاب شده آورده شده است. همانطور که در این جدول مشاهده می‌شود ضریب همبستگی بین توصیف‌کننده‌های انتخاب شده همگی کمتر از ۰/۷۹۰ می‌باشد. لذا نتایج جدول نشان می‌دهد که بین توصیف‌کننده‌های انتخاب شده همبستگی چندانی وجود نداشته و توصیف‌کننده‌های تقریبا مستقل از هم هستند.

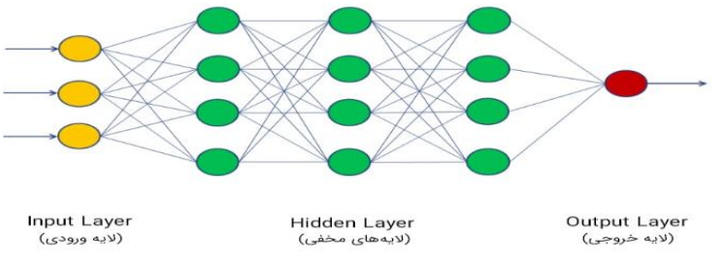
**جدول۲. ماتریس ضرایب همبستگی توصیفگرهای انتخاب شده**

| | MWC09 | RDF020e | HGM | HATS3v | HATS5e |
|---|---|---|---|---|---|
| MWC09 | 1 | | | | |
| RDF020e | 0.242 | 1 | | | |
| HGM | -0.592 | -0.559 | 1 | | |
| HATS3v | -0.237 | -0.456 | 0.790 | 1 | |
| HATS5e | -0.043 | -0.535 | 0.460 | 0.501 | 1 |

### ۲-٦. شبکه های عصبی مصنوعی

شبکه‌های عصبی مصنوعی، سیستم‌ها و روش‌های محاسباتی نوین برای یادگیری، نمایش دانش و در انتها اعمال دانش به دست آمده در جهت پیش‌بینی پاسخ‌های خروجی از سامانه‌های پیچیده هستند. ایده اصلی این گونه شبکه‌ها تا حدودی الهام‌گرفته از شیوه کارکرد سیستم عصبی زیستی برای پردازش داده‌ها و اطلاعات به منظور یادگیری و ایجاد دانش می‌باشد. شبکه عصبی مصنوعی روشی است که دانش ارتباط بین چند مجموعه داده را از طریق آموزش فراگرفته و برای استفاده در موارد مشابه ذخیره می‌کند. یک

شبکه عصبی مصنوعی، از سه لایه ورودی، خروجی و پنهان تشکیل می‌شود. هر لایه شامل گروهی از سلول‌های عصبی (نورون) است که عموماً با کلیه نورون‌های لایه‌های دیگر در ارتباط هستند، مگر این که کاربر ارتباط بین نورون‌ها را محدود کند؛ ولی نورون‌های هر لایه با سایر نورون‌های همان لایه، ارتباطی ندارند. با استفاده از دانش برنامه‌نویسی رایانه می‌توان ساختار داده‌ای طراحی کرد که همانند یک نورون عمل نماید. سپس با ایجاد شبکه‌ای از این نورون‌های مصنوعی به هم پیوسته، ایجاد یک الگوریتم آموزشی برای شبکه و اعمال این الگوریتم به شبکه آن را آموزش داد. نورون، کوچک‌ترین واحد پردازشگر اطلاعات است که اساس عملکرد شبکه‌های عصبی را تشکیل می‌دهد. یک شبکه عصبی مجموعه‌ای از نورون‌هاست که با قرار گرفتن در لایه‌های مختلف، معماری خاصی را بر مبنای ارتباطات بین نورون‌ها در لایه‌های مختلف تشکیل می‌دهند. نورون می‌تواند یک تابع ریاضی غیرخطی باشد، در نتیجه یک شبکه عصبی که از اجتماع این نورون‌ها تشکیل می‌شود، نیز می‌تواند یک سامانه کاملاً پیچیده و غیرخطی باشد. در شبکه عصبی هر نورون به‌طور مستقل عمل می‌کند و رفتار کلی شبکه، برآیند رفتار نورون‌های متعدد است. به عبارت دیگر، نورون‌ها در یک روند همکاری، یکدیگر را تصحیح می‌کنند[۲۵-۲۲]. شکل ۲ نمایی از یک شبکه عصبی مصنوعی را نشان می دهد




**شکل ۲. نمایی از یک شبکه عصبی مصنوعی**

## ۳. نتایج و بحث

### ۳-۱. مدل‌سازی به روش رگرسیون خطی چندگانه(MLR)

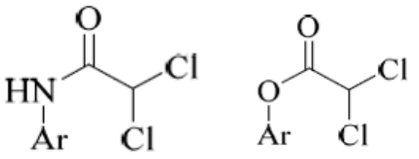
پس از انتخاب مناسب‌ترین توصیف‌کننده‌ها توسط روش مرحله‌ای، مرحله بعدی، ایجاد مدل میان توصیف‌کننده‌های انتخاب شده و $pIC_{50}$ می‌باشد. بین توصیف کننده‌ها و فعالیت دارویی ترکیبات برای سری آموزش با استفاده از روش MLR رابطه زیر به عنوان مدل خطی بدست آمد:

$$\mathbf{pIC_{50}} = 5.859 + 2.828(\mathbf{MWC09}) - 0.080(\mathbf{RDF020e}) - 0.197(\mathbf{HGM}) + 7.289\ (\mathbf{HATS3v}) - 2.152\ (\mathbf{HATS5e})$$

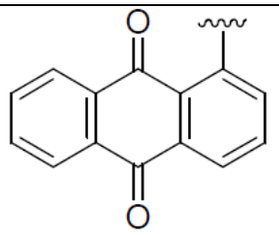
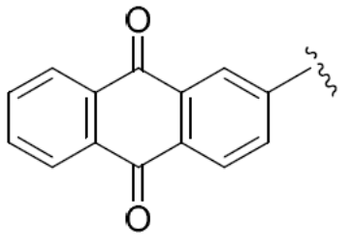
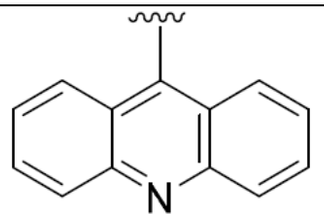
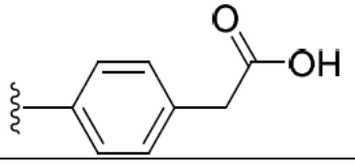
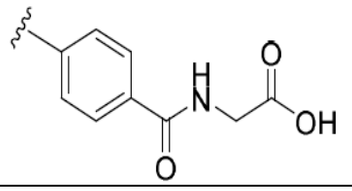
سپس از معادله بدست آمده برای پیش‌بینی فعالیت دارویی ترکیبات سری تست استفاده گردید. مقادیر تجربی و پیش‌بینی شده $pIC_{50}$ برای کلیه ترکیبات مجموعه آموزش و تست در جدول (۳) آورده شده است. شکل(۳) نمودار مقادیر پیش بینی شده بر حسب مقادیر تجربی را نشان می دهد. در این شکل نزدیکی نتایج به خط راست قدرت پیش بینی مدل را نشان می دهد. شکل(۴)

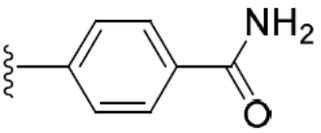
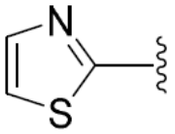
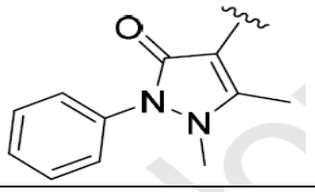
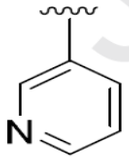
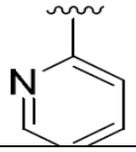
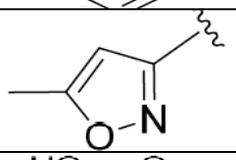
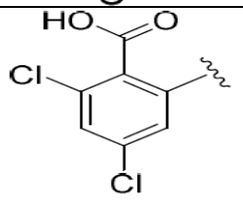
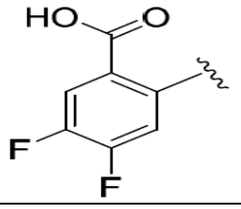
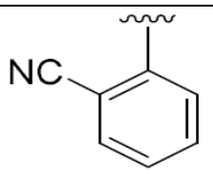
نمودار مقادیر باقیمانده ها(اختلاف مقادیر پیش بینی شده و مقادیر تجربی) را بر حسب مقادیر تجربی نشان می دهد. در این شکل پراکندگی نقاط در حول خط صفر نشان می دهد که خطای سیستماتیک در مدل وجود ندارد.

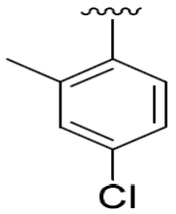
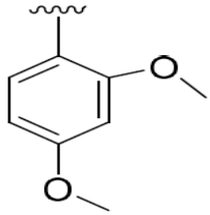
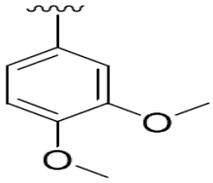
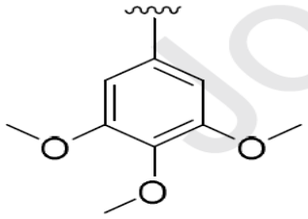
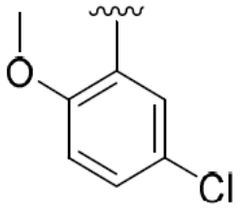
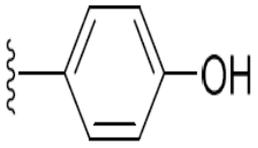
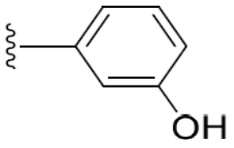
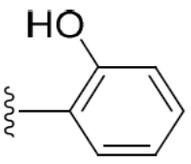
**جدول۳. مقادیر تجربی و محاسبه شده $pIC_{50}$ ترکیبات مختلف برای مجموعه‌های آموزشی و پیش‌بینی در مدل‌های** SW-MLR, SW-ANN

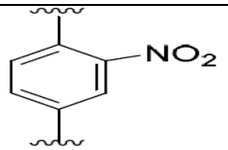
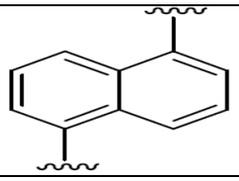
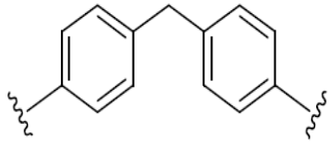
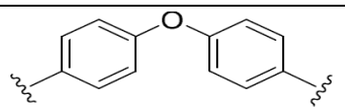
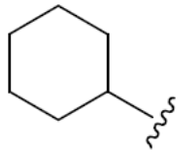
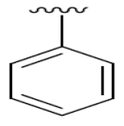
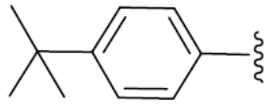
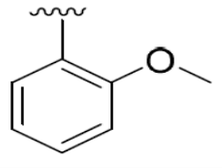
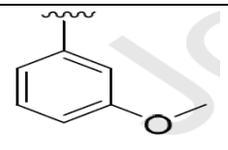
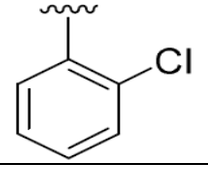


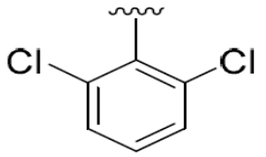
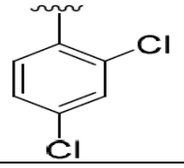
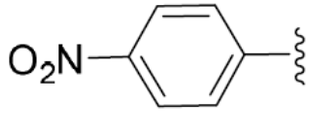
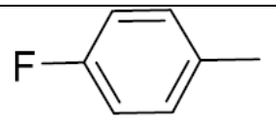
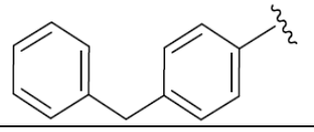

**f1-f26** **f27-f36**

| Name | Ar | $IC_{50}$ | $pIC_{50}$ | SW-MLR | SW-ANN |
|---|---|---|---|---|---|
| F1 |  | 9.46 ± 1.23 | 5.02 | 4.96 | 5.02 |
| F2 |  | 11.47 ± 1.72 | 4.94 | 4.94 | 4.93 |
| F3 |  | 6.03 ± 0.93 | 5.22 | 5.26 | 5.22 |
| F4 |  | 53.67 ± 2.54 | 4.27 | 4.21 | 4.26 |
| F5 |  | 72.94 ± 1.63 | 4.14 | 4.31 | 4.13 |

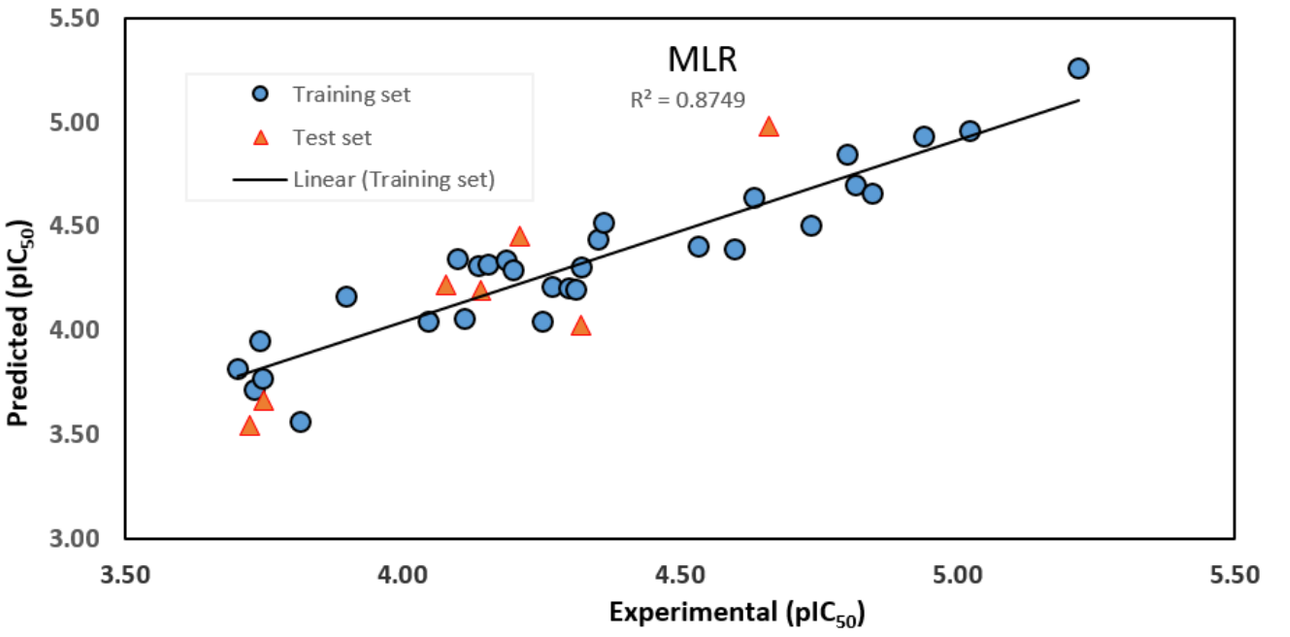
| | | | | | |
|---|---|---|---|---|---|
| F6 |  | 44.32 ±3.27 | 4.35 | 4.44 | 4.43 |
| F7 |  | 184.4 ± 8.80 | 3/73 | 3/72 | 3.72 |
| F8 |  | 125.99 ± 6.33 | 3.89 | 4.16 | 3.90 |
| F9* |  | 188.07 ± 5.72 | 3.72 | 3.55 | 3.68 |
| F10 |  | 197.82 ± 9.18 | 3.70 | 3.82 | 3.79 |
| F11 |  | 47.52 ± 2.44 | 4.32 | 4.30 | 4.29 |
| F13 |  | 49.98 ± 3.37 | 4.30 | 4.21 | 4.27 |
| F14* |  | 83.32 ± 2.66 | 4.08 | 4.22 | 4.17 |
| F15 |  | 178.59 ± 6.59 | 3.75 | 3.77 | 3.74 |

| | | | | | |
|---|---|---|---|---|---|
| F16 |  | 180.97 ±5.98 | 3.74 | 3.95 | 3.75 |
| F17* |  | 21.84 ± 2.16 | 4.66 | 4.99 | 4.73 |
| F18 |  | 15.77 ± 2.50 | 4.80 | 4.85 | 4.80 |
| F19 |  | 70.05 ± 4.92 | 4.15 | 4.32 | 4.11 |
| F20 |  | 18.29 ± 1.64 | 4.74 | 4.51 | 4.73 |
| F21 |  | 55.90 ± 3.51 | 4.25 | 4.05 | 4.23 |
| F22 |  | 65.01 ± 2.94 | 4.187 | 4.34 | 4.20 |
| F23 |  | 77.24 ± 1.67 | 4.11 | 4.06 | 4.09 |

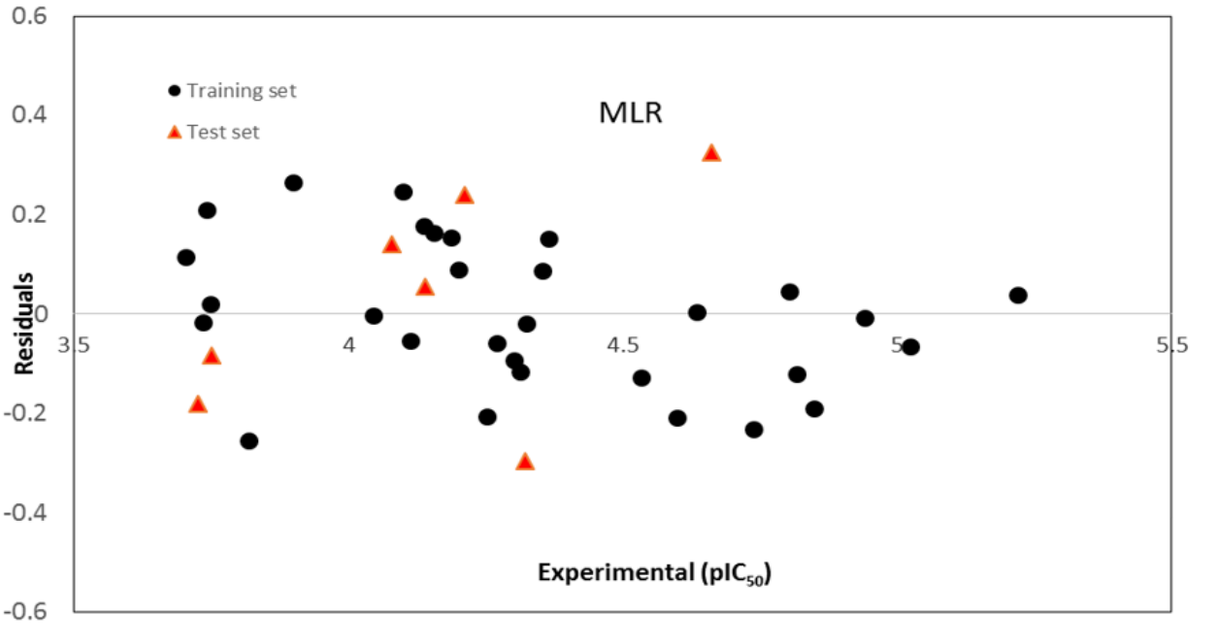
| | | | | | |
|---|---|---|---|---|---|
| F24 |  | 29.25 ± 3.43 | 4.53 | 4.41 | 4.538 |
| F25 |  | 23.24 ± 2.91 | 4.63 | 4.64 | 4.633 |
| F26 |  | 14.20 ± 1.46 | 4.85 | 4.66 | 4.84 |
| F27* |  | 47.74 ± 3.41 | 4.32 | 4.03 | 4.23 |
| F28 |  | 152.37 ± 4.83 | 3.82 | 3.56 | 3.82 |
| F29* |  | 61.31 ± 2.72 | 4.21 | 4.45 | 4.07 |
| F30 |  | 43.25 ± 1.95 | 4.36 | 4.25 | 4.48 |
| F31* |  | 72.43 ± 2.49 | 4.14 | 4.20 | 4.036 |
| F32 |  | 48.75 ± 1.75 | 4.31 | 4.20 | 4.24 |
| F33 |  | 63.08 ± 2.89 | 4.20 | 4.29 | 4.27 |

| | | | | | |
|---|---|---|---|---|---|
| F34* |  | 177.64 ± 4.13 | 3.75 | 3.67 | 3.81 |
| F35 |  | 15.27 ± 1.52 | 4.82 | 4.70 | 4.81 |
| F36 |  | 89.95 ± 1.41 | 4.05 | 4.04 | 4.04 |
| F37 |  | 79.62 ± 2.12 | 4.10 | 4.34 | 4.09 |
| F38 |  | 25.21 ± 1.18 | 4.60 | 4.39 | 4.44 |

P: Used as test (prediction) set
v: Used as validation set




**شکل۳. نمودار مقادیر پیش‌بینی شده $pIC_{50}$ بر حسب مقادیر تجربی برای سری آموزش و تست به روش SW-MLR**




**شکل٤. نمودار تغییرات باقیمانده ها برحسب مقادیر تجربی برای مقادیر pIC50 براساس مدل SW-MLR در دو مجموعه آموزشی و تست**

### ۳-۲. مدل سازی و پیش بینی توسط شبکه های عصبی مصنوعی

در این روش توصیف کننده‌های انتخاب شده وارد شبکه عصبی مصنوعی می‌شوند. پردازش داده‌ها در محیط ویندوز۱۰و با استفاده از نرم افزار MATLAB انجام شد. یک شبکه سه لایه با تابع انتقال سیگموئیدی برای نرون ها طراحی شده است. مقادیر اولیه وزن ها بطور تصادفی از بازه [1 , 0] بوده و قبل از عمل آموزش مقادیر ورودی و خروجی در فاصله [0.9 , 0.1] نرمال شده است. بهینه‌سازی و بهنگام کردن وزنها و بایاس‌ها بوسیله الگوریتمBP [1] انجام شده است. مجموعه داده‌ها به سه گروه تقسیم شده است: مجموعه آموزش، مجموعه ارزیابی و مجموعه تست. مجموعه آموزش(۵۰٪ داده ها) جهت آموزش دادن شبکه عصبی مصنوعی، مجموعه ارزیابی (۲۰٪ داده ها) برای ارزیابی مدل در طی آموزش دادن شبکه و ایجاد مدل مناسب و مجموعه پیش‌بینی (۳۰٪ داده ها) برای تست مدل ایجاد شده به کار رفت.

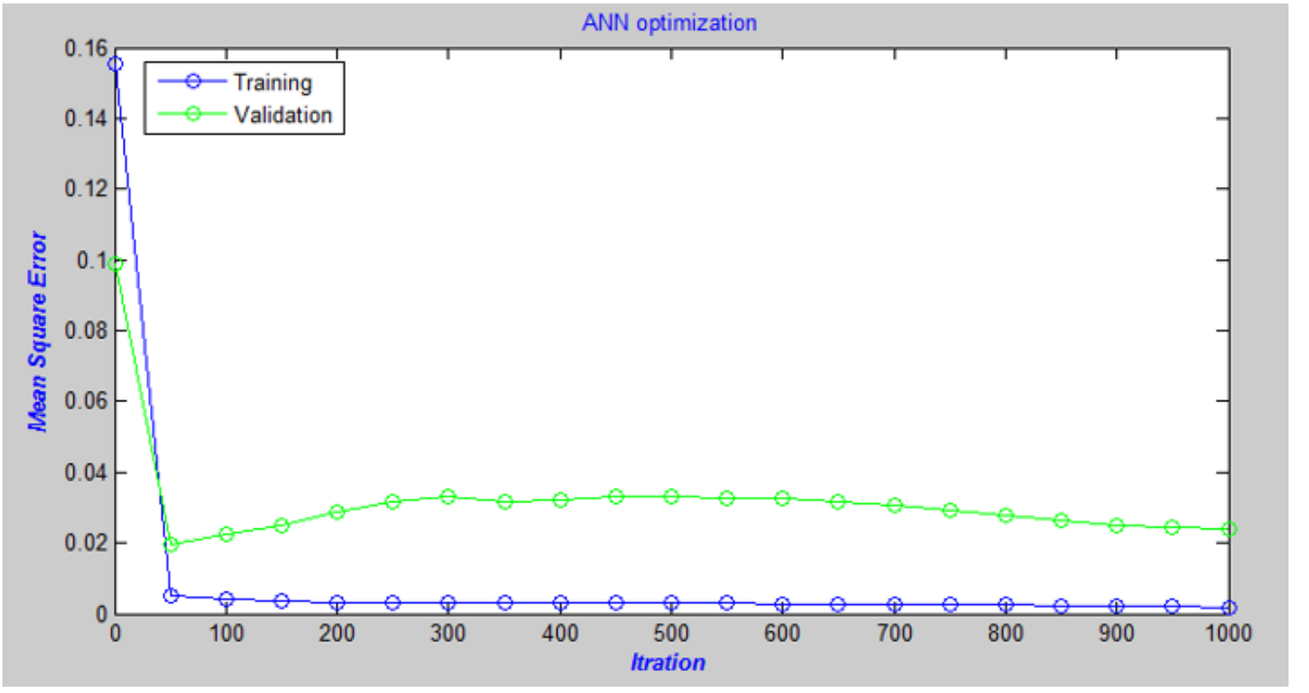
تعداد نرون‌ها در لایه ورودی با تعداد توصیف کننده‌های وارد شده به شبکه‌های عصبی مصنوعی برابر است. به ازای هر تعداد توصیف کننده وارد شده به شبکه عصبی، تعداد نرون‌ها در لایه مخفی بهینه می‌شود. بدین ترتیب که به ازای هر مدل ANN، تعداد نرون‌ها در لایه مخفی از ۱ تا ۱۰ تغییر داده شده و مقادیر RMSE برای مجموعه‌های آموزشی و پیش‌بینی محاسبه گردید. از رسم مقادیر RMSE برحسب تعداد نرون‌ها در لایه مخفی، تعداد نرون‌های لایه مخفی بهینه شد. سپس بقیه پارامترها از جمله وزنها و بایاس‌ها، سرعت یادگیری و مومنتوم نیز بهینه گردید. جدول ۴ مشخصات شبکه عصبی مصنوعی بهینه شده را نشان می‌دهد.

[1]Back-propagation

**جدول ٤. ساختار و مشخصات ANN تولید شده**

| | |
|---|---|
| تعداد گره ها (نرون ها) در لایه ورودی | ۵ |
| تعداد گره ها (نرون ها) در لایه مخفی | ۴ |
| تعداد گره ها (نرون ها) در لایه خروجی | ۱ |
| سرعت آموزش | ۰/۴ |
| مومنتوم | ۰/۳ |
| تعداد تکرارها یا چرخه های آموزشی | ۱۰۰۰ |
| تابع انتقال | تابع سیگموئیدی |

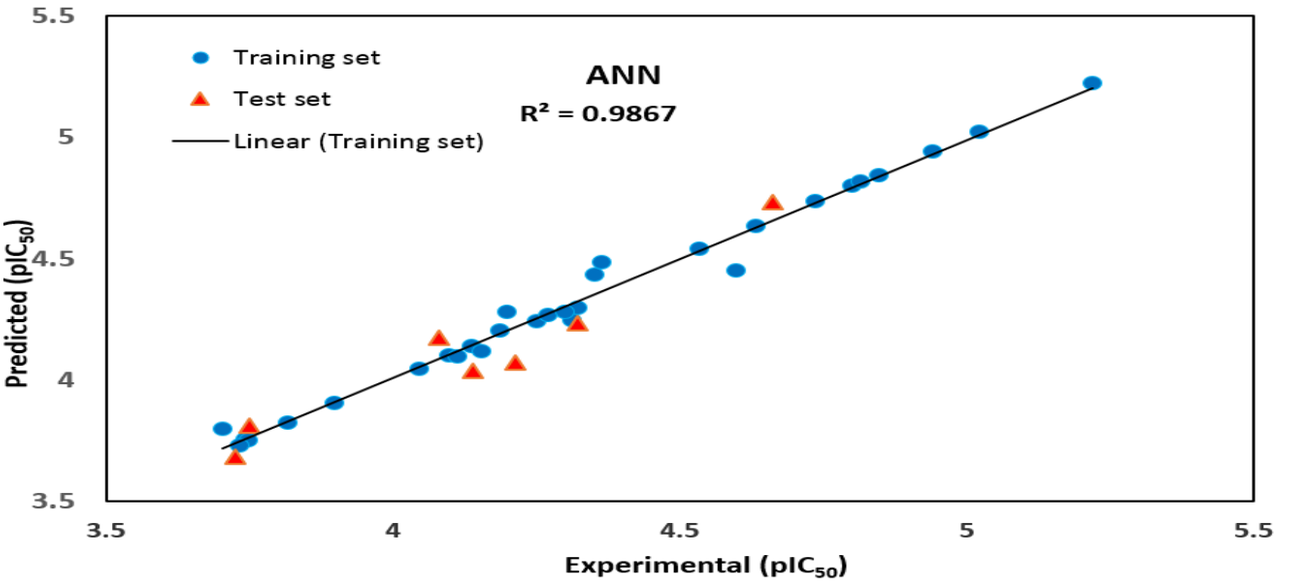
برای جلوگیری از Overfitting در طول آموزش مقادیر RMSE بعد از هر ۵۰ بار چرخه آموزشی، محاسبه و ثبت گردید. شکل ۴ نمودار میزان خطا بر حسب تعداد دورها برای این داده‌ها را نشان می‌دهد. همانطور که ملاحضه می‌شود میزان خطا برای سری آموزش همواره در حال کاهش است اما برای سری ارزیابی در ۵۰ چرخه آموزش، کمترین خطا مشاهده میشود و بعد از آن افزایش می‌یابد. بنابراین این مقدار به عنوان مقدار بهینه تعداد چرخه‌های آموزش انتخاب شد.



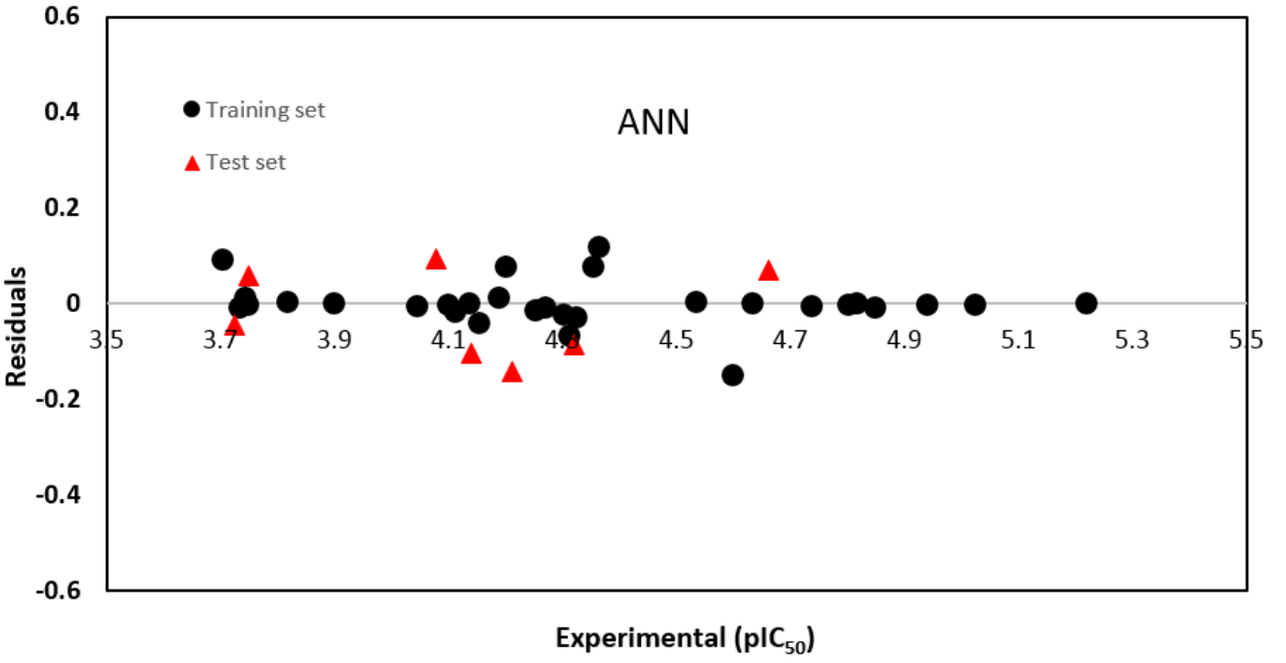

**شکل ٤. مقادیر RMSE برای مجموعه‌های آموزشی و ارزیابی برحسب تعداد چرخه‌های آموزش**

با استفاده از مدل ANN بهینه شده مقادیر فعالیت‌های بازدارندگی($pIC_{50}$) ترکیبات مورد نظر در مجموعه آموزشی، ارزیابی و پیش‌بینی (تست) مورد محاسبه قرار گرفت و درجدول(۳) نشان داده شده است. درشکل (۵) نیز مقادیر محاسبه شده $pIC_{50}$ ترکیبات مورد نظر در مجموعه‌های مختلف برحسب مقادیر تجربی رسم شده‌اند. شکل(۶) نمودار مقادیر باقیمانده ها(اختلاف مقادیر پیش

بینی شده و مقادیر تجربی) را بر حسب مقادیر تجربی نشان می دهد. در این شکل پراکندگی نقاط در حول خط صفر نشان می دهد که خطای سیستماتیک در مدل وجود ندارد.




**شکل۵. مقادیر $pIC_{50}$ محاسبه شده براساس مدل SW-ANN در سه مجموعه آموزشی، ارزیابی و تست برحسب مقادیر تجربی**




**شکل٦. نمودار تغییرات باقیمانده ها برحسب مقادیر تجربی برای مقادیر $pIC_{50}$ براساس مدل SW-ANN در دو مجموعه آموزشی و تست**

### ۳-۳. ارزیابی مدل با استفاده از پارامترهای آماری

اعتبار و اهمیت مدل‌های ساخته شده وقتی مشخص می‌گردد که فعالیت مولکول‌هایی که در سری تست هستند را به خوبی و بطور رضایت بخش و قابل قبول پیش‌بینی کند. در این‌کار چندین روش به منظور ارزیابی و مقایسه توانایی مدل‌های ارائه شده در

پیش‌بینی مقادیر $pIC_{50}$ ذکر شده است. مطابق جدول (۵) پنج پارامتر آماری، جهت ارزیابی توانایی پیش‌بینی مدل‌های ساخته شده به روش‌های MLR و ANN به کار گرفته شد. این نتایج نشان دهنده آن است که شبکه عصبی مصنوعی نسبت به روش MLR توانمندی بیشتری جهت پیش بینی فعالیت دارویی ترکیبات را دارد.

**جدول ۵. پارامترهای آماری برای مدل‌های انتخاب شده**

| | | SW-MLR | SW-ANN |
|---|---|---|---|
| $R^2$ | سری آموزش | ۰/۸۷۵ | ۰/۹۸۶ |
| | سری تست | ۰/۸۴۳ | ۰/۹۲۰ |
| RMSE | سری آموزش | ۰/۲۷۲ | ۰/۰۶۸ |
| | سری تست | ۰/۴۴۱ | ۰/۲۶۹ |
| REP | سری آموزش | ۳/۵۲۸ | ۰/۸۸۴ |
| | سری تست | ۵/۷۰۹ | ۳/۴۸۳ |
| SEP | سری آموزش | ۰/۲۷۲ | ۰/۰۶۸ |
| | سری تست | ۰/۴۴۱ | ۰/۲۶۹ |
| AARD | سری آموزش | ۲/۹۹۲ | ۰/۸۹۰ |
| | سری تست | ۵/۳۹۲ | ۲/۶۷۷ |

### ۳-۳-۱. ارزیابی مدل‌ها توسط روش رد مرحله‌ای تک تک و گروهی

به منظور بررسی بیشتر قدرت پیش‌بینی مدل‌های خطی و غیر خطی، تکنیک رد مرحله‌ای تک تک و گروهی مورد استفاده قرار گرفت. در روش رد مرحله‌ای تک تک، هر بار یکی از ترکیبات به طور تصادفی از سری داده‌ها حذف شدند و در روش رد مرحله-ای 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، پیش‌بینی شدند. این فرایند برای تمام اعضای سری داده‌ها تکرار شد. نتایج حاصل از رد مرحله‌ای و گروهی در جدول(۶) ارائه شده است.

**جدول٦. پارامترهای آماری برای مدل‌های انتخاب شده**

| | | SW-MLR | SW-ANN |
|---|---|---|---|
| $Q^2_{LOO}$ | کل داده ها | ۰/۸۲۴ | ۰/۸۹۲ |
| $Q^2_{LGO}$ | کل داده ها | ۰/۷۸۳ | ۰/۸۴۴ |

### ۳-۳-۲. ارزیابی مدل‌های ارائه شده با استفاده از آزمون Y-تصادفی

این تکنیک ارزیابی مدل، با هدف بررسی هر گونه ارتباط تصادفی بین داده‌ها انجام شد. در این آزمون، متغیر وابسته بطور تصادفی بهم ریخته شد. مدل QSAR جدید با استفاده 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(۷) نشان داده شده است. مقادیر کوچک ضریب تعیین ($R^2$) بیانگر عدم ارتباط شانسی در مدل توسعه یافته توسط رگرسیون خطی چندگانه می‌باشد.

**جدول۷. نتایج حاصل از چندین بار اجرای آزمون Y-تصادفی**

| تکرار | | ۱ | ۲ | ۳ | ٤ | ٥ | ٦ | ۷ | ۸ | ۹ | ۱۰ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $R^2_{test}$ | MLR | ۰/۰۳٦ | ۰/۰۱۰ | ۰/۰۸٦ | ۰/۰٤۹ | ۰/۱۰٥ | ۰/۰٦۸ | ۰/۲۱۳ | ۰/۱۲۰ | ۰/۰٦٦ | ۰/۱۷٥ |
| | ANN | ۰/۱۲٥ | ۰/۱۹۰ | ۰/۰٥٦ | ۰/۰۳۲ | ۰/۱۸۱ | ۰/۰۸٥ | ۰/۰۱٦ | ۰/۰۱۳ | ۰/۱۰٥ | ۰/۲٥۰ |

## ٤. نتیجه گیری

در این تحقیق از دو روش مختلف برای پیش‌بینی فعالیت ترکیبات دارویی($pIC_{50}$) مشتقات N-آریل۲و۲-دی کلرواستامید و آریل۲و۲-دی کلرواستات به عنوان بازدارنده برای درمان سرطان استفاده شد. از آنجایی که اندازه‌گیری فعالیت دارویی بیشتر ترکیبات به صورت تجربی با صرف هزینه، زمان و پیچیدگی زیاد همراه است، دست یابی به مقادیر $pIC_{50}$ با روش‌های تجربی مقرون به صرفه نیست. بنابراین، پیش‌بینی آن با استفاده از روش های محاسباتی از اهمیت بالایی برخوردار است. برای انتخاب توصیف‌کننده‌های مناسب از روش رگرسیون مرحله ایی استفاده شد. سپس این توصیف‌کننده‌ها برای مدل سازی خطی و غیرخطی مورد استفاده قرار گرفتند. نتایج نشان داد که از بین دو روش استفاده شده، روش ANN روش مناسب‌تری برای پیش‌بینی فعالیت این ترکیبات دارویی است. همچنین بررسی ارتباط توصیف‌کننده‌های وارد شده در مدل با اثر بازدارندگی یا $pIC_{50}$ مورد بررسی قرار گرفت. با توجه به نتایج به دست آمده در مدل، توصیف‌کننده‌های انتخاب شده توسط روش رگرسیون مرحله‌ای شامل HATS5e, HATS3v, HGM, RDF020e MWC09 می‌باشند. این توصیف کننده ها تاثیر الکترونگاتیوی، ساختار هندسی و حجم مولکولی را نشان می دهند. نتایج مدل سازی نشان می دهد که مقدار $pIC_{50}$ با الکترونگاتیوی رابطه عکس دارد یعنی با کاهش الکترونگاتیوی ترکیبات می‌توان $pIC_{50}$ را افزایش داد. از طرف دیگر مدل ساخته شده نشان می دهد که حجم مولکولی ترکیبات با $pIC_{50}$ رابطه مستقیم دارد یعنی با افزایش حجم مولکول، $pIC_{50}$ نیز افزایش می یابد. بنابراین گروه های حجیم تر باعث افزایش بیشتری در مقدار $pIC_{50}$ می گردند. از این تحقیق استنباط می شود که با تغییر الکترونگاتیوی گروه های عاملی و تغییر حجم مولکولی ترکیبات، می توان ترکیباتی طراحی و سنتز کرد که فعالیت دارویی موثرتری داشته باشند.

## ۵. مراجع

[1] Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design-a review. *Current topics in medicinal chemistry*, *10*(1), 95-115.

[2] Kubinyi, H. (1997). QSAR and 3D QSAR in drug design Part 1: methodology. *Drug discovery today*, *2*(11), 457-467.

[3] Tandon, H., Chakraborty, T., & Suhag, V. (2019). A concise review on the significance of QSAR in drug design. *Chemical and Biomolecular Engineering*, *4*(4), 45-51.

[4] Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug discovery today*, *21*(8), 1291-1302.

[5] Achary, P. G. (2020). Applications of quantitative structure-activity relationships (QSAR) based virtual screening in drug design: A review. *Mini Reviews in Medicinal Chemistry*, *20*(14), 1375-1388.

[6] Tandon, H., Chakraborty, T., & Suhag, V. (2019). A concise review on the significance of QSAR in drug design. *Chemical and Biomolecular Engineering*, *4*(4), 45-51.

[7] Zivkovic, M., Zlatanovic, M., Zlatanovic, N., Golubović, M., & Veselinović, A. M. (2020). The application of the combination of Monte Carlo optimization method based QSAR modeling and molecular docking in drug design and development. *Mini Reviews in Medicinal Chemistry*, *20*(14), 1389-1402.

[8] Haghshenas, H., Kaviani, B., Firouzeh, M., & Tavakol, H. (2021). Developing a variation of 3D-QSAR/MD method in drug design. *Journal of Computational Chemistry*, *42*(13), 917-929.

[9] Rosell-Hidalgo, A., Young, L., Moore, A. L., & Ghafourian, T. (2021). QSAR and molecular docking for the search of AOX inhibitors: a rational drug discovery approach. *Journal of Computer-Aided Molecular Design*, *35*, 245-260.

[10] Brown, N., Ertl, P., Lewis, R., Luksch, T., Reker, D., & Schneider, N. (2020). Artificial intelligence in chemistry and drug design. *Journal of Computer-Aided Molecular Design*, *34*, 709-715.

[11] Ewees, A. A., Abualigah, L., Yousri, D., Algamal, Z. Y., Al-Qaness, M. A., Ibrahim, R. A., & Abd Elaziz, M. (2022). Improved Slime Mould Algorithm based on Firefly Algorithm for feature selection: A case study on QSAR model. *Engineering with Computers*, *38*(3), 2407-2421.

[12] Mozafari, Z., Arab Chamjangali, M., Arashi, M., & Goudarzi, N. (2021). Performance of smoothly clipped absolute deviation as a variable selection method in the artificial neural network-based QSAR studies. *Journal of Chemometrics*, *35*(5), e3338.

[13] Vahedi, Nafiseh, Majid Mohammadhosseini, and Mehdi Nekoei. "QSAR Study of PARP Inhibitors by GA-MLR, GA-SVM and GA-ANN Approaches." *Current Analytical Chemistry* 16, no. 8 (2020): 1088-1105.

[14] Triolascarya, K., Septiawan, R. R., & Kurniawan, I. (2022). QSAR Study of Larvicidal Phytocompounds as Anti-Aedes Aegypti by using GA-SVM Method. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(4), 632-638.

[15] Ai, H., Wu, X., Zhang, L., Qi, M., Zhao, Y., Zhao, Q., ... & Liu, H. (2019). QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. *Ecotoxicology and Environmental Safety*, *179*, 71-78.

[16] Fereidoonnezhad, M., Tabaei, S. M. H., Sakhteman, A., Seradj, H., Faghih, Z., Faghih, Z., ... & Rezaei, Z. (2020). Design, synthesis, molecular docking, biological evaluations and QSAR studies of novel dichloroacetate analogues as anticancer agent. *Journal of Molecular Structure*, *1221*, 128689.

[17] Quadri, T. W., Olasunkanmi, L. O., Akpan, E. D., Fayemi, O. E., Lee, H. S., Lgaz, H., ... & Ebenso, E. E. (2022). Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors. *Materials Today Communications*, *30*, 103163.

[18] Abdolmaleki, A., & Ghasemi, J. B. (2019). Inhibition activity prediction for a dataset of candidates' drug by combining fuzzy logic with MLR/ANN QSAR models. *Chemical Biology & Drug Design*, *93*(6), 1139-1157.
[19] Abdous, B., Sajjadi, S. M., & Bagheri, A. (2022). Predicting the aggregation number of cationic surfactants based on ANN-QSAR modeling approaches: understanding the impact of molecular descriptors on aggregation numbers. *RSC advances*, *12*(52), 33666-33678.
[20] Fereidoonnezhad, M., Tabaei, S. M. H., Sakhteman, A., Seradj, H., Faghih, Z., Faghih, Z., ... & Rezaei, Z. (2020). Design, synthesis, molecular docking, biological evaluations and QSAR studies of novel dichloroacetate analogues as anticancer agent. *Journal of Molecular Structure*, *1221*, 128689.
[21] Consonni, V., & Todeschini, R. (2010). Molecular descriptors. *Recent advances in QSAR studies: methods and applications*, 29-102.
[22] Hisaki, T., née Kaneko, M. A., Hirota, M., Matsuoka, M., & Kouzuki, H. (2020). Integration of read-across and artificial neural network-based QSAR models for predicting systemic toxicity: A case study for valproic acid. *The Journal of Toxicological Sciences*, *45*(2), 95-108.
[23] Chen, Y., Song, L., Liu, Y., Yang, L., & Li, D. (2020). A review of the artificial neural network models for water quality prediction. *Applied Sciences*, *10*(17), 5776.
[24] Escrivá-Escrivá, G., Álvarez-Bel, C., Roldán-Blay, C., & Alcázar-Ortega, M. (2011). New artificial neural network prediction method for electrical consumption forecasting based on building end-uses. *Energy and Buildings*, *43*(11), 3112-3119.
[25] Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software*, *119*, 285-304.

# Prediction of anticancer activity of some N-aryl 2,2-dichloroacetamide and aryl 2,2-dichloroacetate derivatives using linear and non-linear quantitative structure-activity relationship (QSAR) methods

**Mehdi Nekoei[1]*, Majid Mohammadhosseini[1], Tayyebeh Javaheri[1]**

*[1]Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran*

Submited: 24 June 2023, Revised: 16 September 2023, Accepted: 04 October 2023

**Abstract**

Quantitative structure-activity relationship (QSAR) study to predict the pharmacological activity (IC50) of some N-aryl 2,2-dichloroacetamide and aryl 2,2-dichloroacetate derivatives using multiple linear regression (MLR) and artificial neural networks (ANN) methods as Linear and non-linear methods were used to treat cancer. At first, the structure of desired compounds was drawn and optimized by Hypercam software. The drawn compounds were entered into Dragon software to calculate molecular descriptors and the number of 1481 descriptors was calculated for each molecule. Stepwise regression method was used to select the most suitable descriptors. After selecting the most appropriate descriptors, two methods, MLR and ANN, were used as linear and non-linear methods to study the quantitative-structure-activity relationship for modeling and predicting the medicinal activity of compounds. The performance of each model was evaluated by several statistical parameters. The obtained results show the superiority of ANN method over MLR.

***Keywords:*** *quantitative structure-activity relationship, N-aryl 2,2-dichloroacetamide-aryl, 2,2-dichloroacetate, multiple linear regression, artificial neural networks***.**

***Corresponding author :** Mehdi Nekoei

**Address:** Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran

**Tel:** 02332394289 **E-mail:** m_nekoei1356@yahoo.com