دانشگاه آزاد اسلامی، واحد اراک
شاپا ٤٦٦٨-٢٠٠٨
http://jer.iau-arak.ac.ir

فصلنامه تخصصی تحقیقات حشره‌شناسی
(علمی- پژوهشی)
جلد ۱۳، شماره۱، سال ۱٤۰۰، (۸۲-٦۷)

پیش بینی LD_{50} در مشتقات کربوکسیلیک اسید با مدل های رگرسیون خطی چندگانه و شبکه عصبی مصنوعی

فهیمه محمایی[۱]، عصمت محمدی نسب [۲]*

۱- دانشجوی دکتری شیمی فیزیک، گروه شیمی، واحد اراک، دانشگاه آزاداسلامی، اراک، ایران
۲- استادیار شیمی فیزیک، گروه شیمی،واحد اراک، دانشگاه آزاداسلامی، اراک، ایران

چکیده

در این تحقیق، از طریق مطالعه رابطه ساختار-فعالیت به پیش بینی مقادیر سمیت مشتقات کربوکسیلیک اسید پرداخته شده است. ابتدا مقادیر LD_{50} برای مجموعه ای از ترکیبات مورد مطالعه با استفاده از منابع علمی معتبر استخراج گردید و ساختار آنها به کمک نرم‌افزار گوس ویو05 رسم شده و با نرم‌افزار گوسین09 به روش هارتری فاک و سری پایه ۲۱G-۳ بهینه شدند. سپس با استفاده از نرم افزار دراگون توصیف‌گرهای مولکولی استخراج گردیدند. به کمک ژنتیک الگوریتم و روش برگشتی توصیف‌گرهای نامناسب حذف شده و بهترین آنها برای مدل‌های رگرسیون خطی چندگانه و شبکه عصبی مصنوعی مورد استفاده قرار گرفت. دقت پیش بینی مدل نهایی توسط ضرایب آماری مورد بحث قرار گرفت. اعتبارسنجی تقاطعی و نیز اعتبارسنجی خارجی مدل های پیش بینی همبستگی بسیار بالا را بین مقادیر تجربی و مقادیر پیش بینی گروه های آموزش آزمون و اعتبارسنجی در روش شبکه عصبی مصنوعی نشان داد. مشخص گردید که روش شبکه عصبی مصنوعی با خطای کمتر و ضریب تعیین بالاتر نسبت به روش رگرسیون خطی چندگانه از برتری قابل توجه ای برخوردار می باشد. مدل پیشنهادی می تواند برای پیش بینی $\log(LD_{50})$ ترکیبات جدید کربوکسیلیک اسید مفید واقع گردد.

واژه‌های کلیدی: "سمیت" "روش رگرسیون خطی چندگانه" ، "شبکه عصبی مصنوعی"؛ "مشتقات کربوکسیلیک اسید ".

[*] نویسنده رابط، پست الکترونیکی: *e-mohammadinasab@iau-arak.ac.ir*

تاریخ دریافت مقاله:۹۹/۱۲/۲۰ - تاریخ پذیرش مقاله:۱٤۰۰/۲/۱۲

مقدمه

اسیدهای کربوکسیلیک از گروه های مهم اسیدهای آلی هستند که ترکیبات بسیار ارزشمندی در بدن جانوران محسوب می گردند. اسید فرمیک (جوهر مورچه)، ساده‌ترین عضو گروه اسیدهای کربوکسیلیک است که در نیش مورچه و زنبور یافت می شود و همچنین ترکیب عمده ماده گزش‌زا در برگ گزنه می باشد (Matysiak, et al., 2018). اسید فرمیک بیشتر به عنوان ماده نگهدارنده برای جلوگیری از فاسد شدن و آنتی باکتریال در غذای دام استفاده می‌شود. پاشیدن مقداری از آن روی علف تازه خشک شده، از فساد و پوسیدگی آن جلوگیری کرده و مواد مغذی آن را تا حد بالایی حفظ می‌کند. برای جلوگیری از فساد غذای زمستانی دام‌ها در مجتمع‌های بزرگ دامداری از این ماده استفاده می‌شود و در مرغداری ها به منظور از بین بردن باکتری سالمونلا به غذای مرغ اضافه می گردد. اسید فرمیک به هیدرولیز کلاژن مرغ‌های ضعیف کمک می کند و ممکن است برای تولید پپتیدهای کوچک کلاژن از سایر کلاژن های مهره داران مسن مورد استفاده قرار گیرد (Hong, et al., 2018). همچنین افزودن اسید استیک یا جوهر سرکه به محیط کشت سبب کاهش سرعت رشد قارچ های گیاهان می‌شود (Ghasemian, et al., 2018). استنشاق اسید فرمیک، استیک، پروپیونیک و بوتیریک توسط موشها باعث کاهش سریع سرعت تنفس می شود (Nielsen, 2018). آسیل هالیدها که از مشتقات کربوکسیلیک اسید هستند از واسطه‌های سنتزی مهم به‌شمار می‌روند، زیرا از طریق آن‌ها می‌توان ترکیب‌های مختلفی بدست آورد (Xu, et al., 2005). یک آسیل هالید از جایگزینی هالوژن با OH– یک کربوکسیلیک اسید به دست می‌آید. رشد در کشاورزی و نیاز به کیفیت بهتر کالاهای کشاورزی، تقاضای کود را به دنبال دارد. این، به نوبه خود برای تقویت تولید کلریدهای اسید کربوکسیلیک پیش بینی شده است.

تاکنون تلاش زیادی جهت تعیین خواص فیزیکی–شیمیایی بسیاری از ترکیبات شیمیایی با استفاده از روش‌های محاسباتی، صورت گرفته است. روش های محاسباتی از قابلیت های فراوانی در بررسی و تولید ترکیبات در بدن بدون اثرات جانبی و سمّیت، مطالعه ترکیباتی که از لحاظ فارماکوسینتیکی در سیستم‌های بیولوژیک پایدار هستند، طراحی منطقی مواد پرمصرف در حوزه صنایع مانند رنگ‌هاو مواد شیمیایی، شناسایی مواد سمّی زیان‌آور برای محیط‌زیست، صرفه‌جویی در زمان، کاهش هزینه‌های مربوط به انجام آزمایش‌ها و پیشگویی خواص شیمیایی- فیزیکی مواد شیمیایی جدید با بازدهی بالا برخوردارند (Eriksson, 2003یکی از این تکنیک ها، بررسی ارتباط کمی ساختار- خاصیت [4] (QSPR) و ساختار- فعالیت[5] (QSAR) می‌باشد.

(Polishchuk, 2017; Bagheban Shahri & Niazi, 2016; Mohammaei, 2018; Roy & Mitra, 2011).

در این تکنیک ها، ابتدا سعی می شود که به کمک نرم‌افزارها و برنامه‌های مختلف و با استفاده از انواع توصیف‌گرهای مولکولی، ساختار ترکیبات با فرمول‌های ریاضی، به اعداد تبدیل گردند (Todeschini & Consonni, 2000; Todeschini, 2009). سپس با مدل‌سازی، اقدام به پیشگویی خواص و یا فعالیت ترکیبات می‌شود. تاکنون تحقیقات زیادی در استفاده از روش‌های محاسباتی برای تعیین خواص ترکیبات مختلف انجام‌شده است (Ha, et al., 2019; Asadollahi-Baboli & Dehnavi, 2018; Qin, et al., 2017; Khan, et al., 2019).

نتایج یک مطالعه رابطه کمی ساختار- خاصیت در مورد سمیت ۳۵ اسیدهای کربوکسیلیک آلیفاتیک در محلول آبی نشانگر وابستگی $\log(IGC^{-1}_{50})$ با خواصی مانند ضریب توزیع و ضریب انکسار و قطبش پذیری و دیگر خواص فیزیکی می باشد (Maguna, et al., 2003). اسیدهای کربوکسیلیک آلیفاتیک با وزن مولکولی کم که از تجزیه گیاهان، ترشحات

[4] Quantitative Structure Property Relationship
[5] Quantitative Structure Activity Relationship

ریشه و تجزیه مواد آلی نشات می گیرند، برای چندین فرآیند خاک مانند هوازدگی معدنی، شستشو و سمیت آلومینیوم و آهن، تحرک فلزات سنگین و انحلال مواد مغذی گیاه، مهم تلقی می شوند. (Fox & Comerford, 1990; Fox, 1995; Strobel, et al., 1999) و ضمنا می توانند پروتون را آزاد کنند و به عنوان لیگاند عمل نمایند و از طریق واکنشهای پیچیده ای بر حلالیت و تخلخل فلز تأثیر بگذارند (Tani, et al., 1996; Slattery & Morrison, 1995). سمیت مربوط به ۳۸ اسید کربوکسیلیک آلیفاتیک با استفاده از الگوریتم ژنتیک[6](GA) همراه با روش حداقل مربعات جزئی[7] (PLS) مطالعه شده و ارتباط بین $\log(IGC^{-1}_{50})$ با برخی خصوصیات مولکولی مانند ثابت تفکیک اسیدی مورد مطالعه قرار گرفته است (Kompany-Zareh., 2009). مطالعات QSAR بر روی مشتقات اسید کربوکسیلیک به عنوان مهارکننده های HIV-1 Integrase با استفاده از توصیفگرهای مولکولی سه بعدی با مدل های ماشین بردار پشتیبانی[8] (SVM)، شبکه های عصبی تکثیر برگشتی[9] (BPNN) و رگرسیون خطی چندگانه(MLR)انجام شده است. بهترین مدل QSAR نشان داده است که قطبش پذیری و جرم، موثرترین خواص اتمی تاثیر گذار روی ساختار این ترکیبات می باشند (Cheng, et al., 2010). تعیین فعالیت بیولوژیکی ۸٦ ترکیب از مشتقات کربوکسیلیک اسید و شناسایی بهترین متغیرهای موثر بر فعالیتهای زیستی آنها با استفاده از مطالعات ٤D-QSAR (4)به روش الگوریتم ژنتیکی کنفورماسیون الکترونی و رگرسیون حداقل مربعات غیرخطی انجام شده است (Tuzun, et al., 2018) . تجزیه و تحلیل رابطه کمی ساختار-فعالیت بر روی مشتقات اسید کربوکسیلیک ٤-کینولین انجام شده و یک مدل QSAR توصیفی از طریق محاسبه توصیفگرهای مستقل از هم با استفاده از نرم افزار MOE 2009.10 بدست آمده است. برای یک مجموعه آموزشی از ۲۰ ترکیب، تجزیه و تحلیل به روش حداقل مربعات جزئی، منجر به مدلی با مجذور ضریب همبستگی $R^2= 0.913$ گردیده است (Hajalsiddig & Saeed., 2019).

بررسی ارتباط بین متغیرهای وابسته و مستقل در روش‌های محاسباتی و ارائه مدل‌هایی که بتواند قبل از هرگونه سنتزی، خواص ترکیبات را پیشگویی کند، برای پژوهشگران علوم تجربی از ارزش بسیار بالایی برخوردار می‌باشد.

در این مطالعه متغیرهای مستقل مورد نظر شامل توصیف‌گرهای مولکولی و متغیر وابسته، شاخص LD_{50}[10] می‌باشد. LD_{50} یک معیار آماری است که شامل دوز لازم از یک ترکیب شیمیایی می باشد که ۵۰ درصد حیوانات تحت آزمایش را پس از ١٤ روز می‌کشد و از آن می‌توان برای مقایسه سمیت ترکیبات شیمیایی با یکدیگر استفاده نمود، به‌طوری‌که هرچقدر مقدار LD_{50} کمتر باشد، سمیت ماده بیشتر است (Williams, 2003). ارتباط لگاریتم شاخص سمیت LD_{50} با توصیف‌گرهای مولکولی در این تحقیق برای ۳۷ ترکیب از مشتقات کربوکسیلیک اسید، با کمک روش‌های محاسباتی خطی و غیرخطی، بررسی گردیده و بهترین مدل برای پیش‌بینی کمیت موردنظر ارائه گردیده است.

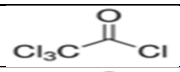
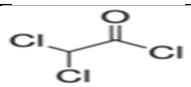
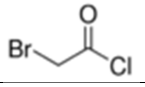
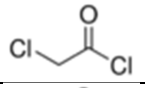
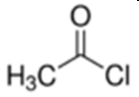
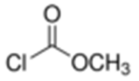
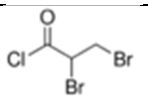
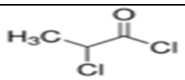
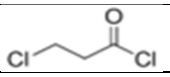
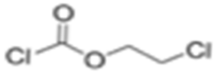
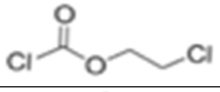
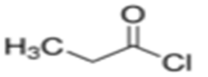
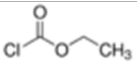
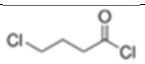
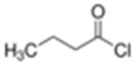
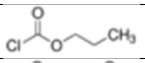
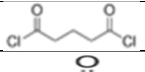
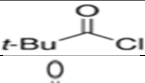
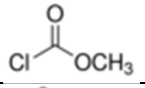
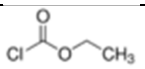
مواد

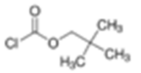
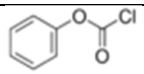
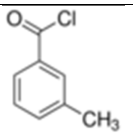
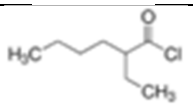
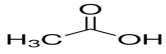
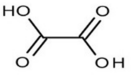
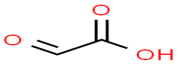
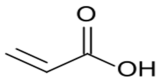
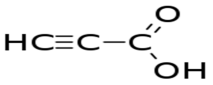
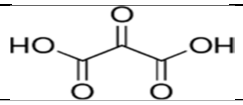
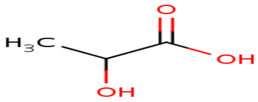
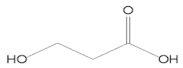
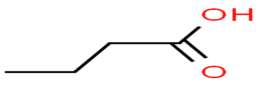
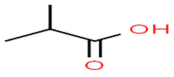
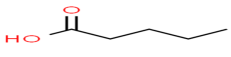
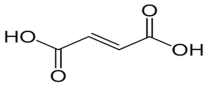
برای انجام این تحقیق، تعداد ۳۷ ترکیب از مشتقات کربوکسیلیک اسید با استفاده از وب‌سایت سیگما آلدریچ (Sigma Aldrich) شناسایی و مقادیر تجربی شاخص سمیت $LD_{50}(mg.kg^{-1})$ برای همه ترکیبات، با استفاده از منبع معتبر علمی (chemidplus) در جدول ۱ مندرج گردیدند.

[6] Genetic Algorithm
[7] *Partial least squares* regression
[8] *Support Vector Machine*
[9] *Back propagation neural network*
[10] Lethal Dose, which causes the death of 50% (one half) of a group of test animals

جدول ۱- نام، ساختار و مقادیر LD_{50} مشتقات کربوکسیلیک اسید مورد مطالعه

Table 1- The name, chemical structure and LD_{50} values of studied carboxylic acids derivatives.

No	Compound	LD_{50} ($mg.kg^{-1}$)	Structure
1	Trichloroacetyl chloride	600	
2	Dichloroacetyl chloride	2,460	
3	Bromoacetyl chloride	3,200	
4	Chloroacetyl chloride	208	
5	Acetyl chloride reagent	910	
6	Methyl chloroformate	60	
7	2,3-Dibromopropionyl chloride	1,200	
8	2-Chloropropionyl chloride	642	
9	3-Chloropropionyl chloride	1,200	
10	1-Chloroethyl chloroformate	470	
11	2-Chloroethyl chloroformate	859	
12	Propionyl chloride	823	
13	Ethyl chloroformate	270	
14	4-Chlorobutyryl chloride	1,350	
15	Butyryl chloride	785	
16	Propyl chloroformate	1,045	
17	Glutaryl chloride	190	
18	Trimethylacetyl chloride	638	
19	Chloroacetic acid	76	
20	2-Chloropropanoic acid	800	

21	6-Chlorohexanoic acid	3,080	
22	Phenyl chloroformate	1,410	
23	m-Toluoyl chloride	3,440	
24	2-Ethylhexanoyl chloride	1,500	
25	methanoic acid	1,100	
26	ethanoic acid	3,310	
27	Oxalic acid	7,500	
28	Glyoxylic acid	3,000	
29	Acrylic acid	2,500	
30	Malonic acid	100	
31	propanedioic acid	1,310	
32	2-hydroxypropanoic acid	3,543	
33	1-hydroxypropanoic acid	2,937	
34	butanoic acid	2,940	
35	2-methylpropanoic	280	
36	pentanoic acid	9,300	
37	Ethyl propionate	7,500	

لازم به ذکر است که همه مقادیر مندرج شده در این مطالعه مربوط به شاخص سمیت تجویز دهانی LD_{50} می‌باشند که به دلیل پراکندگی مقادیر آنها، برای انجام محاسبات، مقادیر لگاریتمی آن‌ها به کار گرفته شده‌اند.

روش کار

پس از جمع‌آوری ترکیبات و داده‌ها، با استفاده از نرم‌افزار گوس ویو[11] ساختار اولیه ۳۷ ترکیب مورد مطالعه ترسیم گردیده و بهینه سازی آنها بوسیله نرم‌افزار گوسین[12] ۰۹ به روش[13] HF با سری پایه 3-21G انجام گردید. نرم‌افزار دراگون[14] به‌منظور محاسبه انواع توصیف‌گرهای مولکولی به کار گرفته شد و به‌منظور غربالگری شاخص‌های به‌دست‌آمده، از برنامه متلب[15] ۲۰۱۰a و الگوریتم ژنتیک (Mirjalili, 2019 ; Mirjalili, et al., 2020) با روش برگشتی استفاده شد (Elisseeff & Guyo, 2003). الگوریتم ژنتیک، یک روش جستجوی هوشمند و تصادفی است که با به‌کارگیری عملگرهای ژنتیک از یک فرایند تکامل تدریجی تبعیت می‌کند. پس از حذف توصیف‌کننده‌های غیرضروری، آن‌هایی که بیشترین میزان همبستگی را با لگاریتم $\log LD_{50}$ داشتند برای مدل سازی با روش‌های MLR و ANN انتخاب شدند (Kutner et al., 2004). در روش‌های رگرسیون خطی چندگانه[16] *(MLR)* و غیرخطی[17] (NLR) که در این مدل‌سازی‌ها استفاده می‌شوند، خواص و یا فعالیت به‌عنوان متغیر وابسته و توصیف‌گرهای مولکولی به‌عنوان متغیرهای مستقل در نظر گرفته می‌شوند (Goodarzi, et al., 2012; Randic & Basak, 2000). در مواقعی که الگوی ساده خطی، ارائه‌دهنده مدل مناسبی برای رسیدن به هدف نیست، روش‌های غیرخطی بکار گرفته می شوند (Fissa, et al., 2019; Roy & Ambure., 2016; Ghamali, et al., 2017) . روش شبکه عصبی مصنوعی[18] ANN که به عنوان یکی از روش‌هایی که به بررسی ارتباط غیرخطی بین متغیرهای مستقل با متغیر وابسته می پردازد، کاربرد گسترده‌ای در حیطه مطالعات ارتباط ساختار–فعالیت دارد (Cross, et al., 1995). روش شبکه عصبی مصنوعی ANN بر اساس نورون‌های عصبی پایه‌گذاری شده و توانایی این را دارد که روش‌های آماری کلاسیک و مدرن را با هم ترکیب کند (Haykin, 1990). در یک شبکه عصبی مصنوعی، برای ایجاد ارتباط بین متغیرهای مستقل و وابسته از لایه‌هایی استفاده می‌شود که این لایه‌ها با یکدیگر در ارتباط هستند و شامل لایه ورودی، لایه خروجی و لایه‌های پنهان یا میانی است. لایه‌های پنهان به این دلیل استفاده می‌شوند که بر ارائه مدل نهایی تأثیرگذار هستند و دلیل بکار بردن یک لایه خروجی به خاطر ارائه یک نتیجه منطقی برای پیش‌بینی خاصیت موردنظر می‌باشد(Gadzuric et al., 2015; Miller & Miller, 2010; Maiellaro et al., 2004).

نتایج و بحث

پس از رسم و بهینه‌سازی ترکیبات مورد مطالعه با نرم‌افزار گوسین، فایل‌های خروجی حاصل از محاسبات کوانتمی برای همه ترکیبات مورد مطالعه، به برنامه دراگون انتقال داده شد تا به محاسبه توصیف گرهای مولکولی پرداخته شود. در نتیجه این محاسبات، ۱۸۵۱ توصیف‌گر مولکولی در ۱۸ گروه مختلف بدست آمد. لذا برای تعیین مناسبترین توصیف گرها از روش الگوریتم ژنتیک برنامه متلب 2010a و روش برگشتی استفاده گردید. از این تعداد توصیفگر اولیه با حذف

[11] Gauss View 05
[12] Gaussian 09
[13] Hartree fock
[14] Dragon
[15] MATLAB 2010a
[16] *Multiple linear regression*
[17] Non*linear regression*
[18] *Artificial Neural Network*

شاخص‌های غیرضروری و توصیف کننده های نامناسب،آنهایی انتخاب شدند که بیشترین میزان همبستگی را با Log LD_{50}برقرار نموده ولی کمترین میزان همبستگی را با توصیفگرهای دیگر داشتند. بنابر این مشخص گردید که تعداد توصیف‌گرهای نهایی برای مدل‌سازی ۱۹ شاخص می باشد. این توصیف‌گرها که در جدول ۲ ارائه‌شده‌اند، برای مشخص شدن الگوی بهتر در روش رگرسیون خطی مورداستفاده قرار گرفتند.

تجزیه و تحلیل در روش خطی

جهت انجام محاسبات در مدل خطی، ترکیبات به دو دسته آموزش و آزمون (به نسبت ۸۰٪ و ۲۰٪) تقسیم شدند. پارامترهای اعتبارسنجی مربوط به آنالیز ترکیبات آموزش که شامل ضریب تعیین[19] R^2 ، ضریب همبستگی[20] Rو ضریب تعیین تعدیل یافته[21] R^2_{adj} و آماره فیشر[22] F و سطح معناداری[23] Sig و خطای مربعات میانگین[24] (MSE) و جذر خطای میانگین مربعات[25] (RMSE) می باشند در جدول (۲) آورده شده اند.

جدول ۲ توصیف گرها وضرایب آماری به دست آمده مربوط به $\log LD_{50}$

Table 2- List of descriptors and statistical coefficients

Model	Independent variables	R	R^2	R^2_{adj}	RMSE	F	Sig
1	Mor13m, IVDE, Mor06u, EEig01r, DISPv, RDF020u, Mor32u, MAXDN, Mor02m, DISPe, Mor09m, TIE, ESpm03d, Mor07m, BEHv4, RDF020v, ATS2p, DISPp, MA	0.827	0.684	0.33	0.439	1.934	0.08
2	Mor13m, IVDE, Mor06u, EEig01r, DISPv, RDF020u, MAXDN, Mor02m, DISPe, Mor09m, TIE, ESpm03d, Mor07m, BEHv4, RDF020v, ATS2p, DISPp, MAXDP	0.827	0.684	0.367	0.427	2.161	0.056
3	Mor13m, IVDE, Mor06u, EEig01r, DISPv, RDF020u, MAXDN, DISPe, Mor09m, TIE, ESpm03d, Mor07m, BEHv4, RDF020v, ATS2p, DISPp, MAXDP	0.827	0.683	0.4	0.416	2.409	0.033
4	Mor13m, IVDE, EEig01r, DISPv, RDF020u, MAXDN, DISPe, Mor09m, TIE, ESpm03d, Mor07m, BEHv4, RDF020v, ATS2p, DISPp, MAXDP	0.826	0.682	0.428	0.406	2.681	0.019
5	Mor13m, IVDE, EEig01r, DISPv, MAXDN, DISPe, Mor09m, TIE, ESpm03d, Mor07m, BEHv4, RDF020v, ATS2p, DISPp, MAXDP	0.825	0.681	0.453	0.397	2.991	0.011
6	Mor13m, IVDE, EEig01r, DISPv, MAXDN, DISPe, Mor09m, TIE, ESpm03d, Mor07m, RDF020v, ATS2p, DISPp, MAXDP	0.823	0.677	0.472	0.390	3.295	0.006
7	Mor13m, IVDE, EEig01r, DISPv, MAXDN, DISPe, Mor09m, TIE, ESpm03d, Mor07m, RDF020v, DISPp, MAXDP	0.819	0.671	0.484	0.385	3.602	0.004
8	Mor13m, IVDE, EEig01r, DISPv, MAXDN, DISPe, TIE, ESpm03d, Mor07m, RDF020v, DISPp, MAXDP	0.814	0.663	0.494	0.381	3.935	0.002
9	IVDE, EEig01r, DISPv, MAXDN, DISPe, TIE, ESpm03d, Mor07m, RDF020v, DISPp, MAXDP	0.805	0.648	0.493	0.382	4.182	0.001
10	), IVDE, EEig01r, MAXDN, DISPe, TIE, ESpm03d, Mor07m, RDF020v, DISPp, MAXDP	0.786	0.617	0.470	0.390	4.195	0.002
11	IVDE, EEig01r, MAXDN, TIE, ESpm03d, Mor07m, RDF020v, DISPp, MAXDP	0.765	0.585	0.446	0.399	4.223	0.002
12	IVDE, MAXDN, TIE, ESpm03d, Mor07m, RDF020v, DISPp, MAXDP	0.752	0.565	0.441	0.401	4.552	0.001
13	IVDE, MAXDN, ESpm03d, Mor07m, RDF020v, DISPp, MAXDP	0.743	0.552	0.444	0.400	5.111	0.001
14	IVDE, MAXDN, ESpm03d, RDF020v, DISPp, MAXDP	0.731	0.534	0.441	0.401	5.729	0.000

[19] *Coefficient* of *determination*
[20] Correlation coefficient
[21] R^2 *adjusted*
[22] Fisher
[23] *Significance level*
[24] Mean Squared Error
[25] Root Mean Squared Error

مقادیر مختلف ضرایب آماری در جدول ۲ برای مدل های مختلف نشان داده شده است. با توجه به اینکه یکی از شرایط معادله مناسب وجود تعداد توصیف گر کمتر می باشد، لذا مدل های با ضریب رگرسیون R^2 =۰.٦٨ که ضرایب آماری نسبتا نزدیک به هم دارند در صورتی می توانند (تا حدودی) مورد قبول واقع گردند که حاوی تعداد توصیف گر کمتری باشند. معادله ارتباط نهایی بدست آمده (با توجه به ضریب فیشر بالاتر، خطای کمتر و تعداد توصیف گر کمتر) بین $\log LD_{50}$ و توصیفگرهای نهایی IVDE, MAXDN, ESpm03d, RDF020v, DISPp, MAXDP برای ۳۷ ترکیب مورد مطالعه به همراه ضریب رگرسیون R، ضریب تعیین R^2، ضریب رگرسیون تعدیل‌شده R^2_{adj}، خطای جذر میانگین مربعات RMSE، سطح معناداری Sig، آماره دوربین–واتسون[۲۶] D-W، به صورت زیر بدست آمده است:

$$\text{Log } LD_{50} = 5.011+0.654\ (\text{MAXDN}) +0.888\ (\text{MAXDP})- 1.158\ (\text{IVDE})-1.076\ (\text{ESpm03d}) -0.576(\text{DISPp})+0.423(\text{RDF020v})$$

N=37, R=0.731, R^2=0.534, R^2_{adj}=0.441, Durbin-Watson= 1.998, RMSE=0.401, F=5.729, Sig=0.000

مقدار ضریب رگرسیون R=۰.۷۳۱ در مدل آخر، میزان ارتباط خطی بین مقادیر $\log LD_{50}$ و متغیرهای مستقل را بیان می‌کند. مقدار بدست آمده برای **ضریب تعیین** نشان می‌دهد که فقط ۵۳٪ تغییرات $\log LD_{50}$ به‌وسیله توصیف گرهای وابسته به آن تبیین می‌شوند. R^2_{adj} مقدار **ضریب تعیین** را با توجه به متغیرهای مستقل اضافه‌شده به **خط رگرسیون** و با توجه به عرض از مبدأهای جدید، تعدیل و اصلاح می‌کند. مقدار محاسبه شده برای RMSE بیانگر خطای جذر میانگین مربعات است. مقدار سطح معنی‌داری برابر با صفر، حاکی از معنادار بودن رابطه فوق می باشد. مقدار بهینه برای آماره D-W در بازه بین ۰ تا ٤ می باشد و مقدار D-W=1.998 برای داده‌های آموزش، نشان دهنده عدم خودهمبستگی میان باقیمانده‌هاست.

تجزیه و تحلیل در روش غیر خطی

با توجه به پایین بودن مقادیر رگرسیون و مقدار نسبتا بالای خطای میانگین در مدل ارائه شده به روش خطی، به مطالعه ارتباط کمی ساختار - فعالیت به روش غیرخطی، اقدام گردید. برای مدل سازی در روش غیرخطی، 6 توصیف‌گر بدست آمده در مدل خطی نهایی، به کار گرفته شده اند. به منظور بررسی همبستگی بین توصیف گرها، بایستی ضرایب همبستگی پیرسون[۲۷] (PCC) و ضرایب نفوذپذیری[۲۸] VIF محاسبه شوند که مقادیر آنها برای توصیف گرهای مدل نهایی در جدول ۳ آورده شده اند. هرچه مقدار PCC بین توصیف گرها کمتر از یک و مقدار VIF بین ۱ تا ۱۰ باشد، می توان نتیجه گرفت بین متغیرهای مستقل مورد بررسی همبستگی وجود ندارد و یا این همبستگی کم می باشد. اگر دو متغیر مستقل دارای همبستگی باشند باید یکی از آن ها حذف گردد(معمولا شاخصی که بالاترین PCC و VIF را دارد حذف می شود). در نهایت مدلی که دارای حداقل همبستگی بین توصیفگرها و کمترین مقدار ضریب نفوذپذیری باشد به عنوان بهترین مدل انتخاب می گردد.

[26] Durbin-*Watson*
[27] Pearson's correlation *coefficient*
[28] Variance inflation factor

جدول ۳ ماتریس همبستگی و ضرایب نفوذپذیری بین توصیفگرهای مولکولی در مدل نهایی $\log LD_{50}$

Table 3- The correlation matrix between the molecular descriptors and permeability coefficients in final model

Descriptor	IVDE	MAXDN	ESpm03d	RDF020v	DISPp	MAXDP	Tolerance	VIF
IVDE	1.000						0.764	1.309
MAXDN	-0.053	1.000					0.390	2.564
ESpm03d	0.162	-0.060	1.000				0.452	2.213
RDF020v	-0.107	0.446	-0.252	1.000			0.643	1.556
DISPp	0.070	0.348	-0.242	0.034	1.000		0.805	1.243
MAXDP	-0.381	0.545	-0.066	0.03	0.077	1.000	0.357	2.798

با توجه به نتایج درج شده در جدول۳ معلوم می گردد که توصیفگرهای هندسی[۲۹]،ساختاری[۳۰]، اطلاعاتی[۳۱] و مجاورت[۳۲] نسبت به دیگر توصیفگرهای مولکولی از اهمیت بیشتری برای مدلسازی سمیت مشتقات کربوکسیلیک اسید برخوردار می باشند. تعریف و نوع شش توصیف گر نهایی در جدول ٤ شرح داده شده است.

جدول ٤ شرح و نوع توصیف گرهای نهایی

Table 4- Description and type of final descriptors

Notation	Descriptor	Type of Descriptor
IVDE	mean information content on the vertex degree equality	Information index
MAXDN	mean information content on the vertex degree equality	Information index
MAXDP	maximal electrotopological negative variation	Topological index
ESpm03d	Spectral moment 01 from edge adj. matrix weighted by dipole moments	Edge adjacency index
DISPp	displacement value / weighted by polarizability	Geometrical descriptor
RDF020v	maximal electrotopological positive variation	Topological index

در مدل سازی به روش ANN کل ترکیبات به صورت تصادفی به دسته آموزشی، آزمون و اعتبارسنجی به نسبت های ۷۰٪ و ۱۵% و ۱۵٪ تقسیم بندی شدند، بطوری که برای نمونه آزمایشی ۲۵ ترکیب و برای نمونه های آزمون و ارزیابی هر یک ٦ ترکیب در نظر گرفته شد و سپس اقدام به مدلسازی گردید و ضرایب آماری برای هر یک به تفکیک محاسبه گردیدند.

ارزیابی مدل به دست آمده در روش ANN برای پیش بینی $\log LD_{50}$

در جدول ۵ تعداد ترکیبات بکار گرفته شده و نیز مقادیر $RMSE, MSE, R^2$ بدست آمده در بهترین مدل برگزیده با استفاده از روش ANN برای سه دسته آموزش، آزمون و اعتبارسنجی ارائه شده است.

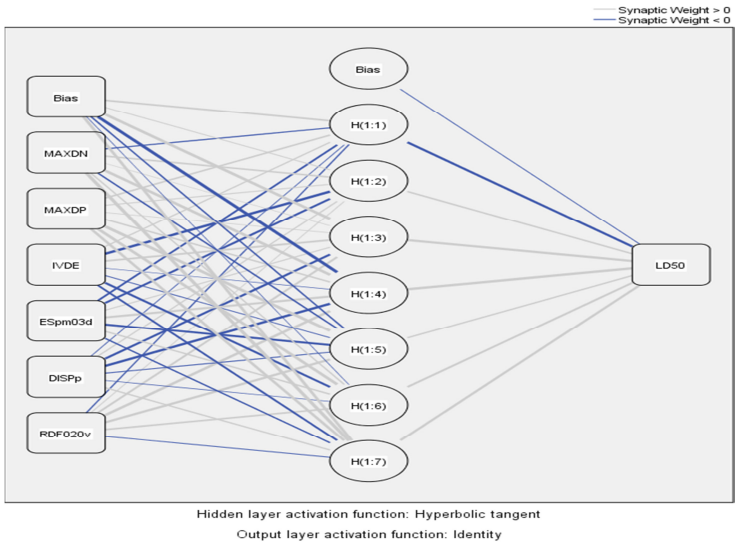
[29] Geometrical
[30] Topological
[31] Information
[32] Edge adjacency

جدول ۵ نتایج مدل سازی برای $\log LD_{50}$ در مجموعه آزمایشی و آزمون و ارزیابی در روش شبکه عصبی مصنوعی

Table-5: The modelling results in train, test and validation sets in ANN method.

parameters	sample	MSE	RMSE	R	R^2
Train	25	0.00183	0.04278	0.97482	0.95027
validation	6	0.02996	0.17309	0.99584	0.99169
test	6	0.02543	0.15948	0.99425	0.98854

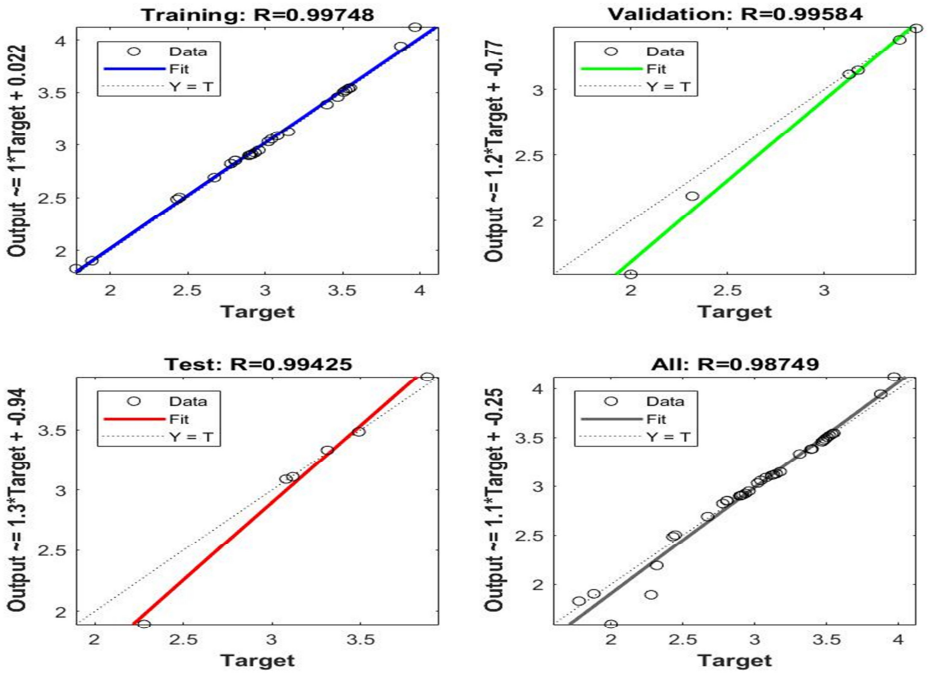
تعداد شاخص های ورودی که درشبکه عصبی به نام لایه های ورودی شناخته می شود، همان ٦ توصیف گر مولکولی مستخرج از مدل نهایی در روش MLR می باشند. تعداد لایه پنهان که برنامه محاسباتی آن را لحاظ کرده است، بعد از انجام عملیات بهینه سازی و تکرار زیاد برای درج بهترین نتیجه، شش لایه و نیز تعداد لایه خروجی نیز یک لایه می باشد. شکل 1 ساختار ANN به کار برده شده در مدل غیر خطی را نشان می دهد.



شکل 1 ساختار ANN برای $\log LD_{50}$

Fig 1 Structure of artificial neural network

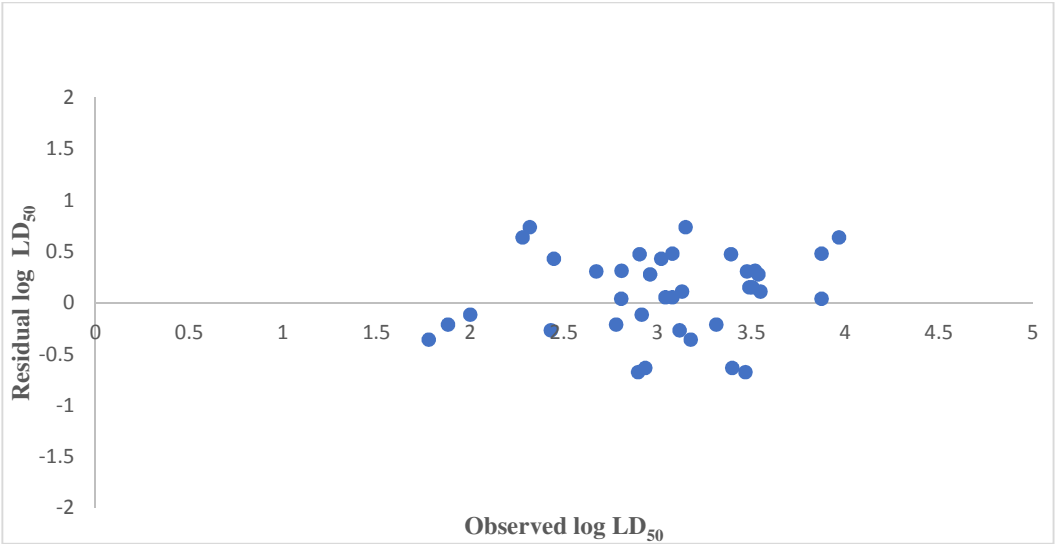
نمودار مقادیر تجربی $\log LD_{50}$ برحسب مقادیر پیش بینی $\log LD_{50}$ در روش ANN برای هر سه دسته آموزش، آزمون و اعتبارسنجی ترکیبات مورد مطالعه در شکل ۲ نشان داده شده است.



شکل ۲ نمودارمقادیرمشاهده شده درمقابل مقادیرپیش بینی شده $\log LD_{50}$ با روش ANN
برای کل داده ها و دسته های آموزش، آزمون واعتبارسنجی

Fig 2 The curves of log LD_{50} observed versus predicted values for all, train, test and validation sets

همچنین در شکل ۳ نمودار تغییرات مقادیر مشاهده‌شده برحسب باقیمانده برای لگاریتم سمیت با روش ANN نشان داده‌شده است.



شکل ۳- نمودار تغییرات مقادیر تجربی $\log LD_{50}$ برحسب مقادیر باقیمانده درروش ANN

Fig 3 The curve of log $LD_{50(exp)}$ versus log $LD_{50(res)}$values in ANN method

تجزیه و تحلیل با استفاده از روش ANN

مقادیر نسبتاً بالای R برای هر سه دسته و هم‌چنین مقدار خطای کم RMSE در آن‌ها برتری روش شبکه عصبی مصنوعی نسبت به روش خطی چندگانه برای پیش‌بینی لگاریتم شاخص سمیت LD_{50} تائید می‌گردد. ضریب رگرسیون R () بالای ۹۷ درصد بیانگر میزان ارتباط خطی بالا بین مقادیر LD50 و توصیف گرهای مولکولی می باشد. با توجه به مقادیر RMSE=0.04278867 وR^2= ۰.۹۵۰۲۷۴، معلوم می شود که بیش از ۹۵٪ از تغییرات شاخص سمیت LD_{50} در ترکیبات آموزش را می توان توسط توصیف‌گرهای مولکولی فوق تبیین نمود. میزان پراکندگی نقاط برای سه دسته آموزش و آزمون و اعتبار سنجی در نمودار مشاهده‌شده برحسب پیش‌بینی‌شده حاکی از همبستگی خوب بین توصیف‌گرهای نهایی با لگاریتم سمیت و درنتیجه کار آیی روش به‌کاررفته در پیش‌بینی مشتقات کربوکسیلیک اسید می‌باشد. همچنین نمودار تغییرات مقادیر مشاهده‌شده برحسب باقیمانده نشان می‌دهد که خطاها در اطراف محور x توزیع یکسانی دارند و این نمایانگر این است که الگوی مذکور برای پیش‌بینی $\log LD_{50}$ در ترکیبات موردمطالعه از دقت بالا و بسیار خوبی برخوردار می‌باشد.

مقادیر تجربی پیش‌بینی و باقیمانده مربوط به لگاریتم شاخص LD_{50} در روش ANN مطابق جدول ٦ گزارش شده است.

جدول 6 مقادیر مشاهده شده، پیش بینی شده و باقیمانده $\log LD_{50}$ برای مشتقات کربوکسیلیک اسید در مدل های MLR و ANN

Table 6The $\log LD_{50}$ observed, predicted and residual values of carboxylic acids derivatives in MLR and ANN

		MLR		ANN	
No	**$\log LD_{50(obs)}$**	**$\log LD_{50(Pred)}$**	**$\log LD_{50(Res)}$**	**$\log LD_{50(Pred)}$**	**$\log LD_{50\ (Res)}$**
1	2.778	3.094	-0.316	2.951	-0.213
2	3.391	3.080	0.310	3.132	0.472
3	3.505	3.054	0.450	2.882	0.150
4	2.318	2.754	-0.433	2.931	0.734
5	2.959	2.843	0.111	2.889	0.274
6	1.778	2.326	-0.548	2.638	-0.361
7	3.079	3.370	-0.291	2.740	0.054
8	2.808	2.731	0.077	2.723	0.311
9	3.079	2.799	0.279	2.869	0.476
10	2.672	2.443	0.228	2.649	0.303
11	2.934	2.826	0.107	2.949	-0.636
12	2.915	2.460	0.455	2.503	-0.118
13	2.431	2.397	0.033	2.553	-0.269
14	3.130	3.127	0.002	3.082	0.109
15	2.895	2.580	0.315	2.505	-0.678
16	3.019	2.569	0.449	2.623	0.428
17	2.279	2.782	-0.503	3.214	0.636
18	2.805	3.045	-0.240	2.937	0.039
19	1.881	2.477	-0.597	2.541	-0.213
20	2.903	3.237	-0.334	3.116	0.472
21	3.489	3.320	0.168	3.016	0.150
22	3.149	3.222	-0.073	2.998	0.734
23	3.537	3.239	0.297	2.801	0.274
24	3.176	3.367	-0.191	2.901	-0.361
25	3.041	3.226	-0.185	3.402	0.054
26	3.520	3.477	0.042	3.465	0.311
27	3.875	3.529	0.345	3.563	0.476
28	3.477	2.755	0.721	3.000	0.303
29	3.398	3.286	0.111	3.094	-0.636
30	2.000	2.690	-0.690	2.636	-0.118
31	3.117	2.802	0.314	3.235	-0.269
32	3.549	3.838	-0.289	3.818	0.108
33	3.468	3.298	0.169	3.359	-0.678
34	2.447	3.248	-0.801	3.125	0.428
35	3.968	3.761	0.206	3.540	0.636
36	3.875	3.455	0.419	3.238	0.039
37	3.312	3.434	-0.122	3.272	-0.213

مقایسه روش‌های رگرسیون خطی با شبکه عصبی، نشان داد که بین مقادیر سمیت مشتقات کربوکسیلیک اسیدها با توصیفگرهای مولکولی، رابطه غیرخطی وجود دارد. مدل رگرسیون خطی بکار گرفته‌شده، علی‌رغم اینکه چارچوب وسیعی را در برمی‌گیرد که با تحلیل‌های زیادی همراه هست، ولی نمی‌تواند الگوی مناسبی برای هدف این تحقیق باشد، زیرا در بعضی مواقع، متغیرهای رگرسیونی با توابع غیرخطی به هم مربوط می‌شوند. بنابر این، شبکه عصبی مصنوعی به‌عنوان یک مدل غیرخطی، می‌تواند روابط غیرخطی پیچیده یا لایه‌های پنهان بین متغیرهای وابسته و مستقل را بیابد و با دقت بهتری نسبت به روش‌های رگرسیونی خطی عمل نماید. روش ANN رابطه خوب بین ورودی و خروجی ایجاد نموده و از حساسیت کمتری نسبت به وجود خطاها در اطلاعات ورودی برخوردار است، لذا از نظر میزان توانایی در پیش‌بینی ارتباط بین متغیر وابسته و متغیرهای مستقل و قابلیت تعمیم نسبت به روش‌های دیگر مدل‌سازی، نتیجه قابل‌قبول‌تری ارائه داده است.

نتیجه گیری

بسیاری از کربوکسیلیک اسیدها، سمی و برای سلامتی بشر مضرند. بسیاری از آنها موادی خورنده و در صورت تماس با پوست باعث سوختگی شیمیایی میگردند و تنفس بخارات آن ها موجب تحریک و سوزش دستگاه تنفسی می شود. اندازه‌گیری سمیت بسیاری از مشتقات کربوکسیلیک اسید از طریق روش‌های آزمایشگاهی به دلایل ذکر شده زیان‌آور می‌باشد. به کارگیری روش های محاسباتی هم باعث صرفه جویی در وقت و هزینه های آزمایشگاهی می شود و هم سبب آسیب نرسیدن به محیط زیست برای ترکیبات شناخته شده و یا فاقد اطلاعات لازم می گردد. به همین منظور در این تحقیق، با مطالعه رابطه ساختار-فعالیت به‌پیش بینی مقادیر لگاریتم سمیت LD_{50} کربوکسیلیک اسیدها با دقت بالا پرداخته‌شده است. از بین تعداد بسیار زیاد توصیف گرهای مولکولی محاسبه شده با روش الگوریتم ژنتیک و برگشتی، معلوم گردید که هفت توصیف گر از گروه های هندسی، ساختاری، مجاورت و اطلاعاتی نسبت به دیگر توصیف گرهای مولکولی از اهمیت بالاتری برخوردار هستند. به منظور مطالعه رابطه ساختار-فعالیت، برای ترکیبات مورد مطالعه از هر دو مدل رگرسیون خطی چندگانه MLR و نیز شبکه عصبی مصنوعی ANN استفاده گردیده است. پارامترهای آماری در مدل نهایی برتری قابل‌توجه مدل ANN را نسبت به مدل MLR، برای پیشگویی لگاریتم سمیت ترکیبات مورد مطالعه نشان دادند. نتایج این تحقیق نشان داد که می‌توان گام‌های بلندی را جهت پیش‌بینی مقادیر سمیت بسیاری از مشتقات جدید کربوکسیلیک اسید با تکیه بر مدل‌های محاسباتی به کمک پارامترهای اعتبارسنجی مناسب برداشت. بدیهی است که چنین مدلی قطعا به منظور شناخت بیشتر ویژگی های این دسته از ترکیبات شیمیایی و حتی کاربرد آنها در طراحی ترکیبات شیمیایی و صنعتی بسیار مفید خواهد بود.

Referance

Asadollahi-Baboli, M. and Dehnavi, S. 2018. Docking and QSAR analysis of tetracyclic oxindole derivatives as α-glucosidase inhibitors. Computational biology and chemistry, 76: 283-292.

Bagheban Shahri, F. and Niazi, A. 2016. Quantitative structure activity relationship study of inhibitory activities of 5-lipoxygenase and design new compounds by different chemometrics methods. Iranian Journal of mathematical Chemistry, 7: 47-59.

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

Cross, S., Harrison, R. F., Kennedy, R. L. 1995. Introduction to neural networks, Lancet, 346: 1075-9.

Elisseeff., A. and Guyon, I. 2003. An introduction to variable and feature selection. J. of Machine Learning Research, 19: 1157-1182.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T D., McDowell, R. M. and Gramatica, P. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification–and regression–based QSARs. Environ Health Perspect. 111(10): 1361–1375.

Fissa, M.R., Lahiouel, Y., Khaouane, L. and Hanini, S. 2019. QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. Journal of Molecular Graphics and Modelling, 87: 109-120.

Fox, T.R. and Comerford, N.B. 1990. Low-molecular-weight organic acids in selected forest soils of the southeastern USA. Soil Science Society of America Journal, 54(4): 1139-1144.

Fox, T.R. 1995. The influence of low-molecular-weight organic acids on properties and processes in forest soils. Carbon forms and functions in forest soils, pp:43-62.

Gadzuric, S. B., Podunavac-Kuzmanovic, S. O., Jokic, A. I., Vranes, M. B., Ajdukovic, N and Kovacevic, S. Z. 2015. Chemometric estimation of post-mortem interval based on Na+ and K+ concentrations from human vitreous humour by linear least squares and artificial neura. Australian, Journal of Forensic Sciences, 46: 166-179.

Ghamali, M., Chtita, S., Ousaa, A., Elidrissi, B., Bouachrine, M. and Lakhlifi, T. 2017. QSAR analysis of the toxicity of phenols and thiophenols using MLR and ANN. Journal of Taibah University for Science, 11(1): 1-10.

Ghasemian, A., Asadollahzadeh, M., Saraeian, A., Resalati, H. and Taherzadeh, M. 2018. Effect of Acetic Acid on Growth and Ethanol Fermentation of Filamentous Fungi Rhizopus oryzae, Mucor indicus, Neurospora intermedia and Aspergilus oryzae. Experimental animal Biology, 7(3): 119-130.

Goodarzi, M., Dejaegher, B. and Heyden, Y. V. 2012. Feature selection methods in QSAR studies. Journal of AOAC International, 95(3): 636-651.

Ha, H., Park, K., Kang, G. and Lee, S. 2019. QSAR study using acute toxicity of Daphnia magna and Hyalella azteca through exposure to polycyclic aromatic hydrocarbons (PAHs). Ecotoxicology, 28(3): 333-342.

Hajalsiddig, T.T.H. and Saeed, A.E.M. 2019. QSAR and molecular docking studies on 4-quinoline carboxylic acid derivatives as inhibition of vesicular stomatitis virus replication. European Journal of Chemistry, 10(1): 45-51.

Haykin, S. 1990. Neural networks -a comprehensive foundation, 2nd ed. New Jersey: Prentice-Hall.

Hong, H., Roy, B.C., Chalamaiah, M., Bruce, H.L. and Wu, J. 2018. Pretreatment with formic acid enhances the production of small peptides from highly cross-linked collagen of spent hens. Food chemistry, 258: 174-180.

Khan, K., Benfenati, E. and Roy, K. 2019. Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. Ecotoxicology and environmental safety, 168: 287-297.

Kompany-Zareh, M. 2009. An improved QSPR study of the toxicity of aliphatic carboxylic acids using genetic algorithm. Medicinal chemistry research, 18(2): 143-157.

Kutner, M. K., Nachtsheim, C. J. and Neter, J. 2004. Applied linear regression models, Boston: McGraw-Hill.

Maguna, F.P., Ninez, M.B., Okulik, N.B. and Castro, E.A. 2003. Improved QSAR analysis of the toxicity of aliphatic carboxylic acids. Russian Journal of General Chemistry, 73(11): 1792-1798.

Maiellaro, P. A., Cozzolongo, R. and Marino, P. 2004. Artificial neural networks for the prediction of response to interferon plus ribavirin treatment in patients with chronic hepatitis C. Cuurent Pharmaceutical Design, 10: 2101-2109.

Matysiak, I., Balcerzak, M. and Michalski, R. 2018. Ion chromatography with conductometric detection for quantitation of formic acid in Polish bee honey. Journal of Food Composition and Analysis, 73: 55-59.

Mirjalili, S. 2019. Genetic algorithm. In Evolutionary algorithms and neural networks, pp: 43-55. Springer, Cham.

Mirjalili, S., Dong, J.S., Sadiq, A.S. and Faris, H. 2020. Genetic algorithm: Theory, literature review, and application in image reconstruction. Nature-inspired optimizers, pp: 69-85.

Mohammaei, F. and Mohammadinasab, E. 2018. Coefficient Partition Prediction of Saturated Monocarboxylic Acids Using the Molecular Descriptors. Journal of the Chilean Chemical Society, 63(3): 4068-4071.

Nielsen, G.D. 2018. Sensory irritation of vapours of formic, acetic, propionic and butyric acid. Regulatory Toxicology and Pharmacology, 99: 89-97.

Polishchuk, P. 2017. ,Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future. J. Chemical Information Model, 57: 2618-2639.

Qin, Z., Wang, M. and Yan, A. 2017. QSAR studies of the bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors by multiple linear regression (MLR) and support vector machine (SVM). Bioorganic & medicinal chemistry letters, 27(13): 2931-2938.

Randic, M. and Basak, S. C. 2000. Multiple regression analysis with optimal molecular descriptors. SAR & QSAR Research, 11: 1-23.

Roy, K. and Ambure, P. 2016. The "double cross-validation" software tool for MLR QSAR model development. Chemometrics and Intelligent Laboratory Systems, 159: 108-126.

Roy, K., and Mitra, I. 2011. On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design. Combinatorial Chemistry & High Throughput Scree, 14: 450-474.

Slattery, W.J. and Morrison, G.R. 1995. Relationship between soil solution aluminium and low molecular weight organic acids in a conservation cropping system. In Plant-Soil Interactions at Low pH: Principles and Management, 589-593.

Strobel, B.W., Bernhoft, I. and Borggaard, O.K. 1999. Low-molecular-weight aliphatic carboxylic acids in soil solutions under different vegetations determined by capillary zone electrophoresis. Plant and Soil, 212(2): 115-121.

Tani, M., Higashi, T. and Nagatsuka, S. 1996. Dynamics of low-molecular-weight aliphatic carboxylic acids (LACAs) in forest soils: II. Seasonal changes of LACAs in an andisol of Japan. Soil science and plant nutrition, 42(1): 175-186.

Todeschini, R. and Consonni, V. 2000. Handbook of Molecular Descriptors. Weinheim: Wiley-VCH.

Todeschini, R. and Consonni, V. 2009. Molecular descriptors for chemoinformatics. Alphabetical listing (2nd ed., Vol. 1). Weinheim: Wiley-VCH.

Tuzun, B., Yavuz, S.C., Sabanci, N. and Saripinar, E. 2018. 4D-QSAR Study of Some Pyrazole Pyridine Carboxylic Acid Derivatives By Electron Conformational-Genetic Algorithm Method. Current computer-aided drug design, 14(4): 370-384.

U.S. National Library of Medicine: (http://chem.sis.nlm.nih.gov/chemidplus/).

Williams, P. L., James, R. C., Roberts. S. M. 2003. Ph.D., Principles of Toxicology: Environmental and Industrial Applications, Second Edition. John Wiley & Sons, Inc.

WWW.sigma-aldrich.com

Xu, X., Du, X., Zheng, M. and XU, Z.Y. 2005. Synthesis of Carboxylic Acid Chlorides from Bis (trichloromethyl) carbonate. Pesticides-shenyang-, 44(6): 265.

Journal of Entomological Research

Volume 13, Issue 1, pages: 67-82

Islamic Azad University, Arak Branch
ISSN 2008-4668
http://jer.iau-arak.ac.ir

Prediction of LD_{50} for carboxylic acid derivatives using multiple linear regression and artificial neural networks models

F. Mohammaei[1], E. Mohammadinasab[2*]

1. PHD student, Department of Chemistry, Arak Branch, Islamic Azad University, P.O. BOX 38135-567, Arak, Iran
2. Assistant Professor, Department of Chemistry, Arak Branch, Islamic Azad University, P.O. BOX 38135-567, Arak, Iran

In this research, Quantitative Structure–Activity Relationship (QSAR) study has been used for prediction of toxicity values of carboxylic acid derivatives. Firstly, the toxicity (LD_{50}) values of data set of studied compounds were taken from the scientific web book and the their structures were drawn with the Gauss view 05 program and optimized at Hartree–Fock level of theory and 3-21G basis set by Gaussian 09 software. Then the dragon software was used for the calculation of molecular descriptors. The unsuitable descriptors were deleted with the aid of the genetic algorithm (GA) and backward techniques, and the best descriptors were used for multiple linear regression (MLR) and artificial neural network (ANN) models. The prediction accuracy of the final model was discussed using the statistical parameters. Leave-one-out cross-validation and external test set of the predictive models demonstrated a high-quality correlation between the observed and predicted toxicity values of all, training, test and validation sets in GA-ANN method. The model by ANN algorithm due to the lower error and higher regression coefficients was clearly superior to those models by MLR algorithm. The proposed model may be useful for predicting log LD_{50} of new compounds of similar class.

Key words: "Toxicity" ; "Multiple linear regression method" ; "Artificial neural network" ; "Carboxylic acid derivatives.

* Corresponding Author, E-mail*: -mohammadinasab@iau-arak.ac.ir*
Received: 10 Mar. 2021 – Accepted: 2 May. 2021