



روشهای محاسباتی و مطالعه ارتباط کمی ساختار- خاصیت جهت پیش‌بینی سمیت برخی از مشتقات فنل‌ها و آنیلین‌ها با استفاده از توصیف‌کننده‌های شیمیایی و مکانیک کوانتومی و شبکه‌های عصبی مصنوعی

مهدی نکویی*، سیدمحسن بابازاده

دانشگاه آزاد اسلامی، واحد شاهرود، دانشکده علوم پایه، گروه شیمی، شاهرود، ایران

تاریخ ثبت اولیه: ۱۳۹۴/۱۰/۱۴، تاریخ دریافت نسخه اصلاح شده: ۱۳۹۴/۱۲/۱۶، تاریخ پذیرش قطعی: ۱۳۹۵/۱/۱۷

چکیده

مشتقات فنل و آنیلین از ترکیبات معمول در فاضلاب صنایع مختلفی همچون نفت و پتروشیمی، آفت کش‌ها، رنگ، تولید مواد شیمیایی آلی، پلاستیک و رزین و غیره است. آلوده شدن منابع آب به این مشتقات یک مشکل جدی و تهدیدی برای سلامتی انسان، بدلیل سمیت بالای آنها محسوب می‌شود. در این تحقیق مطالعه ارتباط کمی ساختار- خاصیت جهت پیش‌بینی سمیت برخی از مشتقات آنیلین‌ها و فنل‌ها با استفاده از توصیف‌کننده‌های شیمیایی و مکانیک کوانتومی و شبکه‌های عصبی مصنوعی انجام شد. در ابتدا ساختار ترکیبات رسم و گروه مناسبی از توصیف‌کننده‌های شیمیایی و مکانیک کوانتومی محاسبه شدند. سپس از روش انتخاب مرحله‌ای برای بدست آوردن بهترین توصیف‌کننده‌ها که بیشترین ارتباط را با سمیت ترکیبات مورد نظر داشتند استفاده گردید. در ابتدا مدل خطی رگرسیون خطی چندگانه ساخته شد. سپس برای به دست آوردن مدل بهتر، از شبکه عصبی مصنوعی استفاده گردید. به همین خاطر، توصیف‌کننده‌های انتخاب شده بوسیله روش مرحله‌ای به عنوان ورودی شبکه عصبی مصنوعی مورد استفاده قرار گرفتند. داده‌های آماری، برتری روش غیرخطی شبکه عصبی مصنوعی را نسبت به روش خطی رگرسیون خطی چندگانه نشان می‌دهد.

واژه‌های کلیدی: ارتباط کمی ساختار- خاصیت، سمیت، رگرسیون خطی چندگانه، شبکه عصبی مصنوعی، توصیف‌کننده‌های مکانیک کوانتومی

۱. مقدمه

امروزه به دلیل گسترش روزافزون صنایع شیمیایی به ویژه بر پایه مواد نفتی و تولید فراورده‌های مربوطه و نیاز گسترده به آب در فرایند تولید، انواع پسابهای خطرناک حاوی ترکیباتی چون هیدروکربورها، چربی‌ها و روغن‌ها، فنل‌ها و سایر ترکیبات آروماتیک به طبیعت وارد

*عهده دار مکاتبات: مهدی نکویی

نشانی: شاهرود، دانشگاه آزاد اسلامی، گروه شیمی

تلفن: ۰۲۳-۳۲۳۹۴۲۸۹-۲۳ پست الکترونیک: E-Mail: m_nekoei1356@yahoo.com

می شوند. در بین ترکیبات شیمیایی موجود در پسابهای صنعتی، فنل، آنیلین و مشتقات آنها از ترکیبات فراگیر هستند که علاوه بر طرق مصنوعی از طریق طبیعی نیز وارد منابع آب شده و همچنین به دلیل ساختمان فیزیکی شان در اکثر ترکیبات شیمیایی و حتی در فاضلابهای شهری نیز وجود دارند و به دلیل پایداری نسبی در محیط، قابلیت انحلال در آب و مشکلات بهداشتی مورد توجه می باشند. فنل و آنیلین و مشتقات آنها در صنایع فولاد، صنایع تولید منسوجات مصنوعی و ساخت رزین و پلاستیک به عنوان کاتالیست، صنایع رنگ، خودروسازی، صنایع آلومینیوم، کودسازی، ساخت آفت کش ها، صنایع دارویی و آرایشی، پتروشیمی و ... کاربرد دارند، لذا به عنوان یک آلاینده مهم در پساب صنایع فوق یافت می شوند [۱].

در سالهای اخیر اثرات فاضلاب های خروجی صنایع بر اکوسیستم به طور گسترده ای مورد توجه قرار گرفته است و محدودیتی در روند توسعه صنایع به دنبال داشته است، زیرا اکثر فاضلابها حاوی یک دامنه وسیعی از آلاینده ها می باشند که اولاً توسط روشهای متداول قابل شناسایی نیست و ثانیاً در صورت شناسایی و تعیین مقدار، حفظ حیات آبزیان در آبهای پذیرنده این نوع فاضلابها به دلیل عدم آگاهی از اثرات همزمان و تشدید کنندگی آلاینده ها امکان پذیر نیست. روش های مختلفی جهت ارزیابی سمیت فاضلاب شهری، شیرابه و فاضلاب صنعتی ارائه شده است [۲-۱]، اما روشهای آزمایشگاهی جهت اندازه گیری سمیت فاضلابها، وقت گیر، هزینه بر و نسبتاً مشکل است. بنابراین اندازه گیری آن با یک روش ساده تر، ارزان تر و سریعتر از اهمیت بالایی برخوردار خواهد بود. یکی از این روشها که امروزه از آن، جهت پیش بینی خواص مختلفی از ترکیبات استفاده می شود روش ارتباط کمی ساختار- خاصیت^۱ (QSPR) است. هدف از مطالعات QSPR پیدا کردن رابطه ای است که بین رفتار فیزیکی شیمیایی یک مولکول با پارامترهای ساختاری آن وجود دارد [۱۰-۳]. نتایج این مطالعات علاوه بر شفاف سازی نحوه ارتباط بین خواص مولکولها و ویژگی های ساختاری آنها به پژوهشگران در پیش بینی رفتار مولکولهای جدید براساس رفتار مولکولهای مشابه کمک می کند. کار اصلی در مطالعات ارتباط کمی ساختار-ویژگی، برقراری رابطه و مدلی بین خواص مولکولی مشخص و توصیف گرهای مولکولی توسط آمار و یا روش های دیگر است. باید توجه شود که مدل ممکن است کاملاً پیچیده باشد، و بنابراین اغلب انواعی از تقریبها به کار می رود. تعدادی از روشهایی که به طور وسیعی در مطالعات مدل سازی ارتباط کمی ساختار-ویژگی استفاده می شوند، روش های خطی هستند که مدل های خطی ساده و قابل تفسیری را می سازند، از قبیل روش حداقل مربعات متداول^۲ (OLS)، رگرسیون اجزاء اصلی^۳ (PCR) و حداقل مربعات جزئی^۴ (PLS). برخی دیگر از روشها مدل های غیرخطی اند، مانند شبکه های عصبی مصنوعی که بطور گسترده از آن استفاده می گردد.

شبکه عصبی مصنوعی^۵ از مجموعه ای از نرونها با ارتباطات داخلی بین یکدیگر تشکیل می شود که قادر است بر اساس اطلاعات و داده های ورودی، جوابهای خروجی را تولید نماید. شبکه های عصبی معمولاً به صورت لایه لایه و منظم ایجاد می گردند. اولین لایه که اطلاعات و داده های ورودی به آن وارد می گردند لایه ورودی است. لایه های میانی، لایه های مخفی و آخرین لایه که جوابهای خروجی از مدل را فراهم می نماید لایه خروجی است [۱۱]. ساده ترین و معمول ترین نوع شبکه عصبی که در بسیاری از علوم از آن استفاده می گردد، شبکه های عصبی چند لایه پیشخورند^۶ که همراه با ناظر از روش پسانتشار خطا برای آموزش کمک می گیرند. در این شبکه تعداد نرونهای لایه ورودی برابر با

1. Quantitative Structure-Property Relationship
2. Ordinary Least Squares
3. Principal Component Regression
4. Partial Least Squares
5. Artificial Neural Networks(ANN)
6. Multi-Layer Perceptron(MLP)

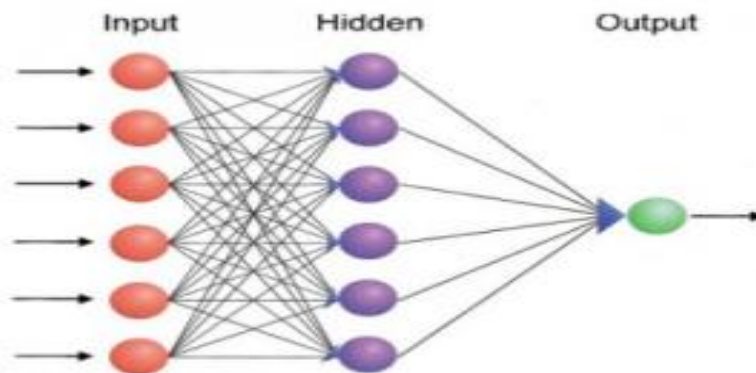
تعداد عناصر بردار ورودی و تعداد نرون‌های لایه خروجی برابر با تعداد عناصر بردار خروجی است. آنالیز دقیق و واقعی برای پیدا کردن تعداد نرون‌های لایه میانی در کل بسیار پیچیده است، اما می‌توان گفت که تعداد نرون‌های لایه میانی تابعی از تعداد عناصر بردار ورودی و همچنین حداکثر تعداد نواحی از فضای ورودی، به طور خطی جداپذیر است. از این رو تعداد نرون‌های لایه مخفی عموماً به طور تجربی به دست می‌آید. هر نرون توسط خروجی اش به نرون‌های لایه بعد متصل بوده ولی با نرون‌های لایه خودش ارتباط ندارد. خروجی هر نرون توسط رابطه زیر تعریف می‌شود [۱۱]:

$$\alpha = f(\sum P_i W_{ij} + b_j) \quad \text{رابطه (۱)}$$

که در این رابطه W_{ij} مقدار وزن اتصال بین نرون j ام لایه مذکور با نرون i ام لایه قبل است که بیانگر اهمیت ارتباط بین دو نرون در دو لایه متوالی است، b_j وزن مربوط به بایاس (Bias) برای نرون j ام، P_i مقدار خروجی از نرون i ام لایه قبل، α مقدار خروجی از نرون j ام و f تابع آستانه نرون j ام هستند. توابع زیادی وجود دارند که در انتقال اعداد از لایه قبل به لایه بعد ممکن است مورد استفاده قرار گیرند که از جمله آنها میتوان توابع سیگموئیدی، گوسی، تانژانت هیپربولیک و ... را نام برد. ولی تابع سیگموئید بیشترین استفاده را در علوم مختلف دارد. این تابع به صورت زیر است:

$$f(z) = 1 / (1 + \exp(-z)) \quad \text{رابطه (۲)}$$

شبکه‌های عصبی مصنوعی شامل دو مرحله هستند. مرحله اول آموزش شبکه است که بر اساس تعدادی از داده‌های اندازه‌گیری شده (حدود ۷۰ درصد کل داده‌ها) انجام می‌پذیرد. در مرحله دوم مدل با حدود ۲۰ تا ۳۰ درصد مقادیر اندازه‌گیری شده صحت سنجی می‌گردد. در مدل‌سازی به کمک شبکه‌های عصبی مصنوعی بسته به متغیرهای ورودی و ارتباط آن‌ها با پارامتر مورد نظر که مدل می‌گردد، حدود ۲۰ درصد اطلاعات را برای تست و ارزیابی مدل به کار می‌برند. البته این داده‌ها نیابستی در مراحل آموزش و صحت سنجی استفاده شده باشند. کاربرد شبکه‌های عصبی مصنوعی جهت مطالعات QSPR بسیار زیاد می‌باشد [۱۵-۱۲]. شکل زیر ساختار یک شبکه عصبی مصنوعی را نشان می‌دهد.



شکل ۱. ساختار یک شبکه عصبی مصنوعی.

سهم اصلی برای گسترش استفاده از مدل‌های ارتباط کمی ساختار-ویژگی مربوط است به توسعه توصیف‌گرهای ساختاری و معادلات ریاضی که خواص فیزیکی و شیمیایی را به ساختار شیمیایی مرتبط می‌کنند. موفقیت ارتباط کمی ساختار-ویژگی را می‌توان توسط بینشی که به تخمین ساختاری خواص شیمیایی و امکان تخمین خواص ترکیبات شیمیایی جدید می‌انجامد توضیح داد. یک فرض اساسی در روش ارتباط کمی ساختار-ویژگی این است که همه خواص (فیزیکی - شیمیایی) از یک ماده شیمیایی به ساختار مولکولی مرتبط است. بر اساس این فرض، تحقیقات ارتباط کمی ساختار-ویژگی در دهه‌های اخیر توسعه بسیاری پیدا کرده‌اند. هدف از مطالعه حاضر، پیش بینی میزان سمیت برخی از مشتقات فنل و آنیلین با استفاده از توصیف‌کننده‌های شیمیایی و مکانیک کوانتمی و شبکه‌های عصبی مصنوعی است.

۲. بخش محاسباتی

۲-۱. دستگاه و نرم افزارهای مورد استفاده

یک دستگاه کامپیوتر پنتیوم IV (CPU = 3.06 GHz) با ویندوز 7 برای انجام محاسبات مورد استفاده قرار گرفت. برای رسم و بهینه سازی شکل مولکولها از نرم افزار هایپرکم^۱ (نسخه ۷/۰) استفاده گردید. برای محاسبه توصیف کننده های مکانیک کوانتومی و شیمیایی از نرم افزارهای هایپرکم و دراگون^۲ استفاده شد. مشخصات برخی از توصیف کننده های محاسبه شده توسط نرم افزار هایپرکم در جدول ۱ آورده شده است. از نرم افزار SPSS برای انتخاب بهترین توصیف کننده ها و از برنامه ANN در محیط MATLAB برای پیش بینی سمیت ترکیبات و به دست آوردن سایر پارامترهای آماری استفاده گردید.

جدول ۱. مشخصات برخی از توصیف کننده های محاسبه شده.

توصیف کننده ها	نماد	اختصار	توصیف کننده ها	نماد	اختصار
توصیف کننده های مکانیک کوانتومی	Highest Occupied Molecular Orbital	HOMO	توصیف کننده های مکانیک کوانتومی	Hardness [$\eta=1/2 (HOMO+LUMO)$]	H
	Lowest Unoccupied Molecular Orbital	LUMO		Softness ($S=1/\eta$)	S
	difference between LUMO and HOMO	E GAP		Electro negativity [$\chi=-1/2 (HOMO-LUMO)$]	χ
	Molecular Polarizability	MP		Electrophilicity ($\omega=\chi^2/2\eta$)	Ω
ویژگی های شیمیایی	Molecule surface area (Approx)	SA (A)	ویژگی های شیمیایی	Partition Coefficient	Log P
	Molecule surface area (Grid)	SA (G)		Hydration Energy	HE
	Mass	M		Refractivity	REF
	Molecule volume	V			

۲-۲. سری داده ها

سری داده ها مربوط به میزان سمیت ۵۷ ترکیب از مشتقات فنل و آنیلین ها است که بطور تجربی اندازه گیری شده است [۱۶]. میزان سمیت این ترکیبات با پارامتر EC_{50} نشان داده شده است. EC_{50} غلظت یا دزی از سم است که اگر یک جمعیت از موجودات زنده در معرض آن قرار گیرند ۵۰ درصد جمعیت، تلف می شوند و اگر به یک موجود زنده منفرد (مثلا یک پستاندار) خورنده شود احتمال مرگ آن موجود ۵۰ درصد خواهد بود. EC_{50} در واقع مقدار سمی است که قادر است ۵۰٪ حیوانات مورد آزمایش را بکشد.

در ابتدا ترکیبات به صورت تصادفی به سه سری شامل سری آموزش، سری ارزیابی و سری پیش بینی تقسیم شدند (جدول ۱). سری آموزش شامل ۳۵ ترکیب (۶۰٪ داده ها)، سری ارزیابی شامل ۱۱ ترکیب (۲۰٪ داده ها) و سری پیش بینی یا تست نیز شامل ۱۱ ترکیب (۲۰٪ داده ها) است. لازم به ذکر است در روش MLR ترکیبات فقط به دو سری آموزش و تست تقسیم شدند و در اینجا سری ارزیابی بکار نرفت، اما داده های تست در هر دو روش یکسان بودند. مقادیر سمیت (EC_{50}) به عنوان متغیر وابسته و توصیف کننده ها به عنوان متغیر مستقل انتخاب شدند.

۲-۳. محاسبه توصیف کننده ها

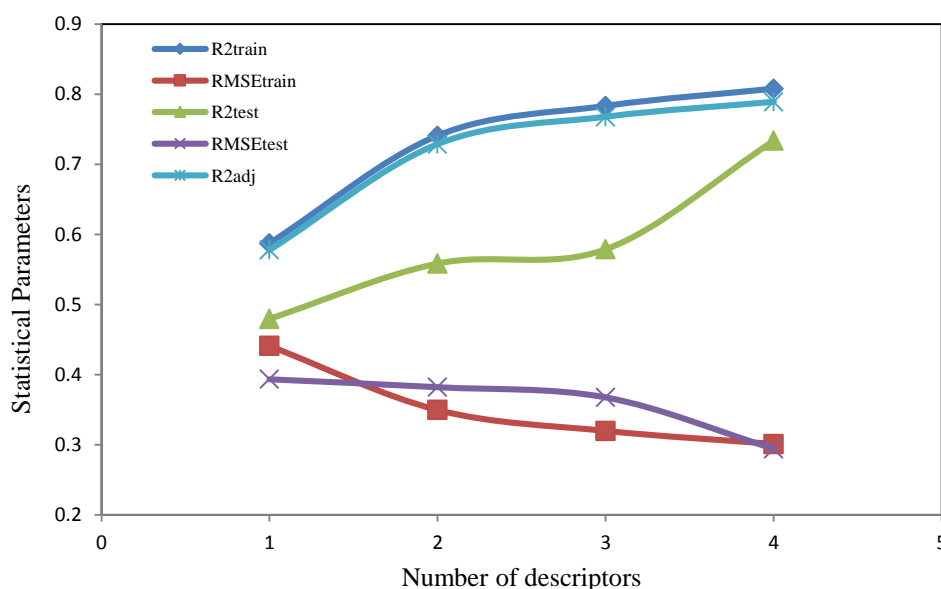
1. HyperChem
2. Dragon

در اولین قدم باید مجموعه آزمایشی (مشتقات فنل و آنیلین) و ساختارهای متناظر جمع آوری شده و به صورت قابل پردازش برای کامپیوتر درآید. در مورد داده های آزمایشی که مقادیر عددی هستند مشکلی نداریم، مساله مهم بیان ساختارهای شیمیایی به شکل قابل پذیرش برای کامپیوتر است. ابتدا ساختارهای مولکولی به کمک نرم افزار هایپرکم رسم و به وسیله روش AM1 بهینه گردید. سپس محاسبه توصیف کننده ها به کمک این نرم افزار و نرم افزار دراگون انجام شد.

۳. نتایج و بحث

۳-۱. مدل سازی با روش رگرسیون خطی چندگانه

برای ساختن مدلی که بیانگر ارتباط ساختاری ترکیبات مورد بررسی با EC_{50} آنها باشد، از روش رگرسیون خطی چندگانه (MLR) استفاده شد. در ابتدا توصیف کننده های محاسبه شده، به عنوان متغیرهای مستقل و مقادیر EC_{50} ترکیبات مورد نظر به عنوان متغیرهای وابسته، به عنوان ورودی به نرم افزار SPSS وارد شدند. در نهایت با استفاده از منوی آنالیز، گزینه ی رگرسیون خطی و روش مرحله ای انتخاب و نهایتاً چندین مدل مختلف بطور جداگانه به دست آمد، که با توجه به خصوصیات آماری آنها از جمله ضریب رگرسیون (R)، آماره F و خطای استاندارد و پس از رسم مقادیر R^2_{train} ، R^2_{test} ، $RMSE_{train}$ و $RMSE_{test}$ بر حسب تعداد توصیف کننده ها، بهترین مدل که دارای بیشترین مقدار R^2 و F کم ترین مقدار خطای استاندارد و شامل توصیف کننده های تا حد امکان قابل توجیه باشد، به عنوان مدل نهایی برای ارتباط EC_{50} ترکیبات با ساختار آنها انتخاب شد. با این روش چهار مدل بدست آمد که مدل چهارم با تعداد ۴ توصیف کننده به عنوان مناسب ترین آنها انتخاب و توسط روش MLR مدل سازی و مورد ارزیابی قرار گرفت. شکل ۲ تاثیر تعداد توصیف کننده ها بر مقدار پارامترهای آماری نشان می دهد. فهرست توصیف کننده های انتخاب شده توسط نرم افزار SPSS به همراه توصیف مختصری از آنها در جدول ۲ آورده شده است.



شکل ۲. تاثیر تعداد توصیف کننده ها بر پارامترهای آماری

جدول ۲. توصیف کننده های انتخاب شده با SPSS و توصیف آنها.

توصیف کننده	نوع توصیف کننده	ضرایب	MF	VIF
Homo	Mechanic quantum descriptors	0.078	0.086	1.089
MATS1e	Electronic descriptors	2.579	-0.223	1.033
RDF030u	RDF descriptors	0.144	0.098	1.019
L1s	WHIM descriptors	0.627	1.037	1.115
Constant	-	-0.031	-	-

پس از انتخاب مناسب ترین توصیف کننده ها توسط روش مرحله ای با استفاده از SPSS، مرحله بعد، ایجاد مدل بین توصیف کننده های انتخاب شده و EC_{50} ترکیبات است. از نرم افزار SPSS برای این منظور استفاده گردید و بین توصیف کننده ها و EC_{50} ترکیبات سری آموزش با استفاده از روش MLR رابطه زیر به دست آمد:

$$EC_{50} = -0.031 + 0.078 (\text{Homo}) + 2.579 (\text{MATS1e}) + 0.144 (\text{RDF030u}) + 0.627 (\text{L1s})$$

$$R^2_{\text{train}}=0.807, \quad F_{\text{train}}=43.043, \quad R^2_{\text{test}}=0.733, \quad F_{\text{test}}=5.235, \quad R^2_{\text{adj}}=0.788, \quad Q^2_{\text{LOO}}=0.759, \quad Q^2_{\text{LGO}}=0.732$$

سپس از معادله به دست آمده برای پیش بینی EC_{50} سری پیش بینی (تست) استفاده گردید. با توجه به پارامترهای آماری فوق، مشاهده می گردد که قدرت پیش بینی توسط روش رگرسیون خطی چندگانه مطلوب نیست. بنابراین برای حصول نتایج بهتر، از شبکه عصبی مصنوعی برای پیش بینی EC_{50} ترکیبات استفاده شد.

۲-۳. مدل سازی با استفاده از شبکه عصبی مصنوعی

برای این کار، ابتدا یک شبکه عصبی مصنوعی سه لایه با تابع انتقال سیگموئیدی و الگوریتم آموزشی لوب-مارکوات طراحی شد. مقادیر اولیه وزن ها بطور تصادفی در بازه $[-1, 0]$ بوده و قبل از عمل آموزش مقادیر ورودی و خروجی در فاصله $[0.1, 0.9]$ نرمالیز گردید. مجموعه داده ها بطور تصادفی به سه سری آموزش، ارزیابی و پیش بینی (تست) تقسیم گردید. برای آموزش دادن شبکه از سری آموزش، برای ارزیابی شبکه، از سری ارزیابی و برای پیش بینی نتایج از سری پیش بینی استفاده شد. ترکیبات سری آموزش، ارزیابی و پیش بینی در جدول ۳ آورده شده است.

برای شبکه عصبی مصنوعی، ابتدا پارامترهایی مانند تعداد نرون ها در لایه مخفی، سرعت آموزش، مومنتم و تعداد دورها بهینه گردید. تعداد نرون ها در لایه ورودی نیز برابر تعداد توصیف کننده های وارد شده به شبکه عصبی مصنوعی می باشد. برای بهینه سازی تعداد نرون ها در لایه مخفی، تعداد نرون ها از ۱ تا ۱۰ تغییر داده شد و مقدار متوسط مربعات خطا (RMSE) برای مجموعه های آموزش و ارزیابی محاسبه گردید و با توجه به مقدار RMSE برای سری ارزیابی برای ۲۰۰۰ دور، تعداد نرون ها در لایه مخفی بهینه شد. سپس سرعت آموزش، مومنتم و تعداد دورها نیز بهینه گردید. جدول ۴ مشخصات شبکه عصبی مصنوعی بهینه شده را نشان می دهد.

جدول ۴: ساختار و مشخصات شبکه عصبی مصنوعی بهینه شده.

۷	تعداد نرون ها در لایه مخفی
۰/۶	سرعت آموزش
۰/۴	مومنتم
۱۲۰۰	تعداد دورها

شبکه ساخته شده دارای چهار نرون در لایه ورودی (تعداد توصیف کننده ها)، ۷ نرون در لایه مخفی و یک نرون در لایه خروجی می باشد. بنابراین ساختار شبکه عصبی مصنوعی بهینه شده بصورت ۴-۷-۱ می باشد. باید توجه نمود که آموزش شبکه زمانی متوقف می شود که RMSE برای سری ارزیابی شروع به افزایش نماید زیرا Overfitting اتفاق می افتد و این اتفاق، قدرت پیش بینی شبکه را کاهش خواهد داد. برای جلوگیری از Overfitting در طول آموزش مقادیر RMSE بعد از هر ۲۰ سیکل برای سری آموزش و ارزیابی محاسبه و ثبت گردید و بدین طریق شبکه بهینه گردید.

با استفاده از مدل ANN بهینه شده، مقادیر EC_{50} ترکیبات مورد نظر در مجموعه آموزش، ارزیابی و پیش بینی مورد محاسبه قرار گرفت که نتایج آن در جدول ۳ نشان داده شده است.

جدول ۳. مقادیر تجربی و محاسبه شده سمیت (pEC_{50}) برای ترکیبات مختلف برای مجموعه های آموزش، ارزیابی و پیش بینی در مدل SW-ANN همراه با مقادیر خطای نسبی.

No.	Compound	Exp ^a (pEC_{50})	ANN ^b (pEC_{50})	Residual ^c
Training set				
1	Aniline	0.31	0.41	0.10
2	Phenol	0.78	1.10	0.32
3	3-Methylaniline	1.04	0.94	-0.10
4	2,4-Dimethylaniline	1.11	1.21	0.10
5	3,5-Dimethylaniline	1.14	1.04	-0.10
6	2-Chlorophenol	1.16	1.26	0.10
7	2-Ethylaniline	1.24	1.34	0.10
8	2,6-Dimethylphenol	1.26	1.33	0.07
9	2-Chloroaniline	1.37	1.27	-0.10
10	3-Ethylaniline	1.38	1.63	0.25
11	3,5-Dimethylphenol	1.38	1.48	0.10
12	2-Ethylphenol	1.4	1.50	0.10
13	3,5-Dichloroaniline	1.45	1.35	-0.10
14	2,4,6-Trimethylphenol	1.45	1.55	0.10
15	3-Chlorophenol	1.49	1.59	0.10
16	2,3,6-Trimethylphenol	1.77	1.87	0.10
17	2,4-Dichloroaniline	1.78	1.86	0.08
18	2,5-Dichloroaniline	1.78	1.88	0.10
19	2,3-Dichloroaniline	1.85	1.71	-0.14
20	2,3-Chlorophenol	1.87	1.97	0.10
21	2,6-Dichloroaniline	1.89	1.79	-0.10

22	2,3,4-Trichloroaniline	1.98	2.00	0.02
23	2,5-Chlorophenol	1.99	2.09	0.10
24	4-Chlorophenol	2.01	2.11	0.10
25	3,4-Dimethylaniline	2.16	1.84	-0.32
26	3-Ethylphenol	2.16	2.26	0.10
27	2,6-Dimethylaniline	2.26	2.20	-0.06
28	3,4-Dichloroaniline	2.35	2.25	-0.10
29	3,4-Chlorophenol	2.44	2.54	0.10
30	2,4,6-Trichlorophenol	2.44	2.54	0.10
31	3,4-Dimethylphenol	2.51	2.41	-0.10
32	4-Ethylaniline	2.83	2.65	-0.18
33	2,3,4-Trichlorophenol	3.05	2.95	-0.10
34	4-Ethylphenol	3.37	3.27	-0.10
35	2,3,5-Trichlorophenol	3.43	2.90	-0.53
Validation set				
1	2,3--Dimethylaniline	0.93	1.03	-0.10
2	2,6-Dimethylaniline	1.11	1.01	0.10
3	3-Chloroaniline	1.19	1.29	-0.10
4	2-Methylphenol	1.41	1.31	0.07
5	2,5-Dimethylphenol	1.56	1.63	-0.01
6	2,6-Chlorophenol	1.78	1.77	0.10
7	3,4,5-Trichloroaniline	1.93	2.03	0.10
8	2,4,5-Trichloroaniline	2.1	2.20	-0.10
9	4-Methylphenol	2.33	2.23	-0.30
10	3,5-Chlorophenol	2.58	2.28	-0.10
11	2,4,5-Trichlorophenol	3.25	3.15	-0.10
Test set				
1	2-Methylaniline	0.73	0.80	0.07
2	2,4,6-Trimethylaniline	1.05	1.46	0.41
3	2,5-Dimethylaniline	1.18	1.18	0.00
4	4-Methylaniline	1.38	1.22	-0.16
5	3-Methylphenol	1.44	1.49	0.05
6	2,3,5-Trimethylphenol	1.68	1.75	0.07
7	4-Chloroaniline	1.81	1.92	0.11
8	2,3-Dimethylphenol	1.96	1.67	-0.29
9	2,4-Chlorophenol	2.15	2.29	0.14
10	2,4-Dimethylphenol	2.31	1.87	-0.44
11	2,3,6-Trichlorophenol	2.48	2.57	0.09

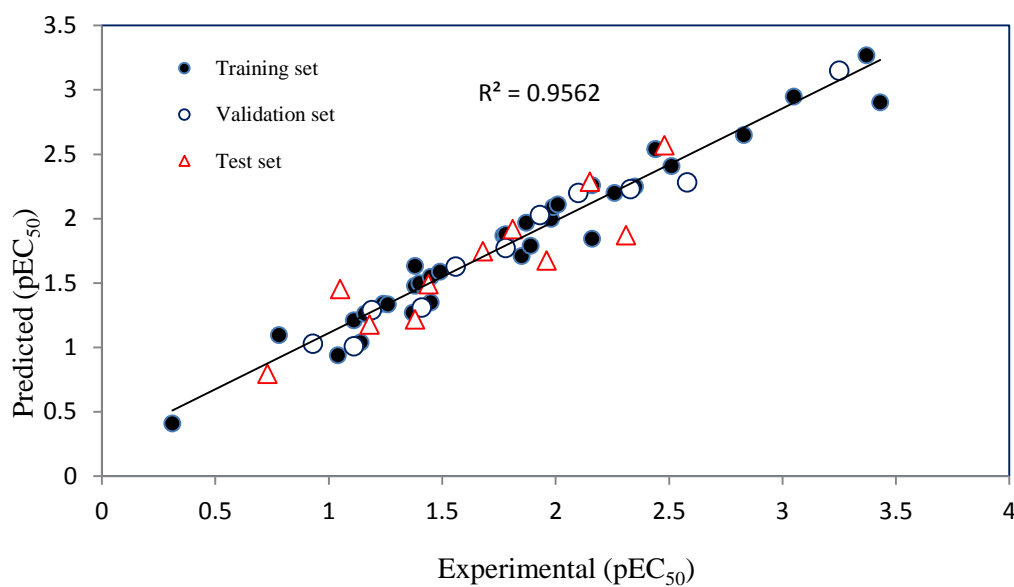
(a) مقادیر تجربی سمیت (pEC₅₀)

(b) مقادیر پیش بینی شده توسط شبکه عصبی مصنوعی

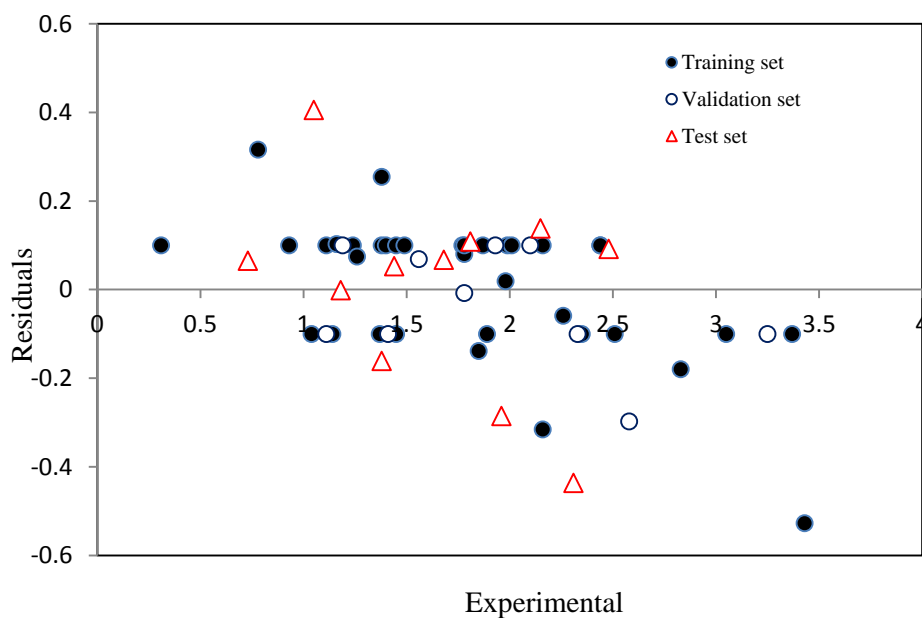
(c) باقیمانده (اختلاف بین مقدار پیش بینی شده و مقدار واقعی)

شکل ۳ مقادیر پیش بینی شده pEC₅₀ ترکیبات مورد نظر توسط شبکه عصبی مصنوعی در مجموعه های آموزش، ارزیابی و پیش بینی را بر حسب مقادیر تجربی نشان می دهد. در این شکل، میزان نزدیکی داده ها به خط راست قدرت پیش بینی مدل را نشان می دهد. شکل ۴ مقادیر باقیمانده

خطاها را نسبت به مقادیر تجربی نشان می دهد. میزان پراکندگی خطاها در اطراف محور، نشان دهنده این است که خطای سیستماتیک در مدل وجود ندارد.



شکل ۳. نمودار مقادیر pEC₅₀ محاسبه شده با کمک مدل SW-ANN برای مجموعه‌های آموزش، ارزیابی و پیش‌بینی بر حسب مقادیر تجربی.



شکل ۴. نمودار تغییرات باقیمانده‌ها بر حسب مقادیر تجربی برای مقادیر pEC₅₀ محاسبه شده بر اساس مدل SW-ANN در سه مجموعه آموزش، ارزیابی و پیش‌بینی.

۴. نتیجه گیری

ارتباط کمی بین ویژگی و ساختار ترکیباتی از مشتقات فنل و آنیلین‌ها با استفاده از روش انتخاب متغیر به صورت مرحله‌ای-رگرسیون خطی چندگانه (SW-MLR) و شبکه عصبی مصنوعی جهت پیش‌بینی میزان سمیت (pEC₅₀) مورد مطالعه قرار گرفت. در ابتدا توصیف کننده

های مولکولی با کمک نرم افزار هایپرکم و دراگون محاسبه و توصیف کننده های مهم با استفاده از روش مرحله ای انتخاب و با دو روش MLR و ANN مدل سازی و پیش بینی انجام شد. پارامترهای آماری نشان می دهد که مدل SW-ANN توانایی پیش بینی میزان سمیت (pEC_{50}) مولکول ها را با قدرت و توانایی بالاتری نسبت به روش SW-MLR دارد.

۵. مراجع

- [۱] ا. ملکی، ا.ح. محوی و ک. ندافی، آب و فاضلاب، ۶۶ (۱۳۸۷) ۱۹.
- [2] N.L. Nemerow, *Industrial and hazardous waste treatment*, 2nd Ed., Van Nostrand Reinhol, New York, (1991).
- [3] A.R. Katritzky, V.S. Lobanov, M. Karelson, *Chemical Society Reviews*, 24(4) (1995) 279.
- [4] M. Tareq Hassan Khan, *Current drug metabolism*, 11(4) (2010) 285.
- [5] M.G. Freire, C.M. Neves S.P. Ventura, M.J. Pratas, I.M. Marrucho, J. Oliveira, A.M. Fernandes, *Fluid Phase Equilibria*, 294(1) (2010) 234.
- [6] A.A. Melnikov, V.A. Palyulin, N.S. Zefirov, *J. Chem. Inf. Model.*, 47(2007) 2077.
- [7] F. Gharagheizi, A. Eslamimanesh, A.H. Mohammadi, D. Richon, *Chemical engineering science*, 66(13) (2011) 2959.
- [8] A.R. Katritzky, I.B. Stoyanova-Slavova, D.A. Dobchev, *J Mol Graph Model.*, 26 (2007) 529.
- [9] M. Fernandez, T.K. Woo, C.E. Wilmer, R.Q. Snurr, *The Journal of Physical Chemistry C*, 117(15) (2013) 7681.
- [10] M. Shamsipur, A. Sirouejad, B. Hemmateenejad, A. Abbaspour, H. Sharghi, K. Alizadeh, S. Arshadi, *J. Electroanal Chem.*, 600 (2007) 345.
- [۱۱] م. منهاج، مبانی شبکه های عصبی، انتشارات دانشگاه صنعتی امیرکبیر، ایران، ۷۱۵ (۱۳۷۹).
- [12] J. Xu, L. Wang, L. Wang, X. Shen, W. Xu, *Journal of computational chemistry*, 32(15) (2011) 3241.
- [13] D.T. Pham and X. Liu, *Neural networks for identification. predication and control*, Springer-Velag London Limited. (1999).
- [14] A.J. Wilczyńska-Piliszek, S. Piliszek, J. Falandysz, *Journal of Environmental Science and Health, Part B*, 47(2) (2012) 111.
- [15] S. Piliszek, A.J. Wilczyńska-Piliszek, J. Falandysz, *Journal of Environmental Science and Health, Part A*, 47(3) (2012) 450.
- [16] M.C. Guimarães, E.G. da Mota, D.G. Silva, M.P. Freitas, *Chemometrics and Intelligent Laboratory Systems*, 134 (2014) 53.