مجله شیمی کوانتومی و اسپکتروسکوپی
۱۰:۳۳ (۱۳۹۹) ۴۹-۶۳
Journal of Quantum Chemistry and Spectroscopy (JQCS)

مدل سازی و مطالعه ارتباط کمی ساختار-خاصیت جهت پیش بینی نیمه عمر بی فنیل های پلی کلرینه با استفاده از رگرسیون خطی چند متغیره و شبکه های عصبی مصنوعی

سکینه بهرامی نسب[۱*]، مهدی نکوئی[۲]، سیدعباس طاهری[۲]

[۱]گروه علم اطلاعات و دانش شناسی، دانشگاه شهید بهشتی، تهران، ایران

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

تاریخ ثبت اولیه:۱۳۹۹/۰۳/۱۲، تاریخ دریافت نسخه اصلاح شده:۱۳۹۹/۰۶/۱۹ ، تاریخ پذیرش قطعی:۱۳۹۹/۰۷/۲۸

چکیده

مطالعه ارتباط کمی ساختار-خاصیت(QSPR) جهت پیش بینی زمان نیمه عمر برخی مشتقات بی فنیل های پلی کلرینه با استفاده از روشهای رگرسیون خطی چند متغیره(MLR) و شبکه های عصبی مصنوعی(ANN) انجام شد. در ابتدا ساختار ترکیبات، رسم و گروه مناسبی از توصیف کننده‌ها محاسبه شدند. سپس از روش انتخاب مرحله‌ای برای بدست آوردن بهترین توصیف کننده‌ها که بیشترین ارتباط را با نیمه عمر ترکیبات مورد نظر داشتند استفاده گردید. با این روش ۶ توصیف کننده شامل Lop, GATS5m, GATS8m, LDip, RDF020u, R2v+ که از انواع توصیف کننده های توپولوژیکی، بار، نمایش سه بعدی مولکول بر اساس پراش الکترونی و تابع توزیع شعاعی هستند انتخاب گردید. در ابتدا مدل خطی MLR ساخته شد. سپس برای به دست آوردن نتایج بهتر از شبکه عصبی مصنوعی استفاده گردید. مقادیر ضریب تعیین (R^2) و ریشه میانگین مربعات خطا (RMSE) برای سری تست به ترتیب برابر ۰/۷۱۶ و ۰/۰۵۰ برای مدل خطی MLR و ۰/۸۹۶ و ۰/۰۳۰ برای مدل غیرخطی ANN بدست آمد. داده های آماری، برتری روش ANN را نسبت به روش MLR نشان می دهد.

واژه های کلیدی: ارتباط کمی ساختار- خاصیت، بی فنیل های پلی کلرینه ، زمان نیمه عمر ، رگرسیون خطی چند گانه، شبکه عصبی مصنوعی

۱. مقدمه

بی فنیل های پلی کلرینه (Polychlorinated biphenyls-PCBs) یک خانواده از ترکیبات سمی و خطرناک برای موجودات و محیط زیست به شمار می آیند. این ترکیبات دارای یک تا ۱۰ اتم کلر بوده که بر اساس تعداد و جایگاه اتمهای کلر دارای ۲۰۹ عضو می باشند. استخلاف های کلر بر حلقه های بی فنیل تعیین کننده میزان سمیت می باشند، به گونه‌ای که هرچه تعداد اتمهای کلر بیشتر و جایگاه‌های پارا و متا بیشتر اشغال شده باشند، ترکیب خطرناک تر خواهد بود. به دلیل ویژگی های منحصر به فردی چون

***عهده دار مکاتبات:** سکینه بهرامی نسب

نشانی: گروه علم اطلاعات و دانش شناسی، دانشگاه شهید بهشتی، تهران، ایران

تلفن:۰۲۳۳۲۳۹۴۲۸۹ **پست الکترونیک:**E-mail:S.bahraminassab@gmail.com

ظرفیت حرارتی بالا این ترکیبات به عنوان روان کننده های هیدرولیکی و کاهنده های حرارتی در صنایع مختلف به ویژه صنایع الکترونیک استفاده می‌شوند. لازم به ذکر است که بی فنیل های پلی کلرینه، آبگریز بوده و این خاصیت موجب تجمع بیولوژیکی آن ها در بدن موجودات زنده می گردد[۲-۱]. هرچند که با کشف خاصیت بیماری‌زایی بی فنیل های پلی کلرینه، تولید و مصرف آنها از اواخر دهه ۸۰ میلادی در بسیاری از کشورها متوقف شد اما به دلیل راهیابی این ترکیبات از کارخانجات تولیدکننده و صنایع مصرف کننده به محیط زیست و از طرف دیگر زیست تجزیه پذیری محدود این مواد مقاوم همچنان در محیط زیست علی الخصوص در محیط های آبی و خاک وجود داشته و تجمیع شده‌اند. بسته به نوع ترکیب، جذب بر روی ذرات معلق در آتمسفر، تجمع در رسوبات دریایی و خاک، همچنین انتقال به آبهای سطحی و سفره های زیرزمینی می تواند اشکال مختلف حضور و سرنوشت محتوم این ترکیبات زیانبار در محیط باشد. تحقیقات سالهای اخیر نشان می‌دهد که بی فنیل های پلی کلرینه قادر به ایجاد سرطان زایی به ویژه در زنان می باشد. این ترکیبات در بافت‌های چربی تجمع یافته و در طی زنجیره غذایی به جانداران اکوسیستم های مختلف و جوامع متعدد از جمله جوامع انسانی وارد می شوند. این ترکیبات کلردار از طریق مکانیسم های مختلفی از جمله ایجاد استرس اکسیداتیو و تولید گونه‌های واکنش‌پذیر همچنین تغییر در بیان ژن‌ها منجر به ایجاد توده‌های سرطانی می گردند. علاوه بر سرطان زایی، آسیب های مغزی- عصبی از دیگر آثار زیانبار مواجهه با چنین ترکیباتی است. مطالعات نشان می‌دهد که این ترکیبات می‌توانند روند ابتلا به بیماری‌هایی چون پارکینسون را تسریع نماید. نظر به آنچه گفته شد امروزه شناسایی و بررسی این ترکیبات یکی از مسائل قابل توجه در علوم بهداشتی و زیست محیطی می باشد[۵-۳]. یکی از راههای بررسی پایداری این ترکیبات در طبیعت، اندازه گیری میزان نیمه عمر این ترکیبات است. نیمه عمر[1]($t_{1/2}$) یک ترکیب مدت زمانی است که طول می کشد تا آن ترکیب به نصف مقدار اولیه خودش تجزیه شود. نیمه عمر هر ترکیب نشان می دهد که آن ترکیب تا چه اندازه پایدار است. هرچه نیمه عمر یک ترکیب بیشتر باشد نشان دهنده پایداری بیشتر آن ترکیب در طبیعت است. بنابراین اندازه گیری نیمه عمر ترکیبات بسیار حائز اهمیت می باشد. اما از آنجاییکه این بررسیها و اندازه گیری‌ها معمولا وقت گیر و هزینه بر می باشند استفاده از روشهایی برای پیش بینی آنها ضروری به نظر می رسد. برای پیش بینی خواص ترکیبات مختلف، ارتباط کمی ساختار - خاصیت[2] (QSPR) روش مطمئن و مناسبی بوده که امروزه بطور وسیع از آن استفاده می گردد[۱۰-۶]. QSPR یک رابطه آماری و ریاضی است که ساختار ترکیبات را به خاصیت مورد مطالعه مرتبط می سازد. روشهای مختلفی از جمله رگرسیون خطی چندگانه(MLR)، کمترین مربعات جزیی(PLS)، شبکه های عصبی مصنوعی(ANN) و ماشین بردار پشتیبان(SVM) در مدلسازی های QSPR مورد استفاده قرار گرفته است[۱۵-۱۱].

هدف از انجام این تحقیق پیش بینی زمان نیمه عمر دسته ایی از بی فنیل های پلی کلرینه با استفاده از روش های رگرسیون خطی چندگانه(MLR) و شبکه های عصبی مصنوعی(ANN) می باشد.

[1] Half-life ($t_{1/2}$)
[2] Quantitative structure property relationships (QSPRs)

۲. روش‌های محاسباتی

۲-۱. انتخاب سری داده‌ها

در این کار تعداد۶۲ ترکیب از مشتقات بی فنیل های پلی کلرینه مورد بررسی قرار گرفت[۱۶]. در ابتدا این ترکیبات به صورت تصادفی به دو گروه سری آموزش و سری پیش‌بینی تقسیم شده است، سری آموزش شامل ۵۰ مولکول(۸۰٪ داده ها) و سری پیش‌بینی شامل ۱۲ مولکول (۲۰٪ داده ها) می‌باشد. لگاریتم مقادیر زمان های نیمه عمر ($log\, t_{1/2}$) به عنوان متغیر وابسته و توصیف کننده‌ها به عنوان متغیر مستقل انتخاب شدند. سری آموزش جهت ایجاد یک مدل مناسب و سری پیش‌بینی جهت ارزیابی مدل مورد استفاده قرار گرفت. در شبکه عصبی مصنوعی علاوه بر سری آموزش و تست از سری ارزیابی نیز استفاده گردید.

۲-۲. محاسبه توصیف‌کننده‌ها، غربالگری و گزینش بهترین آنها

توصیف کننده‌ها مقادیر عددی هستند که خصوصیات مختلفی از مولکول را بیان می‌کنند. در حال حاضر بیش از ۵۰۰۰ توصیف کننده مولکولی وجود دارد که بعد از کاهش توصیف کننده های غیرمفید و سپس ارزیابی و پیدا کردن مناسب‌ترین آنها می‌توانند جهت پیش‌بینی خاصیت مورد نظر مورد استفاده قرار گیرند.

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 بدهند، قبل از ورود متغیر جدید از مدل خارج می‌شود. به این ترتیب داده‌های $log\, t_{1/2}$ به عنوان متغیر وابسته و توصیفگرها به عنوان متغیر مستقل در نظر گرفته شده و تکنیک رگرسیون مرحله‌ای انجام شد. همانطور که می‌دانیم روش رگرسیون مرحله ایی تعداد زیادی مدل ارائه می‌کند. که مدل اول شامل یک توصیفگر، مدل دوم شامل دو توصیفگر و ... می‌باشد. با افزایش تعداد توصیفگرها بالطبع مقدار R^2 افزایش و RMSE[2](خطای جذر میانگین مربعات)کاهش می‌یابد. اما بدلیل پیچیدگی مدل، نمی‌توانیم تعداد زیادی توصیفگر را جهت مدلسازی انتخاب کنیم. بدین منظور و جهت انتخاب تعداد توصیفگرهای مناسب، نمودار پارامترهای مختلف آماری از جمله $RMSE_{train}$, $RMSE_{test}$،R^2_{train}, R^2_{test} برحسب تعداد توصیفگرها رسم و بر طبق آن، تعداد ۶ توصیفگر به عنوان توصیفگرهایی که بیشترین ارتباط را با $log\, t_{1/2}$ ترکیبات دارند، انتخاب شدند. این ۶ توصیفگر به همراه طبقه آنها در جدول۱ ارائه شده است.

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

جدول ۱. توصیفگرهای انتخاب شده توسط رگرسیون خطی چندگانه مرحله به مرحله

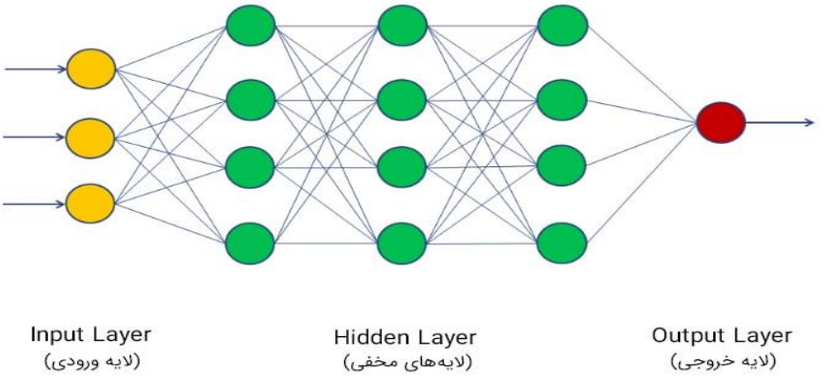
نشانه توصیف کننده	معنی توصیف کننده	نوع توصیف کننده
Lop	Lopping centric index.	Topological
GATS5m	Geary autocorrelation-lag 5/weighted by atomic masses.	2D autocorrelations
GATS8m	Geary autocorrelation-lag 8/weighted by atomic masses.	2D autocorrelations
LDip	Local dipole index.	Charge
RDF020u	Radial Distribution Function-2.0/unweighted.	RDF
$R2v^+$	R maximal autocorrelation of lag 2/weighted by atomic van der Waals volumes.	GETAWAY

۲-۳. شبکه های عصبی مصنوعی

شبکه‌های عصبی مصنوعی مجموعه‌ای از نورون‌ها هستند که از الگوریتم‌های منحصر به فردی پیروی می‌کنند. این مجموعه که از مغز انسان الگوبرداری و الهام گرفته‌شده‌است، با هدف شناسایی الگوها طراحی می‌شوند و مورد استفاده قرار می‌گیرند. به طور کلی می‌توان گفت که شبکه عصبی شامل الگوریتم‌هایی‌ست برای یادگیریِ ماشین، که منجر به طبقه‌بندی کردن داده‌های ورودی و ارائه خروجی مطلوب می‌گردد. به همین دلیل است که می‌توان شبکه‌های عصبی را به عنوان جزئی از فرایند یادگیری ماشین در نظر گرفت.

شبکه‌های عصبی مصنوعی، سیستم‌ها و روش‌های محاسباتی نوین برای یادگیری، نمایش دانش و در انتها اعمال دانش به دست آمده در جهت بیش‌بینی 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 با اعمال این الگوریتم به شبکه، آن را آموزش داد. شکل ۱ نمایی از یک شبکه عصبی مصنوعی را نشان می دهد[۲۰-۱۷].



شکل ۱. نمایی از یک شبکه عصبی مصنوعی

یکی از پایه‌ای‌ترین مدل‌های عصبی موجود، مدل پرسپترون چند لایه[1] است که عملکرد انتقالی مغز انسان را شبیه‌سازی می‌کند. در این نوع شبکه عصبی، بیشتر رفتار شبکه‌ای مغز انسان و انتشار سیگنال در آن مد نظر بوده است و از این رو، گهگاه با نام شبکه‌های پیش‌خورد[2] نیز خوانده می‌شوند. هر یک از سلول‌های عصبی مغز انسان، موسوم به نورون[3]، پس از دریافت ورودی (از یک سلول عصبی یا غیر عصبی دیگر)، پردازشی روی آن انجام می‌دهند و نتیجه را به یک سلول دیگر (عصبی یا غیر عصبی) انتقال می‌دهند. این رفتار تا حصول نتیجه‌ای مشخص ادامه دارد، که احتمالاً در نهایت منجر به یک تصمیم، پردازش، تفکر و یا حرکت خواهد شد.

۳. نتایج و بحث

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

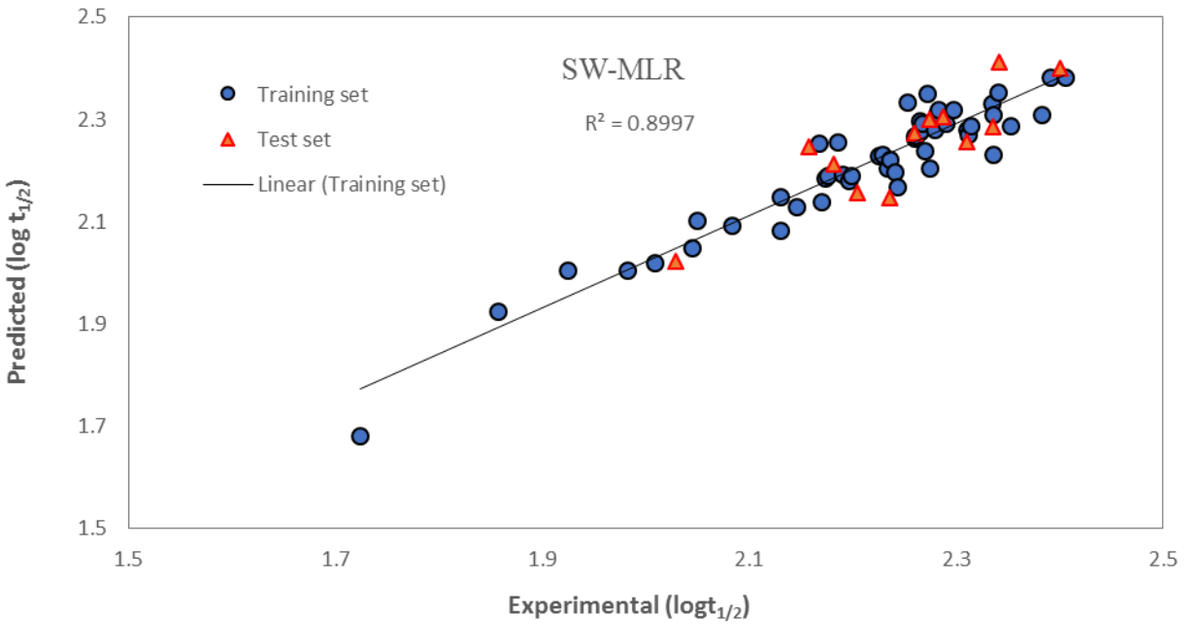
پس از انتخاب مناسب‌ترین توصیف‌کننده‌ها توسط روش مرحله‌ای، مرحله بعدی، ایجاد مدل میان توصیف‌کننده‌های انتخاب شده و $log\, t_{1/2}$ می‌باشد. بین توصیف کننده‌ها و $log\, t_{1/2}$ ترکیبات برای سری آموزش با استفاده از روش MLR رابطه زیر به عنوان مدل خطی بدست آمد:

$$log\, t_{1/2} = -1.167 + 4.509\,(Lop) - 24.172\,(GATS5m) + 8.327\,(GATS8m) - 0.848\,(LDip) + 0.043\,(RDF020u) + 1.378\,(R2v+)$$

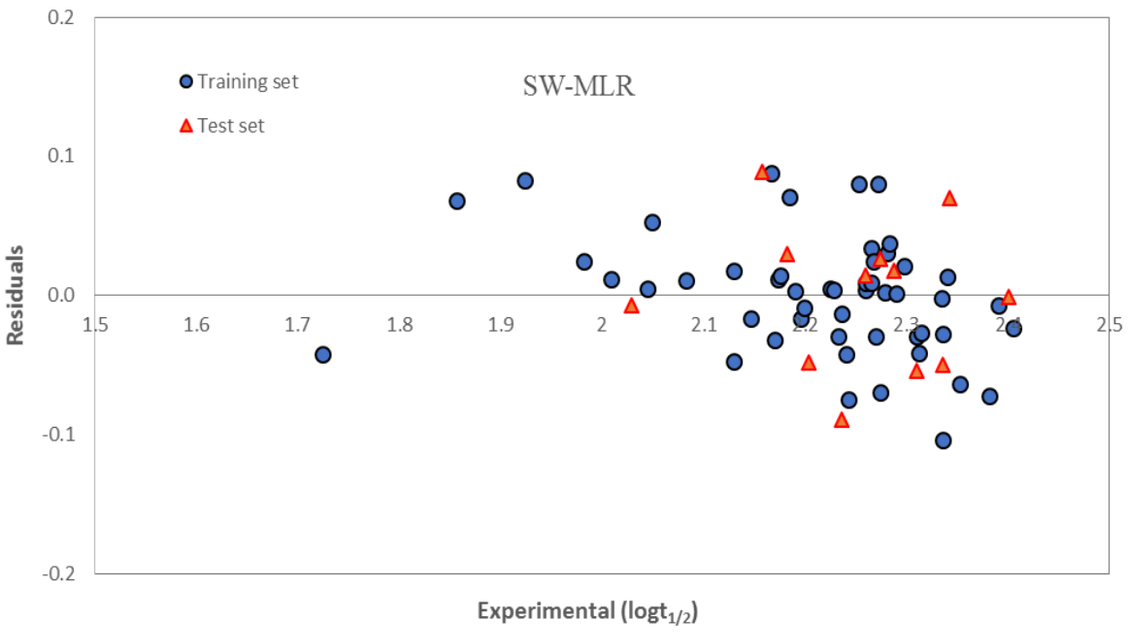
سپس از معادله بدست آمده برای پیش‌بینی $log\, t_{1/2}$ ترکیبات سری تست استفاده گردید. مقادیر تجربی و پیش‌بینی شده $log\, t_{1/2}$ برای کلیه ترکیبات مجموعه آموزش و تست در جدول (۲)آورده شده است. شکل(۲) نمودار مقادیر پیش بینی شده بر

[1] Multi-Layer perceptron (MLP)
[2] Feedforward Networks
[3] Neuron

حسب مقادیر تجربی را نشان می دهد. شکل(۳) 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.



شکل ۲. نمودار مقادیر پیش‌بینی شده $log\ t_{1/2}$ بر حسب مقادیر تجربی برای سری آموزش و تست به روش SW-MLR



شکل۳. نمودار تغییرات باقیمانده ها برحسب مقادیر تجربی برای مقادیر $log\ t_{1/2}$ براساس مدل SW-MLR در دو مجموعه آموزشی و تست

جدول۲. مقادیر تجربی و پیش بینی شده بی فنیل های پلی کلرینه توسط روش MLR و ANN

شماره	نام ترکیبات بی فنیل های پلی کلرینه	log $t_{1/2}$ (Exp)	log $t_{1/2}$ (MLR)	log $t_{1/2}$ (ANN)
سری آموزش				
۱	2,3'-Dichlorobiphenyl	۱/۷۲۴	۱/۶۸۲	۱/۷۲۹
۲	2,2',5- Trichlorobiphenyl	۱/۹۲۸	۲/۰۰۶	۱/۹۹۲
۳	2,2',6- Trichlorobiphenyl	۱/۹۲۴	۲/۰۰۶	۱/۰۰۳
۴	2,4',6- Trichlorobiphenyl	۲/۰۴۵	۲/۰۵۰	۲/۰۴۸
۵	2,2',3,4'- Tetrachlorobiphenyl	۲/۱۴۶	۲/۱۳۰	۲/۱۳۹
۶	2,2',3,5- Tetrachlorobiphenyl	۲/۲۷۴	۲/۲۰۴	۲/۲۷۱
۷	2,2',3,5'- Tetrachlorobiphenyl	۲/۲۴۳	۲/۱۶۹	۲/۲۱۳
۸	2,2',3,6- Tetrachlorobiphenyl	۲/۰۸۳	۲/۰۹۴	۲/۰۹۴
۹	2,2,3,6'- Tetrachlorobiphenyl	۲/۱۹۶	۲/۱۸۰	۲/۱۹۳
۱۰	2,3,3',4- Tetrachlorobiphenyl	۲/۱۳	۲/۱۴۸	۲/۱۳۸
۱۱	2,3,3',6- Tetrachlorobiphenyl	۲/۳۳۶	۲/۳۰۸	۲/۳۳۱
۱۲	2,3,4',6- Tetrachlorobiphenyl	۲/۲۹	۲/۲۹۲	۲/۲۸۳
۱۳	2,3',4',5- Tetrachlorobiphenyl	۲/۳۹۱	۲/۳۸۴	۲/۳۹۶
۱۴	2,2',3,3',5- Pentachlorobiphenyl	۲/۳۱	۲/۲۸۱	۲/۳۴
۱۵	2,2',3,4,5'- Pentachlorobiphenyl	۲/۳۵۲	۲/۲۸۸	۲/۳۵۱
۱۶	2,2',3,4',5'- Pentachlorobiphenyl	۲/۳۱۲	۲/۲۷۱	۲/۲۹۹
۱۷	2,2',4,4',5- Pentachlorobiphenyl	۲/۱۶۷	۲/۲۵۵	۲/۱۸۶
۱۸	2,2'4,5,5'- Pentachlorobiphenyl	۲/۲۸۳	۲/۳۲۱	۲/۲۸۴
۱۹	2,3,3'4,4'- Pentachlorobiphenyl	۲/۴۰۵	۲/۳۸۲	۲/۴۱۸
۲۰	2,3'4,4',5- Pentachlorobiphenyl	۲/۲۶۵	۲/۲۹۹	۲/۲۶۱

۲۱	2,2',3,3',4,4'- Hexachlorobiphenyl	۲/۳۳۴	۲/۳۳۲	۲/۳۳۱
۲۲	2,2',3,3',4,6'- Hexachlorobiphenyl	۲/۱۸۵	۲/۲۶۵	۲/۱۹۶
۲۳	2,2',3,3',5,6'- Hexachlorobiphenyl	۲/۲۶	۲/۲۶۴	۲/۲۵۷
۲۴	2,2',3,3',6,6'- Hexachlorobiphenyl	۲/۲۵۳	۲/۳۳۳	۲/۲۵۷
۲۵	2,2',3,4,4',5'- Hexachlorobiphenyl	۲/۲۲۸	۲/۲۳۲	۲/۲۳۱
۲۶	2,2',3,4,5',6- Hexachlorobiphenyl	۲/۲۷	۲/۲۴۰	۲/۲۷۶
۲۷	2,2',3,4',5,5'- Hexachlorobiphenyl	۲/۳۱۴	۲/۲۸۷	۲/۳۱۸
۲۸	2,3',4,4',5,5'- Hexachlorobiphenyl	۲/۲۷۹	۲/۲۸۲	۲/۲۷۳
۲۹	2,2',3,3',4,4',6- Heptachlorobiphenyl	۲/۲۳۳	۲/۲۰۴	۲/۲۱۴
۳۰	2,2',3,3'4,5,5'- Heptachlorobiphenyl	۲/۲۴۱	۲/۱۹۹	۲/۲۶۱
۳۱	2,2',3,3',4,5,6'- Heptachlorobiphenyl	۲/۲۳۶	۲/۲۲۳	۲/۲۳۴
۳۲	2,2',3,3',4,5',6- Heptachlorobiphenyl	۲/۲۲۵	۲/۲۳۰	۲/۲۱۶
۳۳	2,2',3,3',4,5',6'-Heptachlorobiphenyl	۲/۱۷	۲/۱۳۸	۲/۱۵۴
۳۴	2,2',3,4,4',5,6'- Heptachlorobiphenyl	۲/۱۹	۲/۱۹۴	۲/۱۹
۳۵	2,2',3,4,4',5',6- Heptachlorobiphenyl	۱/۸۵۷	۱/۹۲۶	۲/۸۱۹
۳۶	2,2,3,4',5,5',6- Heptachlorobiphenyl	۲/۰۰۹	۲/۰۲۰	۲/۰۰۵
۳۷	2,3,3',4,4',5,5'- Heptachlorobiphenyl	۲/۱۳	۲/۰۸۲	۲/۱۲۵
۳۸	2,2',3,3',4,4',5,5'-Octachlorobiphenyl	۲/۰۴۹	۲/۱۰۲	۲/۰۴۲
۳۹	2,2',3,3'4,5',6,6'- Octachlorobiphenyl	۲/۱۷۶	۲/۱۹۰	۲/۱۷۳
۴۰	2,3,3',4,4',5,5',6- Octachlorobiphenyl	۲/۳۳۶	۲/۲۳۲	۲/۳۳۹
۴۱	2,2',3,3',4,4',5,5',6-Nonachlorobiphenyl	۲/۲۷۲	۲/۳۵۲	۲/۲۶۷
۴۲	2,2',3,3',4,4',5,6,6'- Nonachlorobiphenyl	۲/۲۸۱	۲/۳۱۱	۲/۲۸۶
۴۳	2,2',3,3',4,5,5',6,6'- Nonachlorobiphenyl	۲/۳۴	۲/۳۵۳	۲/۳۴۸

۴۴	2,2',3- Trichlorobiphenyl	۲/۲۶	۲/۲۶۹	۲/۲۶۴
۴۵	2,3',4- Trichlorobiphenyl	۲/۳۸۲	۲/۳۱۰	۲/۳۸
۴۶	2,3',5- Trichlorobiphenyl	۲/۲۹۷	۲/۳۱۹	۲/۲۹۹
۴۷	2,2',3,3'- Tetrachlorobiphenyl	۲/۲۶۵	۲/۲۷۵	۲/۲۵۸
۴۸	2,2',4,6'- Tetrachlorobiphenyl	۲/۱۹۹	۲/۱۹۰	۲/۲۰۱
۴۹	2,3,4',5- Tetrachlorobiphenyl	۲/۲۶۷	۲/۲۹۲	۲/۲۶۹
۵۰	2,4,4',5- Tetrachlorobiphenyl	۲/۱۷۳	۲/۱۸۵	۲/۱۸۶
سری تست				
۵۱	2,2',3,4',6- Pentachlorobiphenyl	۲/۰۲۹	۲/۰۲۲	۲/۰۰۶
۵۲	2,3,3',4',6- Pentachlorobiphenyl	۲/۱۵۸	۲/۲۴۷	۲/۱۷۶
۵۳	2,3,4,4',5- Pentachlorobiphenyl	۲/۱۸۲	۲/۲۱۲	۲/۱۲۶
۵۴	2,2',3,4,4',5- Hexachlorobiphenyl	۲/۲۰۴	۲/۱۵۶	۲/۲۰۱
۵۵	2,2',4,4',5,5'- Hexachlorobiphenyl	۲/۲۳۶	۲/۱۴	۲/۱۷۹
۵۶	2,3,3',4,4',5- Hexachlorobiphenyl	۲/۲۶	۲/۲۷۴	۲/۲۸۷
۵۷	2,3,3',4,4',6- Hexachlorobiphenyl	۲/۴۰۱	۲/۴۰۰	۲/۴۰۱
۵۸	2,3,3',4',5,6- Hexachlorobiphenyl	۲/۳۴۲	۲/۴۱۲	۲/۳۴۸
۵۹	2,2',3,4,5,5',6- Heptachlorobiphenyl	۲/۳۳۶	۲/۲۸۶	۲/۳۳۳
۶۰	2,2',3,3',4,4',5,6- Octachlorobiphenyl	۲/۳۱	۲/۲۵۶	۲/۳۱۶
۶۱	2,2',3,3',4,5,5',6- Octachlorobiphenyl	۲/۲۸۸	۲/۳۰۶	۲/۲۹۴
۶۲	2,2',3,3',4,4',5,5',6,6'-Octachlorobiphenyl	۲/۲۷۴	۲/۳۰۰	۲/۲۱۸

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

در قسمت دوم این کار، برای حصول نتایج بهتر، توصیف کننده های انتخاب شده توسط روش مرحله ایی، به شبکه عصبی مصنوعی وارد شدند. پردازش داده‌ها در محیط ویندوز ۱۰ و با استفاده از نرم افزار MATLAB انجام شد. یک شبکه سه لایه با تابع

انتقال سیگموئیدی برای نرون ها طراحی شد. 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 داده ها 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 (۳۰٪ داده ها) برای ارزیابی مدل در طی آموزش دادن شبکه و ایجاد مدل مناسب و مجموعه پیش‌بینی (۲۰٪ داده ها) برای تست مدل ایجاد شده به کار رفت.

تعداد نرون‌ها در لایه ورودی با تعداد توصیف کننده های 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 شدند. 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 ۳ مشخصات شبکه عصبی مصنوعی بهینه شده را نشان می‌دهد.

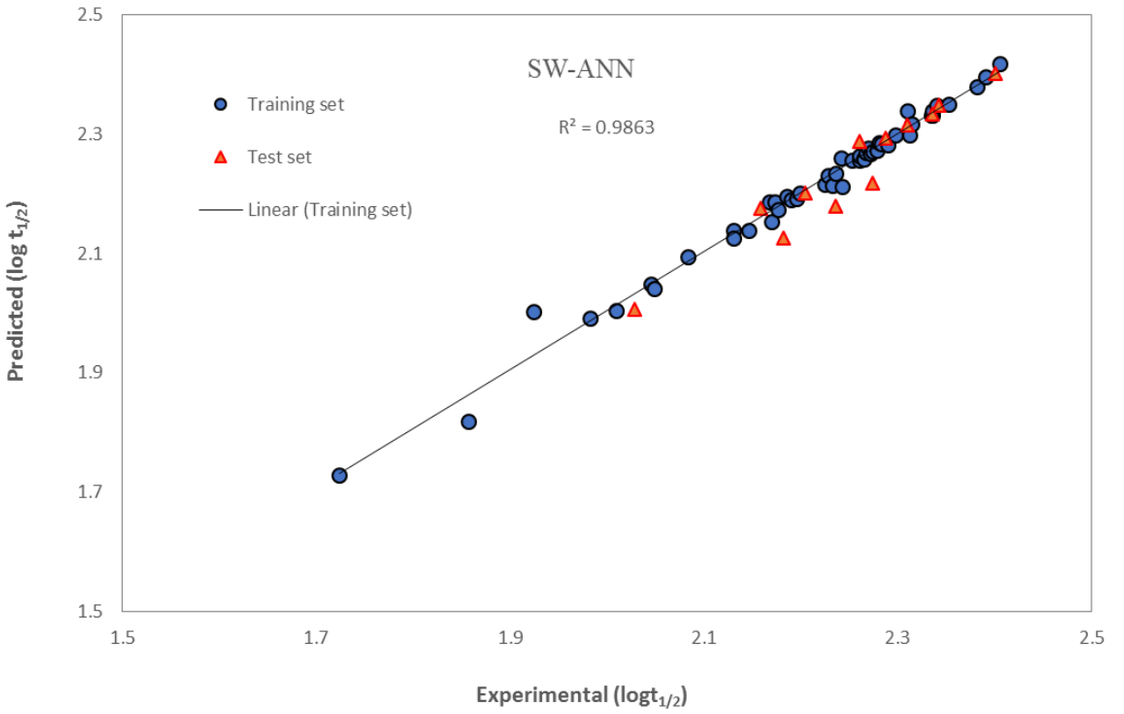
جدول ۳. ساختار و مشخصات ANN تولید شده

No. of nodes in the input layer	۶
No. of nodes in the hidden layer	۵
No. of nodes in the output layer	۱
Learning rate	۰/۴
Momentum	۰/۲
Number of epochs	۱۰۰۰
Transfer function	Sigmoid

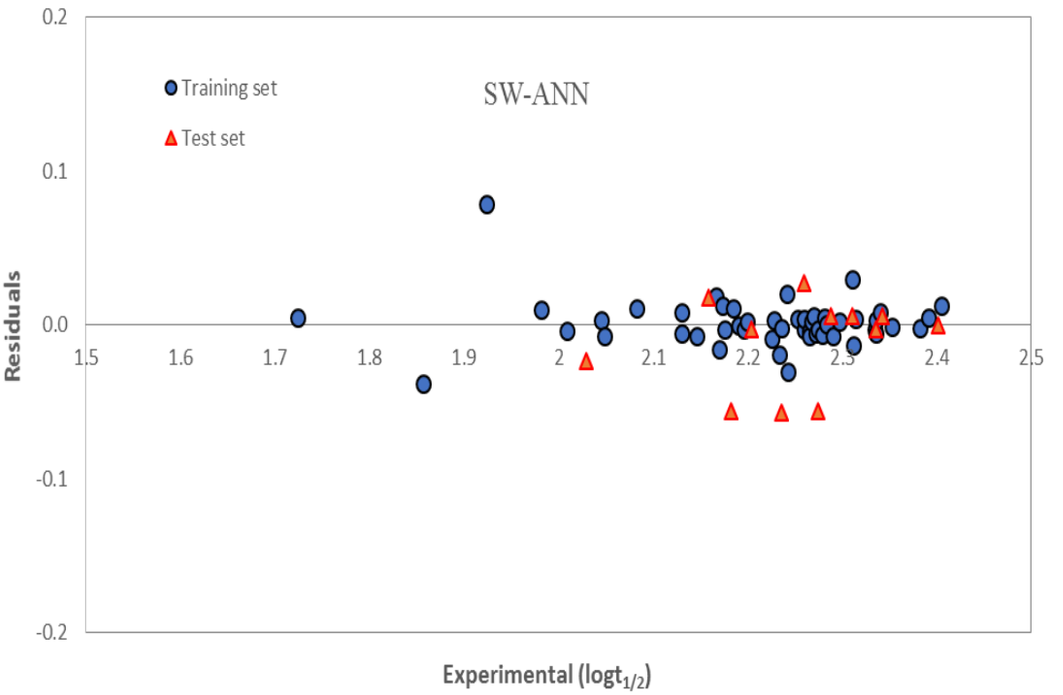
با استفاده از مدل ANN بهینه شده مقادیر $log\ t1_{/2}$ ترکیبات مورد نظر در مجموعه آموزشی، ارزیابی و پیش‌بینی (تست) مورد محاسبه قرار گرفت و در جدول(۲) نشان داده شده است. در شکل(۴) نیز مقادیر محاسبه شده $log\ t1_{/2}$ ترکیبات مورد نظر در مجموعه‌های مختلف برحسب مقادیر تجربی رسم شده‌اند.

به‌منظور بررسی مقبولیت مدل ارائه شده برای مقادیر $log\ t1_{/2}$ محاسبه شده، نمودار تغییرات مقدار خطا برحسب مقادیر تجربی $log\ t1_{/2}$ در شکل(۵) رسم شده است. همانگونه که مشاهده می‌شود توزیع خطا کاملاً تصادفی بوده و مدل ارائه شده از نظر آماری قابل قبول می‌باشد

[1]Back-propagation



شکل ٤. مقادیر $log\ t_{1/2}$ محاسبه شده براساس مدل SW-ANN در دو مجموعه آموزشی و تست برحسب مقادیر تجربی



شکل۵. نمودار تغییرات باقیمانده ها برحسب مقادیر تجربی برای مقادیر $log\ t_{1/2}$ براساس مدل SW-ANN در دو مجموعه آموزشی و تست

۳-۳. ارزیابی مدل با استفاده از پارامترهای آماری

مطابق جدول ۴ چهار پارامتر آماری، جهت ارزیابی توانایی پیش‌بینی مدل های ساخته شده به روش‌های MLR, ANN به کار گرفته شد. همانطور که در این جدول مشاهده می شود تمام پارامترهای آماری برای روش غیرخطی شبکه عصبی مصنوعی بهتر از روش خطی رگرسیون خطی چندگانه است.

جدول ٤. پارامترهای آماری برای مدل‌های انتخاب شده

		SW-MLR	SW-ANN
R^2	سری آموزش	۰/۸۹۹	۰/۹۸۶
	سری تست	۰/۷۱۶	۰/۸۹۶
RMSE	سری آموزش	۰/۰۴۲	۰/۰۱۵
	سری تست	۰/۰۵۰	۰/۰۳۰
REP	سری آموزش	۱/۹۲۷	۰/۷۱۵
	سری تست	۲/۲۴۹	۱/۳۵۸
AARD	سری آموزش	۱/۴۹۷	۰/۴۴۴
	سری تست	۱/۸۴۲	۰/۹۸۱

٤. نتیجه گیری

در این تحقیق از دو روش رگرسیون خطی چندگانه و شبکه های عصبی مصنوعی جهت مدل سازی و پیش‌بینی زمان های نیمه عمر برخی مشتقات بی فنیل های پلی کلرینه استفاده شد. توسط روش مرحله‌ای، ۶ توصیف‌کننده‌ایی که بیشترین ارتباط را با $log\, t_{1/2}$ داشتند انتخاب شدند. جهت مدلسازی از دو روش رگرسیون خطی چندمتغیره(MLR) و شبکه های عصبی مصنوعی(ANN) استفاده گردید. نتایج نشان از برتری روش ANN نسبت به MLR دارد. بنابراین از این روش می توان برای پیش بینی زمان های نیمه عمر ترکیبات مشابه استفاده کرد. یعنی با محاسبه توصیف کننده های یک ترکیب شیمیایی جدید که با نرم افزار دراگون قابل انجام است و وارد کردن آنها به شبکه عصبی بهینه شده می توان مقدار زمان نیمه عمر آن را محاسبه نمود.

٥. مراجع

[1] Zheng, X., Dupuis, K. T., Aly, N. A., Zhou, Y., Smith, F. B., Tang, K., ... & Baker, E. S. (2018). Utilizing ion mobility spectrometry and mass spectrometry for the analysis of polycyclic aromatic hydrocarbons, polychlorinated biphenyls, polybrominated diphenyl ethers and their metabolites. *Analytica chimica acta*, *1037*, 265-273.

[2] Pessah, I. N., Lein, P. J., Seegal, R. F., & Sagiv, S. K. (2019). Neurotoxicity of polychlorinated biphenyls and related organohalogens. *Acta neuropathologica*, *138*(3), 363-387.

[3] Müller, M. H. B., Polder, A., Brynildsrud, O. B., Karimi, M., Lie, E., Manyilizu, W. B., ... & Lyche, J. L. (2017). Organochlorine pesticides (OCPs) and polychlorinated biphenyls (PCBs) in human breast milk and associated health risks to nursing infants in Northern Tanzania. *Environmental research*, *154*, 425-434.

[4] Raffetti, E., Donato, F., Speziani, F., Scarcella, C., Gaia, A., & Magoni, M. (2018). Polychlorinated biphenyls (PCBs) exposure and cardiovascular, endocrine and metabolic diseases: a population-based cohort study in a North Italian highly polluted area. *Environment international*, *120*, 215-222.

[5] Covaci, A., Gheorghe, A., Voorspoels, S., Maervoet, J., Redeker, E. S., Blust, R., & Schepens, P. (2005). Polybrominated diphenyl ethers, polychlorinated biphenyls and organochlorine pesticides in sediment cores from the Western Scheldt river (Belgium): analytical aspects and depth profiles. *Environment international*, *31*(3), 367-375.

[6] Li, M., Yu, H., Wang, Y., Li, J., Ma, G., & Wei, X. (2020). QSPR models for predicting the adsorption capacity for microplastics of polyethylene, polypropylene and polystyrene. *Scientific reports*, *10*(1), 1-11.

[7] Sepehri, B. (2020). A review on created QSPR models for predicting ionic liquids properties and their reliability from chemometric point of view. *Journal of Molecular Liquids*, *297*, 112013.

[8] Borhani, T. N., García-Muñoz, S., Luciani, C. V., Galindo, A., & Adjiman, C. S. (2019). Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. *Physical Chemistry Chemical Physics*, *21*(25), 13706-13720.

[9] Zhu, T., Gu, L., Chen, M., & Sun, F. (2021). Exploring QSPR models for predicting PUF-air partition coefficients of organic compounds with linear and nonlinear approaches. *Chemosphere*, *266*, 128962.

[10] Yan, F., Shi, Y., Wang, Y., Jia, Q., Wang, Q., & Xia, S. (2020). QSPR models for the properties of ionic liquids at variable temperatures based on norm descriptors. *Chemical Engineering Science*, *217*, 115540.

[11] Haiying, Y., Meiping, L., & Junsheng, H. (2020). A Comparative QSPR Study on Photolysis Half-lives of Polychlorinated Dibenzofurans by MLP-ANN and SVM. *Asian Journal of Ecotoxicology*, (4), 240-247.

[12] Moussaoui, M., Laidi, M., Hanini, S., Abdallah, A. E. H., & Hentabli, M. (2021). Critical Properties and Acentric Factors of Pure Compounds Modelling Based on QSPR-SVM with Dragonfly Algorithm. *Kemija u industriji: Časopis kemičara i kemijskih inženjera Hrvatske*, *70*(7-8), 375-386.

[13] Zarei, K., Atabati, M. and Ebrahimi, M., 2007. Quantitative structure-property relationship study of the solvent polarity using wavelet neural networks. *Analytical Sciences*, *23*(8), pp.937-942.

[14] Quang, N. M., Mau, T. X., Nhung, N. T. A., An, T. N. M., & Van Tat, P. (2019). Novel QSPR modeling of stability constants of metal-thiosemicarbazone complexes by hybrid multivariate technique: GA-MLR, GA-SVR and GA-ANN. *Journal of Molecular Structure*, *1195*, 95-109.

[15] Maouz, H., Khaouane, L., Hanini, S., Ammi, Y., Hamadache, M., & Laidi, M. (2020). QSPR studies of carbonyl, hydroxyl, polyene indices, and viscosity average molecular weight of polymers under photostabilization using ANN and MLR approaches. *Kemija u industriji: Časopis kemičara i kemijskih inženjera Hrvatske*, *69*(1-2), 1-16.

[16] Liao, S., Yu, X., Chen, J. and Huang, X., (2019). Prediction of the half-lives of polychlorinated biphenyls based on the IEF-PCM calculations. *Journal of Theoretical and Computational Chemistry*, *18*(07), 1950033.
[17] Salah, M., Altalla, K., Salah, A., & Abu-Naser, S. S. (2018). Predicting Medical Expenses Using Artificial Neural Network. *International Journal of Engineering and Information Systems (IJEAIS)*, *2*(20), 11-17.
[18] Marugán, A. P., Márquez, F. P. G., Perez, J. M. P., & Ruiz-Hernández, D. (2018). A survey of artificial neural network in wind energy systems. *Applied energy*, *228*, 1822-1836.
[19] Pang, Z., Niu, F., & O'Neill, Z. (2020). Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons. *Renewable Energy*, *156*, 279-289.
[20] 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.

Modeling and quantitative structure-property relationship studying to predict the half-life of polychlorinated biphenyls using multivariate linear regression and artificial neural networks

Sakineh Bahraminasab*[1], Mehdi Nekoei [2], S.Abbas Taheri [2]

[1]*Department of Information Science, Shahid Beheshti University, Tehran, Iran*

[2]*Department of Chemistry, Faculty of Science, Shahrood Branch, Islamic Azad University, Shahrood, Iran*

Submited: 01 June 2020, Revised: 09 September 2020, Accepted: 19 October 2020

Abstract

Quantitative structure-property relationship (QSPR) study was performed to predict the half-life of some polychlorinated biphenyl derivatives using multivariate linear regression (MLR) and artificial neural networks (ANN). First, the structure of the compounds, the drawing, and the appropriate group of descriptors were calculated. Then, the step-wise method was used to obtain the best descriptors that were most related to the half-life of the compounds. With this method, 6 descriptors including Lop, GATS5m, GATS8m, LDip, RDF020u, R2v + were selected from the types of topological descriptors, charge, three-dimensional representation of molecules based on electron diffraction and radial distribution function. First, a multiple linear regression linear model was constructed. Then, artificial neural network was used to obtain better results. The values of coefficient of determination (R^2) and root mean square error (RMSE) for the test series were equal to 0.716 and 0.050 for the MLR linear model and 0.896 and 0.030 for the nonlinear ANN model, respectively. Statistical data show the superiority of ANN method over MLR method.

Keywords: *Quantitative structure-property relationship, polychlorinated biphenyls, half-life, multiple linear regression, artificial neural network.*

***Corresponding author :** Sakineh Bahraminasab

Adress: Department of Information Science, Shahid Beheshti University, Tehran, Iran

Tel: 02332394289 **E-mail:** S.bahraminassab@gmail.com