



## بررسی تاثیر فیلتر های مختلف کاهش نویز بر عملکرد تقریب نرخ نفوذ حفاری توسط مدل های مبتنی بر آموزش ماشین

احسان برنجکار

کارشناسی ارشد مهندسی نفت، دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران

ebrenjkar@gmail.com

ارسال : بهمن ماه ۹۹ پذیرش: فروردین ۱۴۰۰

### چکیده

مدلسازی و تخمین دقیق نرخ نفوذ حفاری (ROP) مقدمه ای برای ایجاد یک برنامه ریزی بهتر در جهت کاهش زمان عملیات حفاری و کنترل هزینه های مصرفی میباشد. با اینحال تخمین این فاکتور کلیدی به سهولت امکان پذیر نیست و دلیل اصلی آن ارتباط پیچیده میان نرخ نفوذ و متغیر های حفاری میباشد. از سوی دیگر، وجود نویز در داده ها، سبب افزایش زمان آموزش مدل ها شده و دقت آنها را به شدت کاهش میدهد. در این مطالعه، روش شبکه های عصبی پرسپترون چند لایه (MLP) و روش حداقل مربعات رگرسیون بردار پشتیبان (LSSVR) به همراه چهار فیلتر کاهش نویز داده به منظور تخمین نرخ نفوذ استفاده شده است. داده های مورد استفاده به منظور تغذیه مدلها، از واحد نمودار گیری گل (MLU) و گزارش نهایی یک چاه حفاری شده واقع در جنوب غربی ایران جمع اوری شده است که پس از فرایند انتخاب ویژگی از طریق الگوریتم تیرید شبیه سازی شده (SA) و شبکه های عصبی مصنوعی، یازده متغیر از کل متغیرها انتخاب گردید. در ادامه، نقاط خارج از محدوده حذف شدند و نویز کلی داده ها بوسیله فیلترهای میانه، ساویتزکی گولی، میانگین متحرک وزن دار و سیگنال ویولت کاهش یافت. در پایان، عملکرد مدل های توسعه داده شده و فیلترهای به کار رفته، توسط شاخص های آماری مختلفی مورد بررسی و مقایسه قرار گرفت و ثابت گردید که روش LSSVR با کرنل پایه شعاعی و دیتای خروجی از فیلتر میانگین متحرک وزن دار، بهترین نتایج را در دو بخش آموزش و از مون مدل نمایش میدهد.

واژگان کلیدی: نرخ نفوذ حفاری، شبکه های عصبی پرسپترون چند لایه، حداقل مربعات رگرسیون بردار پشتیبان، انتخاب ویژگی، حذف نویز، الگوریتم تیرید شبیه سازی شده.

### ۱- مقدمه

توسعه بهینه میادین نفتی، مستلزم برداشتن موانعی است که سبب افزایش هزینه های مصرفی شده و بازدهی عملیات حفاری را کاهش میدهند [۱]. بهینه سازی عملیات حفاری را میتوان با افزایش سرعت حفاری بدست آورد، درحالی که الزامات مربوط به محیط زیست و محدودیت های تجهیزات نیز در نظر گرفته شده باشند [۲]. اولین چاه حفاری شده در یک میدان، همواره بیشترین هزینه ها را در مقایسه با چاههای بعدی در همان میدان به خود اختصاص میدهد. اما با افزایش آشنایی با خصوصیات میدان و سازند های حفاری

شده، میتوان برنامه ریزی های دقیق تری را برای بهینه سازی عملیات حفاری چاههای آینده تدارک دید. بر این اساس میتوان بهینه سازی و کاهش هزینه های مربوط به چاههای بعدی را تا رسیدن به یک نقطه پایدار، ادامه داد [۳]. ثابت شده که عدم بهینه سازی پارامترهای حفاری ممکن از منجر به طولانی شدن زمان عملیات و افزایش هزینه ها شود. این موضوع در خصوص عملیات حفاری چاه های دریایی بیشتر حائز اهمیت است [۴].

بهینه سازی حفاری، با کاهش زمان عملیات انجام می شود، زیرا زمان همیشه برابر با مصرف سرمایه است. مفهوم زمان صرف شده برای هر عملیات حفاری را می توان بر حسب نرخ نفوذ حفاری (ROP) بیان کرد. بنابر این تخمین نرخ نفوذ از ان جهت که سبب انتخاب بهینه پارامترهای حفاری و کاهش هزینه های مصرفی میشود، بسیار حائز اهمیت است [۵]. سریع ترین نرخ نفوذ الزاما به کم ترین هزینه در هر فوت از چاه منتج نمی شود و ممکن است عوامل دیگری نظیر فرسودگی سریع مته و شکست تجهیزات باعث شوند تا در مقابل هزینه های عملیات حفاری به طور چشمگیری افزایش یابند [۶]. متغیر های تاثیر گذار بر نرخ نفوذ تا کنون به درستی درک نشده اند و از طرفی دیگر عدم قطعیت بالایی در مورد پارامترهای اثر گذار بر این فاکتور کلیدی وجود دارد [۷]. همچنین تعداد زیادی از متغیرها، به صورت نامشخص بر فرایند حفاری اثر می گذارند که رابطه آنها به صورت غیر خطی و پیچیده میباشد [۸].

تا کنون، محققان بسیاری به بررسی تاثیر متغیرهای حفاری بر نرخ نفوذ و چگونگی ارتباط آنها با یکدیگر پرداخته اند. در طول چند دهه گذشته، مدل های ریاضی مختلفی به منظور تخمین نرخ نفوذ ارائه شده است که به انتخاب یک شرایط عملیاتی بهتر کمک میکنند [۹]. مدل های نرخ نفوذ، یک رابطه ریاضی هستند که متغیر نرخ نفوذ (به عنوان یک متغیر وابسته) را به سایر متغیر های عملیات حفاری (به عنوان متغیر های مستقل) مرتبط می سازند. نرخ نفوذ از ان جهت به عنوان یک پارامتر وابسته نامیده میشود که پیش بینی ان تحت تاثیر پارامترهای بسیاری مانند هیدرولیک حفاری، وزن روی مته، سرعت چرخش مته، نوع مته و ویژگی های سازند قرار دارد [۱۰]. تحقیقات قابل توجهی در اواخر نیمه اول قرن بیستم با هدف ارائه مدل های تخمین گر نرخ نفوذ آغاز گردید، که غالب معادلات ارائه شده بر پایه دو متغیر وزن روی مته و سرعت چرخش بودند. پس از ان زنجیره تغییرات و تحقیقات صورت پذیرفت تا با در نظر گرفتن تاثیر سایر متغیرها، دقت تقریب مدل های نرخ نفوذ افزایش یابد [۱۱]-[۱۵].

در این مطالعه به منظور توسعه مدل های تخمین گر نرخ نفوذ، از داده های واحد نمودار گیری گل<sup>۲</sup> و گزارش نهایی یک چاه حفاری شده واقع در جنوب غربی ایران، استفاده گردید. انتخاب متغیر های ورودی یکی از چالش های پیش رو در توسعه مدل میباشد، به طوری که انتخاب تعداد ورودی های بالا سبب افزایش پیچیدگی مدل شده و از طرفی انتخاب تعداد ورودی های پایین سبب یک مدل سازی ضعیف از دانش موجود میگردد [۱۶]. فرایند انتخاب ویژگی<sup>۳</sup>، با ترکیب الگوریتم تبرید شبیه سازی شده<sup>۴</sup> (SA) و شبکه عصبی مصنوعی بهینه شده با الگوریتم پس انتشار خطا<sup>۵</sup>، انجام گرفت و یازده متغیر ورودی از ۱۶ متغیر جمع اوری شده به منظور توسعه مدل های تخمین گر نرخ نفوذ انتخاب شدند.

داده های خام معمولا دارای داده های نویز هستند که سبب پوشیده شدن داده های مفید میشوند [۱۷]، [۱۸]. نویز داده ها ناشی از عواملی همچون عدم دقت مناسب سیستم های ثبت کننده و خطاهای انسانی میباشد که سبب ایجاد اختلال در فرایند آموزش مدل شده و فرایند آموزش را به یک پروسه زمان بر و دشوار تبدیل میکند. بنابر این غلبه بر داده های نویز برای دستیابی به مدلهایی قابل اعتماد و دقیق تر، بسیار ضروری است [۱۹]، [۲۰]. در این مطالعه، به منظور بهبود فرایند آموزش مدلها و افزایش دقت تخمین نرخ نفوذ، از چهار روش کاهش نویز مختلف استفاده گردید.

<sup>1</sup> Rate of penetration (ROP)

<sup>2</sup> Mud logging unit (MLU)

<sup>3</sup> Feature selection

<sup>4</sup> Simulated annealing (SA)

<sup>5</sup> Back propagation algorithm (BP)

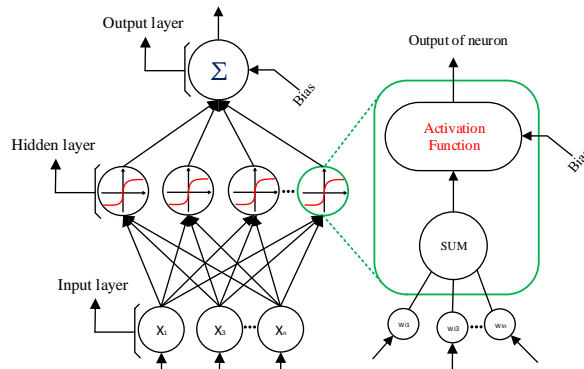
پس از فرایند پیش پردازش داده ها، از دو روش شبکه های عصبی پرسپترون چند لایه<sup>۱</sup> (MLP) و حد اقل مربعات رگرسیون بردار پشتیبان (LSSVR)، به منظور تخمین نرخ نفوذ استفاده گردید و پس از آموزش مدلها بوسیله ۸۰ درصد از داده ها، ارزیابی عملکرد مدل ها با ۲۰ درصد باقی مانده مورد بررسی قرار گرفت. ارزیابی خطا در دو بخش آموزش و آزمون مدل، توسط شاخص های آماری مختلفی مورد بررسی قرار گرفت که شامل میانگین درصد نسبی خطای مطلق، میانگین مربعات خطا، جذر میانگین مربعات خطا و ضریب رگرسیون میباشد. نتایج این مطالعه نتایج بسیار ارزشمندی در خصوص انواع فیلتر های کاهش نویز ارائه مینماید. همچنین، یکی از تفاوت های این مطالعه با مطالعات پیشین، مقایسه طیف وسیعی از فیلتر های کاهش نویز به صورت عملیاتی میباشد که از این طریق میتوان کارآمدترین فیلتر را در جهت کاهش نویز انتخاب کرده و از آن برای توسعه هر چه دقیقتر مدل های مبتنی بر آموزش ماشین بهره برد.

## ۲- کلیات روشها

### ۲-۱- شبکه عصبی پرسپترون چند لایه (MLP)

شبکه های عصبی مصنوعی، یک سیستم پردازش اطلاعات است که سعی در تقلید از ویژگی های عملکرد سیستم عصبی انسان دارد. این سیستم به عنوان یک مدل رایانه ای سازگار به شمار میرود که می تواند ارتباطات، الگوها یا نگاشت هایی را بین اشیا یا داده ها ایجاد کند [۲۱]. این سیستم همچنین محبوب ترین تکنیک هوش مصنوعی برای شناخت الگوی داده ها میباشد که از عناصر اساسی به نام نورون ساخته شده است. در این تکنیک، از معادلات غیرخطی برای بدست آوردن نتایج مربوط به مسائل پیش بینی استفاده میشود، که در قسمتهای اساسی شبکه که همان نورون ها هستند، قرار دارند. برای آموزش شبکه های عصبی، از پارامترهای ورودی و پارامترهای خروجی متناظر استفاده می شود تا بتوان در نهایت به یک تقریب از تابع مورد نظر دست یافت [۲۲]. در سیستم شبکه های عصبی مصنوعی، در طی یک فرایند یادگیری که در آن اتصال بین سلولهای عصبی تنظیم میشوند، اختلاف بین مقادیر پیش بینی شده و مقادیر حقیقی به حد اقل میرسد [۲۱].

شبکه عصبی پرسپترون چند لایه (MLP) یک نوع خاص از شبکه عصبی است که دارای یک لایه ورودی، یک یا چند لایه مخفی و یک لایه خروجی میباشد. ساختار یک MLP در شکل ۱ به تصویر کشیده شده است.



شکل ۱- ساختار یک شبکه عصبی با معماری سه لایه ای با تابع فعال ساز سیگموئید

ساختار یک شبکه عصبی MLP متشکل از سه لایه میباشد که شامل: لایه ورودی، لایه خروجی و لایه پنهان می باشد. لایه ورودی توسط پیوندهایی که متشکل از مقادیر وزن مشخصی میباشد به توابع فعال ساز موجود در لایه پنهان متصل شده اند. همچنین خروجی لایه پنهان نیز توسط اتصالات وزن دار مشخصی به لایه خروجی وصل شده است [۲۳]. در لایه پنهان، پردازش اطلاعات با استفاده از توابع فعال ساز مستقر در نورون ها انجام میگردد که معروفترین این توابع نوع سیگموئید (رابطه ۲) میباشد.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

<sup>۱</sup> Multiple layer perceptron (MLP)

لایه خروجی، مقادیر تجمیع شده از نورون های لایه پنهان را بوسیله یک تابع خطی به مقادیر خروجی تبدیل میکند و در انتها مقادیر تولید شده برای مقایسه کارایی شبکه عصبی با خروجی مورد نظر مقایسه می شود [۲۱]. در صورتی که  $\Pi$  ورودی به شبکه وارد شده و  $k$  نورون در لایه پنهان مستقر باشد، رابطه زیر تشکیل میگرد:

$$y = f^o \left( \omega^b + \sum_{i=1}^k \omega_{ij}^o f^h \left( \omega_j^b + \sum_{i=1}^n \omega_{ji}^h x_i \right) \right) \quad (2)$$

که در رابطه فوق،  $f^o$  تابع فعال ساز در لایه خروجی،  $f^h$  تابع فعال ساز در هر یک از نورون های لایه مخفی،  $\omega^b$  مقدار بایاس اضافه شده به لایه خروجی و  $\omega_j^b$  مقدار بایاس اضافه شده به لایه پنهان می باشد [۵]. این نوع شبکه امروزه به طور گسترده ای در تقریب توابع دلخواه مورد استفاده قرار می گیرد زیرا تعمیم خوب و دقت بالایی را فراهم می کند. تقریباً در غالب مسائل پیشینی، یک پرسپترون چند لایه فقط با یک لایه پنهان برای تقریبی هر مساله دلخواهی با دقت دلخواه کافی است زیرا قابلیت های تقریب یک پرسپترون چند لایه بیشتر از تعداد لایه های پنهان به تعداد نورون ها وابسته است. همچنین، در هنگام تصمیم گیری در مورد تعداد نورون های موجود در لایه های پنهان، بایستی نهایت دقت را انجام داد زیرا نورون های زیاد در لایه پنهان باعث افزایش تعداد پارامترها (وزن اتصال) می شود و از نظر محاسباتی شبکه را ناکارآمد می کند، در حالی که تعداد بسیار کم نورون در لایه پنهان باعث می شود که شبکه از دقت کمتری برخوردار باشد [۲۴].

## ۲-۲- حد اقل مربعات رگرسیون بردار پشتیبان (LSSVR)

ماشین بردار پشتیبان<sup>۱</sup>، یکی از روش های آموزش ماشین است که توسط وپنیک و کورتس<sup>۲</sup> ارائه شده است و بر مبنای یادگیری اماری<sup>۳</sup> استوار می باشد [۲۵]. از دیدگاه ماشین بردار های پشتیبان، کمینه سازی ریسک ساختاری<sup>۴</sup> به عنوان تابع هدف، به جای کمینه سازی ریسک تجربی<sup>۵</sup> در نظر گرفته میشود، که این مساله سبب دستیابی به یک بهینه جهانی و عدم رخ دادن بیش برآزش<sup>۶</sup> می گردد [۲۶]. علاوه بر مسائل طبقه بندی<sup>۷</sup>، می توان از ماشین بردار های پشتیبان در مسائل رگرسیون نیز استفاده نمود [۲۷].

رگرسیون بردار پشتیبان، یک نوع خاص از ماشین های بردار پشتیبان است که با هدف تقریب توابع غیرخطی و پیچیده ارائه شده است و تفاوت آن با روش های دسته بندی، در پیوسته بودن مجموعه هدف می باشد [۲۸]. در نوع رگرسیون، ابتدا داده های آموزش به فضای ویژگی<sup>۸</sup> نگاشت میشوند. در این حالت، توابع کرنل عمل نگاشت داده های ورودی را به فضای ویژگی انجام داده و به عنوان واسطه ای میان متغیرهای ورودی و متغیر هدف عمل می کنند.

در صورت در اختیار داشتن داده های آموزشی  $\{(x_i, y_i), i=1:N, x_i \in R^N, y_i \in R\}$ ، در حالی که  $x_i$  نشان دهنده  $i$  امین داده ورودی از فضای  $N$  باشد و  $y_i$  خروجی واقعی متناظر با این مجموعه ورودی باشد، مدل رگرسیون بردار پشتیبان را می توان به صورت رابطه ۳ نوشت.

$$y_i = f(x_i) = W^T \phi(x_i) + b \quad (3)$$

در حالی که،  $f(x_i)$  خروجی مدل یا مقدار تقریب زده شده برای  $y_i$  بوده،  $\phi(x_i)$  تابع ویژگی بردار ورودی  $x_i$  را نمایش می دهد.  $W$  و  $b$  نیز، مقادیر وزن ها و بایاس های مدل بوده که در حین فرایند آموزش، بهینه و تنظیم می شوند. تنظیم این پارامترها توسط به کارگیری یک تابع تخطی<sup>۹</sup> انجام می شود. این تابع تخطی، به صورت روابط ۴ و ۵ تعریف می شود [۲۹].

<sup>1</sup> Support vector machine

<sup>2</sup> Vapnik and cortes

<sup>3</sup> Statistical learning theory

<sup>4</sup> Structural risk minimization principle(SRM)

<sup>5</sup> Empirical risk minimization principle(ERM)

<sup>6</sup> Over fitting

<sup>7</sup> Clustering

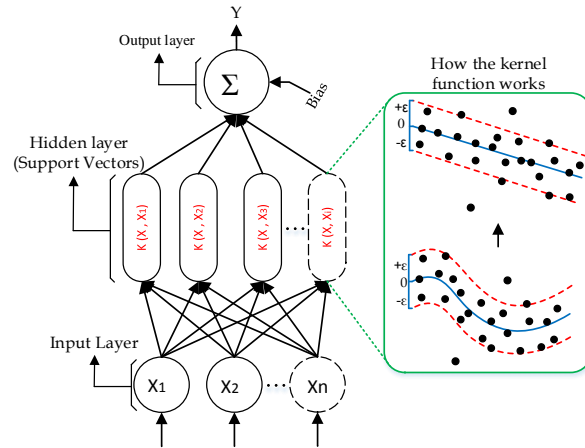
<sup>8</sup> Feature space

<sup>9</sup> Loss function

$$R(C) = \frac{1}{2} \|w\|^2 + c \times \frac{1}{n} \sum_{i=1}^n |y_i - f(x_i)|_{\epsilon} \quad (4)$$

$$|y_i - f(x_i)|_{\epsilon} = \begin{cases} 0 & |y_i - f(x_i)| \leq \epsilon \\ |y_i - f(x_i)| - \epsilon & \text{Otherwise} \end{cases} \quad (5)$$

در حالی که، C ضریب تخطی بوده و ε بیشینه خطای مورد قبول میباشد. تازمانی که اختلاف خروجی مورد نظر کمتر از ε باشد، تابع خطا برابر با مقدار صفر خواهد بود و در غیر این صورت مقدار تابع خطا بر اساس میزان تخطی از ناحیه مجاز، به صورت خطی افزایش میابد [۲۵]. در شکل ۲، یک ساختار کلی از مدل رگرسیون بردار پشتیبان و شمایی از نحوه عملکرد یک تابع کرنل در لایه میانی، نمایش داده شده است.



شکل ۲- ساختار شبکه رگرسیون بردار پشتیبان (SVR)

مسئله کمینه کردن مقدار تابع تخطی را میتوان به صورت رابطه ۶ بیان نمود [۲۹].

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \frac{1}{n} \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (6)$$

به شرط این که رابطه زیر برقرار باشد:

$$\begin{cases} -y_i + w^T \varphi(x_i) + b \leq \epsilon + \xi_i & i=1,2,\dots,m \\ y_i + w^T \varphi(x_i) - b \leq \epsilon + \xi_i^* & i=1,2,\dots,m \\ \xi_i, \xi_i^* \geq 0 & i=1,2,\dots,m \end{cases} \quad (7)$$

در این مسئله بهینه‌سازی،  $\xi_i$  و  $\xi_i^*$  رعایت محدودیت‌ها را تضمین می‌کنند، C تعادل میان میزان پیچیدگی مدل و خطای آموزش را برقرار کرده و ε برای تعیین بازه تغییرات مناسب استفاده می‌شود. با استفاده از معادله لاگرانژ، مسئله بهینه‌سازی دوگانه مدل به صورت رابطه ۸ حاصل می‌شود [۲۸].

$$\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) k(x_i, x_j) + \dots \quad (8)$$

$$\frac{1}{2} \sum_{i=1}^n ((\epsilon - y_i) \alpha_i + (\epsilon + y_i) \alpha_i^*)$$

به شرط این که رابطه زیر برقرار باشد:

$$\sum_{j=1}^n (\alpha_i - \alpha_i^*) = 0 \quad (9)$$

$$0 \leq \alpha_i, \alpha_i^* \leq C$$

$$i = 1, 2, \dots, m$$

برای حل مساله بالا، مدل رگرسیون بردار پشتیبان در نهایت به فرم زیر تعریف می‌شود:

$$\sum_{i=1}^n (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (10)$$

که در این رابطه،  $\alpha_i$  و  $\alpha_i^*$  ضرایب لاگرانژ بوده و  $k(x_i, x)$  تابع کرنل مورد استفاده در مدل SVR میباشد. رایج ترین کرنل های مورد استفاده در ادبیات تحقیق شامل سه کرنل گوسی یا پایه شعاعی<sup>۱</sup> (رابطه ۱۱)، کرنل چند جمله ای<sup>۲</sup> (رابطه ۱۲) و کرنل خطی<sup>۳</sup> (رابطه ۱۳) می باشند [۲۷]، [۳۰].

$$G(x_i, x_j) = \exp\left(-0.5 \frac{\|x_i - x_j\|^2}{\sigma^2}\right) \quad (11)$$

$$G(x_i, x_j) = (ax_i x_j + b)^q \quad (12)$$

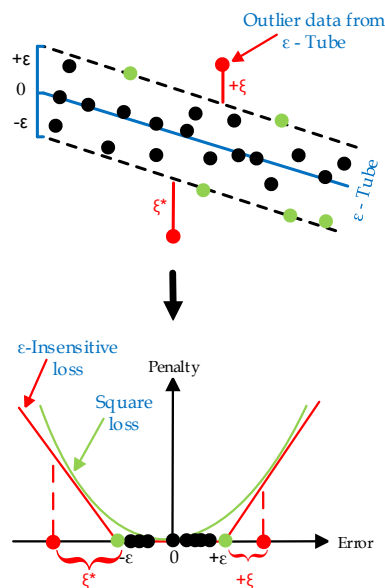
$$G(x_i, x_j) = (x_i, x_j) \quad (13)$$

که در این روابط،  $\delta$  پهنای تابع پایه شعاعی،  $q$  درجه تابع چند جمله ای،  $a$  و  $b$  ثوابت هستند. از آنجایی که مقادیر این پارامترها در ساختار فضای ویژگی و راه حل نهایی موثر هستند، لذا تعیین مقادیر مناسب برای این پارامترها در توسعه مدل های SVR، بسیار حائز اهمیت میباشد [۳۱].

تفاوت روش حد اقل مربعات ماشین بردار پشتیبان (LSSVR) با روش کلاسیک، در تعریف تابع تخطی میباشد [۳۲]. در این روش، به جای یک تابع تخطی با مفهوم استانه  $\epsilon$  (روابط ۴ و ۵)، یک تابع تخطی مربعی تعریف می شود که سبب بوجود آوردن یک معادله لاگرانز، با حل مستقیم سیستم خطی کوهن تا کر [۳۳]، مطابق رابطه زیر میشود.

$$\begin{bmatrix} 0 & I_n^T \\ I_n & K + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b_0 \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (14)$$

که در این رابطه،  $I_n$  یک بردار واحد  $[n \times 1]$  است،  $T$  به معنی ترانهاد<sup>۴</sup> ماتریس یا بردار است،  $b$  بردار وزن،  $b_0$  مقدار انحراف<sup>۵</sup> مدل میباشد. عملکرد دو تابع تخطی نوع کلاسیک و تابع تخطی حد اقل مربعات، در شکل ۳ نمایش داده شده است.



شکل ۳- نحوه عملکرد دو تابع تخطی کلاسیک و حد اقل مربعات در تعیین مقادیر خطای خروج داده ها از محدوده مشخص شده

راه حل معادله قبل را میتوان بوسیله روشهای استاندارد حل معادلات خطی یافت. بنابر این روش LSSVR، به عنوان یک زیر مجموعه بسیار قوی از روش SVR، میتواند از پیچیده شدن مدل جلوگیری کرده و اهداف تقریب تابع را به خوبی برآورده سازد [۳۴]. از طرف

<sup>1</sup> Gaussian kernel or radial basis function(RBF)

<sup>2</sup> Polynomial

<sup>3</sup> Linear

<sup>4</sup> Transpose

<sup>5</sup> Offset

دیگر تعداد پارامترهای تنظیمی مدل در تعریف تابع خطا، کمتر از روش SVR کلاسیک میباشد که همین مساله سبب گردیده تا حل مسائل LSSVR نسبت به مسائل SVR ساده تر باشد. [۳۵]

### ۳-۲- الگوریتم تبرید شبیه سازی شده (SA)

الگوریتم تبرید شبیه سازی شده (SA)، یک روش بهینه سازی اکتشافی با رویکرد تصادفی میباشد. مزیت اصلی این الگوریتم در مقایسه با روش های بهینه سازی قدیمی تر، توانایی آن در فرار از حداقل محلی<sup>۱</sup> میباشد. ایده اصلی این روش، انتخاب یک حرکت تصادفی به جای انتخاب بهترین حرکت (بهترین همسایگی) در میان حرکت های ممکن در هر تکرار از الگوریتم میباشد، به طوری که اگر راه حل جدید هزینه را بهتر کند (هزینه را کاهش دهد)، به عنوان حالت بعدی پذیرفته شده و اگر این راه حل هزینه را افزایش دهد، فقط با احتمال P پذیرفته می شود. P به عنوان احتمال متروپلیس-هیستینگر<sup>۲</sup> نام برده شده و به صورت زیر تعریف می شود [۳۶]:

$$P(\Delta E) = e^{-\frac{\Delta E}{T}} \quad (15)$$

که در این رابطه،  $\Delta E$  نشان دهنده تغییر در انرژی (ارزش تابع هزینه) ناشی از تغییر حالت است و T متغیر دما یا متغیری شبیه به آن است که این احتمال را کنترل میکند. شبیه سازی SA در دمای بالا شروع می شود که به آن نقطه شروع یا دمای اولیه<sup>۳</sup> ( $T_0$ ) میگویند. در هر دما، تلاشهای مختلف (حرکت) تازمانی انجام می شود که سیستم به یک حالت پایدار از درجه حرارت برسد. سپس مقدار T بسته به فرایند تبرید، تادمایی بسیار پایین و نزدیک به صفر کاهش میابد. بر این اساس، احتمال پذیرش حرکت هایی که سبب افزایش تابع هزینه می گردد، بالا است و با کاهش دما این احتمال کاهش میابد [۳۶]، [۳۸].

زمان بندی تبرید در SA، در حقیقت بر اساس زنجیره مارکوف<sup>۴</sup> است که در یک دمای ثابت رخ میدهد. برای تعریف این زنجیره به چند پارامتر نیاز داریم که عبارت هستند از دمای اولیه ( $T_0$ )، چگونگی تغییر دما در حین اجرای الگوریتم، طول و تعداد زنجیره های مارکوف یا تعداد تلاشها در هر دما، تعریف یک معیار توقف و یا دمای نهایی.

دو پارامتر دمای اولیه و چگونگی تغییر دما در حین اجرای الگوریتم، بیشترین تاثیر را بر کند کردن الگوریتم SA دارد. ثابت شده است که اگر روند کاهش دما به اندازه کافی کند باشد، SA به سمت بهینه جهانی<sup>۵</sup> با احتمال یک همگرا می شود. اما کاهش سریع دما، روند بهینه سازی الگوریتم را با پیدا کردن حداقل یک کمینه محلی متوقف می کند.

### ۳- جمع اوری و پیش پردازش داده ها

داده های مورد نیاز به منظور توسعه مدل های تخمین گر نرخ نفوذ، از واحد نمودار گیری گل و گزارش نهایی یک چاه حفاری شده واقع در جنوب غربی ایران جمع اوری شده است. این چاه به صورت عمودی حفاری شده و دارای یک مونتاژ پایین حفره مرسوم میباشد. داده های جمع اوری شده از این چاه شامل ۱۸۰۴ نمونه با ۱۶ متغیر مستقل (پارامترهای حفاری) و یک متغیر وابسته (نرخ نفوذ) میباشد. در جدول ۱، پارامترهای آماری حداقل، حد اکثر، میانگین و انحراف استاندارد هر یک از این متغیر ها ارائه شده است. از آنجایی که داده های خام همواره دارای نویز و نقاط خارج از محدوده هستند، نیاز است تا کیفیت این داده ها با هدف استخراج داده های موثر، بهبود یابد. همچنین فرایند انتخاب ویژگی، سبب از بین بردن ریسک اضافه کردن متغیرهایی با همبستگی پایین به مدل میشود، که در نهایت پیچیدگی محاسبات را کاهش داده و زمان آموزش را بهبود میبخشد. همچنین، به منظور آموزش بهتر مدل های مبتنی بر آموزش ماشین از رابطه ۱۶ استفاده گردید تا داده ها در بازه -۱ و +۱ نرمال شوند.

$$x_i = 2 \times \frac{X - X_{min}}{X_{max} - X_{min}} - 1 \quad (16)$$

<sup>1</sup> Local minimum

<sup>2</sup> Metropolis-Hastings

<sup>3</sup> Initial temperature

<sup>4</sup> Markov chain

<sup>5</sup> Global optimum

که در این رابطه،  $X_i$  مقدار نرمال شده،  $X$  مقدار واقعی،  $X_{min}$  کمترین مقدار و  $X_{max}$  بیشترین مقدار از پارامتر مورد نظر می باشد [۲۲].

جدول ۱- جزئیات اماری متغیرهای جمع اوری شده به منظور توسعه مدل های تخمین گر نرخ نفوذ

پارامتر	کد پارامتر	واحد	حد اقل	حد اکثر	میانگین	انحراف استاندارد
عمق	A1	m	۱۵۲.۰۰	۲۳۳۵.۰۰	۱۲۷۰.۵۴	۶۵۵.۶۶
وزن روی مته	A2	klbf	۲۰۰	۲۷.۴۶	۹.۶۳	۵.۲۷
وزن روی قلاب	A3	klbf	۹۲.۰۰	۲۳۸.۴۲	۱۷۳.۸۳	۳۵.۷۳
سرعت چرخش مته	A4	rpm	۷۴.۵۳	۲۰۰.۶۳	۱۶۶.۸۶	۲۵.۰۶
قطر مته	A5	in	۸.۲۵	۱۷.۵۰	۱۳.۱۴	۳.۳۵
گشتاور	A6	lbf.ft	۱۱۹۹۴.۲۰	۴۵۸۹۶.۲۸	۲۹۴۳۰.۴۱	۶۵۸۱.۷۲
فشار پمپاژ گل	A7	psi	۳۰۰.۸۷	۲۴۶۰.۶۹	۱۳۵۳.۲۵	۵۶۹.۵۳
نرخ جریان گل ورودی	A8	gpm	۲۹۸.۷۰	۹۹۱.۹۲	۶۶۸.۴۴	۱۹۵.۷۳
نرخ جریان گل خروجی	A9	gpm	۴۲۹.۱۷	۴۲۹.۲۰	۴۲۹.۱۹۱	۰.۰۰۴
زمان بالا آمدن کنده ها از چاه	A10	min	۱۶.۲۳	۵۳.۶۹	۳۴.۷۷	۶.۷۴
وزن گل ورودی	A11	pcf	۶۱.۳۷	۱۲۰.۴۳	۸۵.۹۸	۱۹.۷۶
وزن گل خروجی	A12	pcf	۶۱.۹۶	۱۲۰.۷۹	۸۶.۱۸	۱۹.۷۷
دمای گل ورودی	A13	°C	۱۴.۷۳	۵۴.۵۹	۳۹.۴۶	۹.۰۳
دمای گل خروجی	A14	°C	۲۹.۰۳	۱۰۵.۵۶	۵۸.۸۴	۱۹.۱۰
چگالی گل در حال چرخش	A15	Pcf	۶۱.۶۲	۱۲۰.۶۰	۸۳.۶۱	۱۸.۸۷
مدت زمان فعالیت مته	A16	hr	۷۵۲.۲۳	۱۵۰۲.۹۰	۱۰۹۸.۱۶	۲۲۳.۹۴
نرخ نفوذ مته	-	min/m	۲.۵۲	۴۰.۴۸	۱۵.۵۶	۷.۱۴

### ۱-۳- انتخاب ویژگی

فرایند انتخاب ویژگی، در جهت بهبود زمان و عملکرد سیستم بسیار اهمیت دارد. رویکرد انتخاب متغیرهای ورودی، بایستی بر اساس اهمیت تاثیر هر یک از متغیرها بر خروجی هدف باشد تا علاوه بر دستیابی به حد اقل میزان خطا، از پیچیده شدن محاسبات و دشوار شدن فرایند آموزش جلوگیری به عمل آید [۳۹]. از طرفی دیگر، پیش بینی نرخ نفوذ با تمام پارامترهای حفاری امکان پذیر نیست و ضروری است که پارامترهای ورودی، بر پایه تحلیل و بررسی اهمیت تاثیر هر یک از متغیرها تعیین شوند [۱۶]. روش های مختلفی برای انتخاب موثرترین ویژگی ها وجود دارد. برخی از این روش ها شامل فیلترها<sup>۱</sup>، روش های دسته بندی<sup>۲</sup>، روش توکار<sup>۳</sup>، روش های یادگیری بدون نظارت<sup>۴</sup>، روش های ترکیبی<sup>۵</sup> و روش های اکتشافی<sup>۶</sup> میباشند [۴۰]-[۴۳]. روش های انتخاب ویژگی که مبتنی بر جستجوی ترتیبی یا اکتشافی هستند، معمولاً به راه حل های قوی تری دست می یابند [۳۹]. در این مطالعه، با کمک روش شبکه های عصبی و الگوریتم بهینه سازی تبرید شبیه سازی شده (SA)، بهترین متغیرهای ورودی برای توسعه مدلها تعیین شدند. ساختار

<sup>1</sup> Filters

<sup>2</sup> Wrapper

<sup>3</sup> Embedded

<sup>4</sup> Unsupervised learning technique

<sup>5</sup> Sequential

<sup>6</sup> Heuristic

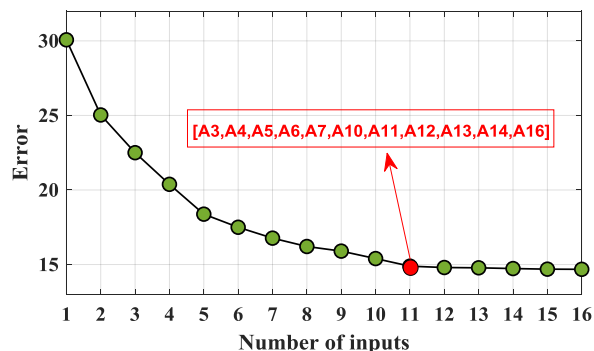


شبکه عصبی در این بخش، شامل یک لایه پنهان با ۱۵ نورون حاوی تابع فعالسازی سیگموئید و یک نورون در لایه خروجی با تابع فعال ساز خطی<sup>۱</sup> میباشد.

به منظور انتخاب بهترین متغیر های ورودی از میان متغیر های تصمیم، یک مساله انتخاب ویژگی چند هدفه، در غالب یک مساله تک هدفه تجزیه شده<sup>۲</sup> تعریف گردید، که در ان یک قید ظرفیت، وظیفه ی انتخاب تعداد مشخصی از متغیر های ورودی را از میان تمام متغیر های مستقل بر عهده دارد. بنابر این در هر چرخه، بهترین متغیر های ورودی در هر ظرفیت تعیین میشود. در جدول ۲، پارامتر های تنظیمی الگوریتم SA، ارائه شده است. همچنین در شکل ۴، مقدار MSE در مقابل بهترین متغیر های ورودی انتخاب شده در هر ظرفیت توسط الگوریتم SA، نمایش داده شده است. همانطور که در شکل ۴ مشاهده میشود، کاهش مقدار خطا از تعداد ۱۱ ورودی به بعد، به طور چشمگیری کاهش میابد و نمودار یک روند تقریباً ثابت را دنبال میکند. این مشاهده بیان کننده عدم ضرورت اضافه کردن یک ورودی جدید به منظور بهبود مقدار خطای مدل سازی، میباشد.

جدول ۲- پارامتر های تنظیمی الگوریتم SA

مقدار	پارامتر
۲۰	تعداد تکرار های الگوریتم
۱۰	حد اکثر تعداد تکرار های فرعی
۱۰	دمای اولیه
۰.۹۹	نرخ کاهش دما
۵	تعداد تکرار هر جمعیت



شکل ۴- نمودار حد اقل میزان خطای شبکه در مقابل ورودی های موثر انتخاب شده توسط الگوریتم SA

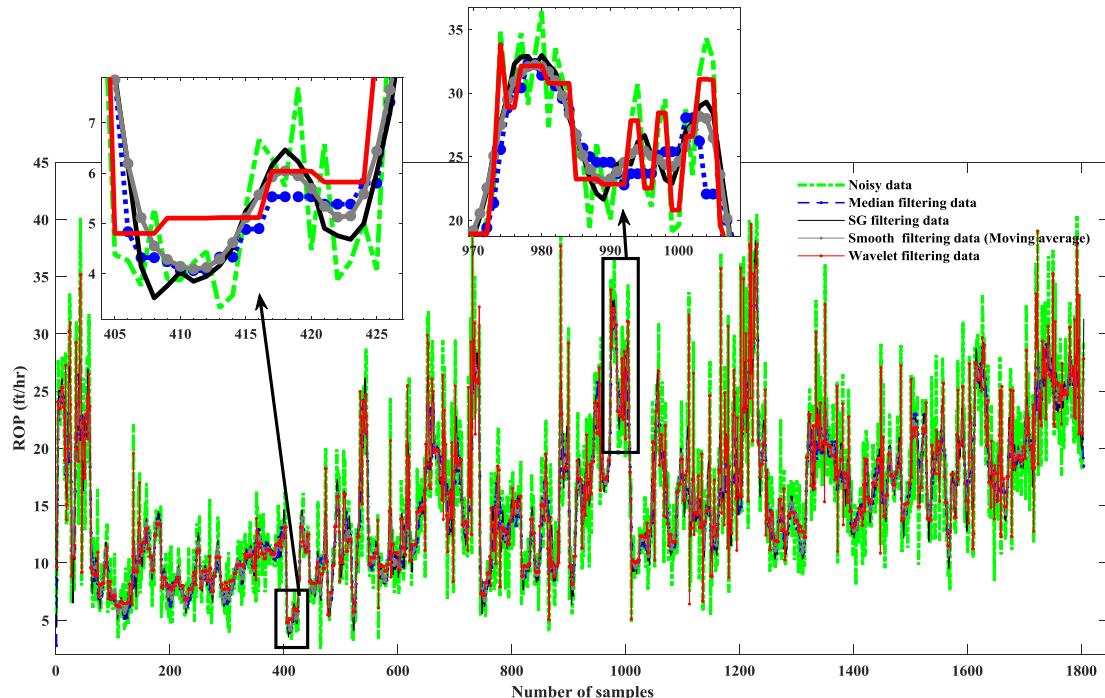
### ۲-۳- کاهش نویز

داده های خام معمولاً دارای داده های نویز هستند که سبب پوشیده شدن داده های مفید میشوند [۱۷]، [۱۸]. نویز داده ها ناشی از عواملی همچون عدم دقت مناسب سیستم های ثبت کننده و خطاهای انسانی میباشد که سبب ایجاد اختلال در فرایند آموزش مدل شده و این فرایند را به یک پروسه زمانبر و دشوار تبدیل میکند، که در ان عملکرد الگوریتم کاهش پیدا کرده است. بنابر این غلبه بر داده های نویز برای دستیابی به مدلهایی قابل اعتماد و دقیق تر، بسیار ضروری میباشد [۱۹]، [۲۰]. ایجاد نویز در داده های فرایند حفاری میتواند دلایل متفاوتی از قبیل تغییرات سازند، ناپایداری رشته حفاری و ... داشته باشد [۳۹].

<sup>1</sup> Linear function (Purelin)

<sup>2</sup> Decomposition

روش های متفاوتی برای غلبه بر داده های نویز وجود دارد که هر کدام دارای عملکرد متفاوتی در کاهش نویز داده ها هستند. در این مطالعه، از چهار روش کاهش نویز مختلف شامل فیلتر میانه<sup>۱</sup>، ساویتزکی گولی<sup>۲</sup>، میانگین متحرک<sup>۳</sup> و سیگنال ویولت<sup>۴</sup> استفاده گردید. پیاده سازی این روشها، در محیط نرم افزار متلب انجام گرفت و این فرایند بر روی تمام متغیرهای ورودی اعمال گردید. به طور نمونه در شکل ۵، نحوه عملکرد این چهار روش بر روی متغیر نرخ نفوذ حفاری نمایش داده شده است. همچنین در جدول ۳، توضیحاتی در مورد نحوه عملکرد این روشها در ادبیات تحقیق، به همراه توابع استفاده شده در محیط متلب ارائه شده است.



شکل ۵- مقایسه نرخ نفوذ حفاری، قبل و بعد از فرایند حذف نویز توسط روش های مورد استفاده در این مطالعه

جدول ۳- نحوه عملکرد روش های حذف نویز استفاده شده در این مطالعه و توابع آنها در محیط نرم افزار متلب

نام روش	نحوه عملکرد	نام تابع
فیلتر میانه	اعمال یک فیلتر میانه یک بعدی [۴۴].	" medfilt1"
حذف نویز ساویتزکی گولی	اعمال یک تابع چند جمله ای، با درجه و سائز پنجره مشخص [۴۵].	" sgolayfilt"
میانگین متحرک وزندار	ایجاد یک پنجره با طول ثابت و محاسبه مقدار میانگین متحرک وزندار در هر پنجره [۴۶].	" smoothdata"
سیگنال ویولت	بر اساس خاصیت فشرده سازی انرژی سیگنال در چند ضریب از ضرایب موجک و انتخاب یک مقدار آستانه مناسب برای ضرایب [۴۷].	" wdenoise"

#### ۴- مدل سازی و تخمین نرخ نفوذ

همانطور که در بخش های پیشین عنوان گردید، از ۱۱ متغیر ورودی شامل وزن روی قلاب، سرعت چرخش مته، قطر مته، گشتاور مته، فشار پمپاژ گل، زمان بالا آمدن کننده های حفاری از چاه، وزن گل ورودی به چاه، وزن گل خروجی از چاه، دمای گل ورودی به چاه، دمای گل خروجی از چاه و ساعات فعالیت مته، به منظور مدل سازی نرخ نفوذ استفاده شده است. این داده ها پس از طی فرایند کاهش نویز و انتخاب ویژگی، به صورت تصادفی به دو زیر مجموعه داده های آموزشی شامل ۱۴۴۳ نقطه داده (۸۰ درصد از

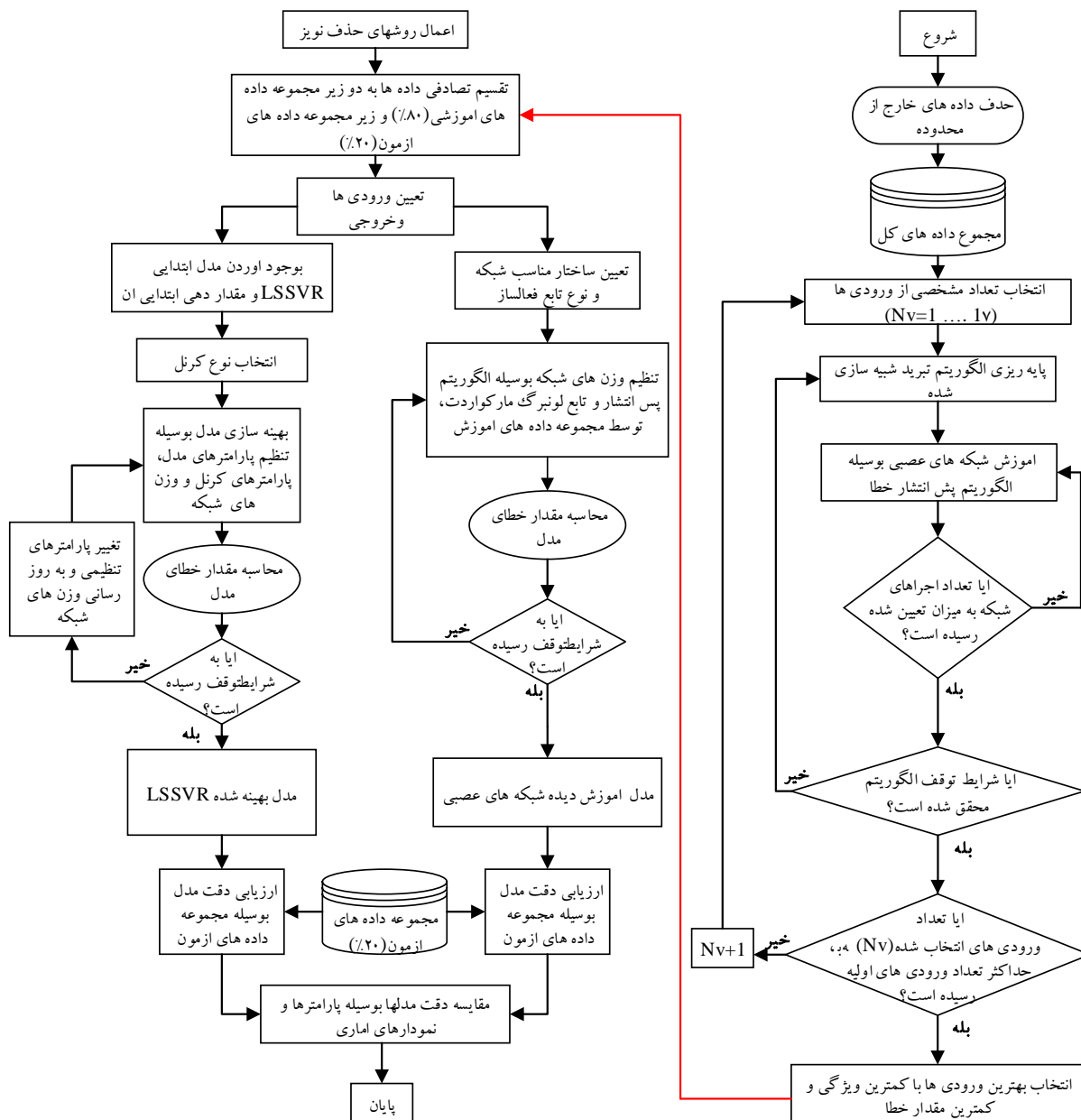
<sup>1</sup> Median filter

<sup>2</sup> Savitzky-Golay (SG)

<sup>3</sup> Moving average

<sup>4</sup> Wavelet signal

داده ها) و مجموعه داده های آزمون شامل ۳۶۱ نقطه داده (۲۰ درصد از داده ها) تقسیم شدند و سپس در فرایند مدل سازی مورد استفاده قرار گرفتند. در شکل ۶، فلوجارت مدل سازی با روش های مورد استفاده در این مطالعه به طور خلاصه شرح داده شده است. همچنین در ادامه این بخش، به طور کامل به تشریح فرایند مدل سازی توسط روش های مورد مطالعه در این مقاله پرداخته شده است.



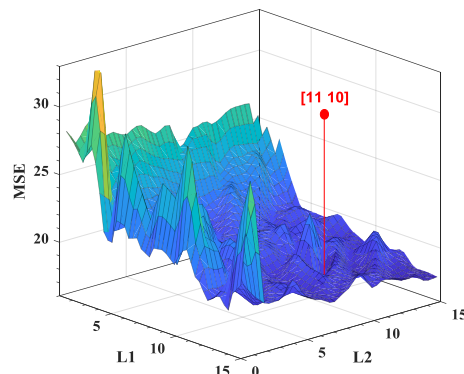
شکل ۶- فلوجارت روش های استفاده شده به منظور تخمین نرخ نفوذ

#### ۴-۱- توسعه مدل شبکه های عصبی پرسپترون چندلایه با الگوریتم پس انتشار خطا (BPMLP)

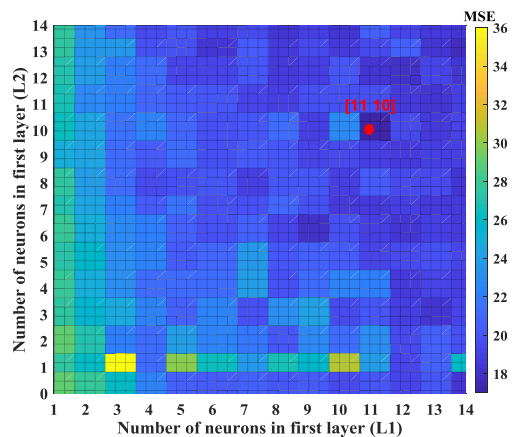
یافتن تعداد بهینه لایه ها، نورون ها و نوع توابع انتقال یکی از چالش های پیش روی مدل سازی با روش شبکه های عصبی می باشد. در این مطالعه، از روش اعتبار سنجی متقابل کی فولد<sup>۱</sup> با تعداد ۱۰ تقسیم برای دستیابی به تعداد بهینه واحد های لایه مخفی استفاده گردید. در این روش اعتبار سنجی، ابتدا کل داده ها به ۱۰ بخش تقسیم میشوند. سپس یک بخش به عنوان مجموعه داده های آزمون انتخاب شده و سایر بخش های باقی مانده، به عنوان مجموعه داده آموزشی در نظر گرفته می شوند. این مراحل تا جایی ادامه میابد که تمام بخش ها به عنوان داده های آزمون انتخاب شوند. در پایان مقدار خطا برای تمام ۱۰ مجموعه تست به طور میانگین محاسبه

<sup>۱</sup> K-fold cross validation

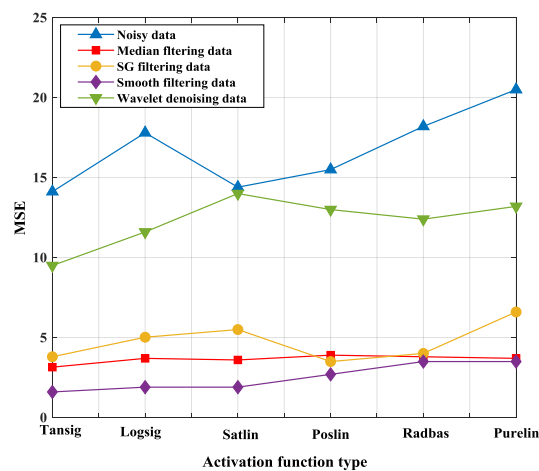
شده و به عنوان عملکرد مربوط به ساختار مورد نظر گزارش میشود. با در نظر گرفتن دو لایه پنهان و حد اکثر ۱۴ نورون در هر لایه، شبکه عصبی در مجموع ۲۱۰۰ مرتبه اجرا گردید و ۲۱۰ مقدار میانگین خطا برای هر ساختار بدست آمد که نتایج آن به صورت نمودار های سه بعدی و کانتوری، به ترتیب در شکل های ۷ و ۸ نمایش داده شده است. بر اساس نتایج بدست آمده، یک ساختار دو لایه ای با ۱۱ نورون در لایه نخست و ۱۰ نورون در لایه دوم، عملکرد مناسبی را برای مدلسازی نرخ نفوذ ارائه میدهد. پس از دستیابی به تعداد واحد های بهینه در لایه پنهان، به ارزیابی تاثیر نوع توابع فعالساز بر دقت خروجی بخش آزمون مدل پرداخته شد. بدین منظور، شش تابع فعالساز مختلف در محیط متلب مورد ارزیابی قرار گرفت. در شکل ۹، نتایج بدست آمده از این ارزیابی نمایش داده شده است. در جدول ۳ نیز، خلاصه ای از جزئیات ساختار نهایی شبکه عصبی مورد نظر برای تخمین نرخ نفوذ، ارائه شده است.



شکل ۷- نمودار سه بعدی مقدار میانگین مربعات خطای بخش آزمون مدل در مقابل تعداد نورون های لایه پنهان



شکل ۸- نمودار کانتور مقدار میانگین مربعات خطای بخش آزمون مدل در مقابل تعداد نورون های لایه پنهان



شکل ۹- عملکرد توابع فعالساز مختلف، بر اساس مقدار خطای بخش آزمون مدل، برای داده های اولیه و کاهش نویز یافته

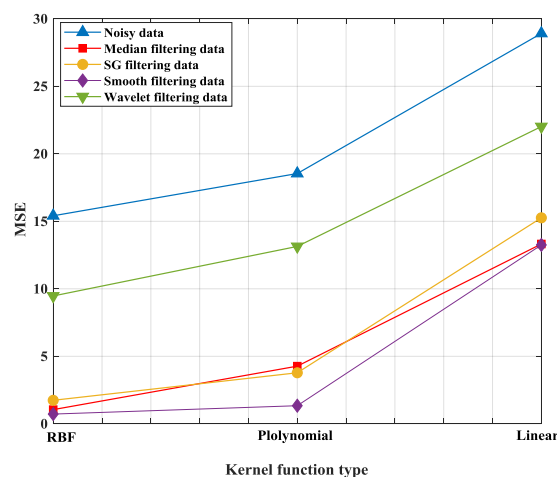
جدول ۳- پارامترهای تنظیم شده برای مدل شبکه های عصبی مصنوعی (MLP) به منظور مدل سازی و تخمین نرخ نفوذ

تنظیمات شبکه	مقدار یا نوع
الگوریتم و تابع آموزشی	پس انتشار خطا با تابع لونیگ-مارکواردت
نوع شبکه	" Feed forward "
تعداد لایه های پنهان	۲
تعداد نورون های لایه پنهان	{۱۱ ۱۰}
تابع فعال ساز لایه پنهان	"sigmoid"
تابع فعالساز لایه خروجی	"purelin"
تعداد ورودی ها	۱۱
تعداد خروجی ها	۱

## ۲-۴- توسعه مدل حد اقل مربعات ماشین بردارهای پشتیبان (LSSVR)

همانطور که در بخش های پیشین اشاره گردید، انتخاب یک کرنل مناسب، تاثیر به سزایی بر دقت تقریب مدل دارد. در این مطالعه، عملکرد سه کرنل گوسی، چند جمله ای و خطی در ساختار مدل LSSVR، مورد بررسی قرار گرفت، که نتایج آن در شکل ۱۰ نمایش داده شده است. همانطور که از نتایج پیداست، کرنل پایه شعاعی یا گوسی دارای بالاترین دقت نسبت به دو کرنل خطی و چند جمله ای میباشد و بهترین تخمین را برای نرخ نفوذ حفاری، نسبت به دو کرنل خطی و چند جمله ای ارائه نموده است. در جدول ۵، پارامترهای تنظیمی مدل LSSVR و کرنل پایه شعاعی برای هر یک از مجموعه داده های پیش پردازش شده و داده های اولیه، ارائه شده است.

در این بخش نیز، نخست آموزش مدل LSSVR توسط ۸۰ درصد از داده ها انجام شد و پس از تنظیم وزنها و پارامترهای مدل، ارزیابی دقت مدل توسط ۲۰ درصد از داده های باقی مانده صورت گرفت. این فرایند برای مجموعه داده های اولیه و مجموعه داده های خروجی از روشهای حذف نویز، تکرار گردید و نتایج در غالب معیار های آماری مختلف بدست آمد.



شکل ۱۰- عملکرد توابع کرنل مختلف، بر اساس مقدار خطای بخش آزمون مدل، برای داده های اولیه و کاهش نویز یافته

جدول ۴- پارامترهای تنظیمی مدل LSSVR با کرنل پایه شعاعی (RBF)

پارامتر $\gamma$	پارامتر تابع کرنل	مجموعه داده
۱۰.۸۷	$10.87\sigma^2 =$	داده های اولیه
۱۴۷.۲۴	$0.34\sigma^2 =$	داده های فیلتر میانه
۳۳۸.۹۲	$0.8\sigma^2 =$	داده های فیلتر ساویترکی گولی
۴۳۸۹.۹۹	$0.56\sigma^2 =$	داده های فیلتر میانگین وزندار متحرک
۱۰.۲۴	$0.36\sigma^2 =$	داده های فیلتر سیگنال ویولت

## ۵- بحث و نتایج

به منظور تجزیه و تحلیل خطای هر یک از مدل ها، آنالیز آماری بوسیله پارامترهای میانگین درصد نسبی خطای مطلق (AAPRE)، خطای میانگین مربعات (MSE)، خطای جذر میانگین مربعات (RMSE) و ضریب تعیین ( $R^2$ ) مورد بررسی قرار گرفت، که نتایج بدست آمده، در جدول ۵ گزارش شده است. این پارامترهای آماری به صورت روابط زیر تعریف شده اند:

میانگین درصد نسبی خطای مطلق (AAPRE):

$$AAPRE = \frac{1}{n} \sum_{i=1}^n \left| \frac{(ROP)_{field} - (ROP)_{predict}}{(ROP)_{field}} \right| \times 100 \quad (17)$$

خطای میانگین مربعات (MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^n \left( (ROP)_{field} - (ROP)_{predict} \right)^2 \quad (18)$$

خطای جذر میانگین مربعات (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left( (ROP)_{field} - (ROP)_{predict} \right)^2} \quad (19)$$

ضریب تعیین ( $R^2$ ):

$$R^2 = 1 - \frac{\sum_{i=1}^n \left( (ROP)_{field} - (ROP)_{predict} \right)^2}{\sum_{i=1}^n \left( (ROP)_{field} - \left( \frac{1}{n} \sum_{i=1}^n (ROP)_{field} \right) \right)^2} \quad (20)$$

جدول ۵- نتایج آماری بدست آمده به منظور ارزیابی خطای مدلها توسعه داده شده در این مطالعه

نام مدل	مجموعه داده	بخش ارزیابی مدل	AAPRE (%)	MSE	RMSE	$R^2$
LSSVR	داده های اولیه	آموزش	۱۳.۷۴	۷.۳۳	۲.۷۰	۰.۸۵
		آزمون	۲۱.۷۲	۱۸.۷۸	۴.۳۳	۰.۶۵
	داده های فیلتر میانه	آموزش	۲.۷۹	۰.۳۳	۰.۵۸	۰.۹۸
		آزمون	۶.۰۷	۱.۳۷	۱.۱۷	۰.۹۶
	داده های فیلتر ساویترکی گولی	آموزش	۴.۲۴	۰.۷۴	۰.۸۶	۰.۹۸
		آزمون	۷.۷۲	۲.۵۰	۱.۵۸	۰.۹۳
	داده های فیلتر میانگین وزندار متحرک	آموزش	۰.۸۹	۰.۰۴	۰.۲۱	۰.۹۹

۰.۹۸	۰.۶۷	۰.۴۵	۲.۶۷	آزمون	داده های فیلتر سیگنال ویولت	BPMLP
۰.۹۰	۲.۰۰	۴.۰۳	۷.۹۴	آموزش		
۰.۷۱	۳.۴۶	۱۲.۰۱	۱۲.۳۲	آزمون		
۰.۷۵	۳.۵۱	۱۲.۳۴	۱۹.۳۷	آموزش	داده های اولیه	
۰.۶۹	۳.۹۳	۱۵.۵۰	۲۵.۵۸	آزمون		
۰.۹۱	۱.۶۰	۲.۵۸	۸.۴۲	آموزش	داده های فیلتر میانه	
۰.۸۹	۲.۰۴	۴.۱۷	۱۰.۱۲	آزمون		
۰.۸۹	۱.۹۴	۳.۷۷	۱۰.۶۷	آموزش	داده های فیلتر ساویترکی گولی	
۰.۸۲	۲.۶۲	۶.۹۰	۱۴.۱۱	آزمون		
۰.۹۱	۱.۶۹	۲.۸۷	۹.۱۵	آموزش	داده های فیلتر میانگین وزندار متحرک	
۰.۸۹	۱.۸۸	۳.۵۵	۹.۴۴	آزمون		
۰.۸۴	۳.۴۹	۱۲.۱۹	۱۶.۹۵	آموزش	داده های فیلتر سیگنال ویولت	
۰.۸۰	۳.۸۴	۱۴.۷۶	۱۸.۲۰	آزمون		

بر اساس نتایج ارائه شده در جدول ۱۰، روش LSSVR با داده های فیلتر میانگین وزندار متحرک بالاترین دقت را در تخمین نرخ نفوذ با میانگین درصد نسبی خطای مطلق ۲.۶۷ و ضریب تعیین ۰.۹۸ برای بخش آزمون مدل، ارائه داده است. پارامترهای میانگین درصد نسبی خطای مطلق و ضریب تعیین، از جمله پارامترها مهم در ارزیابی عملکرد یک مدل میباشند [۴۸]. به منظور تحلیل دقیقتر عملکرد مدلها، یک آنالیز تحلیل خطای گرافیکی مورد استفاده قرار گرفت. شکل ۱۱، مقادیر خروجی هدف را در مقابل مقادیر پیشبینی شده نشان میدهد. بررسی این نمودارها و ارقام ضریب رگرسیون آنها، نشان میدهد که مدل LSSVR با داده های فیلتر میانگین وزندار متحرک، دارای بالاترین دقت در تخمین نرخ نفوذ میباشد و خط برازش شده بر روی داده ها، فاصله بسیار نزدیکی به خط با شیب یک دارد. بر اساس مطالعه ادبیات تحقیق، میتوان بیان داشت که مقدار ضریب رگرسیون میان ۰.۷ تا ۰.۸ نشان دهنده یک مدل قابل قبول است و مقدار ضریب رگرسیون بالاتر از ۰.۹ نشان دهنده عملکرد بسیار رضای کننده از مدل میباشد [۱۶]. نمودارهای جعبه ای\*، یکی دیگر از ابزارهای بسیار کارآمد در تحلیل دقت خطای مدل میباشند [۴۹]. نمودارهای جعبه ای در سال ۱۹۷۷ توسط جان توکی<sup>†</sup> ارائه شده است، که یک روش استاندارد برای نمایش توزیع داده ها براساس پنج شاخص اماری حد اقل، حد اکثر، میانه، چارک اول<sup>‡</sup> و چارک سوم<sup>§</sup> میباشد. از طریق این نمودارها میتوان علاوه بر توزیع داده ها، وجود داده های دورافتاده و تقارن در داده ها را بررسی نمود [۵۰]. به منظور تحلیل دقیق تر خطای مدلها، نمودارهای جعبه ای بخش آموزش و آزمون دو مدل توسعه داده شده در این مطالعه، به ترتیب در شکل های ۱۲ و ۱۳ نمایش داده شده است. همچنین در شکل های ۱۴ و ۱۵، مقادیر پیشبینی شده در دو بخش آموزش و آزمون مدل، با خروجی هدف مقایسه شده است.

## ۶- نتیجه گیری و پیشنهاد

بهینه سازی فرایند حفاری مستقیماً با افزایش نرخ نفوذ ارتباط دارد، چرا که میتوان از این طریق سرعت حفاری را تا حد مجازی افزایش داده و در هزینه های مصرفی صرفه جویی نمود. بهینه سازی فرایند حفاری مستلزم درک ارتباط میان پارامترهای مختلف موثر بر این فرایند میباشد، تا بتوان با یک تخمین دقیق از فاکتور نرخ نفوذ، برنامه ریزی بهتری را برای چاههای آینده فراهم نمود. در

\* Box plot

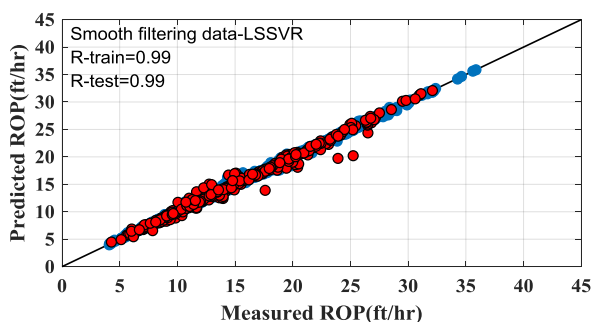
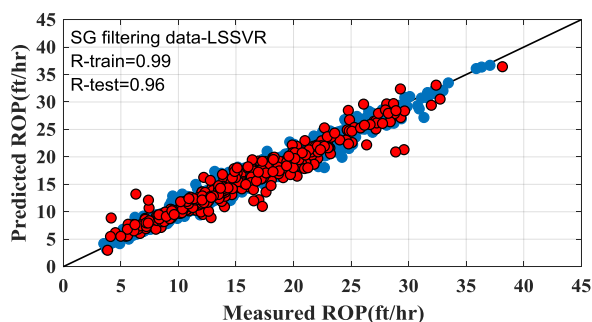
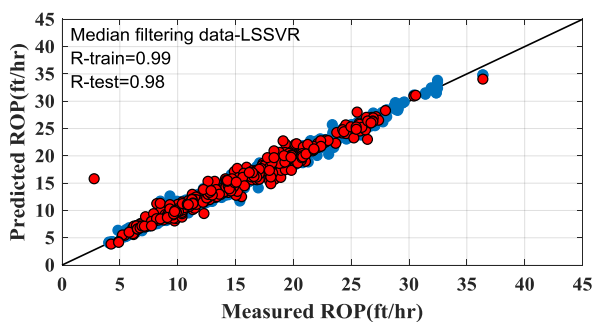
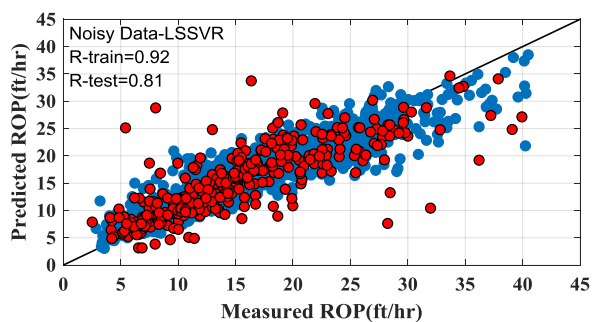
† John Tukey

‡ First quartile

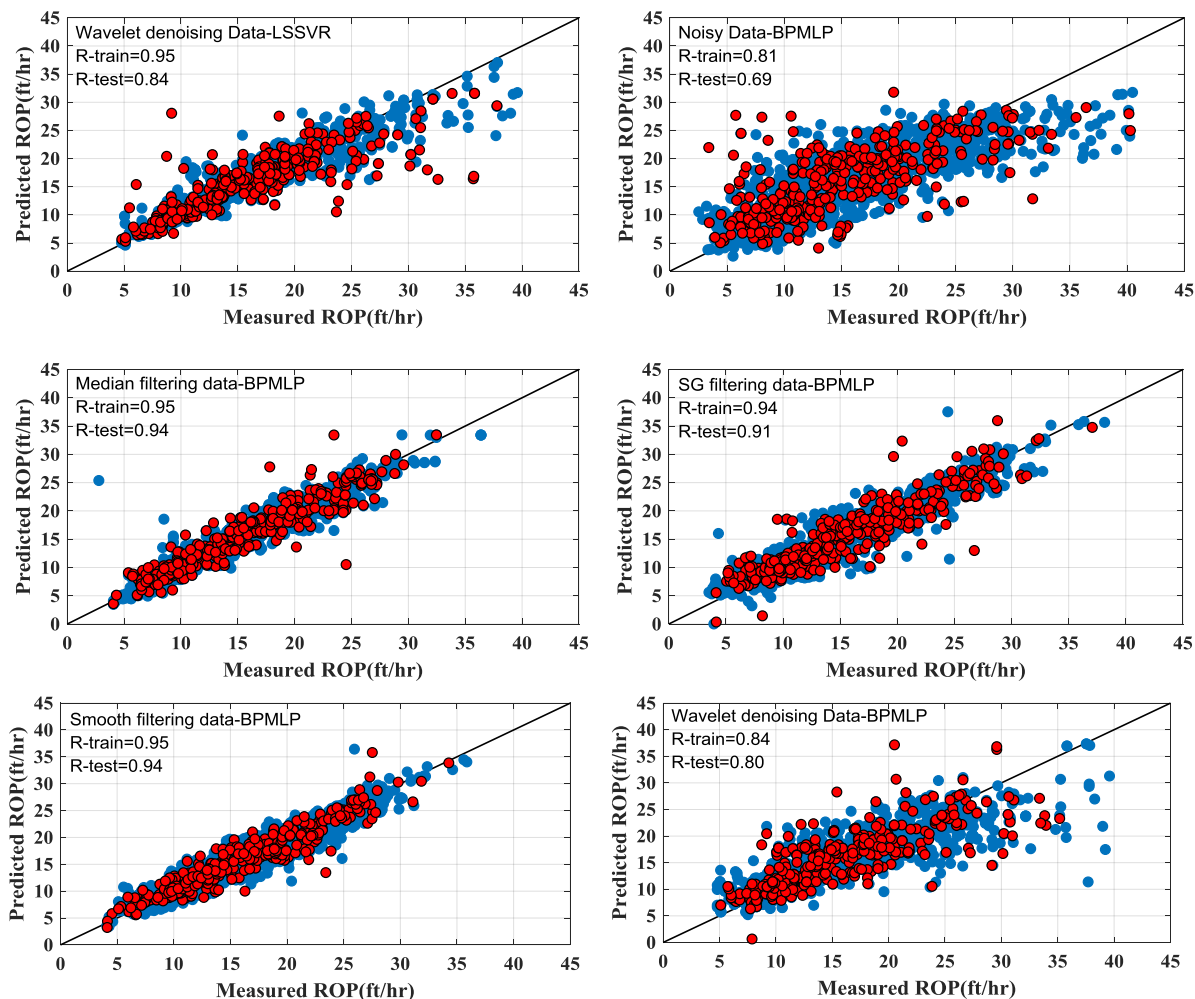
§ Third quartile

این مطالعه، از از دو روش حد اقل مربعات رگرسیون بردار پشتیبان (LSSVR) و شبکه های عصبی پرسپترون چند لایه بهینه شده با الگوریتم پس انتشار خطا (BPMLP)، به منظور مدل سازی و تخمین نرخ نفوذ استفاده گردید. داده های مورد نیاز برای توسعه مدلها، از یک چاه حفاری شده واقع در جنوب غربی ایران جمع اوری شده است. از آنجایی که داده های خام دارای نویز بالایی هستند، کاهش نویز داده ها قبل از فرایند مدل سازی ضروری است، که اینکار توسط چهار فیلتر مختلف مورد ارزیابی قرار گرفت. در ادامه به تشریح نتایج بدست آمده از این مطالعه پرداخته شده است، که میتواند در برنامه ریزی عملیات حفاری چاههای نفت و گاز، مورد توجه قرار گیرد.

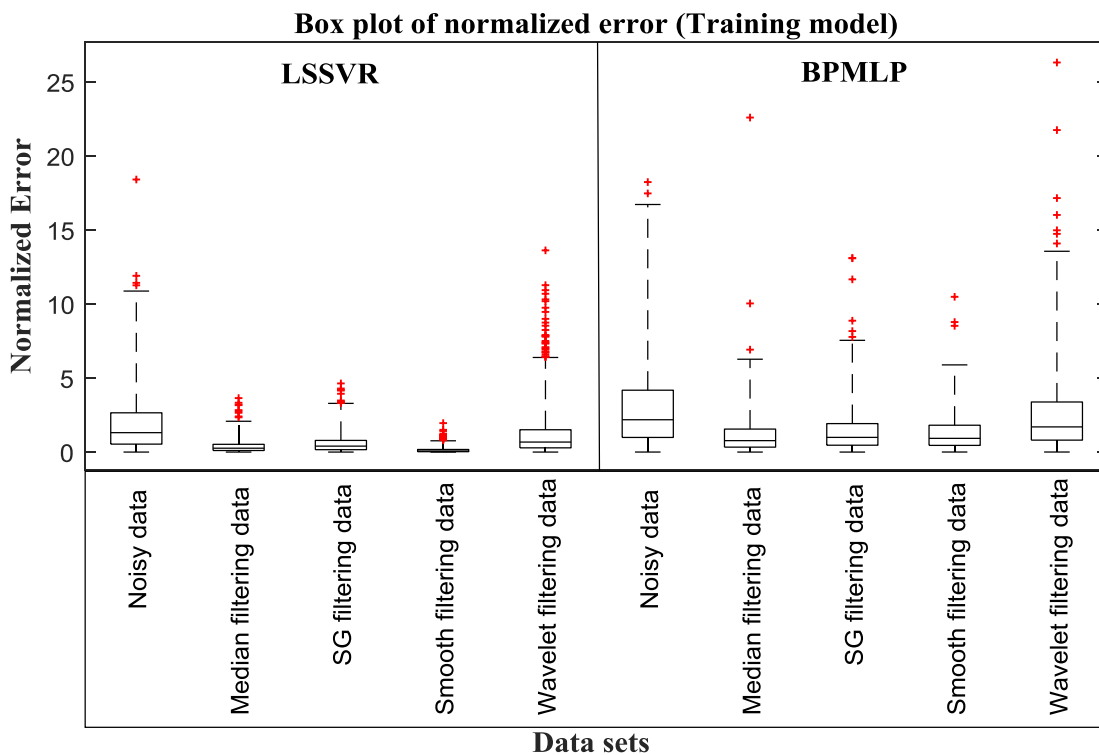
- پیش پردازش داده ها یکی از مهمترین اقداماتی است که میبایست قبل از فرایند مدل سازی انجام گیرد. این موضوع سبب افزایش درجه اطمینان و بهبود فرایند آموزش مدل میشود. بر اساس نتایج ثابت گردید که فیلتر میانگین وزندار متحرک، بهترین عملکرد را در کاهش نویز داده های حفاری و افزایش عملکرد آموزش مدل دارا میباشد.
- انتخاب متغیر های ورودی مناسب یکی دیگر از جنبه های مهم در مدل سازی میباشد، که از طریق آن میتوان ابعاد مساله را کاهش داده و در زمان و هزینه ها صرفه جویی نمود. در این مطالعه، با استفاده از ترکیب شبکه های عصبی و الگوریتم تیرید شبیه سازی شده (SA)، ۱۱ متغیر مستقل (پارامترهای فرایند حفاری) و یک متغیر وابسته (نرخ نفوذ) به عنوان ورودی های موثر مدل انتخاب گردید.
- در بخش توسعه مدل شبکه های عصبی، تابع فعال ساز سیگموئید بهترین عملکرد را به عنوان تابع مستقر در لایه پنهان، برای تخمین نرخ نفوذ حفاری ارائه نمود. علاوه بر این، نشان داده شد که روش کی فولد، یک روش کار آمد برای دستیابی به یک ساختار بهینه از شبکه های عصبی میباشد، که از طریق آن میتوان یک ساختار بهینه را با هدف کاهش ابعاد شبکه و افزایش دقت تخمین، مشخص نمود.
- بر اساس نتایج اماری بدست آمده مشخص گردید که روش حد اقل مربعات بردار پشتیبان با کرنل پایه شعاعی، خطای کمتری نسبت به روش شبکه های عصبی بهینه شده با الگوریتم پس انتشار دارد.



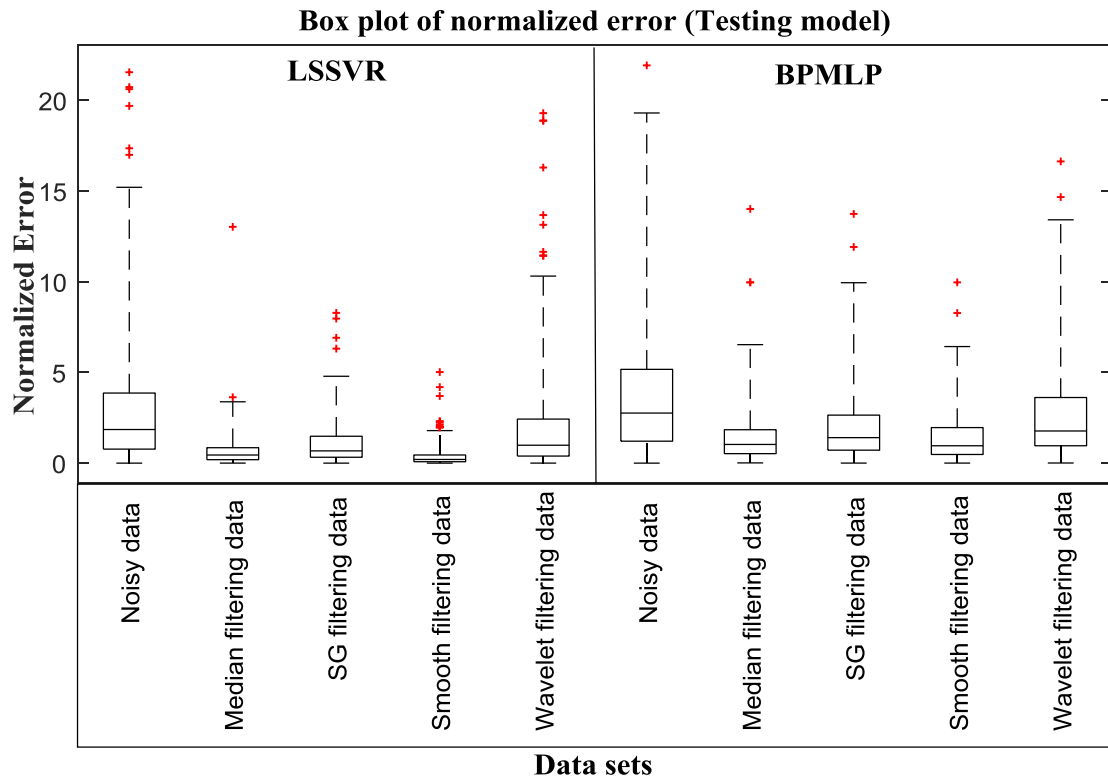




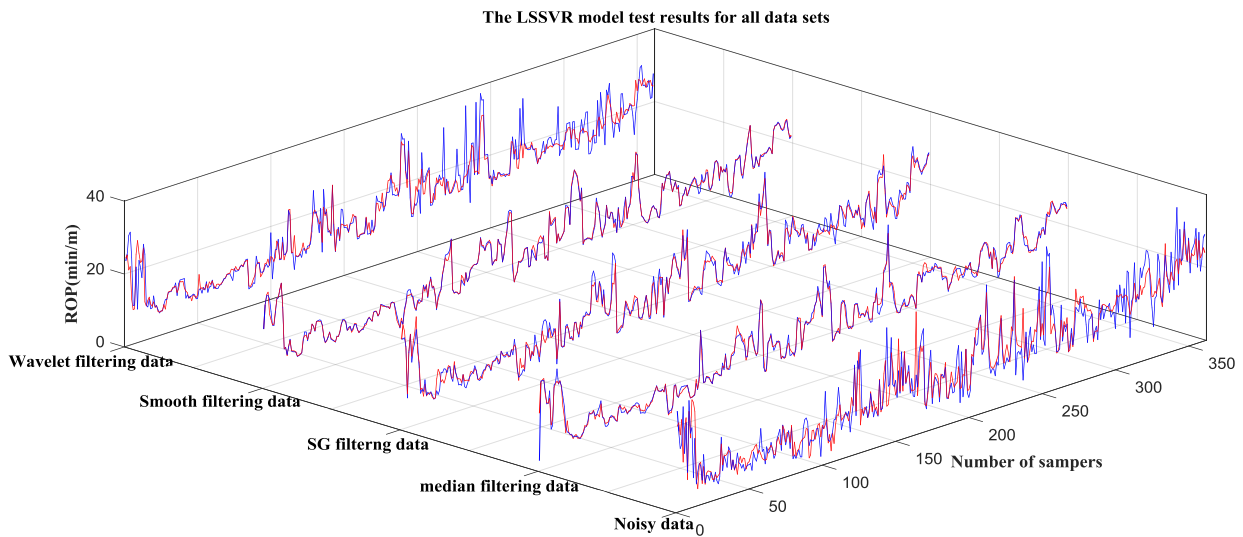
شکل ۱۱- نمودار های رگرسیون مقدار هدف در مقابل مقدار تخمین زده شده از نرخ نفوذ



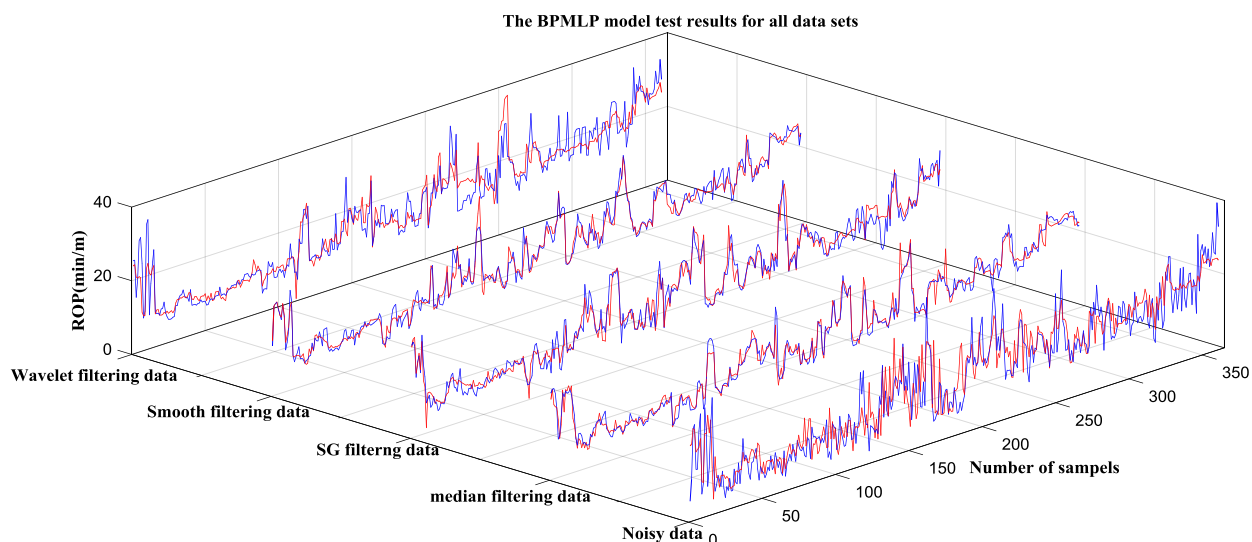
شکل ۱۲- نمودارهای جعبه ای مقدار خطای مطلق بخش آموزش مدل‌های توسعه داده شده



شکل ۱۳- نمودارهای جعبه ای مقدار خطای مطلق بخش آزمون مدل‌های توسعه داده شده



شکل ۱۴- مقایسه نرخ نفوذ واقعی و پیش بینی شده توسط مدل حد اقل مربعات رگرسیون بردار پشتیبان (LSSVR) در بخش آزمون مدل



شکل ۱۵- مقایسه نرخ نفوذ واقعی و پیشبینی شده توسط مدل شبکه های عصبی بهینه شده با الگوریتم پس انتشار (BPMLP) در بخش آزمون مدل

#### ۷- مراجع

1. T. Eren and M. E. Ozbayoglu, "Real time optimization of drilling parameters during drilling operations," in SPE Oil and Gas India Conference and Exhibition 2010, OGIC, 2010, vol. 2010-Janua, doi: 10.2118/129126-ms.
2. M. A. Alum and F. Egbon, "Semi-analytical models on the effect of drilling fluid properties on rate of penetration (ROP)," 2011.
3. S. Irawan, A. Rahman, and S. Tunio, "Optimization of weight on bit during drilling operation based on rate of penetration model," Res. J. Appl. Sci. Eng. Technol., vol. 4, no. 12, pp. 1690–1695, 2012.
4. A. Nascimento, D. Tamas Kutas, A. Elmgerbi, G. Thonhauser, and M. Hugo Mathias, "Mathematical modeling applied to drilling engineering: an application of Bourgoyne and young ROP model to a presalt case study," Math. Probl. Eng., vol. 2015, 2015.
5. E. Brenjkar, E. Biniiaz, and D. Kasra, "Prediction of penetration rate in drilling operations: a comparative study of three neural network forecast methods," J. Pet. Explor. Prod. Technol., no. 0123456789, 2021, doi: 10.1007/s13202-020-01066-1.
6. A. M. Paiaman, M. K. Al-Askari, B. Salmani, B. D. Alanazi, and M. Masihi, "Effect of drilling fluid properties on rate of Penetration," Nafta, vol. 60, no. 3, pp. 129–134, 2009.
7. G. Irma, "Postgraduate Division of the Faculty of Engineering, UNAM and Lorena Berumen, Postgraduate Division of the Faculty of Economy, UNAM, 2009. Optimization model for an oil well drilling program: Mexico case," Oil Gas Bus., 2009.
8. J. R. P. Mendes, T. C. Fonseca, and A. Serapião, "Applying a genetic neuro-model reference adaptive controller in drilling optimization," World oil, pp. 29–36, 2007.
9. E. Wiktorski, A. Kuznetcov, and D. Sui, "ROP Optimization and Modeling in Directional Drilling Process," SPE Bergen One Day Semin., no. April, Apr. 2017, doi: 10.2118/185909-ms.
10. [ F. Akgun, "Drilling rate at the technical limit," Int. J. Pet. Sci. Technol., vol. 1, no. 1, pp. 99–119, 2007.
11. E. M. Galle and H. B. Woods, "Best Constant Weight and Rotary Speed for rotary Rock Bits," Drilling and Production Practice. American Petroleum Institute, p. 26, Jan. 01, 1963, Accessed: Jun. 18, 2019. [Online]. Available: <https://www.onepetro.org/conference-paper/API-63-048>.
12. M. G. Bingham, "A new approach to interpreting rock drillability," Pet. Publ. Co., 1965.
13. W. C. Maurer, "The" perfect-cleaning" theory of rotary drilling," J. Pet. Technol., vol. 14, no. 11,

pp. 1–270, 1962.

14.A. T. Bourgoyne and F. S. Young, “A Multiple Regression Approach to Optimal Drilling and Abnormal Pressure Detection,” *Soc. Pet. Eng. J.*, vol. 14, no. 04, pp. 371–384, Aug. 1974, doi: 10.2118/4238-PA.

15.T. M. Warren, “Drilling model for soft-formation bits,” *J. Pet. Technol.*, vol. 33, no. 06, pp. 963–970, 1981.

16.X. Shi, G. Liu, X. Gong, J. Zhang, J. Wang, and H. Zhang, “An Efficient Approach for Real-Time Prediction of Rate of Penetration in Offshore Drilling,” *Math. Probl. Eng.*, vol. 2016, pp. 1–13, 2016, doi: 10.1155/2016/3575380.

17.R. Y. Wang, V. C. Storey, and C. P. Firth, “A Framework for Analysis of Data Quality Research,” *IEEE Trans. Knowl. Data Eng.*, vol. 7, no. 4, pp. 623–640, 1995, doi: 10.1109/69.404034.

18.D. F. Nettleton, A. Orriols-Puig, and A. Fornells, “A study of the effect of different types of noise on the precision of supervised learning techniques,” *Artif. Intell. Rev.*, vol. 33, no. 4, pp. 275–306, Apr. 2010, doi: 10.1007/s10462-010-9156-z.

19.J. R. Quinlan, “The effect of noise on concept learning,” *Mach. Learn. An Artif. Intell. approach*, vol. 2, pp. 149–166, 1986.

20.X. Zhu, X. Wu, and Q. Chen, “Eliminating Class Noise in Large Datasets,” 2003.

21.S. e Z. Lashari, A. Takbiri-Borujeni, E. Fathi, T. Sun, R. Rahmani, and M. Khazaeli, “Drilling performance monitoring and optimization: a data-driven approach,” *J. Pet. Explor. Prod. Technol.*, vol. 9, no. 4, pp. 2747–2756, Dec. 2019, doi: 10.1007/s13202-019-0657-2.

22.K. Al-Azani, S. Elkatatny, A. Ali, E. Ramadan, and A. Abdulaheem, “Cutting concentration prediction in horizontal and deviated wells using artificial intelligence techniques,” *J. Pet. Explor. Prod. Technol.*, vol. 9, no. 4, pp. 2769–2779, Dec. 2019, doi: 10.1007/s13202-019-0672-3.

23.F. Cordoni, “A comparison of modern deep neural network architectures for energy spot price forecasting,” *Digit. Financ.*, vol. 2, no. 3–4, pp. 189–210, Dec. 2020, doi: 10.1007/s42521-020-00022-2.

24.S. Panghal and M. Kumar, “Multilayer Perceptron and Chebyshev Polynomials Based Neural Network for Solving Emden–Fowler Type Initial Value Problems,” *Int. J. Appl. Comput. Math.*, vol. 6, no. 6, p. 157, Dec. 2020, doi: 10.1007/s40819-020-00914-2.

25.C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1023/A:1022627411411.

26.M. S. Ahmad, S. M. Adnan, S. Zaidi, and P. Bhargava, “A novel support vector regression (SVR) model for the prediction of splice strength of the unconfined beam specimens,” *Constr. Build. Mater.*, vol. 248, p. 118475, 2020, doi: 10.1016/j.conbuildmat.2020.118475.

27.K. Cheng, Z. Lu, Y. Wei, Y. Shi, and Y. Zhou, “Mixed kernel function support vector regression for global sensitivity analysis,” *Mech. Syst. Signal Process.*, vol. 96, pp. 201–214, 2017, doi: 10.1016/j.ymsp.2017.04.014.

28.H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, “Support vector regression machines,” in *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference*, 1997, vol. 9, pp. 155–161, Accessed: Jun. 18, 2019. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.5909>.

29.M. S. AL-Musaylh, R. C. Deo, Y. Li, and J. F. Adamowski, “Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting,” *Appl. Energy*, vol. 217, pp. 422–439, May 2018, doi: 10.1016/j.apenergy.2018.02.140.

- 30.W.-C. Hong, "Chaotic particle swarm optimization algorithm in a support vector regression electric load forecasting model," *Energy Convers. Manag.*, vol. 50, no. 1, pp. 105–117, Jan. 2009, doi: 10.1016/j.enconman.2008.08.031.
- 31.E. El-Sebakhy, T. Sheltami, S. Al-Bokhitan, Y. Shaaban, P. Raharja, and Y. Khaeruzzaman, "Support Vector Machines Framework for Predicting the PVT Properties of Crude-Oil Systems," *Mar.* 2007, doi: 10.2523/105698-MS.
- 32.J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999, doi: 10.1023/A:1018628609742.
- 33.H. W. Kuhn and A. W. Tucker, "Nonlinear programming," *Traces Emerg. Nonlinear Program.*, no. x, pp. 247–258, 2014, doi: 10.1007/978-3-0348-0439-4\_11.
- 34.F. Chauchard, R. Cogdill, S. Roussel, J. M. Roger, and V. Bellon-Maurel, "Application of LS-SVM to non-linear phenomena in NIR spectroscopy: Development of a robust and portable sensor for acidity prediction in grapes," *Chemom. Intell. Lab. Syst.*, vol. 71, no. 2, pp. 141–150, 2004, doi: 10.1016/j.chemolab.2004.01.003.
- 35.Y. Li, X. Shao, and W. Cai, "A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples," *Talanta*, vol. 72, no. 1, pp. 217–222, 2007, doi: 10.1016/j.talanta.2006.10.022.
- 36.S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," *Science* (80-. ), vol. 220, no. 4598, pp. 671–680, May 1983, doi: 10.1126/science.220.4598.671.
- 37.L. Ingber, "Simulated annealing: Practice versus theory," *Math. Comput. Model.*, vol. 18, no. 11, pp. 29–57, Dec. 1993, doi: 10.1016/0895-7177(93)90204-C.
- 38.L. Ingber, "Very fast simulated re-annealing," *Math. Comput. Model.*, vol. 12, no. 8, pp. 967–973, 1989, doi: 10.1016/0895-7177(89)90202-1.
- 39.S. B. Ashrafi, M. Anemangely, M. Sabah, and M. J. Ameri, "Application of hybrid artificial neural networks for predicting rate of penetration (ROP): A case study from Marun oil field," *J. Pet. Sci. Eng.*, vol. 175, no. December 2018, pp. 604–623, Apr. 2019, doi: 10.1016/j.petrol.2018.12.013.
- 40.G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.
- 41.M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, 2004, doi: 10.1109/TPAMI.2004.71.
- 42.E. Davoudi and B. Vaferi, "Applying artificial neural networks for systematic estimation of degree of fouling in heat exchangers," *Chem. Eng. Res. Des.*, vol. 130, pp. 138–153, 2018, doi: 10.1016/j.cherd.2017.12.017.
- 43.E. Gholami, B. Vaferi, and M. A. Ariana, "Prediction of viscosity of several alumina-based nanofluids using various artificial intelligence paradigms - Comparison with experimental data and empirical correlations," *Powder Technol.*, vol. 323, pp. 495–506, 2018, doi: 10.1016/j.powtec.2017.10.038.
- 44.T. S. Huang, G. J. Yang, and G. Y. Tang, "A Fast Two-Dimensional Median Filtering Algorithm," *IEEE Trans. Acoust.*, vol. 27, no. 1, pp. 13–18, 1979, doi: 10.1109/TASSP.1979.1163188.
- 45.A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.
- 46.A. C. Bovik and S. T. Acton, "Basic Linear Filtering with Application to Image Enhancement," in *The Essential Guide to Image Processing*, Elsevier, 2009, pp. 225–239.
- 47.E. Sugawara and H. Nikaido, *Wavelets and Statistics*, vol. 103, no. 12. New York, NY: Springer

New York, 1995.

48.A. Karkevandi-Talkhoonchah, S. Hajirezaie, A. Hemmati-Sarapardeh, M. M. Husein, K. Karan, and M. Sharifi, "Application of adaptive neuro fuzzy interface system optimized with evolutionary algorithms for modeling CO<sub>2</sub>-crude oil minimum miscibility pressure," *Fuel*, vol. 205, pp. 34–45, Oct. 2017, doi: 10.1016/j.fuel.2017.05.026.

49.C. Hegde, H. Daigle, H. Millwater, and K. Gray, "Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models," *J. Pet. Sci. Eng.*, vol. 159, pp. 295–306, Nov. 2017, doi: 10.1016/j.petrol.2017.09.020.

50.S. H. Kramer and R. Rosenthal, "Meta-analytic Research Synthesis," in *Comprehensive Clinical Psychology*, Elsevier, 1998, pp. 351–368.

# Investigating the Effect of Different Noise Reduction Filters on the Rate of Penetration Approximation by Machine Learning Based Models

**Ehsan Brenjkar**

Master's degree of Petroleum Engineering, Science and Research Branch, Islamic Azad University, Tehran-Iran

## Abstract

Accurate modeling and estimation of drilling rate of penetration (ROP) is a prelude to better planning to reduce drilling time and control costs. However, estimating this key factor is not easy and the main reason is the complex relationship between penetration rate and drilling variables. On the other hand, the presence of noise in the data increases the training time of the models and drastically reduces their accuracy. In this study, the multilayer perceptron neural network (MLP) method and the least squares support vector regression (LSSVR) method along with four data noise reduction filters have been used to estimate the penetration rate. The data used in order to feed the models, collected from mud logging unit (MLU) and final report of a drilled well that located in south-west Iran. After the feature selection process with simulated annealing (SA) optimization algorithm and artificial neural networks, eleven variables were selected from all of the variables. Out-of-range points were then removed and the overall noise of the data was reduced by median filter, savitzky-golay, moving average and signal wavelet. Finally, the performance of the developed models and filters used were evaluated and compared by different statistical indicators and it was proved that the LSSVR method with radial basis kernel and output data from moving average filter shows the best results in the both training and testing section of the model

**Keywords:** Drilling rate of penetration, Multilayer perceptron neural network, Least square support vector regression, Feature selection, Noise reduction, Simulated annealing algorithm.