EFFECTIVE POROSITY PREDICTION FROM WELL LOG DATA USING SUPPORT VECTOR MACHINE (SVM)

Sudarmaji Saroji[1*] , Muhammad Fajrul Haqqi[2], Suryo Prakoso [3]

[1]Department of Physics, Universitas Gadjah Mada 55281, Indonesia
[2] Department of Computer Science, Universitas Gadjah Mada 55281, Indonesia
[3] Petroleum Engineering, Faculty of Earth and Energy Technology, Universitas Trisakti, Indonesia
e-mail : ajisaroji@ugm.ac.id

Abstrak. Algoritma *Support Vector Machine* (SVM) merupakan metode pembelajaran mesin yang dikenal memiliki akurasi tinggi dan efisiensi komputasi yang baik untuk aplikasi prediksi dan klasifikasi. Dalam studi ini, SVM diterapkan untuk memprediksi prosistas efektif dari data log sumur. Model prediksi dioptimalkan menggunakan modul *GridsearhCV* (GS CV) dan telah diterapkan pada tujuh sumur dari lapangan 'Mentari', Indonesia. Dilakukan enam variasi kombinasi pelatihan dan pengujian untuk mengevaluasi performa prediksi porositas efektif. Hasil terbaik diperoleh dari konfigurasi empat sumur pelatihan dan tiga sumur pengujian, dengan akurasi mencapai 71% dan waktu latihan 1,98 detik. Hasil menunjukkan bahwa peningkatan jumlah data pelatihan dapat meningkatkan akurasi, meskipun dengan waktu komputasi yang lebih lama. Studi ini mengonfirmasi bahwa SVM memiliki kemampuan prediksi porositas efektif yang baik serta berpotensi menjadi alat bantu dalam mendukung dan menyederhanakan proses interpretasi geologi, khususnya pada tahap awal analisis dengan ketersediaan data pelatihan yang mencukupi.
Kata Kunci: petrofisika; pembelajaran mesin; porositas efektif; prediksi; *Support Vector Machine* (SVM)

Abstract.
Support Vector Machine (SVM) algorithm is a machine learning method renowned for its high accuracy and computational efficiency in prediction and classification tasks. In this study, SVM was applied to predict effective porosity from well log data. The prediction model was optimized using GridsearhCV (GS CV) module and tested on seven wells from the 'Mentari' field, Indonesia. Six variations of training-testing configuration were evaluated to assess the prediction performance. The best ere acjieved using four training wells and three testing wells, yielding an accuracy of 71% with training time of 1.98 seconds. The analysis revealed that increasing the volume of training data improves accuracy, albeit with longer computational time. This study confirms that SVM demonstrates strong predictive capability for effective porosity and has the potential to serve as a supporting tool in simplifying geological interpretation, particularly during the initial analysis stage when sufficient training data is available.
***Key word**: petrophysics; machine learning; effective porosity; prediction; Support Vector Machine (SVM)*

INTRODUCTION

Predicting effective porosity from well log data using statistical methods has been extensively studied in recent years. The initial approach, the multivariate linear regression (MLR) method, was introduced by Wendt et al. in 1986. Subsequent studies expanded upon this by incorporating multivariate non-linear regression (MNLR), which improved prediction accuracy and enabled the estimation of a broader range of data. These statistical techniques have been continously refined to enhance the reliability of well log data analysis, ensuring closer alignment between predictions and actual field measurements. One of the key objectives in this field of study is to accurately estimate effective porosity using well log data.

Recent advancements in predictive modeling have led to the development of intelligent systems capable of autonomous learning. Machine learning techniques utilize computational methods to extract knowledge from data through supervised or unsupervised learning processes. The rapid evolution of machine learning algorithms and artificial intelligence (AI) methodologies has been driven by the

DOI : http://dx.doi.org/10.12962/j25023659.v11i1.1936

Artikel Masuk : 31-07-2024 Artikel Diulas 05-08-2024 Artikel Diterima : 21-04-2025

availability of large-scale datasets (big data) and the increasing accessibility of low-cost computing resources. Among various machine learning models, Support Vector Machine (SVM) has demonstrated high effectiveness in predicting effective porosity with improved accuracy.

The SVM method was originally introduced by V.N. Vapnik and A.Y. Chervonenkis in 1964 as part of the "Generalized Portrait Method." The mathematical formulation of this approach was later detailed by Vapnik and Lerner (1963). SVM is a supervised machine learning technique that identifies the optimal hyperplane to separate data clusters into two or more classes while maximizing the margin between the hyperplane and the nearest data points, known as support vectors. In this study, Support Vector Regression (SVR), a regression-based variant of the SVM algorithm, was employed (Vapnik and Lerner,1963). SVR operates by segmenting the dataset into two classes using a hyperplane and establishing boundary lines that act as threshold values to encompass as many data points as possible (Che and Wang, 2014. Data points located near these threshold boundaries are considered potential support vectors, with the most comprehensive coverage yielding the most effective model. The optimization process within this study aimed to minimize the margin (ε) and soft margin or slack variable (ξ) between the hyperplane and observed data. To optimize the model, the GridSearchCV module was utilized to determine the best hyperparameters, such as C and epsilon, from a predefined grid (Figure 1).

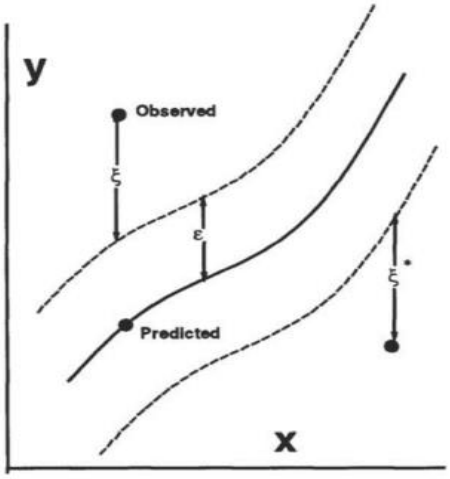


Figure 1. The parameters for the support vector regression (Aizerman and Braverman,1964)

Several researchers have utilized the Support Vector Machine (SVM) algorithm for well log analysis and processing (Unpingco, 2019) . Hall (2016) applied SVM for facies classification based on well log data, while Halotel et al. (2020) implemented SVM to automate the facies classification process. Saroji et al. (2021) applied SVM for lithofacies classification using multi-well log data from an Indonesian oil field. Haqqi et al. (2023) used the XGBoost for effective porosity estimation. In a recent study, Nugroho et al. (2024) used xgboost and random forest to estimate effective porosity and permeability on well log data in fajar field, south sumatera basin,indonesia. This study aims to estimate effective porosity from well log data in the Mentari Field, Indonesia, using SVM method, while also assessing its validation and scalability. SVM algorithm is selected due to its stability and reliability in handling regression problems.

RESEARCH METHOD

This study utilizes well log data from seven wells (SM-01A, SM-03A, SM-10A, SM-11A, SM-22A, SM-24A and SM-25A) in the Mentari field, South sumatera Basin, Indonesia. The primary logs used for predicting effective porosity(PHIE) include gamma ray (GR) , density (RHOB), resistivity

DOI : http://dx.doi.org/10.12962/j25023659.v11i1.1936
Artikel Masuk : 31-07-2024 Artikel Diulas 05-08-2024 Artikel Diterima : 21-04-2025

(ILD) and neutron porosity (NPHI) (Rider, 2002). SVM implementation was evaluated using six distinct training-validation data configurations (Table 1).

Table 1. Variations in the distribution of sample data groups for training and evalution

Variation	SM-25A	SM-01A	SM-03A	SM-10A	SM-11A	SM-22A	SM-24A	Training	Testing
var 1	**5%**	**95%**						452	24632
var 2	**20%**		**80%**					4481	20603
var 3	**33%**			**67%**				11468	17715
var 4	**50%**				**50%**			17399	13616
var 5	**69%**					**31%**		17399	7685
var 6	**83%**						**17%**	20017	5067
n_data	452	4029	2888	4099	5931	2618	5067		

As addition for feature input data, we generated and calculated onic log values (DT) and shale volume from others well log data. We used the following Gardner's Relationship for sonic log data generation from density log data (Gardner et al. 1974)

$$\rho = \alpha V_p^{\,0.25} \tag{1}$$

The equation (1) represents the relationship between the velocity of the P wave (V_p) and the density (ρ). The conversion constant value (α) is 0.31 when using velocity in meters per second (m/s) or 0.23 when expressed in feet per second (ft/s).

The shale volume log is derived from the interpretation of gamma-ray log data, which indicates that gamma-ray values are higher in shale formations compared to sandstone lithology. The quantitative determination of shale volume is based on the assumption that the maximum gamma-ray log value within each well represents 100% shale, while the minimum value corresponds to the absence of shale, indicating sandstone lithology (Asquith and Gibson, 1982). The midpoint of this 0-100% scale serves as the boundary between these two lithologies and can be determined using a simple linear calculation, as outlined below.

$$V_{sh} = \frac{GR_{log} - GR_{min}}{GR_{max} - GR_{min}} \tag{2}$$

The formulation pertains to rock formation, which is categorized into clastic (e.g., Cibulakan and Talang Akar formations) and carbonate (e.g., Baturaja formation). Based on these two categories, each sample is classified into different groups, as shown in Table 2 and Figure 2.

Table 2. Classification of sample data according to shale volume and lithological formations.

Criteria	Label
Vsh < 0.5 and not Baturaja Fm	Low GR (clastic)
Vsh > 0.5 and not Baturaja Fm	High GR (clastic)
Vsh < 0.5 and Baturaja Fm	Low GR (carbonate)
Vsh > 0.5 and Baturaja Fm	High GR (carbonate)

DOI : http://dx.doi.org/10.12962/j25023659.v11i1.1936

Artikel Masuk : 31-07-2024 Artikel Diulas 05-08-2024 Artikel Diterima : 21-04-2025

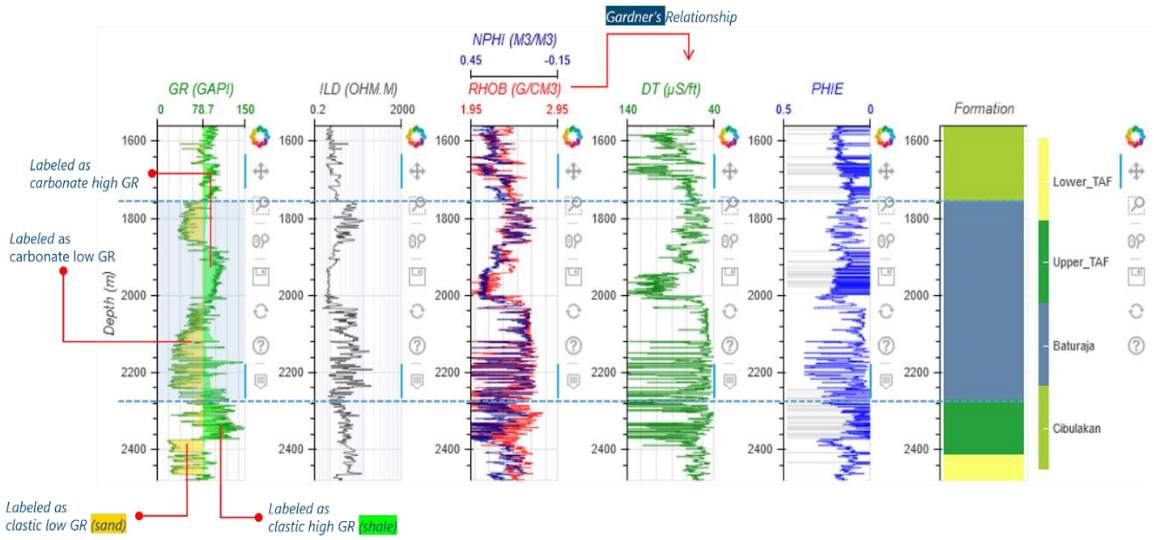


Figure 2. Interpretated well log data of SMR-24A.

Figure 3 illustrates six variables that were analyzed using the SVM model to predict the effective porosity log. These variables (ILD, NPHI, DT, GR, RHOB, and VSH) are numerical data with the correlations between each feature depicted in the figure. Additionally, another variable is based on shale volume and rock formation.

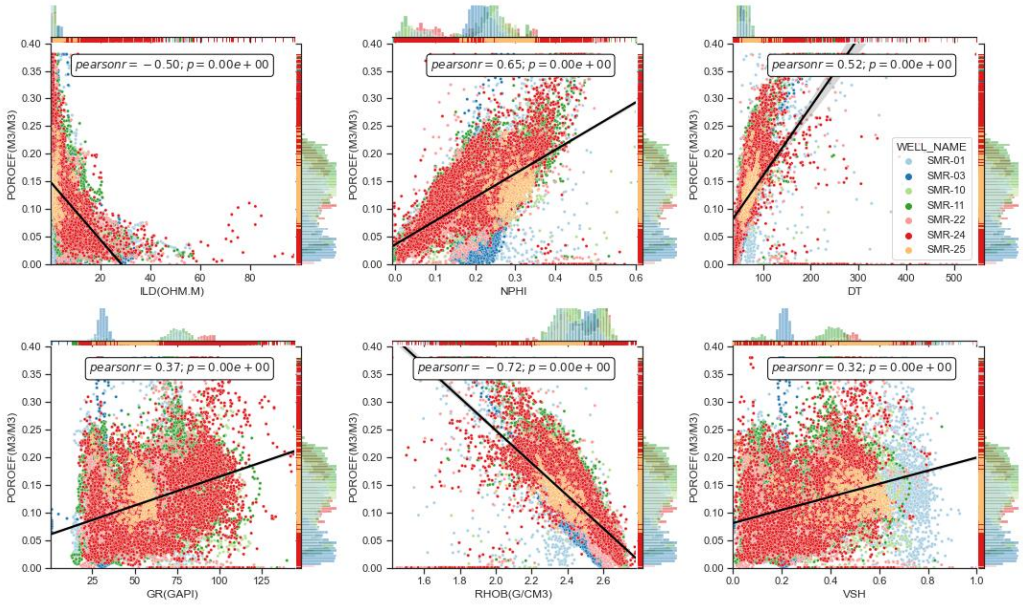


Figure 3. the cross-correlation among six distinct features derived from the complete well log dataset.

The basic idea of SVM (regression) algorithm to calculate flatness function defined by the following equation (Drucker et al., 1996).

$$f(x) = \langle w, x \rangle + b \qquad w \in X, b \in R \tag{3}$$

Where w is weight, x is support vector, and b is coefficient. The equation aims to minimize w value and can be calculated by minimizing to the form $||w||^2$. This can be solved by convex optimization using following equation.

$$\frac{1}{2}||w||^2 + C\sum_{i=0}^{\ell}(\xi_i + \xi_i^*) \tag{4}$$

With the provision of

DOI : http://dx.doi.org/10.12962/j25023659.v11i1.1936

Artikel Masuk : 31-07-2024 Artikel Diulas 05-08-2024 Artikel Diterima : 21-04-2025

$$y_i - \langle w, x\rangle - b \leq \varepsilon + \xi_i \quad (5)$$
$$\langle w, x\rangle + b - y_i \leq \varepsilon + \xi_i \quad (6)$$
$$\xi_i \xi_i^* \geq 0 \quad (7)$$

Eq.(…) is evaluated by a loss function where the prediction can be subject to a penalty or not. The simplest loss function is the ε-insensitive loss function which has the following equation.

$$L_\varepsilon(y) = \begin{cases} 0, & for\ |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, & for\ others \end{cases} \quad (8)$$

All deviation limits greater than ε will be penalized by C, where C is a positive constant value. The optimal solution to optimize hyperplane is using Lagrange multiplier so that the equation w can be simplified to

$$w = (a_i - a_i^*)x_i \quad (9)$$

So, the optimization on the hyperplane can be written to.

$$f(x) = (a_i - a_i^*)\langle x_i, x\rangle + b \quad (10)$$

With $a_i - a_i^*$ is Lagrange multiplier calculation from weights. In this research, SVR algorithm implemented on non-linear dataset so that the values of x_i and x transformed into a space feature in high dimensions by mapping the vectors x_i and x into the kernel function. Finally, the optimization equation becomes.

$$f(x) = (a_i - a_i^*)K\langle x_i, x\rangle + b \quad (11)$$

With K is kernel functions. The principle of SVR algorithm determined by the type of kernel function to be used and the kernel parameter settings (12). In this study, the authors use the Radial Basis Function (RBF) kernel with the following equation.

$$K\langle x_i, x\rangle = exp\frac{\|x_i - x\|^2}{2\sigma^2} \quad (12)$$

Based on these equations, there are three important hyperparameters that must be initialized during optimization such as constraint violation (C), epsilon, and gamma. The value of C shows the tradeoff of the complexity in decision-making rules and calculated loss function. Epsilon shows the smoothness effect of SVR and generalizability and complexity of the models. And gamma related to minimize the occurrences of underfitting and overfitting. In this studied, authors just tunning C and epsilon value using GridSearchCV module

Table 3. The pseudo code of SVM algorithm

Algorithm. Support Vector Machine (Regression)
1. Initialize $a_i = 0$, $a_i^* = 0$, and matrix calculate with $R_{ij} = (K(x_i, x) + \lambda^2)$ for $i, j = 1, \ldots n$
2. For every training data calculate:
a. $E_i = y_i - \sum_{j=1}^{1}(a_j^* - a_j)\, R_{ij}$
b. $\delta a_i^* = min\{max[\gamma(E_i - \epsilon), -a_i^*], C - a_i^*\}$ $\delta a_i = min\{max[\gamma(E_i - \epsilon), -a_i], C - a_i\}$
c. $a_i^* = a_i^* + \delta a_i^*$ and $a_i = a_i + \delta a_i$
3. Calculate using regression function: $f(x) = (a_i - a_i^*)K\langle x_i, x\rangle + \lambda^2$
4. with K is RBF (Radial Basis Function) kernel function: $K\langle x_j, x\rangle = exp\frac{\|x_i - x\|^2}{2\sigma^2}$

DOI : http://dx.doi.org/10.12962/j25023659.v11i1.1936

Artikel Masuk : 31-07-2024 Artikel Diulas 05-08-2024 Artikel Diterima : 21-04-2025

RESULTS AND DISCUSSION

The Support Vector Machine (SVM) algorithm requires the configuration of several hyperparameters before training, a process known as hyperparameter tuning. The main objective of this process is to tailor the model to the variability inherent in the dataset—well log data in this case. Proper hyperparameter tuning is essential for enhancing the model's ability to predict effective porosity accurately.

Hyperparameter tuning was performed manually using the GridSearchCV (GS) module. This module systematically evaluates all possible combinations of hyperparameter values within the predefined grid and identifies the set that yields the optimal performance. The results for various hyperparameters of the SVM algorithm are presented in Table 4.

Table 4. Hyperparameter value tunning using GridSearchCV (GS)

Hyperparameter	grid	Default	GS-SVM
C	[0.1, 1, 10, 100]	1.0	0.1
epsilon	[0.01, 0.02, 0.04, 0.06, 0.08, 0.1]	0.1	0.02

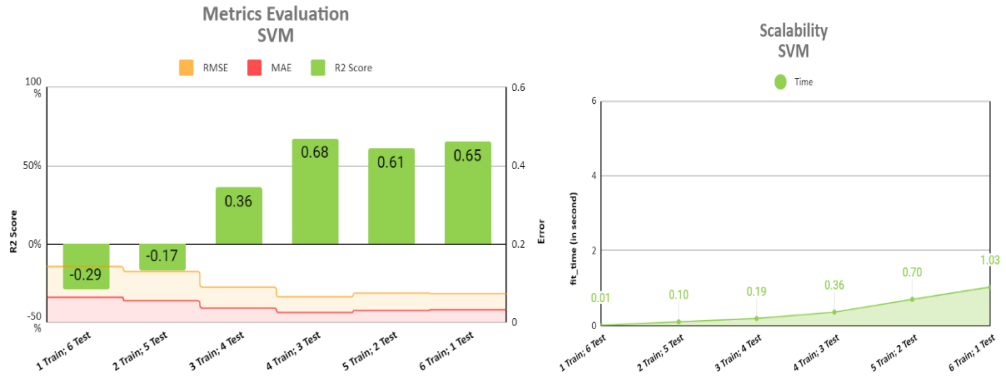
To assess performance of the GridSearchCV in the SVM model, three evaluation metrics were used including: the R^2 score as the scoring metric, and RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) as error metrics. These three indicators are defined by the following equations (Karunasingha, 2022).

$$R^2(y,p) = 1 - \frac{\sum_{i=1}^{n}(y_i - p_i)^2}{\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)^2} \quad (13)$$

$$e_{rmse}(y,p) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - p_i)^2} \quad (14)$$

$$R^2(y,p) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - p_i)^2 \quad (15)$$

In this context, y_i represents the actual effective porosity value, while p_i denotes the predicted effective porosity obtained from the application of the SVM method. The results of the model evaluation, based on the three indicators and their respective training times, along with six variations of the training-evaluation process, are presented in the following figures



(a) (b)

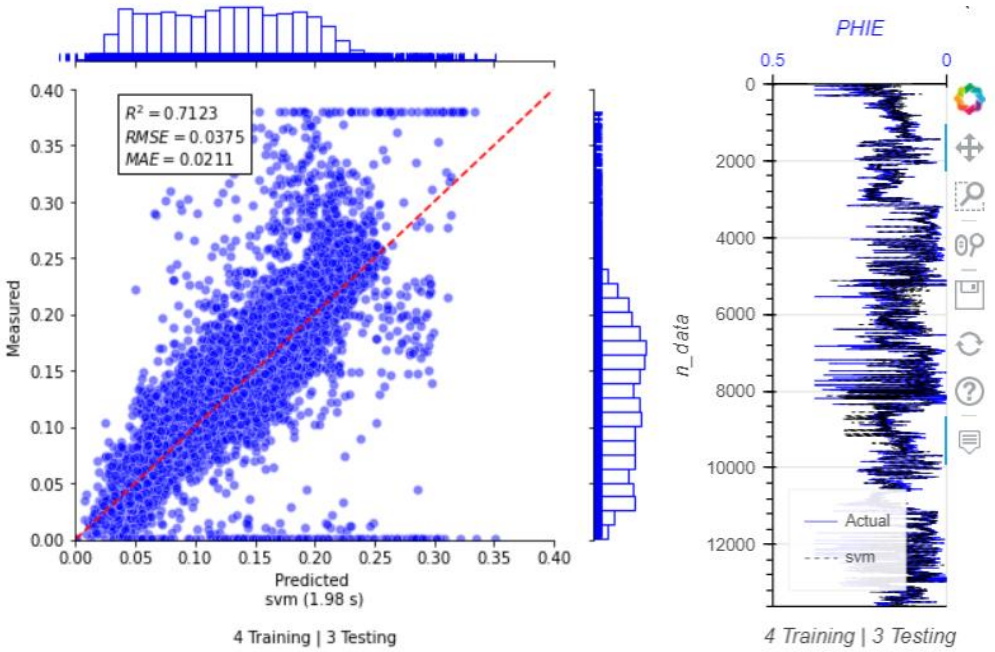
Figure 4. The R2 score value from the model evaluation using 3 indicators (a) and consuming training time (b)

DOI : http://dx.doi.org/10.12962/j25023659.v11i1.1936

Artikel Masuk : 31-07-2024 Artikel Diulas 05-08-2024 Artikel Diterima : 21-04-2025

Quantitative results, as shown in Figure 4, indicate that the model's performance significantly improves as the amount of training data increases. The worst performance was observed in the first training scenario, which involved testing on six wells and resulted in a high error rate and a negative R^2 score of -11%. The R^2 score becomes positive when he model is trained on data from more than two wells, with the best performance achieved using six wells for training and one for testing, yielding an accuracy of approximately 75%.

While model accuracy generally increases with the number of training wells, this comes at the cost of longer training time, ranging from 0.01 to 5.74 seconds, demonstrating the model's scalability. With about 20,000 training samples from six wells, the SVM model reached an effective porosity prediction accuracy of 75.07%, with a training time of 5.74 seconds. Moreover, the model achieved a slightly lower accuracy of 71.23% when trained with four wells, but with a training time that was three times faster. However, training with five wells led to a significant drop in accuracy, indicating that a larger dataset does not always guarantee better performance. The most balanced and stable performance was observed when the model was trained on four wells and tested on three, making this configuration the most optimal. Under these conditions, the SVM model demonstrated reliable accuracy with a reduced training time of 1.98 seconds compared to the maximum of 5.74 seconds when using six wells.



(a) (b)

Figure 4. Cross plot of predicted-actual data of effective porosity log using SVM algorithm (a) and product of blind test(b).

Figure 4 presents a cross plot that illustrates how the predicted values (represented by the red line) closely align with the actual effective porosity data. The majority of the effective porosity values fall within the range of 0.01 to 0.25, indicating a strong correlation between the predicted and actual data. However, outside this range, some data points deviate from the red line, reflecting poorer estimation accuracy. This pattern is consistent with the blind well test results, where inaccuracies were noted for samples in the 6000-8000 range. It is clear that the model faces difficulties in accurately predicting effective porosity values above 0.25, likely due to a lack of sufficient data for this range.

CONCLUSION

Effective porosity prediction can be achieved using the Support Vector Machine (SVM) algorithm, combined with the GridSearchCV module for hyperparameter tuning. The best evaluation

DOI : http://dx.doi.org/10.12962/j25023659.v11i1.1936

Artikel Masuk : 31-07-2024 Artikel Diulas 05-08-2024 Artikel Diterima : 21-04-2025

result yielded an accuracy of approximately 71% and a training time of around 1.98 seconds. Model performance improves with increased training data, peaking at four training wells, after which accuracy fluctuates, indicating potential overfitting. While the model performs well in predicting porosity values between 0.01 and 0.25, it struggles to accurately predict values exceeding 0.25.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the Mathematics and Natural Sciences Faculty at Universitas Gadjah Mada for the financial support, as well as to the Geophysical Laboratory for their comprehensive assistance.

REFERENCES

Aizerman, M. A., and Braverman, E. M. (1964). "Theoretical foundations of the potential function method in pattern recognition learning", *Automation and Remote Control,*Vol. *25*, pp.821–837.

Asquith, G. B., and Gibson, C. R. (1982). Basic well log analysis for geologists 3rd Printing. American Association of Petroleum Geologists.

Che, J., and Wang, J. (2014). "Short-term load forecasting using a kernel-based support vector regression combination model", *Applied Energy,* Vol.132, pp.602–609.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapnik, V. N. (1996). "Support Vector Regression Machines", *Advances in Neural Information Processing Systems*.

Gardner, G. H. F., Gardner, L. W., and Gregory, A. R. (1974). "Formation velocity and density: The diagnostic basics for stratigraphic traps", *Geophysics,* Vol.39, pp.770–780.

Halotel, J., Demyanov, V., and Gardiner, A. (2020). "Value of geologically derived features in machine learning facies classification", *Math Geosci*, Vol.52, pp.5–29.

Hall, B. (2016). "Facies classification using machine learning", *The Leading Edge,* Vol.35, pp.906–909.

Haqqi, M. F., Sudarmaji,. and Prakoso, S. (2023). "An Implementation of Xgboost Algorithm to Estimate Effective Porosity on Well Log Data", J*ournal of Physics: Conference Series,* 2498.

Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences,* Vol. 585, pp.609–629.

Nugroho, I. D. R., Trisna, M. D., and Sudarmaji. (2024). "An implementation of XgBoost and Random Forest Algorithm to Estimate Effective Porosity and Permeability on Well Log Data at Fajar Field, South Sumatera Basin, Indonesia", *Indonesian Journal of Applied Physics,* Vol.14, No.2, pp.271–280.

Rider, M. (2002). The geological interpretation of well logs. Rider-French Consulting Ltd.

Saroji, S., Winata, E., and Hidayat, P. P. W. (2021). "The Implementation of Machine Learning in Lithofacies Classification Using Multi-Well Logs Data", *Aceh International Journal of Science and Technology*, Vol.10, No.1, pp.9–17.

Unpingco, J. (2019). Python for Probability, Statistics, And Machine Learning. Springer International Publishing.

Vapnik, V. N., and Lerner, A. Y. (1963). "Recognition of patterns with help of generalized portraits", *Avtomatika i Telemekhanika,* Vol. 24, No.6, pp.774–780.

Vapnik, V. N., Boser, B. E., and Guyon, I. M. (1992). A Training Algorithm For Optimal Margin Classifier. Annual Workshop on Computational Learning Theory.

Wendt, W. A., Sakurai, S. T., and Nelson, P. H. (1986). Permeability prediction from well logs using multiple regression. In *Characterization* (pp. 181–221).

DOI : http://dx.doi.org/10.12962/j25023659.v11i1.1936

Artikel Masuk : 31-07-2024 Artikel Diulas 05-08-2024 Artikel Diterima : 21-04-2025