The Application of Support Vector Machine to Estimate Synthetic Shear Sonic Log

Maman Rohaman [1*], Ignatius Sonny Winardhi [2], Yody Rizkianto [3]

[1] Geophysical Engineering, Universitas Pembangunan Nasional Veteran Yogyakarta

[2] Geophysical Engineering, Institut Teknologi Bandung

[3] Geological Engineering, Universitas Pembangunan Nasional Veteran Yogyakarta

* corresponding email: maman.rohaman@upnyk.ac.id

Abstract – Rock physics modelling is commonly applied to characterize the subsurface. Sonic log provides the elastic properties in advanced petrophysics modelling or rock physics modelling. Although it is very important, to obtain shear sonic measurement results is very expensive. Therefore, empirical and artificial intelligence allow some solutions to estimate synthetic shear sonic log. This study applicate PCA as feature selection and SVM as the regressor with TAF as the target interval for well NEGF1P. The results of feature selection are GR, DTC, and MSF log as selected features. GS optimizes the SVM kernel parameter using selected features. The best parameters for each kernel (linear and rbf) and selected feature are the input to estimate synthetic shear sonic log. The estimation result using linear kernel has R^2 0.845 and root mean square error (RMSE) 15.132 and using rbf kernel has R^2 0.886 and RMSE 12.989. The estimation results construe that rbf kernel estimates the synthetic sonic log with more precision than the linear kernel and indicates the linear relation between the estimated and origin log. The three other wells apply SMV with rbf kernel best parameters and selected features to estimation the synthetic shear sonic in similar interval and younger interval (GUF).

Keywords: shear sonic; rock physics; petrophysics; svm; pca.

INTRODUCTION

Rock physics modelling is the approach to characterize the subsurface condition. It affects to the exploration and production aspect (Kim, 2022; Maleki et al., 2014). In this modelling, several elastic rock properties can be generated according to data availability. Rock physics modelling can distribute elastic properties representing the lithology or fluid content using petrophysical properties. In other words, petrophysics is an essential property in modelling rock physics.

Core, log, and seismic data are required to model rock physics or petrophysics. Core data provide elastic properties using direct measurement. The measurement is carried out on a rock sample. This measurement needs a lot of time due to step by step procedure for generating rock properties. The result of this measurement is discrete data depending on the number of rock samples (Kim, 2022). It's different from core data, logging is the indirect method that measures the rock properties vertically. Nevertheless, in deviated well, the log data has a lateral resolution. This method is more efficient compared to coring data. This method provides high-resolution data vertically in the depth domain. This method must be performed in different well locations to cover lateral distribution (Kim, 2022). Another indirect measurement is seismic acquisition. The seismic data is preferable in lateral coverage but vertically has a lower frequency compared to log data.

The standard log tools measure natural gamma radiation, formation density, resistivities in different penetrations, and neutron porosity. These log data provide petrophysics and lithology information (Kim, 2022). In addition, acoustic logging measures compressional sonic log. This log data provides the elastic properties in advanced petrophysics modelling or rock physics modelling. The measurement of compressional sonic utilize monopole logging tool. This tool has limitations in measuring shear sonic. Therefore, the dipole logging tool is used to acquire shear sonic and compression sonic simultaneously. This measuring principle is the reason that dipole logging tool is the more expensive cost compared to monopole logging tool. As a consequences, the shear sonic log is more limited available compared to the compressional sonic log (Kim, 2022).

Referring to this limited log data, several approaches have been developed. Empirical rock physics models are the approach to model missing elastic data (Rohaman, 2017). These models are generally based on calibration coefficients and less geological consistent parameters (Greenberg and Castagna, 1992). The other models, such as theoretical and heuristic models, use continuum mechanics-based approximations of the relationship between elastic and poroelastic properties of rocks (Avseth et al., 2005). The general challenge in the empirical model is model fitting. The assumptions in model fitting are generally not the universal model that fit all rock type so the parameters have to be calibrated by the available data set. This complexity issue non-uniqueness solution of the model using that approach (Lorentzen et al., 2022). Therefore, in the last two decades, artificial intelligence, especially machine learning is employed to predict missing log. This approach is increasingly developed in so many aspects of industries, not only in oil and gas industry. In 2014, Maleki et al. used supervised machine learning Support Vector Regressor (SVR) and Back-Propagate Neural Network (BPNN) to predict shear wave velocity. In 2021, Lorentzen et al. performed Linear Regressor (LR), Random Forest Regressor (RFR), SVR and Artificial Neural Network (ANN) to estimate shear sonic log in heterogeneous and fractured Lower Cretaceous of the Danish North Sea. In 2021, Mordekhai et al. apply the parametric and non-parametric regression to predict the missing log data using Principal Component Analysis (PCA) for feature selection and using LR, Support Vector Machine (SVM), and Gaussian Process Regression (GPR) to estimate missing log data. In 2022, Kim J. utilized hybrid machine learning to generate a sonic log synthetic. The approach combines unsupervised machine learning for data clustering and Particle Swarm Optimation (PSO) to determine hyperparameters.

The focused area of the study is Jambi Sub Basin, South Sumatera Basin, and located on west Bukit Barisan Mountain. South Sumatera Basin is categorized into four sub-basin, South Palembang, North Palembang, Central Palembang, and Jambi Sub-Basin. South Sumatera Basin is a back-arc basin formed in range Pre-Tertiary to Early Tertiary. Extensional back-arc stresses along the subduction zone in the Early Tertiary produced a series of grabens across Sumatra and Java and extended onto the adjacent Sunda Shelf to the north. The subduction zone was formed when the oceanic crush of the Indian Ocean was subducted northward under Sundaland Craton (Sutjiningsih et al., 2007).

The target of this study is Talang Akar Formation (TAF) and Gumai Formation (GUF). TAF is believed to be the dominant source of commercial hydrocarbons in the South Sumatra Basin (Ginger and Fielding, 2005). TAF is classified into two sub-formations, Lower Talang Akar (LTAF) and Upper Talang Akar (UTAF). The deposition of LTAF involves a transgressive cycle, which commenced in Late Oligocene. The deposition is syn-rift transgressive fluvio-deltaic. LTAF filled in the half-graben and occasionally deposited in the basement high. The other interval, UTAF deposition is deltaic to marine sedimentation. It consists of interbedded sandstone, shale and coals eventually accompanied by carbonates streaks (Sutjiningsih et al., 2007).

The sedimentary succession of GUF Formation comprises transgressive cycle, which commenced in Early Miocene. Marine conditions were eventually established during the continuing transgression sequence with deposition of an open marine facies in the Gumai Formation, which consists of marine shales, claystones, marls, and fine to moderate grained thick sandstones with moderate to good porosity and permeability (Sutjiningsih et al., 2007).

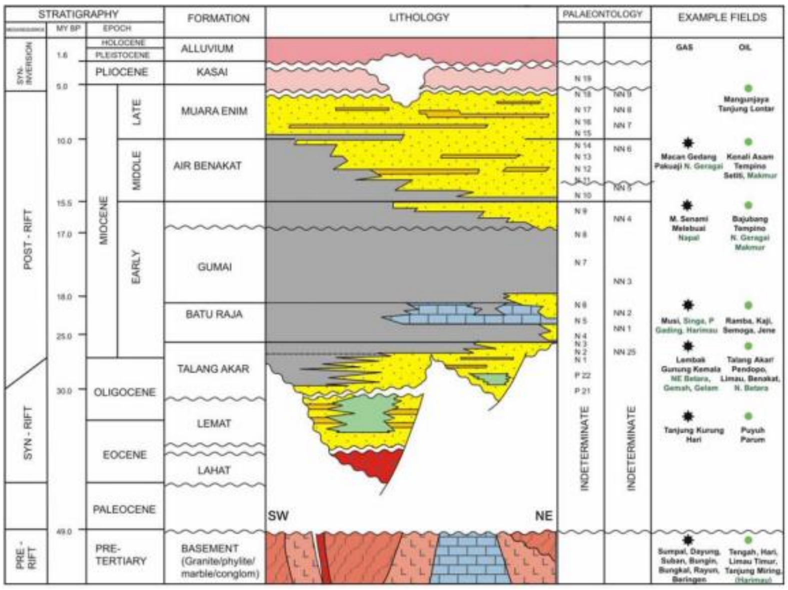

Figure 1. Chronostratigraphic Scheme for South Sumatera Basin
Source: Ginger and Fielding, 2005

METHODS

The objective of the study is to estimate synthetic shear sonic log. That log can be estimated using several wireline logs. Four wells have the available wireline log data, NEGF1P, NEGF2P, NEGF3PU, and NEGF4PU. The DTS log in Well NEGF4PU is available, so this well is the object to train the regression and the validation. The best parameter in that experiment will be applied to the other well in which the DTS log is unavailable. **Table 1** informs the log data availability for each well specifically.

Table 1. Well Log Data Availability

Well Name	CALI	GR	RHOB	DTC	RT	DTS	CNL	MSF	PEF
NEGF1P	√	√	√	√	√		√	√	√
NEGF2P	√	√	√	√	√		√	√	√
NEGF3PU	√	√	√	√	√		√	√	√
NEGF4PU	√	√	√	√	√	√	√	√	√

Principal Component Analysis

Before estimating the synthetic sonic log, the prior step is to determine the input data/features used in the prediction. The method for feature selection is Principal Component Analysis (PCA). PCA is made up of a linear transformation of m original variables into new variables (components) where each new component is a linear combination of the original m variables (Davis, 2006; Flood et al., 2015). PCA is performed in a manner in which the total variance of the dataset is accounted for by each successive new component (Flood et al., 2015). **Figure 2** represents two variables plotted on a 2D axis. Then, these variables are transformed into two principal components. In the component axis, the observation data can be classified clearly.

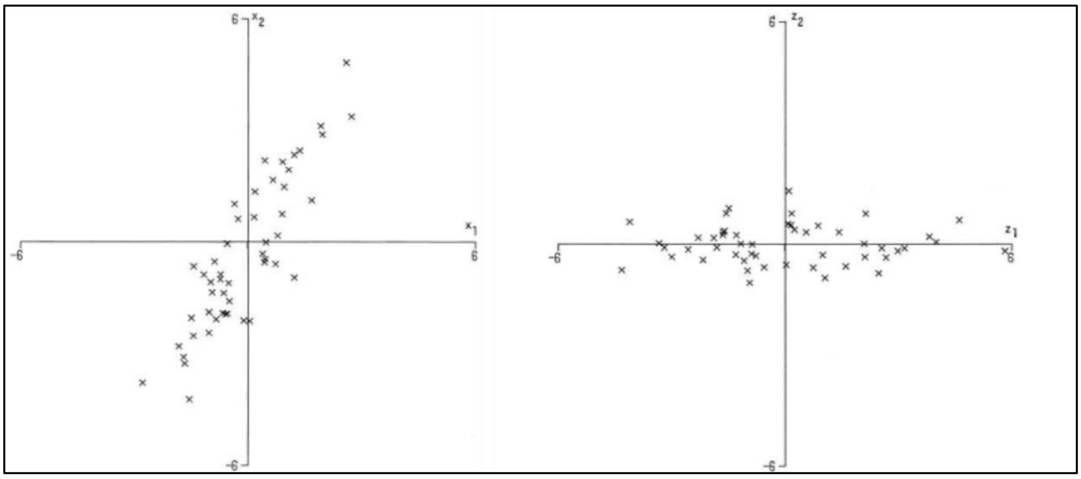

Figure 2. Plot 50 observation of variable x_1 and x_2 (left) and plot Plot 50 observation of variable x_1 and x_2 with respect to their principal component z_1 and z_2 (right).
Source: Jolliffe, 2002

Support Vector Machine

The method for DTS estimation is SVM with various kernels (linear and radial basis function). Equation 3 is the linear kernel function, and **Equation 4** is rbf kernel function (Kecman, 2005). SVM classifies the data using a decision boundary commonly called hyperplane. **Equation 1** shows the discriminant or decision function. Where **x** is the feature vector with *n* numbers, **w** is the weight of each feature, and *b* is a bias. **Figure 3** illustrates the hyperplane of two classes of data. SVM determines the hyperplane by considering the maximum margin classifier. **Equation 2** is the objective function to determine the maximum margin classifier (M) (Kecman, 2005).

$$d(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = \sum_{i=1}^{n} w_i x_i + b \quad (1)$$
$$\mathrm{M} = \frac{2}{\|\mathbf{w}\|} \quad (2)$$
$$\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = (\mathrm{x}^T \mathrm{x}_i) \quad (3)$$
$$\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = e^{\frac{1}{2}[(x - \mathrm{x}_i)^T [\Sigma(x - \mathrm{x}_i)]^{-1}]} \quad (4)$$

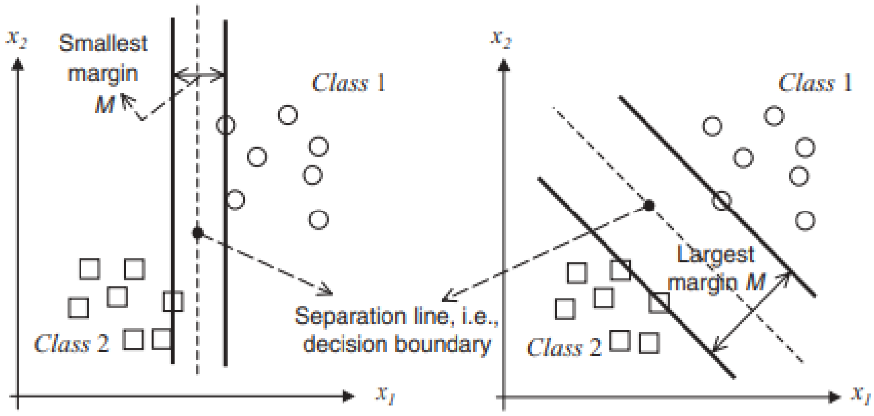

Figure 3. The Decision Boundary of The Data
Source: (Kecman, 2005)

Grid Search

The parameters of SVM consisting of C and sigma are determined by the systematic search method. Grid Search (GS) as the systematic search method is commonly defined as parameter optimization. This method performs to a defined model consisting of several grids. The objective function values of each grid will be calculated and used for considering the best solution of the model. The best solution of this model is grid value with minimum objective values (Grandis, 2009). The workflow for parameter optimization follows the **Figure 4**.

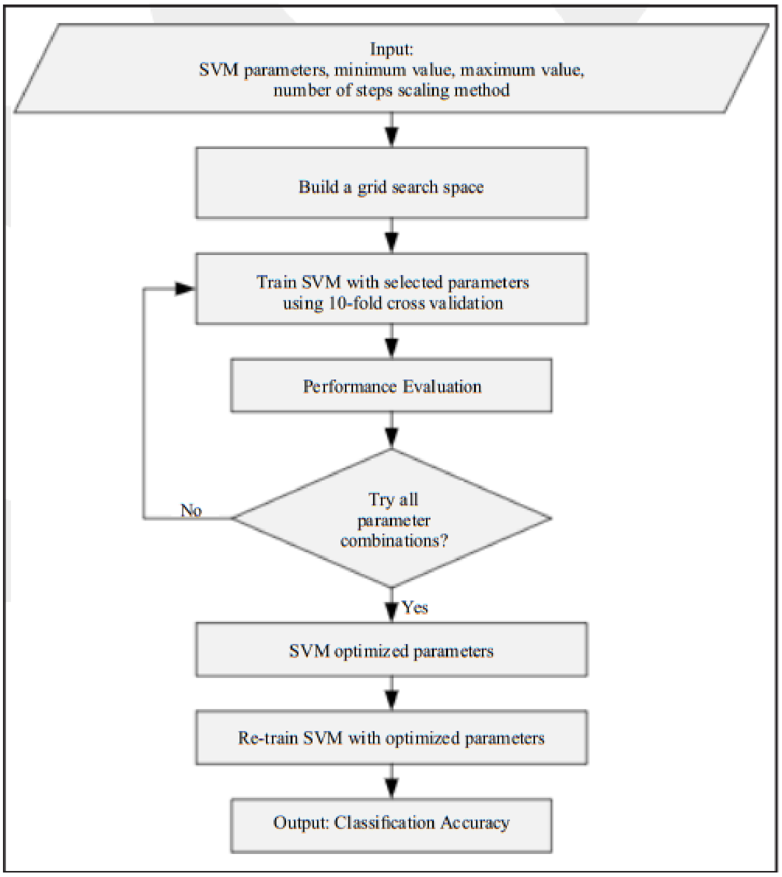

Figure 4. SVM Parameter Optimization using Grid Search.
Source: Syarif, 2016; Mordekhai et al., 2021

Cross Validation

K-Fold is one of the algorithms to cross-validate the prediction results. This algorithm separates the dataset into training and validation subset data. The combination of two subset data will be variated depending on the number of folds. In this study, the number of folds is ten folds (Syarif, 2016).

RESULTS AND DISCUSSION

Feature selection is performed in Well NEGF4PU because this well has a DTS log as a data validator. Log CALI, GR, RHOB, DTC, RT, CNL, MSF, and PEF called data features are the input data for feature selection using PCA. PCA transforms the features into the principal component. This process calculates the variance of each component. The value of variance represents discriminant abilities. PC1 and PC2 are the component results of the feature selection process. **Figure 5** informs us about the variance of the data for each principal component. The explained value ratio of PC1 is 93.18 %, which indicates this component can discriminate the data and has dominant contribution to it.

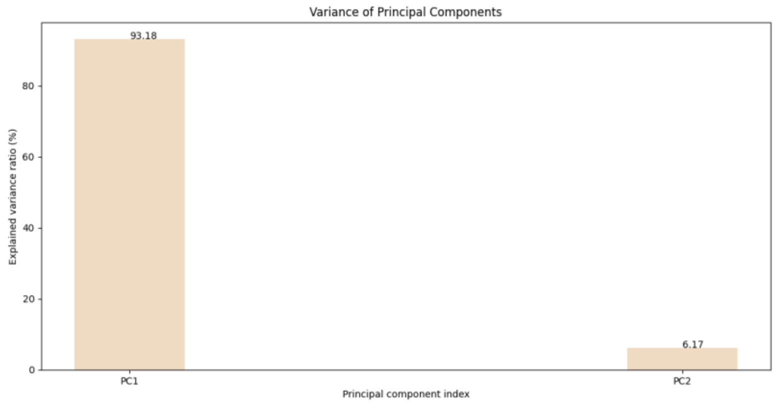

Figure 5. The Variance of the Principal Components

In the transformation of the input data into respective components, PCA calculates the eigenvalue and eigenvector. Each well log as the feature has an eigenvector that contributes to PC1 as the main component. Based on the information in **Figure 6**, GR has the biggest eigenvector compared to other logs. It represents the log data has a dominant contribution for PC1. In spite of that result, DTC and MSF will be included in the synthetic shear sonic log estimation process because it has enough contribution. According to that consideration, the selected features are GR, DTC and MSF log.

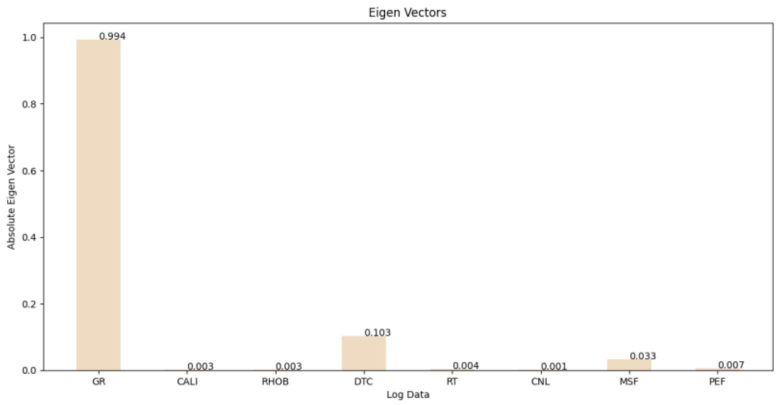

Figure 6. The Eigen Vector of Well Log Data for PC1

Before estimating the synthetic sonic log, a searching parameter is performed to find the best kernel parameter using Grid Search algorithm. The range models of C parameter and gamma, respectively, are 0.1 – 100 and 0.001 – 1. The input data in this process are selected features and DTS log as the target data. **Table 2** shows the best parameter for each kernel using Grid Search.

Table 2. Kernel Parameters

Type of Kernel	C	*gamma*
Linear	0.1	1
Gaussian rbf	100	0.001

SVM estimates the synthetic sonic log using the best kernel parameter and selected features. This estimation is performed in Well NEGF4PU. The result of synthetic sonic log estimation is shown by **Figure 7**. This result using linear kernel has R^2 0.845 and root mean square error (RMSE) 15.132. This result using rbf kernel has R^2

0.886 and RMSE 12.989. Visually, **Figure 7** shows the estimation result of synthetic shear sonic using linear and rbf kernel compared to DTS log as origin data. Statistically, the estimation results construe that rbf kernel estimates the synthetic sonic log with more precision than the linear kernel. **Figure 7** indicates the linear relationship between DTS log and the synthetic shear sonic log predicted by SVM using rbf kernel.

Furthermore, cross-validation is performed to validate the results. This validation uses using K-Fold algorithm. The number of folds is conducted by ten folds. The results of cross-validation using linear kernel are RMSE 3.888 and R^2 0.892. On the other hand, the results of cross-validation using rbf kernel are RMSE 3.809 and R^2 0.841. The results indicate that the prediction using rbf kernel is more precise based on RSME value.

Table 3. Statistical Result of prediction and Validation

Type of Kernel	Prediction		Cross-Validation	
	RMSE	R^2	RMSE	R^2
Linear	15.132	0.845	3.888	0.892
Gaussian rbf	12.989	0.886	3.809	0.841

Based on the experiment on the Well NEGF4PU, the other wells apply SMV with rbf kernel best parameters and selected features to estimation the synthetic shear sonic. The estimation will be performed in similar interval formation and younger interval (GUF). GUF is the main target of these wells. Therefore, the analysis of estimation results is focused on that interval. Synthetic shear sonic as the estimation result using SVM contributes to rock properties transformation, such as Mu-Rho (MR). MR is a rock property that the product of multiplication Lame constant Mu and density. Mu, defined as rigidity, represents rock properties related by rigidity of lithology. In other words, this property can be used to delineate lithology. Compacted sandstone has higher rigidity compared to shale or uncompacted sandstone (Rohaman, 2017).

Figure 8-10 inform the relation of rock properties in Well NEGF1P, NEGF2P, and NEGF4PU, respectively. Crossplot RHOB-GR shows the lithology delineation based on the log data. While crossplot MR- RHOB shows transformed properties indicating the lithology delineation and fluid content delineation. The crossplot MR-RHOB represents that the trend of shale rigidity is to the left of the sandstone trend. Even though the lithologies are not clearly separated, sandstone has higher rigidity compared to shale. In sandstone interval, that crossplot can separate the fluid content that contain high or low water saturation values.

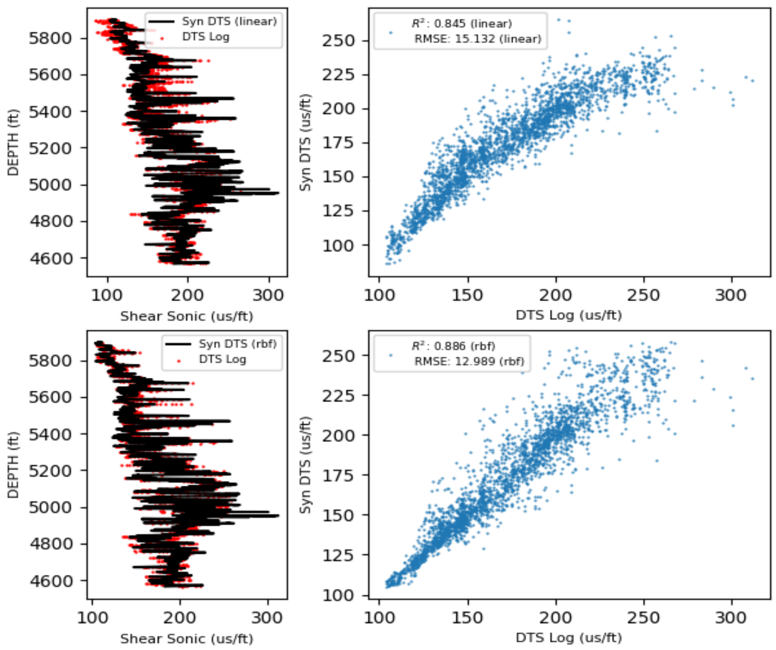

Figure 7. Synthetic Shear Sonic compared to DTS Log.

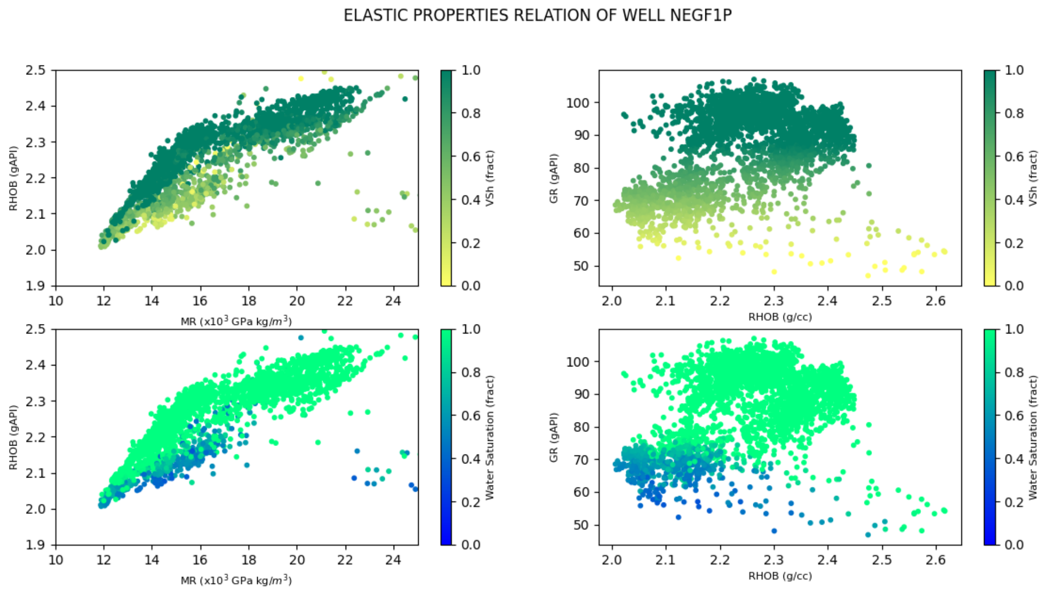

Figure 8. Rock Properties Relation of Well NEGF1P

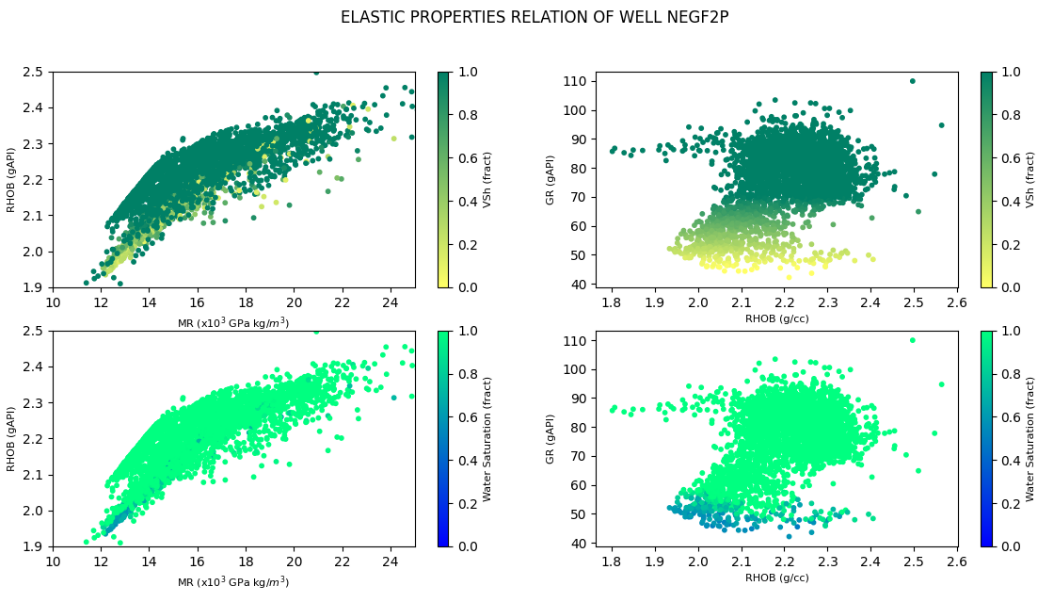

Figure 9. Rock Properties Relation of Well NEGF2P

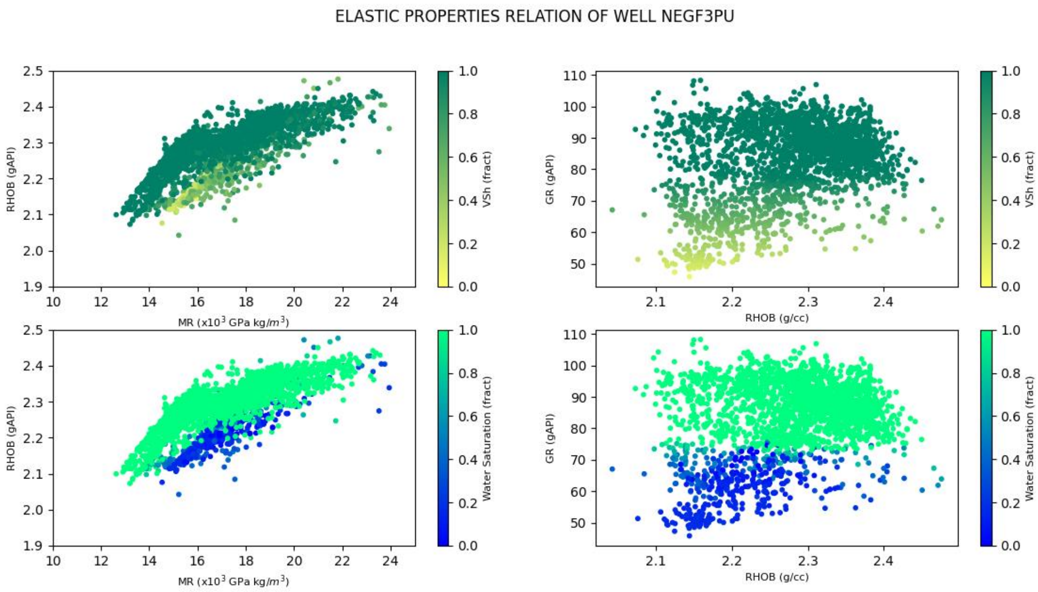

Figure 10. Rock Properties Relation of Well NEGF3PU

CONCLUSION

The estimation of the shear sonic log process includes feature selection, search parameters, regression and cross-validation. PCA as feature selector conduces GR, DTC, and MSF log as selected features. Furthermore, the selected features will be the input to search for the best kernel parameter using Grid Search algorithm. In this process, GS search the best C for linear 0.1 and for rbf 100. Moreover, the best gamma of the linear kernel is 1 and the best one is 0.001. These parameters are employed to estimate synthetic shear sonic log. The estimation result using linear kernel has R^2 0.845 and root mean square error (RMSE) 15.132 and using rbf kernel has R^2 0.886 and RMSE 12.989. The estimation results construe that rbf kernel estimates the synthetic sonic log with more precision than the linear kernel and indicates the linear relation between the estimated and origin log. The other wells apply SMV with rbf kernel best parameters and selected features to estimation the synthetic shear sonic. Afterwards, this estimation result was transformed to MR. Crossplot MR-RHOB can separates the different lithologies and fluid content.

REFERENCES

Avseth, P., Mukerji, T. & Mavko, G. (2005) Quantitative Seismic Interpretation: Applying Rock Physics Tools to Reduce Interpretation Risk. Cambridge, UK: Cambridge University.

Davis, J. C. (1986). Statistics and Data Analysis in Geology. 3nd ed. John Wiley and Sons, Knowledge as value: Clarification of zonation procedure (pp 646). New York, United State of America.

Flood, R.P., Orford, J.D., McKinley, J.M., Roberson, S. (2015). Effective grain size distribution analysis for interpretation of tidal–deltaic facies: West Bengal Sundarbans, 318(2015), 58-74. http://dx.doi.org/10.1016/j.sedgeo.2014.12.
007

Ginger D, Fielding K. (2005). The Petroleum Systems and Future Potential of The South Sumatra Basin, Proceedings Indonesian Petroleum Association, August 2005.

Grandis, H. (2009). An Introduction of Geophysics Inversion Modelling. Jakarta, Indonesia: Himpunan Ahli Geofisika Indonesia (HAGI).

Greenberg, M.L. & Castagna, J.P. (1992) Shear-wave velocity estimation in porous rocks: theoretical formulation, preliminary and applications. Geophysical Prospecting, 40(2), 195–209. https://doi.org/10.1111/j.1365-2478.1992.tb00371.x

Jolliffe, I.T. (2002). Principal Component Analysis, 2nd ed. New York, United State of America: Springer.

Kecman, V. (2005). Support Vector Machine-An Introduction. Springer, Knowledge as value: Support Vector Machine in Classification and Regression (pp 1-47). Berlin, Germany.

Kim, J. (2022). Synthetic shear sonic log generation utilizing hybrid machine learning techniques, 3(2022), 53-70. https://doi.org/10.1016/j.aiig.2022.09.001

Lorentzen, M., Bredesen, K., Mosegaard, K., Nielsen, L. (2022). Estimation of shear sonic logs in the heterogeneous and fractured Lower Cretaceous of the Danish North Sea using supervised learning, 70(8), 1410-1431. https://doi.org/10.1111/1365-2478.13252

Rohaman, M. (2017). Application of Stochastic Inversion to Characterize Sandstone Reservoir of Gumai Formation in MR Field, Master's Program Thesis, Institut Teknologi Bandung.

Maleki, S., Moradzadeh, A., Riabi, R.G., Gholami, R., & Sadeghzadeh, F. (2014). Prediction of shear wave velocity using
empirical correlations and artificial intelligence methods, 3(2014), 70-81. http://dx.doi.org/10.1016/j.nrjag.2014
.05.001

Mordekhai, M., Qudsi, I., Guntoro, J. P. S. (2021). The Application of Parametric and Nonparametric Regression to Predict The Missing Well Log Data, 8(3), 285-399. DOI: 10.17014/ijog.8.3.385-399

Sutjiningsih W, Alamsyah MN, Marpaung LP, Suta IN, Siagian T, Subekti A, Patria P, Fathkan. (2007). Rock Physics Analysis for Feasibility of AVO and Inversion Studies: A Tool for Exploration Strategy in Jabung Block, Proceedings Joint Convention Bali 2007, Bali, November 2007.

Syarif, I., Prugel-Bennett, A., and Wills, G., (2016). SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance, 14 (4), 1502-1509. DOI: 10.12928/telkomnika.v14i4.3956