Spatial Machine Learning Algorithms to Discover Prospective Oil and Gas Wells Locations Based on Surface Driving Factors

Safira, R. A. D.,[1] Nurwatik, N.,[1*] Hariyanto, T.[1] and Lee, S.[2]
[1]Department of Geomatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
E-mail: nurwatik@its.ac.id*
[2]Geoscience Data Center, Korea Institute of Geoscience and Mineral Resources (KIGAM), 124, Gwahak-ro, Yuseong-gu, Daejeon 34132, Republic of Korea
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v21i2.3935

Abstract
Indonesia faces challenges in transitioning its energy sector, aiming to shift from coal to natural gas, achieve net zero emissions with renewable energy, and overcome geographical complexity obstacles, diverse cultural perspectives, and a developing regulatory framework. To address these issues, the government actively studies the Grand National Energy Strategy in enhancing petroleum and fuel refineries. This research aims to expand the academic approach by utilizing spatial technology and machine learning to optimize the new oil and gas well placement determination and meet the high-demand resources. Four algorithms, support vector machine (SVM), random forest (RF), artificial neural network (ANN), and k-nearest neighbor (KNN), with four training and testing splitting scenarios (80:20, 75:25, 60:40, and 50:50) are used to produce probability map of the wells site suitability along with fourteen surface driving factors related to the environmental agreement. The outcome indicates that the 80:20 RF model demonstrated excellence, achieving a 0.95 accuracy, 1.00 sensitivity, 0.90 specificity and Cohen's Kappa, 0.91 precision, and 0.99 area under the curve, showcasing the optimal fit with validation data. The four surface driving factors with the highest important index indicate that the well placement is sensitive to historical disaster, ease of accessibility, and hydrocarbon sourcing.

Keywords: Energy Use, Machine Learning, Oil and Gas Wells, Spatial Technology, Surface Driving Factor

1. Introduction

Indonesia's energy demand is projected to grow by 3.5% annually until 2050, primarily driven by fuel oil and the transportation sector, which makes up 42% of total energy use [1]. Despite a temporary slowdown in 2020 because of the COVID-19 pandemic, fuel oil demand is expected to climb by an average of 2.7% each year. However, the current refinery capacity cannot fulfil the magnitude of this order, and the quality of energy from aging oil and gas wells is diminishing [2]. Hence, Indonesia must upgrade its refining operations to meet domestic needs for higher-quality fuels [3]. Meanwhile, the energy transition approach in Indonesia is currently at the stage of switching from coal to natural gas before exchanging to net zero emission for the long term towards 100% use of renewable energy. This transition faces obstacles due to Indonesia's complex geography and varied political and cultural landscapes, impacting costs and the framework for energy distribution [4]. Governments and stakeholders are studying the Grand National Energy Strategy to boost local production and ensure energy is accessible, quality-assured, affordable, and environmentally friendly. One of the strategies is to optimize natural gas use and enrich the lifting of petroleum and fuel refineries' capacity [1].

In determining the location for constructing new oil and gas wells, it is vital to minimize the risk of unsuccessful explorations, as it can be challenging, costly, and risky [5]. Although exploring the well's location assessment based on geophysical conditions proved accurate, it is considered a conventional and multifaceted strategy. These are supposed to be the most expensive course, with costs ranging from 3-5 thousand USD/km^2 operation [6]. Thus, an additional approach is required to overcome these challenges by considering many factors to generate an accurate area prospect prediction model [7].

Therefore, this research aims to assess the location from a spatial and geological perspective. Band ratio calculation from remote sensing imagery proved excellent to discover heavy oil in North Sumatra by analysing iron oxide, clay mineral, vegetation index, and temperature factors [8] and [9]. Another research utilized a pragmatic strategy to explore the hydrocarbon capacity by employing gravity data from rock density calculation [10]. Subsequently, previous research highlighted the importance of flat terrains with water access and the need to consider land use and proximity to infrastructure for building the wells [11] and [12]. These standards address concerns about ground movement on inclined surfaces, the ease of operational fracturing, vehicle efficiency, and safety deliberations [13].

The need for progressive technology solutions is growing to reduce risk and maintenance expenses, streamline decision-making processes, and improve productivity effectively. Machine learning has emerged as a popular solution to these challenges [14]. Previous research utilized a variant of the SVM approach, namely the Least Square SVM (LSSVM), for positioning the fractional wells [15]. The LSSVM is a simplified approach that uses linear equations instead of quadratic programming. However, it lacks resilience when outliers are present, loses sparseness compared to SVM, and is limited to real-world presentation. On the other hand, nonlinear solutions like neural networks are indeed one of the most potent methods for nonlinear models. They can potentially uncover hidden information in model complexity with maximum learning flexibility [16]. Furthermore, RF algorithm performed well for locating oil and gas exploration in Central Sumatra using collected data from various parameters with 97.3% accuracy and 96.9% kappa [17].

Due to a scarcer investigation into exploring the potential location of oil and gas wells in Central Java Province, Indonesia, this research proposes identifying suitable sites for oil and gas well infrastructure in Blora Regency by evaluating some machine learning performances. This research uses four algorithms (SVM, RF, ANN, and KNN) and tests four different ratios to divide the training and testing data (80:20, 75:25, 60:40, and 50:50). The selection of splitting ratio is significantly crucial in ascertaining the model's accuracy, so offering guidance for improved selection of splitting ratios for the model is invaluable [18]. This research aims to develop straightforward machine-learning models that use surface driving factors.

2. Materials and Methods

2.1 Data

This research employed data in raster and vector format. Although more concentrated on the geomatics and spatial approach, this study considers the potential oil and gas resources beneath the surface by analysing mineral context from imagery. The training-testing samples, highlighting oil and gas well locations up to 2017, were obtained from Indonesia's ESDM One Map. For the driving factors, it utilized imagery data with no specific date in which the images were built from available scene combinations to fill in the empty pixels due to cloud cover masking; the vector data comprising the geology, the polyline river and road networks, and soil type data; the topographical data including the elevation raster-based data; and point-based gravity data. These data will be extracted to generate the fourteen factors, with data source and resolution outlined in Table 1.

Table 1: Data sources of driving factors

Data (Resolution/scale)	Source	Driving factors (Abbreviation)	
Landsat 8 OLI Level-2 and TIRS Level-1 (30 m)	USGS	Normalized difference vegetation index	(NDVI)
		Clay mineral index	(CMI)
		Iron oxide index	(IOI)
		Land surface temperature	(LST)
		Land cover	(LCO)
Lithology and Geostructure Shapefile (1:100,000)	Geological Agency of Indonesia	Geological type	(LIT)
		Distance from the fault	(DFT)
GGMplus (200 m); SRTMgravity (3")	Murray Lab	Complete Bouguer anomaly	(CBA)
National Digital Elevation Model DEMNAS (8 m)	Indonesian Geospatial Portal	Slope	(SLP)
		Aspect	(ASP)
		Elevation	(ELV)
Topographical Map (1:25,000)	Indonesian Geospatial Portal	Distance from the river	(DRV)
		Distance from the road	(DRO)
Soil Type Map (1:50,000)	Ministry of Agriculture	Soil type	(SOI)

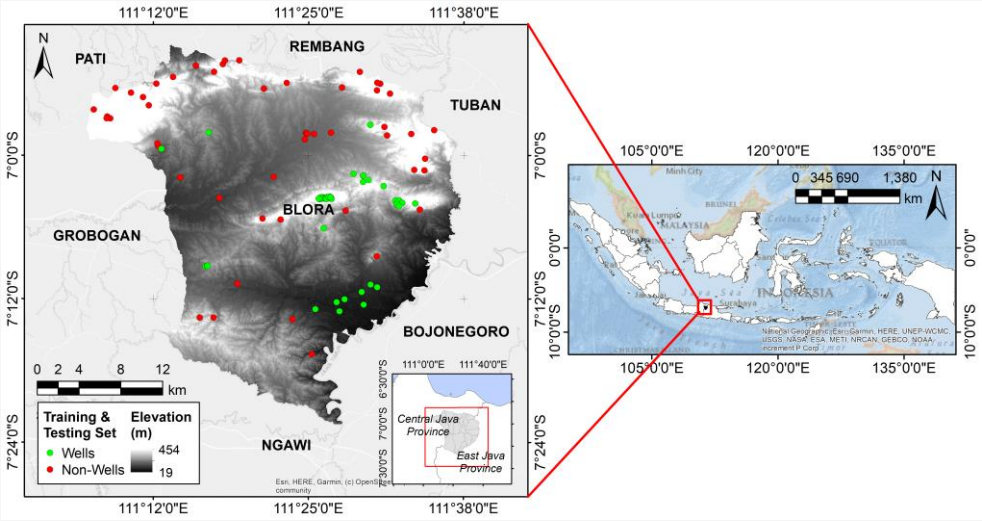

Figure 1: Topography of the research area and the distribution of well and non-well locations

This research plotted oil and gas well and non-well points as the training-testing splitting base. The training set trains the data for model fitting, and the testing set evaluates the model's accuracy. This process is essential to prevent the model from overfitting if the entire oil and gas well point coordinate is involved in fitting, ultimately leading to suboptimal predictions in future scenarios [19]. Furthermore, according to directives, 50 points were randomly selected from inappropriate areas for well placement, comprising steep slopes, flood-prone territories, lakefronts, wetlands, and sites with a high risk of erosion and landslides [13]. Eventually, a collected dataset of 100 points contributed to design training and testing sets based on four splitting ratio scenarios. The well and non-well points are labeled as 1 and 0, respectively. Figure 1 displays these points' distribution.

2.2 Methods

2.2.1 Deriving surface driving factors

DEMNAS data directly derives the SLP, ASP, and ELV factors. This research calculated the slope as a percentage by multiplying 100 by the ratio of rise to run. The eight classes of sloping directions: east, northeast, north, west, northwest, southeast, south, and southwest, plus one flat class define the ASP factor. Hence, the slope and elevation are continuous variables, and the aspect is discrete. Since the modeling requires all data in raster format, vector-sourced polygon data, including geology and soil maps, was rasterized, resulting in discrete factors. Afterward, the distance quantification from polyline data was developed using the Euclidean Distance methods. These data comprise faults, road networks, and river networks. NDVI, CMI, and IOI indices serve as indicators for detecting micro and macro seeps, suggesting the presence of subsurface oil and gas [9] and [20]. Meanwhile, the LST factor plays a pivotal role in the land's physical processes at regional and global levels. Calculating LST using the Single Channel method contains radiometric correction on the TIRS band, calculating at-satellite brightness temperature using TOA radiance data, Gamma (γ) and Delta (δ) parameter, land surface emissivity from the proportion of vegetation and cavity effect, inserting atmospheric correction parameters constant, then calculating LST value sequentially [21]. Further, LCO classification employing SVM algorithm delineates five land cover types (built-up areas, water bodies, agriculture, farm fields, and plantations) with a 95.53% overall accuracy.

The gravity data employed in hydrocarbon exploration is generally Complete Bouguer Anomaly (CBA) as the prior data for rock density determination [22]. GGMplus provides free-air anomaly data, while SRTM2 gravity includes terrain correction data. Calculating the free-air anomaly (Δg_F) to gain CBA value is preceded by calculating the simple Bouguer correction. Equation 1 defines the Bouguer correction (δg_B) where H is the station's height. Next, this research performed terrain corrections to calculate topographical details on the Bouguer plate. Equation 2 defines the terrain correction (δg_T) where R is the Earth's mean radius, ρ is the Bouguer correction density (2670 kg/m^3), and H' is the height of each point at a distance (l) from the station. Hence, the CBA value is defined by Equation 3 [23]. Finally, Figure 2 visualizes the fourteen driving factors.

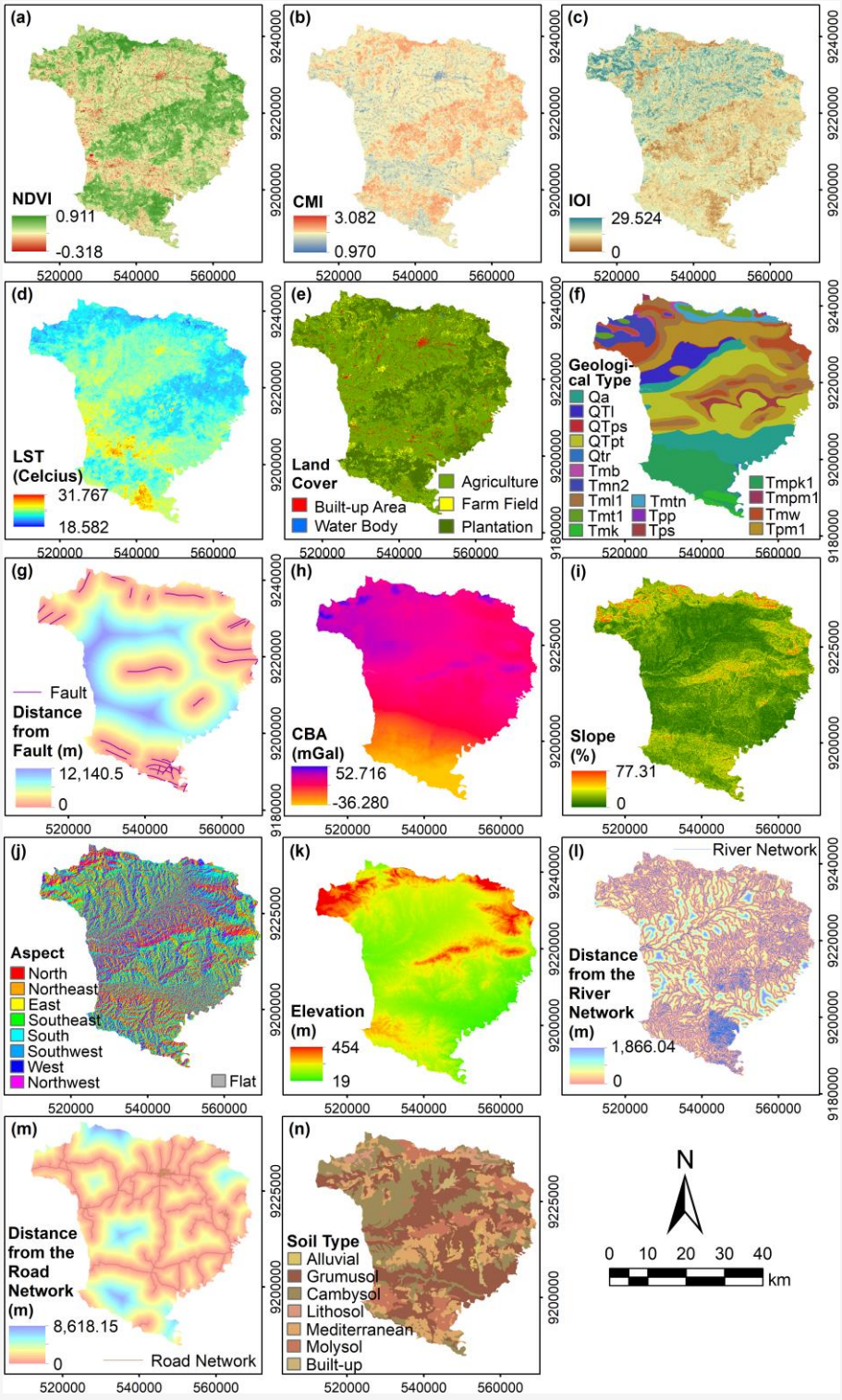

Figure 2: Fourteen surface driving factors:
(a) NDVI; (b) CMI; (c) IOI; (d) LST; (e) LCO; (f) LIT; (g) DFT; (h) CBA;
(i) SLP; (j) ASP; (k) ELV; (l) DRV; (m) DRO; (n) SOI

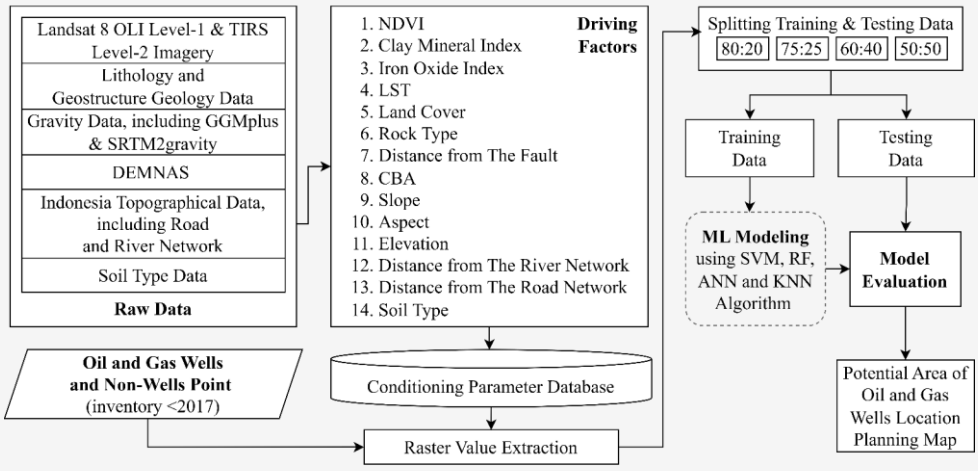

Figure 3: Overall data processing workflow

For visualization and data processing purposes, it projected our spatial data to the Universal Transverse Mercator (UTM) coordinate system, specifically in zone 49S.

$$\delta g_B = 0.1119H$$

Equation 1

$$\delta g_T = \frac{G\rho R^2}{2}\left(\int_{\sigma}\frac{\left(H'-H\right)^2 d\sigma}{l^3} - \frac{3}{4}\int_{\sigma}\frac{\left(H'-H\right)^4 d\sigma}{l^5}\right)$$

Equation 2

$$CBA = \Delta g_F - \delta g_B + \delta g_T$$

Equation 3

2.2.2 Overall workflow

Figure 3 illustrates the workflow of this research. Overall, this research contains three key stages: (1) preprocessing, (2) modeling, and (3) model evaluation and analysis.

2.2.2.1 Preprocessing

Preprocessing involves extracting pixel values to the 50 well and non-well points. Next, this research used feature scaling through a max-min formula to normalize continuous factors and ensure each is equally considered. Then, this research converted the class names to a factor set up for the categorical factors. Finally, sample points are scaled to a range of 0-1 before being divided into training and testing sets [24].

2.2.2.2 Modeling

SVM is a popularly recognized algorithm in supervised ML that effectively tackles regression and classification challenges [25]. By the kernel and structural risk minimization concept, SVM exhibits excellent scalability when presented with large and high dimensional amounts of input, including nonlinearity of real-world domain, and reduces the estimate of a model's generalization error [26]. The radial basis function is widely recognized as the most employed kernel, effectively yielding favorable outcomes in many applications. Equation 4 denotes the formula of the RBF function ($K(x_i, x_j)$), where x_i and x_j indicate training set pairs, while γ is the kernel parameter [27]. RF is an assembly of decision trees that work together to produce an outcome and is favorably effective for large data sets, allowing for accurate data classification [28]. The trees' predictions are put through a voting process where the forest determines the category with the most votes [29]. Equation 5 denotes the RF mathematical model where C_{RF} and C_N are RF outcomes and the n^{th} tree predictive class. Meanwhile, x represents the input variable [30]. ANN has the concept of imitating the human brain to examine large sets of properties and find hidden causal relationships. It is better at modeling convoluted nonlinear relationships than traditional regression methods, fast, reliable, and can handle immense amounts of data [31]. The general equation of ANN (y_i) is denoted in Equation 6, where X_i and y_i are input and output variables respectively, n and B_j indicate the neurons' number and bias in the hidden layer; meanwhile W_{ji} plays as a weight connector [32]. KNN is a potent non-parametric and one of the most straightforward algorithms in the classification process, presented by Cover and Hart in 1968. The system completely circumvents the issue of probability densities [33]. Only two parameters describe the KNN algorithm, namely the distance calculation, which determines the similarity between the samples, and the number of nearby neighbors [34].

$$K\left(x_i, x_j\right) = \exp\left(-\gamma\left(x_i - x_j\right)^2\right), \quad \gamma > 0$$

Equation 4

$$C_{RF} = Majority\ vote\left\{C_N(x)\right\}_{n=1}^{N}$$

Equation 5

$$y_i = f\left(B_j + \sum_{i=1}^{n} w_{ji} X_i\right)$$

Equation 6

2.2.2.3 Model assessment

A confusion matrix compares the model's actual values with the predictions, deriving accuracy (Acc), sensitivity (Sen), specificity (Spe), predictive value (PPV), and Cohen's Kappa (CK) parameters [30]. The four elements that define these parameters consist of the number of True Positive (TP) and True Negative (TN), representing the accurate identification of positive and negative cases among the actual positives and negatives, respectively; False Positive (FP), defining the negatives that are indicated inaccurately as positives; and False Negative (FN), representing the positives that are inaccurately identified as negatives [35]. Acc calculates the ratio of correct classification to all classifications; Sen and Spe aim to measure how agreeably the model classified the true positives and true negatives, respectively; PPV indicates the quality of positive classifications; and CK denotes the level of agreements between two raters [30]. Receiver Operator Characteristic (ROC) curves illustrate how the true positive rate changes about the false positive rate as different thresholds are applied [36]. ROC has a valuable feature in that their areas under the curve (AUC) can be compared quantitatively to analyze and contrast them [37].

2.2.2.4 Autocorrelation analysis

This research applied Getis-Ord Gi* proceeded in Moran's I to gain a profound comprehension of the spatial association mechanisms [38]. Some GIS tools denote the statistic as General G-Statistic to show High/Low Clustering of data by the z-score. Equation 7 depicts the General G-Statistic formula [39]. x_i and x_j represent attribute values for two different features in a dataset, $w_{i,j}$ denotes the spatial weight between these two features, n is the total number of features.

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \quad j \neq i$$

Equation 7

3. Results

3.1 Modeling Results

Before conducting RF modeling, selecting the best 'mtry' value, representing the number of variables considered when building Decision Trees, through cross-validation is crucial [30]. The optimal 'mtry' values for RF models are 10, 24, 36, and 41 for splits of 80:20, 75:25, 60:40, and 50:50, respectively. For KNN modeling, finding the optimal k-value is essential to reduce noise in classification [40]. Here, this research used three kernels: optimal, rectangular, and triangular, and two distance metrics: Euclidean and Manhattan. Cross-validation revealed optimal k-values of 15 for 80:20, 5 for 75:25, and 19 for 60:40 and 50:50 splits. Figure 4 exposes the outcomes of the probability mapping conducted via sixteen models at a 30-m resolution. A probability of 1 indicates highly favorable locations for oil and gas well placement, while 0 indicates the least potential areas. The findings categorize probability values into five levels: "Most Potential" (1.0-0.8), "More Potential" (0.8-0.6), "Moderate Potential" (0.6-0.4), "Less Potential" (0.4-0.2), and "Least Potential" (0.2-0.0). In the SVM modeling, the average probability is 0.32, 0.36, 0.32, and 0.30 for SVM 80:20, 75:25, 60:40, and 50:50, respectively. The "Less Potential" class dominates the "rea and varies from 62 to 82% (1200.02 to 1585.48 km^2). Subsequently, the average probability of the RF modeling is 0.60, 0.70, 0.82, and 0.81 for RF 80:20, 75:25, 60:40, and 50:50, respectively.

The "More Potential" and "Moderate Potential" classes prominently encompass the study area under the RF 80:20 configuration, accounting for 45% (871.14 km^2) and 48% (926.68 km^2) of the study area, respectively. Meanwhile, the classes with higher potential, specifically "More Potential" and "Most Potential," prevail in the coverage area for the remaining RF models. The average probability of the ANN modeling is 0.58, 0.57, 0.76, and 0.73 for ANN 80:20, 75:25, 60:40, and 50:50, respectively. In four scenarios, the "Most Potential" class emerges as the dominant category. The largest proportion of this category accounts for 70% of the study area (1353.41 km^2) in the ANN 50:50. In the KNN 80:20, class coverage is evenly distributed, leading to balanced proportions in each potential class. In contrast, the KNN 75:25 and 60:40 can only model two classes: "Most Potential" and "Least Potential." Conversely, the KNN 50:50 is primarily marked by the "More Potential" class, covering 37% of the study area (712.37 km^2).

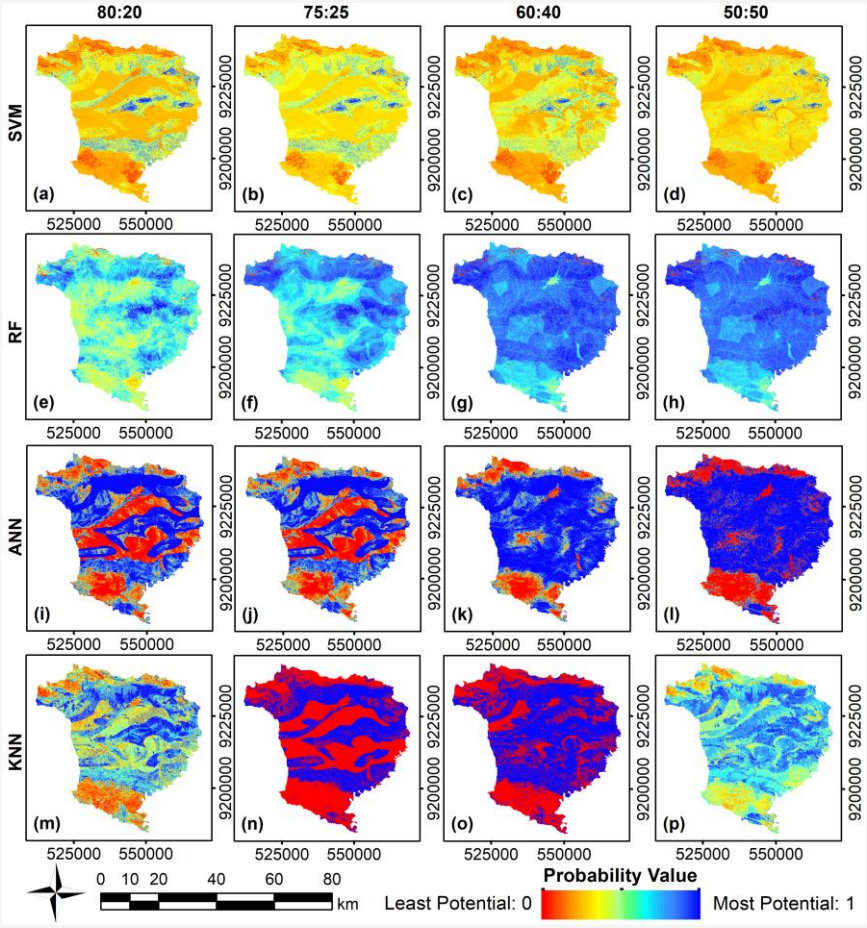

Figure 4: Probability maps: (a-d) SVM models; (e-h) RF models; (i-l) ANN models; (m-p) KNN models; with four splitting ratio scenarios: (a, e, i, m) 80:20, (b, f, j, n) 75:25, (c, g, k, o) 60:40, and (d, h, l, p) 50:50

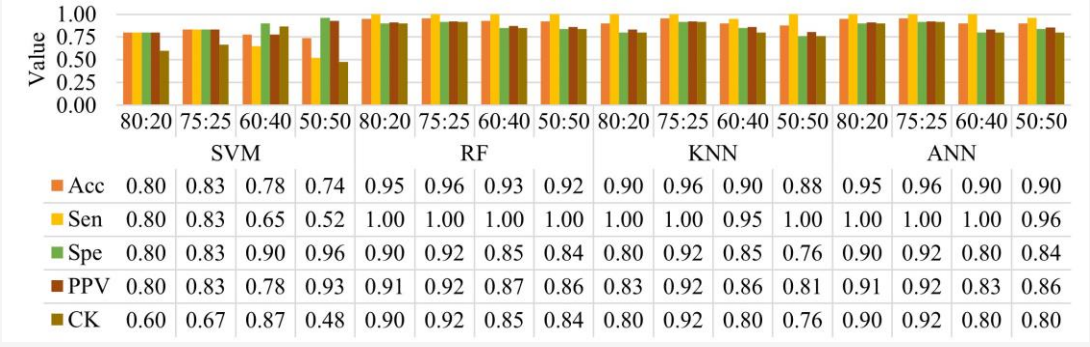

Figure 5: Performance evaluation of testing set from sixteen models

3.2 Model Performance

Figures 5 and 6 show the summary of the model performance evaluation. Among the sixteen models, the RF, KNN, and ANN 75:25 algorithms had the highest accuracy (0.96), while the SVM 50:50 had the lowest (0.74). The average accuracy of the sixteen models was 0.891, with a standard deviation of 0.069. The higher the accuracy value, the better the algorithm's prediction ability [41]. A value of 1 in Sen indicates that all well points (class: yes) are correctly predicted. Meanwhile, the SVM 50:50 had the lowest Sen (0.52). The average Sen of the sixteen models was 0.920, with a 0.146 standard deviation. For the Spe parameter, SVM 50:50 had the highest Spe (0.96), while KNN 50:50 had the lowest value (0.76). The Spe indicates the algorithm's reliability in predicting non-well points (class: no).

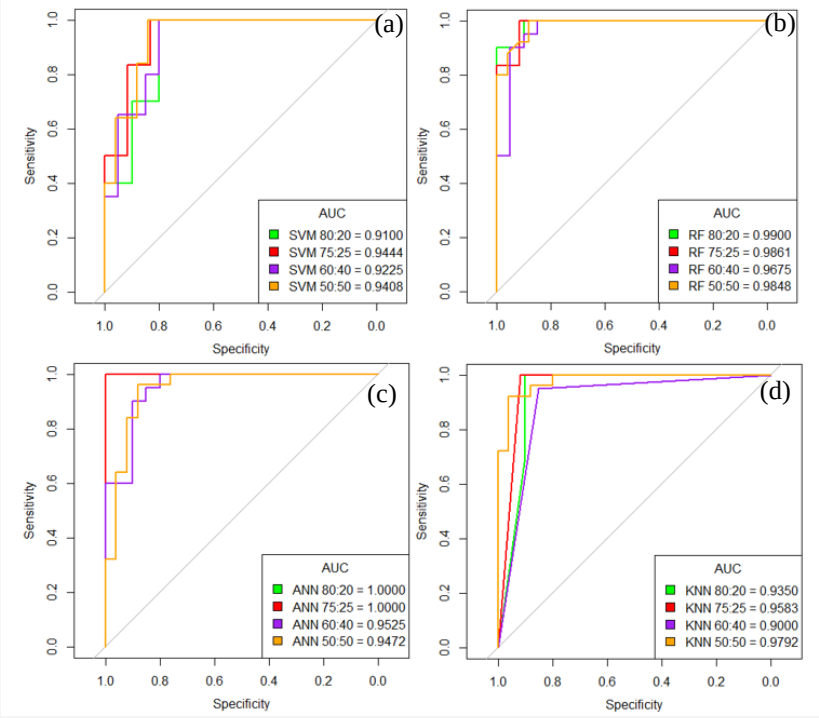

Figure 6: ROC-AUC curves of (a) SVM, (b) RF, (c) ANN, and (d) KNN models

The average Spe of all models was 0.861, with a 0.056 standard deviation. For the PPV parameter, SVM 50:50 and 60:40 had the highest and lowest values, with 0.93 and 0.78, respectively. Models with RF, KNN, and ANN algorithms had PPV performance with an average value of 0.891, 0.857, and 0.881. The average PPV of the sixteen models was 0.866, with a 0.049 standard deviation.

For the CK parameter, the RF, KNN, and ANN 75:25 algorithms had the highest CK (0.92), while the SVM 50:50 had the lowest value (0.48). The CK of 0.48 in the SVM 50:50 falls into the interpretation of a weak level of agreement with 15-35% reliable data. Furthermore, the models with ANN 80:20 and ANN 75:25 have the perfect AUC value, while KNN 60:40 has the lowest value (0.90). Overall, the average AUC value of the sixteen models is 0.957 (strong level) with a 0.031 standard deviation.

Based on the evaluation performance summarized in Figures 5 and 6, the RF, ANN, and KNN models in the 80:20 and 75:25 scenarios demonstrate exceptional results. Subsequently, an overlay analysis was performed to compare the probability values of potential oil and gas well locations from each model with the oil and gas information validation data of Blora Regency in 2008 and 2021, obtained from the ADPMET Discussion of Blora Regency Government 2021. The purpose was to determine the most suitable model.

The overlay results indicate that the RF 80:20 model is the most appropriate. Compared to the latest existing wells in 2021, the eight new wells are situated in areas with probability values of 0.5-0.7, corresponding to the "Moderate Potential" to "More Potential" classes.

4. Discussion

4.1 The Important Index and Spatial Autocorrelation of Surface Driving Factors

To ascertain the significance of the various predictors utilized in the reliability system's modeling, the important degree index (IDI) is employed (Figure 7). This analysis facilitates the selection of superior predictors for modeling endeavors, culminating in heightened accuracy [42]. The SLP showed the highest index (100%) in most models. Meanwhile, based on the IDI of the RF 80:20 as the best result in this study, the rank was continued by the DRO (46.17%), ELV (44.98%), and the DFT (37.56%). Speaking of the lowest IDI in SVM and KNN models, the CBA has a 3.91% and 0.92% index for 80:20 and 75:25 scenarios and LIT has an 11.95% and 12.22% index for 60:40 and 50:50, respectively. However, the lowest IDI for the RF was ASP (5.22% and 2.87%) for the 80:20 and 75:20 scenarios, LST (0.29%) for the 60:40 scenario, and the DRV (0.00%) for the 50:50 scenario.

Besides, the lowest IDI for the ANN are CBA (12.57%), ASP (12.09%), ELV (10.49%), and DFT (2.35%) for the 80:20, 75:25, 60:40, and 50:50 scenarios, respectively. For the spatial autocorrelation analysis, this research considers the finest model (RF 80:20), towards its association with the IDI value. The statistical parameter determines whether data is clustered, random, or evenly distributed.

Moran's I value ranges from -1 to 1, with values closer to -1 indicating even distribution and closer to 1 indicating clustering. A value closer to zero implies a more random distribution [43]. The results from Figure 8 reveal that the SLP factor with the highest IDI does not necessarily have the highest I value. The Moran's statistic varies when its overall relative contribution to modeling is considered [44].

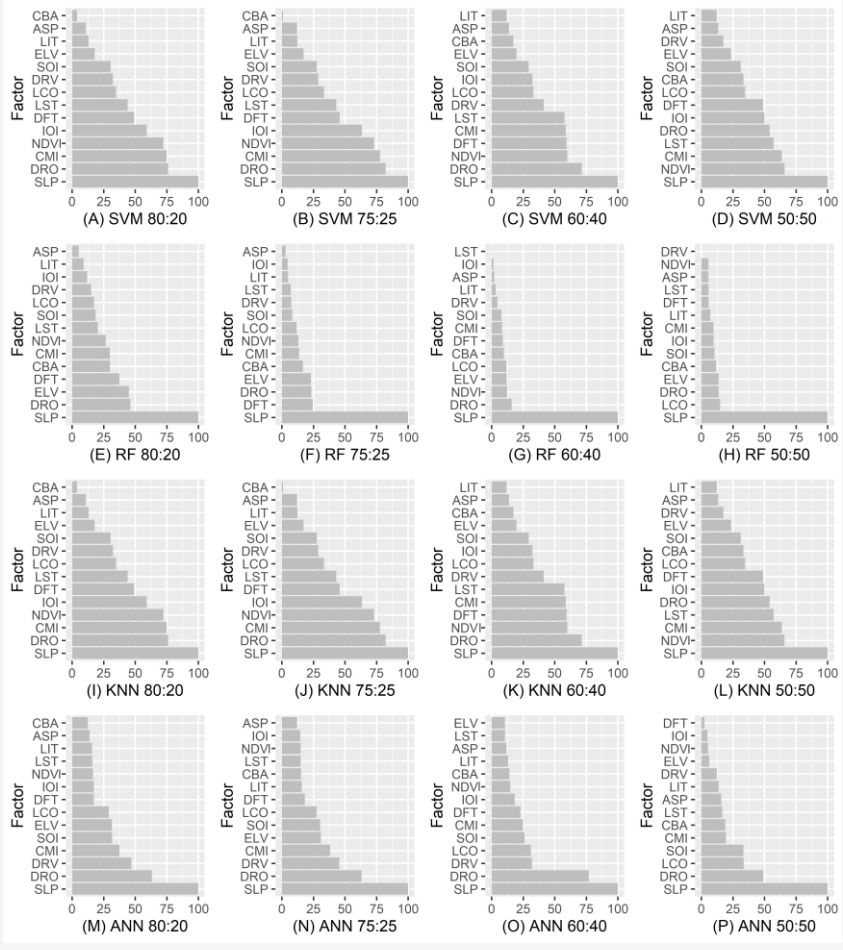

Figure 7: Important Degree Index (IDI) for 16 models

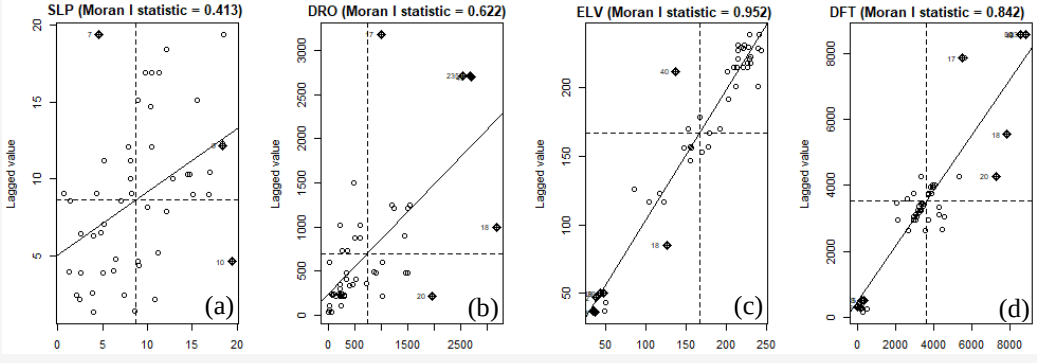

Figure 8: Moran's I Statistic for (a) SLP, (b) DRO, (c) ELV, and (d) DFT factors

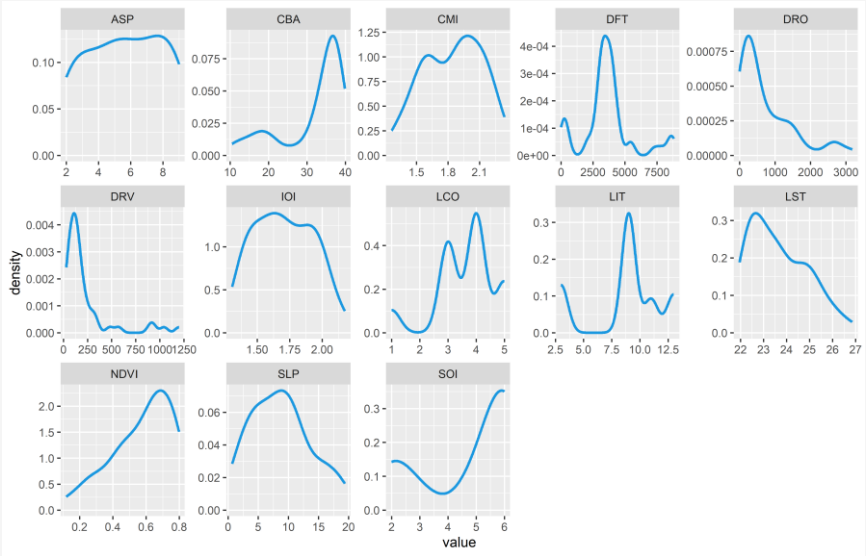

Figure 9: Density distribution of oil and gas well points on a raster value of fourteen factors

4.2 Surface Driving Factors Characteristics of Existing Wells

Figure 9 shows the density of well points in Blora Regency based on their characteristics on the surface factors. The NDVI factor indicates that most wells are in areas with abundant vegetation (index value 0.6-0.8). [20] noted that oil and gas fields often align with anomalies in the vegetation index, likely due to seepage affecting plant growth. This study supports the idea that some wells are in lower vegetation areas. Additionally, some points located in areas with high vegetation density may be ascribed to alterations in land cover conditions at the well sites. Subsequently, LST characteristics vary across oil and gas fields, influenced by land cover changes, resource extraction, urbanization, and green spaces [45]. In a broader context, hydrocarbon seepage areas generally show higher LST values [46]. Moreover, existing wells are found in regions with LST between 22°C and 24°C, while reforestation at some wells may lower LST.

The IOI factor is crucial for identifying hydrocarbon sources on Earth's surface [47] and typically ranges from 1.5 to 1.75 in this study. [48] confirmed that open lands generally have higher iron values than vegetated areas. This study supports this, showing higher IOI at well points than rice fields and surrounding areas. Moreover, the CBA factor shows significant variations at well points, with dominant values between 30-40 mGal. CBA reflects subsurface rock density patterns influenced by sediment thickness and crustal and lithospheric density [49]. Specific CBA patterns in this study remain unclear, and further investigation, including spectral analysis, is needed to understand the area's rock structure [50].

For safety reasons, wells should be over 15 meters from faults to avoid damage zones where water injection can cause casing leaks [51]. In this study, most wells are 2.5 to 4 kilometers from faults (Figure 9). Regarding geological types, most wells are placed over the Wonocolo Formation, primarily at Tmw category, consisting of sedimentary rocks like limestone sandstone, claystone, and glauconite sandstone [52]. Sedimentary rocks, making up about 75% of the Earth's surface, are key in forming oil and natural gas from buried organic material [53]. Moreover, about 25% of the wells have a slope range of 0-10.393%, with most between 5-11%, and are at 200-250 meters elevations. Sixteen wells are below 100 meters, and 34 are between 100-250 meters elevation. This finding matches [11], indicating that oil and gas wells are typically on moderate slopes at lower elevations. Next, the ASP factor varies widely, reflecting the practice of avoiding steep slopes to reduce risks of geohazards, pipeline ruptures, oil spills, and service disruptions [54].

Optimally placing well points for oil and gas exploration is vital. This placement relies on proximity to roads and rivers, land cover, and soil composition. Existing wells are found within 0-500 meters of roads, as observed in recent studies [55], and 0-250 meters of rivers. Accessible water sources are essential for extraction processes [11] and [56]. Regarding land cover, most wells are situated in farm field, which allow easy infrastructure development and are reasonably far from residential zones. Subsequently, the prevalent soil type at the well sites in this study is Mollisol, known for its fertility and use in agriculture due to its aeolian and carbonaceous sediments [57].

Mollisol, often rich in clay, is common in oil and gas fields [58]. However, its organic content necessitates geotechnical assessments for slope stability and bearing capacity before planning well placements.

5. Conclusion

This study uses SVM, RF, KNN, and ANN algorithms to assess potential well locations. Fourteen environmental factors—NDVI, CMI, IOI, LST, LIT, DFT, CBA, SLP, ASP, ELV, DRV, DRO, SOI, and LCO—are analyzed along with four splitting ratios. The models produce distinct values for probability, IDI, and Moran's I Statistic. The RF 80:20 model performs best with an Acc of 0.95, Sen of 1.00, Spe and CK of 0.90, PPV of 0.91, and AUC of 0.99. It indicates that 48.05% (926.69 km²) of Blora Regency has a "Moderate Potential" for wells, followed by 45.17% with a "More Potential" level. Key factors influencing well location include SLP, DRO, ELV, and DFT, which are related to land movement, construction accessibility, and ease of finding hydrocarbons. For future studies, this research recommends using feature selection methods like multicollinearity analysis and reviewing sample point characteristics for accurate results.

Acknowledgments

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2024.

References

[1] Pusat Pengkajian Industri Proses dan Energi (PPIPE), (2021). *Outlook Energi Indonesia 2021 Perspektif Teknologi Energi Indonesia [Indonesia Energy Outlook 2021 Indonesian Energy Technology Perspective]*. Badan Pengkajian dan Penerapan Teknologi, Jakarta, *Technical Report*.

[2] Arkyasa, M., (2023). Minister: Oil Production Decline Because of Old Wells. *Tempo*. [Online]. Available: https://en.tempo.co/read/1685810/minister-oil-production-decline-because-of-old-wells. [Accessed: Dec. 15, 2024].

[3] Asian Development Bank (ADB), (2021). *Indonesia Energy Sector Assessment, Strategy and Road Map Update [Internet]*. Manila. https://doi.org/10.22617/TCS200429.

[4] Fathoni, H. S., Setyowati, A. B. and Prest, J., (2021). Is Community Renewable Energy Always Just? Examining Energy Injustices and Inequalities in Rural Indonesia. *Energy Research & Social Science*, Vol. 71. https://doi.org/10.1016/j.erss.2020.101825.

[5] Rahman, M., Munadi, S., Widarsono, B. and Caryana, Y. K., (2011). Technology Challenges in Indonesia Oil and Gas Development. *Lemigas Scientific Contributions*, Vol. 34(1), 11–17.

[6] Bondar, K. M., Minaev, V. A. and Faddeev, A. O., (2021). Cost Estimate for Exploration of Oil and Gas Fields in the Arctic Zone. *IOP Conference Series: Materials Science and Engineering*, Vol. 1079(6). https://doi.org/10.1088/1757-899X/1079/6/062069.

[7] Sheng, J., Sun, J., Bai, Y., Liu, Z., Wei, H., Li, L., Su, G. and Wang, Z., (2020). Evaluation of Hydrocarbon Potential Using Fuzzy AHP-based Grey Relational Analysis: A Case Study in the Laoshan Uplift, South Yellow Sea, China. *Journal of Geophysics and Engineering*, Vol. 17(1), 189–202. https://doi.org/10.1093/jge/gxz107.

[8] Ahmad, W. A., Ahmed, M. A. and Al-sharia, G. H., (2017). Using Normalized Difference Vegetation Index (NDVI) to Identify Hydrocarbon Seepage in Kifl Oil Field and Adjacent Areas South of Iraq. *Journal of Environment and Earth Science*, Vol. 7(1), 16-27.

[9] Suliantara, S., Susantoro, T. M., Setiawan, H. L. and Firdaus, N., (2021). A Preliminary Study on Heavy Oil Location in Central Sumatra using Remote Sensing and Geographic Information Sytem. *Scientific Contributions Oil and Gas*, Vol. 44(1), 39–54. https://doi.org/10.29017/SCOG.44.1.489.

[10] Alaofin, O. T., (2022). Application of Gravity *Data for Hydrocarbon Exploration Using Machine Learning Assisted Workflow*. Master's Thesis. Louisiana State University and Agricultural and Mechanical College. Available: https://repository.lsu.edu/gradschool_theses/5482/.

[11] Purba, D., Adityatama, D. W., Fadhillah, F. R., Al-asyari, M. R., Ivana, J., Abi, R., Tiyana, Larasati, T., Gumelar, P., Gunawan, A., Shafar, N. A., Anugrah, A. N. M. and Nugraha, R. P., (2021). A Discussion on Oil & Gas and Geothermal Drilling Environment Differences and their Impacts to Well Control Methods. *Proceeding, 47th Workshop on Geothermal Reservoir Engineering*, California. 1-15.

[12] United Nations (UN), (2018). Guidance on Land-Use Planning, the Siting of Hazardous Activities and Related Safety Aspects. *UN-iLibrary*. [Online]. Available: https://doi.org/10.18356/df07526b-en. [Accessed: Dec. 15, 2024].

[13] United States Department of the Interior and United States Department of Agriculture, (2005). Chapter 4-Construction and Main tenance. *Surface Operating Standards and Guidelines for Oil and Gas Exploration and Development*, 4th ed. BLM National Science and Technology Center, Denver. 15–36.

[14] Sircar, A., Yadav, K., Rayavarapu, K., Bist, N. and Oza, H., (2021). Application of Machine Learning and Artificial Intelligence in Oil and Gas Industry. *Petroleum Research*, Vol. 6(4), 379-391.

[15] Ahmadi, M. A. and Bahadori, A., (2015). A LSSVM Approach for Determining Well Placement and Conning Phenomena in Horizontal Wells. *Fuel*, Vol. 153, 276–283. https://doi.org/10.1016/j.fuel.2015.02.094.

[16] Dalatu, P. I., Fitrianto, A. and Mustapha, A., (2017). A Comparative Study of Linear and Nonlinear Regression Models for Outlier Detection. *Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016)*, Herawan T., Ghazali R., Nawi NM., Deris MM., Eds. Bandung: Springer, 2016. 316–326.

[17] Susantoro, T. M., Wikantika, K., Suliantara, S., Setiawan, H. L., Harto, A. B. and Sakti, A. D., (2023). Applying Random Forest to Oil and Gas Exploration in Central Sumatra Basin Indonesia Based on Surface and Subsurface Data. *Remote Sensing Applications: Society and Environment*, Vol. 32. https://doi.org/10.1016/j.rsase.2023.101039.

[18] Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I. and Pham B. T., (2021). Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Shen Y-S, Ed. *Mathematical Problems in Engineering*, 1–15. https://doi.org/10.1155/2021/4832864.

[19] Joseph, V. R. and Vakayil, A., (2022). SPlit: An Optimal Method for Data Splitting. *Technometrics*, Vol. 64(2), 166–176. https://doi.org/10.1080/00401706.2021.1921037.

[20] Susantoro, T. M., Saepuloh, A., Agustin, F., Wikantika, K. and Harsolumakso, A. H., (2020). Clay Mineral Alteration in Oil and Gas Fields: Integrated Analyses of Surface Expression, Soil Spectra, and X-Ray Diffraction Data. *Canadian Journal of Remote Sensing*, Vol. 46(2), 237–251. https://doi.org/10.1080/07038992.2020.1771174.

[21] Thammaboribal, P., (2024). Investigating Land Surface Temperature Variation and Land Use Land Cover Changes in Pathumthani, Thailand (1997-2023) using Landsat Satellite Imagery: A Comprehensive Analysis of LST and Urban Hot Spots (UHS). *International Journal of Geoinformatics*, Vol. 20(2), 27–41. https://doi.org/10.52939/ijg.v20i2.3063.

[22] Zaenudin, A., Dani, I. and Amalia, N., (2020). Delineasi Sub-Cekungan Sorong Berdasarkan Anomali Gaya Berat [Delineation of Sorong Sub-Basin Based on Gravity Anomaly]. *Jurnal Geocelebes*, Vol. 4(1). https://doi.org/10.20956/geocelebes.v4i1.7976.

[23] Anjasmara, I.M., (2020). *Buku Ajar Geodesi Fisik* [Physical Geodesy Textbook]. Surabaya: Departemen Teknik Geomatika, Institut Teknologi Sepuluh Nopember.

[24] Ozsahin, D. U., Taiwo Mustapha, M., Mubarak, A. S., Said Ameen, Z. and Uzun, B., (2022). *Impact of Feature Scaling on Machine Learning Models for the Diagnosis of Diabetes*. 2022 International Conference on Artificial Intelligence in Everything (AIE), August 2-4, 2022, Lefkosa, Cyprus. 87-94. https://doi.org/10.1109/AIE57029.2022.00024.

[25] Shmilovici, A., (2009). Support Vector Machines. *Data Mining and Knowledge Discovery Handbook*, Springer US, Boston, MA. 231–247. https://doi.org/10.1007/978-0-387-09823-4_12.

[26] Dibike, Y. B., Velickov, S., Solomatine, D. and Abbott, M. B., (2001). Model Induction with Support Vector Machines: Introduction and Applications. *Journal of Computing in Civil Engineering*, Vol. 15(3), 208–216. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208)

[27] Shi, H., Xiao, H., Zhou, J., Li, N. and Zhou, H., (2018). *Radial Basis Function Kernel Parameter Optimization Algorithm in Support Vector Machine Based on Segmented Dichotomy*. 2018 5th International Conference on Systems and Informatics (ICSAI), November 10-12, 2018.

[28] Aji, A., Husna, V., and Purnama, S., (2024). Multi-Temporal Data for Land Use Change Analysis Using a Machine Learning Approach (Google Earth Engine). *International Journal of Geoinformatics*, Vol. 20(4), 19–28. https://doi.org/10.52939/ijg.v20i4.3145.

[29] Palczewska, A., Palczewski, J., Marchese Robinson, R. and Neagu, D., (2014). *Interpreting Random Forest Classification Models Using a Feature Contribution Method*. Bouabana-Tebibel, T., Rubin, S. (eds) Integration of Reusable Systems. Advances in Intelligent Systems and Computing, Vol. 263 https://doi.org/10.1007/978-3-319-04717-19.

[30] Nurwatik, N., Ummah, M. H., Cahyono, A. B., Darminto, M. R. and Hong, J. H., (2022). A Comparison Study of Landslide Susceptibility Spatial Modeling Using Machine Learning. *ISPRS International Journal of Geo-Information*, Vol. 11(12). https://doi.org/10.3390/ijgi11120602.

[31] Zou, J., Han, Y. and So, S. S., (2008). Overview of Artificial Neural Networks. *Methods in Molecular Biology*, J. M. Walkers, Ed. Springer, 14-22.

[32] Mehdi, B., Brahmi-Ingrachen, D., Belkacemi, H. and Muhr, L., (2023). Development of a Mathematical Model Based on an Artificial Neural Network (ANN) to Predict Nickel Uptake Data by a Natural Zeolite. *Physical Sciences Forum*, Vol. 6(1). https://doi.org/10.3390/psf2023006004.

[33] Zhang, Y., Cao, G., Wang, B. and Li, X., (2019). A Novel Ensemble Method for K-Nearest Neighbor. *Pattern Recognition*, Vol. 85, 13–25. https://doi.org/10.1016/j.patcog.2018.08.003.

[34] Parvis, H., Alizadeh, H. and Minaei-Bidgoli, B., (2008). MKNN: Modified K-Nearest Neighbor. *Proceedings of the World Congress on Engineering and Computer Science*, https://www.iaeng.org/publication/WCECS2008/WCECS2008_pp831-834.pdf.

[35] Zeng, G., (2020). On the Confusion Matrix in Credit Scoring and its Analytical Properties. *Communications in Statistics - Theory and Methods*, Vol. 49(9), 2080–2093. https://doi.org/10.1080/03610926.2019.1568485.

[36] Muschelli, J., (2020). ROC and AUC with a Binary Predictor: A Potentially Misleading Metric. *Journal of Classification*, Vol. 37(3), 696-708. https://doi.org/10.1007/s00357-019-09345-1.

[37] Clark, R. D. and Webster-Clark, D. J., (2008). Managing Bias in ROC Curves. *Journal of Computer-Aided Molecular Design*, Vol. 22, 141–146. https://doi.org/10.1007/s10822-008-9181-z.

[38] Ord, J. K. and Getis, A., (2010). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, Vol. 27(4), 286–306. https://doi.org/10.1111/j.1538-4632.1995.tb00912.x.

[39] Thammaboribal, P., TRIPATHI, N., Junpha, J., Lipiloet, S., and Wongpituk, K. (2024). Examining the Correlation between COVID-19 Prevalence and Patient Behaviors, Healthcare, and Socioeconomic Determinants: A Geospatial Analysis of ASEAN Countries. International Journal of Geoinformatics, Vol. 20(3), 95–112. https://doi.org/10.52939/ijg.v20i3.3159.

[40] Lubis, Z., Sihombing, P. and Mawengkang, H., (2020). Optimization of K Value at the K-NN Algorithm in Clustering Using the Expectation Maximization Algorithm. *IOP Conference Series: Materials Science and Engineering*, Vol. 725(1). https://doi.org/10.1088/1757-899X/725/1/012133.

[41] Ma, J., Ding, Y., Cheng, J. C. P., Tan, Y., Gan, V. J. L. and Zhang, J., (2019). Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective. *IEEE Access*, Vol. 7, 148059-148072. https://doi.org/10.1109/ACCESS.2019.2946401.

[42] Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T. and Zeileis, A., (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, Vol. 9(1). https://doi.org/10.1186/1471-2105-9-307.

[43] Jackson, M. C., Huang, L., Xie, Q. and Tiwari, R. C., (2010). A Modified Version of Moran's I. *International Journal of Health Geographics*, Vol. 9(1). https://doi.org/10.1186/1476-072X-9-33.

[44] Liu, X., Kounadi, O. and Zurita-Milla, R., (2022). Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features. *ISPRS International Journal of Geo-Information*, Vol. 11(4). https://doi.org/10.3390/ijgi11040242.

[45] Effat, H. A. and Hassan, O. A. K., (2014). Change Detection of Urban Heat Islands and Some Related Parameters Using Multi-Temporal Landsat Images; A Case Study for Cairo City, Egypt. *Urban Climate*, Vol. 10. 171–188. https://doi.org/10.1016/j.uclim.2014.10.011.

[46] Suherman, A., Rahman, M. Z. A. and Busu, I., (2014). Albedo and Land Surface Temperature Shift in Hydrocarbon Seepage Potential Area, Case Study in Miri Sarawak Malaysia. *IOP Conference Series: Earth and Environmental Science*, Vol. 18. https://doi.org/10.1 088/1755-1315/18/1/012148.

[47] Ezzat, A. O., Ali, M. S. and Al-Lohedan, H. A., (2022). Synthesis, Characterization, and Application of Magnetite Nanoparticles Coated with Hydrophobic Polyethyleneimine for Oil Spill Cleaning. *Journal of Chemistry*, Vol. 2022(1). https://doi.org/10.1155/2022/3368298

[48] Arisanty, D., Saputra, A.N., Rahman, A.M., Hastuti, K. P. and Rosadi, D., (2021). The Estimation of Iron Oxide Content in Soil based on Landsat 8 OLI TIRS Imagery in Wetland Areas. *Pertanika Journal of Science and Technology*, Vol. 29(4). 2829–2843. https://doi.org/10.47836/PJST.29.4.32.

[49] Lewerissa, R., Alzair, N. and Lapono, L., (2021). Identification of Ransiki Fault Segment in South Manokwari Regency, West Papua Province, Indonesia based on Analysis of a High-resolution of Global Gravity Field: Implications on the Earthquake Source Parameters. *IOP Conference Series: Earth and Environmental Science*, Vol. 873(1). https://doi.org/10.1088/1755-1315/873/1/012048.

[50] Santoso, D., Tirza, A., Wahyudi, E.J., Alawiyah, S., Kadir, W. G. A. and Sule, R., (2018). An Application of Gravity Method to Estimate a Storage Capacity of Ngrayong Formation for Carbon Capture and Storage (CCS) Pilot Project of Gundih Field, East Java, Indonesia. *SSRN Electronic Journal*,. https://doi.org/10.2139/ssrn.3365938.

[51] Fu, X. F., Lan, X., Meng, L.D., Wang, H. X., Liu, Z. B., Guo, Z. Q. and Chen, Z. H., (2016). Characteristics of Fault Zones and their Control on Remaining Oil Distribution at the Fault Edge: A Case Study from the Northern Xingshugang Anticline in the Daqing Oilfield, China. *Petroleum Science*, Vol. 13(3). 418–433. https://doi.org/10.1007/s12182-016-0116-3.

[52] Prihutama, F. A., Danistya, A. and Widada, S., (2018). Karakteristik Geologi dan Skenario Reservoir Hidrokarbon Sebagai Rencana Pengembangan Zona Prospek Lapangan "Tesseract" Cekungan Jawa Timur Utara pada Zona Rembang, Jawa Timur [Geological Characteristics and Hydrocarbon Reservoir Scenarios as a Development Plan for the "Tesseract" Field Prospect Zone of the North East Java Basin in the Rembang Zone, East Java]. *Seminar Nasional Kebumian ke-11*, 436–454.

[53] Downey, M. W., (2004). Oil and Natural Gas Exploration. *Encyclopedia of Energy*, Elsevier. 549–558.

[54] Lee, E. M., Fookes, P. G. and Hart, A. B., (2016). Landslide Issues Associated with Oil and Gas Pipelines in Mountainous Terrain. *Quarterly Journal of Engineering Geology and Hydrogeology*, Vol. 49(2). 125–131. https://doi.org/10.1144/qjegh2016-020.

[55] Dwi Safira, R. A., Nurwatik, N. and Hariyanto, T., (2023). Identifying Potential Areas for Oil and Gas Well Location Planning Using Support Vector Machine Algorithm. *IOP Conference Series: Earth and Environmental Science*, Vol. 1276(1). https://doi.org/10.1088/1755-1315/1276/1/012068.

[56] Allison, E. and Mandler, B., (2018). *Water in the Oil and Gas Industry*. Petroleum and the Environment. American Geoscience. [Online] Available: https://www.americangeosciences.org/sites/default/files/AGIPEWaterIntrowebfinal.pdf. [Accessed: Dec. 15, 2024].

[57] Eswaran, H. and Reich, P.F., (2005). *World Soil Map.* Encyclopedia of Soils in the Environment, Elsevier, 352–365. https://doi.org/10.1016/B0-12-348530-4/00019-9.

[58] Khan, A. H. and Yousaf, A., (2016). Mollisols Soils Stabilization Using Lime Modified by Salts. *Pakistan Journal of Engineering and Applied Sciences*, Vol. 18. 78–88.