Modern Technologies for Ranking Territories by Hydrocarbon Prospectivity Using Machine Learning

Danis Nurgaliev, Eduard Ziganshin *, Timur Safin, Fagim Garaev

Institute of geology and petroleum technology, Kazan Federal University, Kazan, Russia
erziganshin@kpfu.ru

Abstract. This article presents a geological and geophysical technology designed to assess the oil and gas potential of territories without the use of expensive seismic or other conventional methods. The technology is based on the integration of geological, geophysical, and geochemical data using original mathematical algorithms developed at Kazan Federal University (Russia). Its modular design allows the research program to be adapted to specific geological conditions, the degree of exploration maturity, and available financial resources, thus ensuring an optimal balance between cost and informational value. The novelty of the work lies in the application of machine-learning algorithms for an integrated prospectivity assessment. Feature selection was performed using a correlation matrix that accounts for nonlinear relationships.

Keywords: Magnetic Surveying, Geochemical Hydrocarbon Exploration, Machine Learning, Non-Seismic Methods, Petroleum Prospecting.

1 Introduction

Seismic surveying traditionally plays a key role in the exploration and appraisal of hydrocarbon fields, primarily addressing structural objectives. An alternative approach involves methods that directly or indirectly indicate the presence of hydrocarbons in the subsurface, including geochemical and various non-seismic geophysical techniques. These methods are significantly less expensive compared with seismic surveys and offer the advantage of ranking the prospectivity of a territory for subsequent detailed investigations.

According to current concepts and estimations based on magneto-mineralogical and paleomagnetic methods, the age of most petroleum accumulations does not exceed 10^6–10^7 years [1]. This is largely due to the fact that hydrocarbon accumulations begin to degrade almost immediately after hydrocarbons (HCs) are trapped. Migrating hydrocarbons move through fracture networks, undergo oxidation, and are biodegraded by microorganisms, thereby altering the redox conditions of the surrounding environment. These epigenetic transformations create geophysical and geochemical anomalies above the accumulations, which serve as indicators used in this study to identify potential hydrocarbon reservoirs.

R. Rzayev et al. (eds.), *Proceedings of the International Conference on Current Problems in Engineering and Applied Sciences (ICCPEAS 2025)*, Advances in Engineering Research 299,
https://doi.org/10.2991/978-94-6239-668-5_84

The technology described in this work is based on several fundamental assumptions:

Hydrocarbons in oil and gas forms are generated through the transformation of organic matter in source rocks within the sedimentary succession under the influence of heat and mass transfer from the lower crust and upper mantle.

Multiple cycles of hydrocarbon generation and migration into traps may occur within the same region.

From the moment of formation, hydrocarbon accumulations begin to degrade and develop dispersion halos in overlying strata; they cannot persist over long geological timescales in the sedimentary cover. Their lifetime depends on geodynamic conditions and the efficiency of sealing formations.

Neotectonic processes significantly influence both degradation and potential reformation of hydrocarbon accumulations.

The increasing need to analyze complex datasets and identify nonlinear relationships has led to the widespread adoption of machine-learning algorithms. The objective of this study is to demonstrate the efficiency of the proposed technology and evaluate the applicability of machine learning for prospectivity prediction.

2 Study Area and Methods

Administratively, the study area is located within the Far Eastern Federal District. From the standpoint of tectonic zoning, the territory lies within the Vilyui Syneclise of the East Siberian Platform. The sedimentary cover comprises Riphean, Vendian, Paleozoic, and Mesozoic strata. Basement depth varies from several kilometers beneath uplifts to 5–12 km in depressions. Riphean deposits are likely present only fragmentarily and restricted to the flanks of the hemisyneclise. Vendian deposits unconformably overlie Riphean and basement rocks. Paleozoic and Mesozoic strata are represented by all major systems. In terms of petroleum potential, the area belongs to the Vilyui gas-bearing region of the Lena–Vilyui petroleum province. Four productive complexes are recognized: Upper Permian, Lower Triassic, Middle Triassic–Lower Jurassic, and Middle–Upper Jurassic [2]. The proposed technology consists of several stages.

2.1 Stage 1: Morphometric Analysis of Digital Elevation Models (DEM)

At the first stage, a digital elevation model (DEM) is analyzed to identify areas potentially favorable for the preservation of hydrocarbon accumulations. Key criteria include low neotectonic activity, moderate fracture density in the sedimentary cover, and calibration using the spatial distribution of known fields, productive wells, and dry wells. As a result, areas with high and low prospects for substantial hydrocarbon accumulations were identified.

Morphometric analysis relies on the premise that landforms reflect the combined influence of endogenous (tectonic) and exogenous (erosional) processes during the neotectonic stage. Neotectonic movements strongly affect the formation and destruction of hydrocarbon accumulations. DEM analysis also provides insight into the macroscopic fracture system of the sedimentary cover through the evaluation of lineament density. Surface lineament density primarily reflects the activity of crustal blocks. Where blocks exhibit the highest modern mobility, lineament density tends to be greatest.

2.2 Stage 2: Planning and Acquisition of Geophysical and Geochemical Surveys

The second stage involves planning survey profiles and measurement stations for detailed geophysical and geochemical investigations. The method suite includes tidal gravimetric monitoring, GNSS geodetic monitoring, high-precision magnetic surveying, self-potential (SP) surveying, and geochemical sampling of near-surface sediments (chromatography and isotope analysis). Each method provides prospectivity indicators relevant to the detection of oil and gas accumulations.

Localization Based on Tidal Gravimetry and GNSS Monitoring. The movement of crustal blocks during lunisolar tides can be detected using satellite geodesy (GPS/GNSS). These observations yield information on block size (coordinates within a block respond coherently), geometry of subvertical block boundaries (based on displacement vectors), and macroscopic geomechanical properties such as Poisson's ratio, which varies with gas saturation. Gas-saturated blocks display two key behaviors:

– They exhibit a time lag in reaching tidal maxima and minima; the lag increases with gas saturation and may reach several tens of minutes. This parameter is referred to as the "Shift" parameter.

– Their uplift amplitude during tidal forcing is slightly higher than that of non-gas-saturated blocks. The ratio of theoretical to observed amplitude is termed the "Ratio" parameter.

These features can be detected by seismic block tomography, continuous GNSS monitoring, or quasi-continuous gravimetric measurements. Gravimetry is the most cost-effective and reliable method due to its integral character and low sensitivity to local noise. Ideally, gravimetric monitoring is complemented by GNSS to determine both the tidal lag and amplitude. In this study, the Ratio parameter proved more informative due to high noise levels. Since it characterizes entire blocks (5–10 km in size), it enables spatial delineation of gas-saturated regions. Several such regions were identified, corresponding to areas with elevated compressibility and, consequently, increased gas saturation.

Localization Based on Geochemical Sampling. Geochemical and isotopic analyses are effective tools for evaluating hydrocarbon potential. Isotopic data enable the determination of gas concentrations, origin, and emission intensity. Methane and its homologs provide key diagnostic information. The isotopic composition of methane ($\delta^{13}C$) is a reliable indicator of its genesis [3]. In the study area, methane is predominantly thermogenic, indicating deep hydrocarbon generation processes.

However, high near-surface methane concentrations alone do not confirm a commercial accumulation. They may result from direct emission from source-rock intervals lacking an effective seal. Additional information is provided by biodegradation markers of light alkanes: the ratios $i\text{-}C_4/n\text{-}C_4$ and $i\text{-}C_5/n\text{-}C_5$ reflect microbial oxidation, as microorganisms preferentially consume normal alkanes. Zones of enhanced biodegradation may indicate suppressed hydrocarbon flow due to the presence of a seal and possible accumulation. Minimal biodegradation suggests strong direct fluid discharge.

Comparative biodegradation of butane and pentane yields further diagnostic information: preferential biodegradation of pentane suggests its anomalous enrichment in the deep fluid, indicating potential condensate-rich accumulations.

Although none of these indicators is uniquely diagnostic on its own, their integrated analysis provides valuable evidence to be considered in subsequent interpretation.

Localization Based on High-Precision Magnetic and SP Surveys. High-precision magnetic surveying and SP measurements aim to identify vertical HC migration pathways and confirm geochemical anomalies. Their efficiency stems from epigenetic mineralogical and geochemical transformations caused by upward HC migration.

Hydrocarbon flux modifies redox conditions, shifting the redox boundary (ROB) upward to depths of 100–300 m, and nearly to the surface along faults. Oxidized upper zones contain Fe^{3+} oxides and hydroxides (hematite, goethite, lepidocrocite) with high magnetic susceptibility. Reduced zones contain Fe^{2+} minerals (pyrite, siderite), forming low-magnetization intervals. Ferrimagnetic sulfides (e.g., greigite) may form in near-surface fracture zones. This vertical mineral zonation produces local magnetic minima (10–20 nT amplitude; source depths 500–1000 m) above HC migration zones.

Self-potential anomalies arise from both electrochemical and filtration mechanisms [4-7]. The ROB functions as a natural electrochemical cell; its depth variations generate SP anomalies (ΔU_SP). Electrochemical anomalies are driven by redox reactions in zones with migrating HC fluids. Filtration anomalies reflect the movement of groundwater: discharge zones yield positive ΔU_SP anomalies; recharge zones produce negative ones.

Above HC reservoirs, magnetic and SP anomalies often correlate. Coincidence of magnetic minima, SP anomalies (source depth 200–600 m), and elevated thermogenic methane fluxes indicates zones of through-going hydrocarbon emission. Positive SP anomalies may indicate areas with reduced methane flux and the presence of an effective seal.

The most promising targets are not narrow fault zones but broad areas of large-scale HC emission through extensive rock volumes - readily identified by consistent magnetic and SP anomalies.

2.3 Stage 3: Integration of All Methods

At the third stage, all acquired datasets were integrated for comprehensive interpretation.

3 Results

The ranking of the study area by prospectivity was performed by summing the favorable factors characteristic of hydrocarbon accumulations in the region:
areas with medium to high values of macro-fracturing within the sedimentary cover;

areas of low geodynamic activity during the neotectonic period;
areas demonstrating anomalous behavior in several tide-related parameters (e.g., amplitude — Ratio parameter, phase — Shift parameter);
areas exhibiting a flux of thermogenic methane from depth, identified through near-surface soil sampling;
areas showing significant biodegradation of butane and pentane homologues in soil-gas samples;
areas with "negative" magnetic anomalies of 10–20 nT, with sources located at depths of 500–1000 m (zones of reduced magnetization in rocks);
areas with both negative and positive natural electric field anomalies of at least 25 mV, with sources located at depths of 200–600 m, indicating the position of the redox boundary.

When the indicators described above partially or fully coincide, areas of the study region can be classified according to their varying prospects for hydrocarbon accumulation. The different indicators possess unequal levels of informativeness, which in the conventional approach is accounted for through expert judgment. In this study, quantitative weighting of factors was not introduced, as the low spatial resolution of several datasets and the large extent of the investigation area do not allow robust assignment of universal coefficients.

It should also be noted that some of the applied methods (DEM analysis, gravimetric and GPS monitoring) provide spatially continuous, area-based information, whereas profile-based methods (electrical prospecting, magnetic surveys, and geochemical sampling) delineate narrow anomalous zones along survey lines. Therefore, at the first stage, we outlined regional-scale areas where the prospective features identified from DEM, gravimetric, and geodetic data showed mutual agreement. Subsequently, for these areas, the results of profile-based surveys were analyzed. Based on the combined set of indicators, the territories were conditionally grouped into three classes of prospectivity. The category of "extremely high prospectivity" includes areas where anomalies detected by geophysical (SP, magnetometry) and geochemical methods are consistent with expert interpretation. The "high prospectivity" class comprises areas where no more than three indicators suggest low potential, provided that the profile-based evidence remains supportive. Areas where four indicators characterize the area as non-prospective were assigned to the "moderate prospectivity" class.

However, it should be emphasized that the combined geophysical and geochemical indicators form a highly multidimensional system with partially nonlinear relationships. Under such conditions, expert interpretation remains essential, although its reproducibility and degree of formalization are limited. To verify the identified patterns and quantitatively assess the consistency of heterogeneous indicators, an additional ranking procedure using machine-learning methods was applied at the final stage of the study.

The machine-learning models were constructed using a combined set of geophysical and geochemical parameters. The training dataset consisted of 120 samples (with balanced class distribution), and the test dataset consisted of 30 samples; the total number of observations was 346. This relatively small dataset determined the choice of algorithms capable of operating effectively with limited sample sizes and potentially imbalanced classes.

The classification procedures used Logistic Regression, Support Vector Classifier, k-Nearest Neighbors, and Random Forest algorithms, all of which have proven effective for small and partially imbalanced datasets. Prior to model training, standard data preprocessing was carried out, including handling missing values, feature scaling (where required by the algorithm), and an assessment of inter-feature dependencies. Initial feature selection was performed using the Phik (ϕk) correlation matrix. In contrast to conventional correlation coefficients, this method accounts for both linear and nonlinear relationships between variables (Fig. 1).

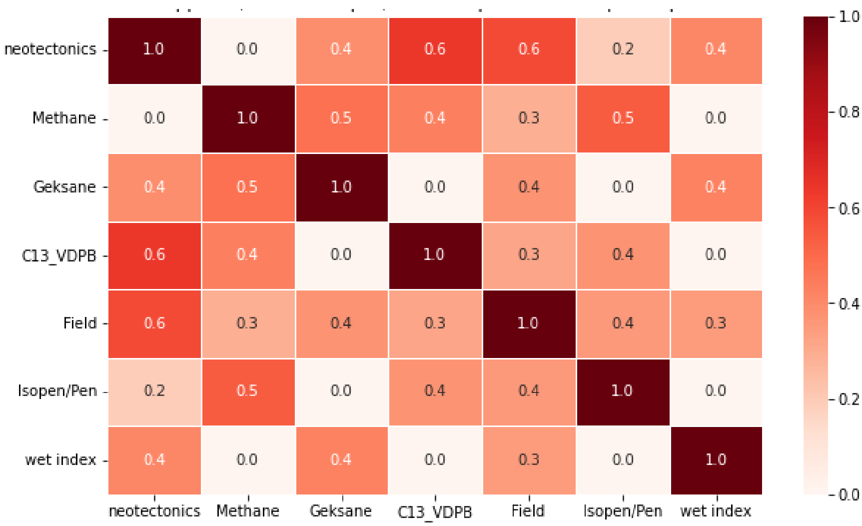

Fig. 1. Correlation matrix of the filtered features.

To evaluate the performance of the models, the ROC curve metric was used alongside additional classification indicators. The robustness of the results was assessed through cross-validation and bootstrap resampling. The RandomForest model demonstrated the best performance, yielding a ROC-AUC value of 0.907 on the test dataset, with an Accuracy of approximately 0.80. The analysis of model stability using repeated stratified cross-validation produced average ROC-AUC values of around 0.81, indicating a moderate sensitivity of the results to the choice of data partitions.

Despite the limited size of the dataset and the wide confidence intervals for several metrics, the model is capable of adequately ranking areas by their prospectivity and reproducing the main patterns identified by expert analysis. Thus, machine-learning methods may serve as an auxiliary tool for the quantitative verification and refinement of expert judgments, enhancing their reproducibility while not replacing expert interpretation.

4 Conclusion

The application of machine-learning methods, in combination with expert interpretation of geophysical and geochemical data, made it possible to quantitatively assess the identified patterns and evaluate the consistency of the indicators used for ranking prospective areas. The obtained results demonstrated that the models are capable of reproducing the key elements of the expert-based classification and can also highlight several additional zones of interest, thereby supporting the validity of the expert approach applied in this study.

It is important to emphasize that, due to the limited size of the dataset, machine-learning algorithms can serve only as an auxiliary tool for verifying expert conclusions and should not be used as an independent basis for decision-making. Their primary role lies in formalizing the relationships between heterogeneous parameters, assessing the stability of the identified patterns, and improving the reproducibility of the ranking procedure.

The RandomForestClassifier demonstrated the best performance (ROC-AUC = 0.907 on the test set), indicating the model's ability to correctly rank areas in terms of their prospectivity. However, the analysis of confidence intervals and stability assessments using RepeatedStratifiedKFold suggests that the statistical reliability remains limited under the current data volume, and the results should therefore be interpreted jointly with expert evaluation.

Thus, the integration of expert analysis and machine-learning algorithms enhances the objectivity and reproducibility of prospectivity assessment by providing an additional means of validating interpretational decisions. The developed approach makes it possible to identify promising areas of regional significance - zones exhibiting elevated hydrocarbon gas emissions and favorable geophysical indicators - which ensures a scientifically grounded ranking of territories and helps optimize the costs of exploration and appraisal work while maintaining high reliability of the results.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Nurgaliev, D. K., et al. (2009). Scientific foundations of modern technology for forecasting the petroleum potential of territories. Proceedings of Kazan University, 151(4), 192–201.
2. Vasiliev, S. A., et al. (2018). Oil and gas potential of the Vilyuy Syneclise and prospects for hydrocarbon deposit exploration. GGDOGF, (12), 14–26.
3. Bernard, B. B., Brooks, J. M., & Sackett, W. M. (1977). A geochemical model for characterization of hydrocarbon gas sources in marine sediments. In Offshore Technology Conference, Houston, Texas.
4. Pirson, S. J. (1978). Physical and chemical signals around oil and gas traps. The Log Analyst, 19(2).

5. Seifullin, R. S., et al. (2003). Petrophysical features of epigenetically altered rocks overlying hydrocarbon reservoirs. Geophysical Journal, 25(5), 61–73.
6. Nurgaliev, D. K., & Shvydkin, E. K. (1999). Prospects of non-seismic methods in assessing the petroleum potential of structures. Georesources, 1(1), 35–37.
7. Shvydkin, E. K., Wasserman, V. A., & Zhelenkova, E. A. (2008). A complex of geophysical and geochemical methods (GGHM) for assessing the prospectivity of structures. Drilling and Oil, (2), 10–13.