International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I9P226
Volume 70 Issue 9, 258-268, September 2022

Original Article

Classification of Metageosystems by Ensembles of Machine Learning Models

Stanislav Yamashkin[1], Anatoliy Yamashkin[2], Milan Radovanović[3,4], Marko Petrović[3], Ekaterina Yamashkina[1]

[1] *Institute of Electronics and Lighting Engineering, National Research Mordovia State University, Mordovia, Russia.*
[2] *Geography Faculty, National Research Mordovia State University, Mordovia, Russia.*
[3] *Geographical Institute "Jovan Cvijić", Serbian Academy of Sciences and Arts, Belgrade, Serbia.*
[4] *Institute of Sports, Tourism and Service, South Ural State University, Chelyabinsk Oblast, Russia.*

[1]*Corresponding Author : yamashkinsa@mail.ru*

Received: 26 July 2022 Revised: 21 September 2022 Accepted: 26 September 2022 Published: 30 September 2022

Abstract *- The article describes geoinformation methods and algorithms for interpreting Earth remote sensing data based on forming an ensemble of shallow classifiers based on the Ensemble Learning methodology. The proposed solution can be used to assess the stability of geosystems and predict natural processes. The difference between the created approach is determined by the new organization scheme of the metaclassifier as a decision-making unit and the use of a geosystem approach to preparing data for automated analysis through deep learning models. The article shows that the use of ensembles built according to the proposed method makes it possible to carry out an automated operational analysis of spatial data for solving the problem of the thematic mapping of metageosystems and natural processes to provide conditions for the sustainable development of regions. At the same time, combining models into an ensemble based on the proposed architecture of the metaclassifier makes it possible to increase the stability of the analyzing system: the accuracy of decisions made by the ensemble tends to the accuracy of the most efficient monoclassifier of the system.*

Keywords *- Machine learning, Deep learning, Artificial neural network, Spatial data, Ensembles, Sustainable development.*

1. Introduction

Progress in the field of technologies for digital mapping and analysis of geospatial data and remote sensing materials of the earth, as well as the development of methodological and algorithmic support for the process of analyzing land structure, have led to an increase in demand for geographic information [1]. The relevance of solving the scientific problem of developing new methods and algorithms for the intelligent analysis of spatial data based on machine learning technologies to support the process of making managerial decisions in the field of analyzing the state and structure of land use systems is determined by the need to transition to advanced digital technologies to ensure an effective solution of strategic tasks of sustainable spatial development and territorial planning [2].

The purpose of the study presented in the article is the development and testing of methods and algorithms for constructing ensembles of machine learning models for solving the problem of analyzing the structure and state of metageosystems. The analysis of scientific publications shows that the methods and algorithms of machine learning can be effectively used to interpret geospatial data, which are characterized by the properties of spatial dependence, spatial heterogeneity and scalability [3, 4]. At the same time, applying methods and algorithms of deep machine learning to geospatial data analysis faces many open problems that require scientifically based solutions. Among the most relevant are the following [5]:

- development of a system of methods and algorithms for integration and preliminary processing of spatial data based on new methods of machine learning and digital processing of data signals;
- formation of methodological, algorithmic and software for building deep learning models that allow interpreting multidimensional arrays of spatial data;
- development of a methodological approach to solving the problem of designing, iterative development and implementation of geoportal systems as access points to distributed arrays of spatial information and optimization, optimized for solving practical problems in land use systems analysis.

As an object of systemic spatial analysis in modern science in the field of spatial data analysis, geosystems are defined as "... the earthly space of all dimensions, where the individual components of nature are in a system connected and, as a certain integrity, interact with the cosmic sphere and human society" [6]. The doctrine of geosystems has been developed not only in studying natural objects and processes

but also in analyzing their interaction with social and economic systems [28]. In such an extended interpretation, geosystems are "metageosystems", the digital models that should be used as the main tool for spatial analysis.

2. Related works, Materials and Methods

Machine learning models used to solve the problem of classification of metageosystems can have different architectures (artificial neural networks, decision trees, support vector machines) and hyperparameters [8, 9]. Moreover, they can successfully train on different data sets about the interpreted territory, which can be multidimensional and multimodel, including information about the dynamic spectral properties of the analyzed spatial area, invariant characteristics of geosystems, features and signs of its spatial organization, other attributive, spatial and temporary information [10-12].

In the past few years, the concept of deep machine learning has taken a significant place in the field of spatial data analysis. Multilayer neural network models are based, among other things, on convolutional neural network components. Achieving high interpretation accuracy, in this case, is possible by extracting complex hierarchical features and non-linear dependencies from spatially distributed information. However, using capacious deep neural networks encounters obstacles that make their implementation much more difficult [13-15].

Firstly, such models can be effectively trained only on large sets of labeled spatial data, the formation of which requires serious time and economic costs, including significant amounts of field research and many hours of post-processing [16, 17].

Secondly, deep neural network models are not a panacea in choosing a tool for interpreting spatial data: representing a black box; they can be subject to the problem of overfitting, poor generalization of information and poor interpretability [18-20].

Finally, training deep convolutional neural network models places high demands on hardware: experimental research on fine-tuning the model can be significantly delayed without using expensive GPUs [21-23].

The solution of the indicated problem points is possible with the simultaneous development of two directions: the design of methods and algorithms for integrating and extracting informative territorial features of reduced dimension and the introduction of lightweight models for their interpretation [24, 25]. Shallow neural networks are not only less demanding on computational resources but also more resistant to generalization and overfitting problems. The solution to the problem of interpreting spatial data on metageosystems should be based on understanding objective territorial characteristics. The invariant properties of geosystems are revealed when studying the morphometric parameters of the area, which change irreversibly over a long period, while the dynamic properties are revealed based on periodic remote monitoring data.

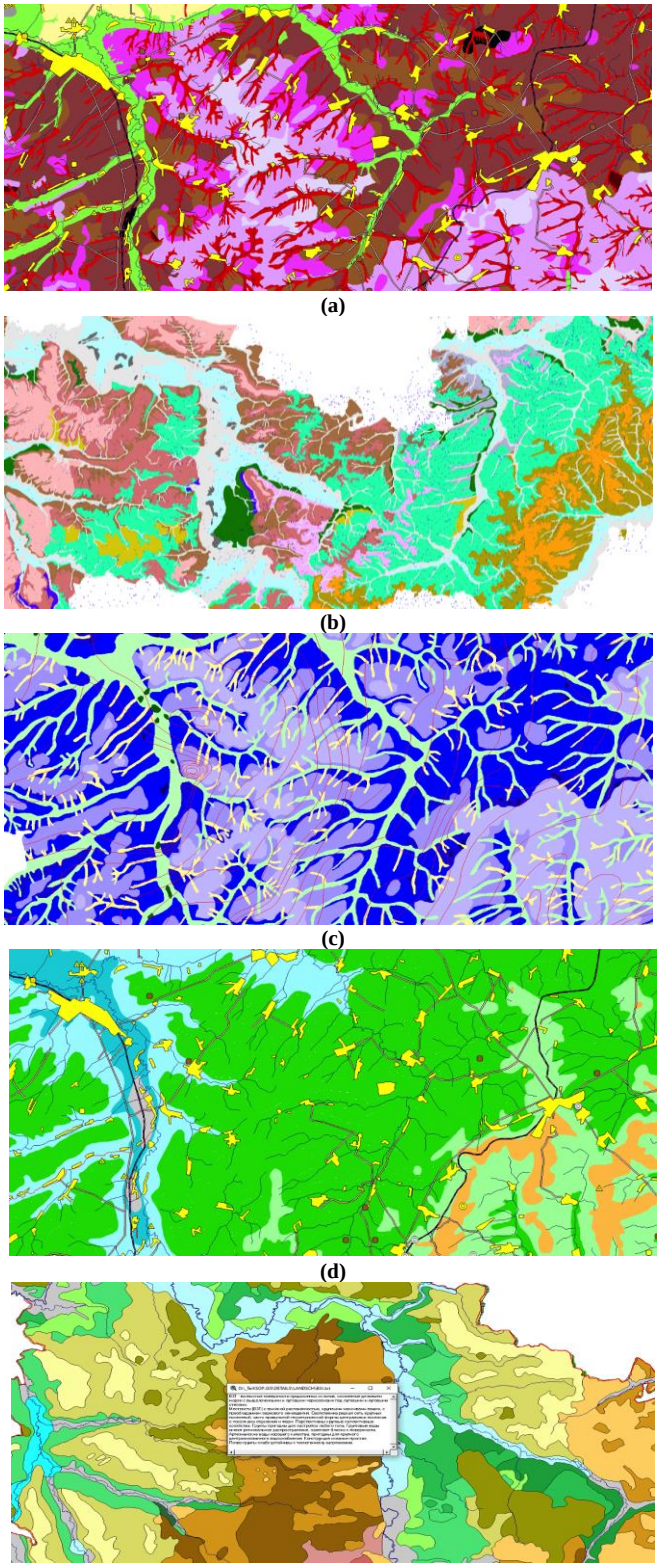
The study presented in this article aims to solve the scientific problem of increasing the efficiency of using machine learning models (primarily artificial neural networks) in solving the problem of classifying metageosystems based on earth remote sensing (ERS) data. In particular, the focus of the study is on solving problems that are open and relevant from the point of view of the current state of research:

- Development of a methodology for data preparation and construction of machine learning models that are protected from the problem of overfitting and weak generalization, showing high accuracy when working with limited sets of labeled data;
- Formation of a system of recommendations for the effective expansion of the labeled data set on metageosystems through the use of automated, consolidated auxiliary data (including satellite imagery, digital elevation models, and digital landscape maps), along with the traditional use of a series of affine transformations;
- Reducing the dimension of the analyzed spatial data and the capacity of machine learning models to solve the problem of increasing the stability of the classifier and weakening the requirements for the hardware platform of its operation;
- Generalization of the transfer learning algorithm in the analysis of geosystems, in which the developed model, being trained on one labelled data, can adapt for reuse on new data sets.

Modeling of the hierarchical system of taxa of geosystems is oriented towards the allocation of categories (distinguished by the features of the macro- and mesoclimate), classes (mapped by orographic features), groups (diagnosed by types of water and geochemical regime), types (determined by soil-biotic features), genera (reflects morphosculptural landforms and their constituent deposits on a regional scale of research) and species.

To solve the problem of analyzing the structure of metageosystems, theEarth remote sensing data are of current importance. At the same time, digital maps systematized in regional and federal geoinformation systems are of great importance. As part of the implementation of the project "Digital spatial data infrastructures and models of territorial metageosystems for sustainable development", thematic maps (Fig. 1) of soil cover (a), aquifer (b), groundwater depth (c), bedrock and sediments (d), as well as an integrated digital landscape map (e).

To solve the problem of automated classification of metageosystems of a territory, it is important to use methods and algorithms of machine learning.



(a) **(b)** **(c)** **(d)** **(e)**

Fig. 1 Thematic maps of regional GIS

An important direction in this area is using ensembles of classifiers built on the Ensemble Learning methodology, combining various models into a system, and increasing the accuracy and stability of machine learning models.

Algorithm for designing ensemble monoclassifiers. Machine learning models used to solve the problem of classifying metageosystems can have different architectures and hyperparameters and be trained on different data sets. The ensemble of machine learning models E is formed based on monoclassifiers M_i (separately trained models) and the metaclassifier MC, which makes the resulting decision when solving the problem of classifying a territorial object X_j to determine whether it belongs to the class of metageosystems. The set of monoclassifiers of the ensemble BC consists of trained models M_i, which perform the function of determining whether a territorial object X_j belongs to the class of metageosystems c_k.

The difference between ensemble monoclassifiers can be their architectural and topological organization and the data used to train the model. When classifying a territorial object X_jThe monoclassifiers of the ensemble form a vector of hypotheses P_j regarding the belonging of this object to a certain class of metageosystems c_k from the nomenclature of classes $\mathbb{C}$ with power K.

At the same time, the trained monoclassifier M_i, when gaining experience in solving the problem of classifying metageosystems based on the quality measure P, returns a set of Bayesian probabilities that determine the degree of confidence of the monoclassifier M_i in the truth of the fact that the territory X_j belongs to the class of metageosystems c_k. The decision Y_{ij} of the monoclassifier M_i about the belonging of a certain territorial object X_j to a specific class of the metageosystem can be made by choosing the class c_i for which the calculated Bayesian probability is maximum.

The resulting hypothesis Y_E about assigning the territory X_j to a specific class of metageosystems c_k is made by the MC metaclassifier of ensemble E. In this case, it is advisable to make the resulting decision based on the output data of deep learning models based on weighted voting, the generalized representation of which has the following form:

$$Y_{Ej} = \underset{k}{\operatorname{argmax}}_{k\in[1,K]} \left(\sum_{i=1}^{N} \varphi(\mathrm{M}_i, k) \cdot \psi(p_{jik}) \right)$$

In this formula, the parameter $\varphi(\mathrm{M}_i, k)$ is a weight coefficient, which is a measure of the efficiency of the monoclassifier M_i in the detection of class k metageosystems. The function φ determines the transformation of the form $\varphi: \mathrm{M}_i \to \Lambda_{ik}$, in which the weight coefficient and measure of efficiency Λ_{ik} are determined by

mathematical transformations of the experimentally obtained data of the error matrix M_i of the monoclassifier M_i.

The following algorithm for calculating the efficiency measure Λ_{ik} is proposed:

1) Construction of the error matrix M_i for each monoclassifier M_i of the system.
2) Calculation of absolute accuracy metrics for the classifier M_i when determining metageosystems of class k: hits (TP_{ik}), true deviations (TN_{ik}), errors I (FP_{ik}) and II (FN_{ik}) types.
3) Calculation of the relative metric $\mathcal{R}_{ik}$, which determines the classification accuracy of class k metageosystems, which makes it possible to carry out an integral assessment of the obtained error matrix M_i with a number in the interval [0;1]. Thus, the estimate F_β, which comprehensively considers the indicators of precision and recall and, consequently, errors of types I and II, as well as the number of correct hits of the monoclassifier. In addition, the metric is tuned by configuring the parameter β, which allows you to emphasize the influence of accuracy and recall on the result.

$$\mathcal{R}_{ik} = F_{i\beta k} = (1+\beta^2)\cdot\frac{precision_{ik}\cdot recall_{ik}}{(\beta^2\cdot precision_{ik})+recall_{ik}} = \frac{(1+\beta^2)\cdot TP_{ik}}{(1+\beta^2)\cdot TP_{ik}+\beta^2\cdot FN_{ik}+FP_{ik}}$$

If necessary, another metric can be designed that satisfies the requirements for building an ensemble.

4) Deactivation of inefficient classifiers to the threshold value ϵ can be carried out according to the following principle:

$$\tilde{\mathcal{R}}_{ik} = (\mathcal{R}_{ik} > \epsilon)?\frac{\mathcal{R}_{ik}-\epsilon}{1-\epsilon}:0$$

As a result of the given conditional ternary operation, two problems are solved: first, the values of efficiency measures that are less than the threshold ϵ are reset to zero, removing inefficient monoclassifiers from the decision-making system; secondly, the resulting value is again normalized in the interval [0;1]. When ϵ=0, the possibility of refusing to use deactivation is implemented.

5) Activation of the metric by implementing an additional non-linear normalized monotonic transformation $\Lambda_{ik} = \theta(\tilde{\mathcal{R}}_{ik})$.

The metric activation operation minimizes or accelerates the metric's growth at its boundary values. The logistic curve can be taken as an activation function θ. Under the identical mapping $\mathrm{id}_{\tilde{\mathcal{R}}_{ik}}$, the possibility of not using activation is realized.

The resulting value of the metric Λ_{ik} can be used to determine the measure of the efficiency of the monoclassifier M_i for the detection of metageosystems of class k.

The function ψ for calculating the measure of the vote of a monoclassifier is a transformation of the form $\psi: p_{jik} \rightarrow \mathbb{Q}_{jik}$In which the Bayesian probability p_{jik}, which determines the degree of confidence of the monoclassifier M_i in the truth of the fact that territory X_j belongs to the class of metageosystems c_k, is transformed into a measure of vote $\mathbb{Q}_{jik}$.

The measure of the vote can be determined by the "winner takes all" principle, in which the monoclassifier M_i puts 1 for the most likely solution and 0 for all the others. With the identical mapping id_ψThe voting will take into account the Bayesian probabilities that the territory X_j belongs to the class of metageosystems k. Finally, the $\mathbb{Q}_{jik}$ the metric can be activated by implementing an additional non-linear normalized monotonic transformation $\theta(\mathbb{Q}_{jik})$, which changes the voting measure at the boundary values.

The key to calculating the informative characteristics of territorial metageosystems analyzed using machine learning models gives an idea of geodiversity, defined as the diversity of the lithogenic basis of landscapes, soil and vegetation features, and the processes occurring in them and implicitly characterizes hydrological and climatic processes. At the same time, landscape diversity is a more complex concept. It determines the systemic organization of Spatio-temporal elements of different levels: classes, groups, types, genera and types of geosystems. The concept of landscape diversity makes it possible to consider a territory as a well-structured system with an organized subordination of natural-territorial complexes.

Landscape diversity can be defined as a form of abstraction of the real world. The properties of territorial objects and processes are defined by numerical variables or qualitative concepts and can be systematized by classes. The calculation of landscape diversity metrics can be based on the identification of simple numerical indicators: color moments and histograms, heterogeneity parameters, and the number of contours or sections within a particular area analyzed based on remote monitoring. Finally, a significant amount of information about the study area is contained in synthetic digital landscape maps, which traditionally represent the final artifact of research activities and can potentially be used as input data for automated analysis.

Thus, the characteristics of the territory's landscape diversity should be considered a complex integral indicator containing information about the hierarchical organization of geosystems and their natural, social, and economic features. The extracted numerical parameters of landscape diversity have a much lower dimension than multidimensional remote monitoring materials and digital maps while maintaining high information content. Consequently, they can be successfully analyzed by robust, lower-capacity machine learning models. With all the advantages, using deep convolutional neural network models leads to contradictions that need to be resolved. First, their sustainable training requires expert labeling of significant training data, which is very time-consuming and resource-consuming. Secondly, deep convolutional models are very demanding on computational resources, which are not always available. The solution to the indicated problems is possible due to the introduction of lighter, for example, wide, fully connected (FC) models of small depth. The power of densely connected layers in the framework of the experiment was 20 and 10 neurons.

These systems are more resilient to overfitting, can be trained efficiently on significantly smaller datasets, and are potentially suitable for reuse based on transfer learning. However, such neural networks can efficiently analyze one-dimensional data vectors rather than multi-channel images of the territory. For this reason, it is necessary to propose an algorithm for extracting information-intensive features of a territory of reduced dimension.

The first group of features includes descriptors of the territory D_{ERS}, which can be calculated based on the territory's image in certain spectral ranges (for example, in the visible spectrum). Let us present an important limitation to the calculated parameters: they must have the property of visibility - cartograms built on their basis must be informative for specialists in data analysis and geosciences.

Based on satellite imagery of the territory, the following descriptors of the DERS group were calculated in the experiment.

1. Landscape metrics of heterogeneity based on the calculation of the informational entropy of the territory E and the spread Δ (characterizing the change in spectral brightness relative to the average value):

$$H = \langle E, \Delta \rangle = \langle \sum_{i=1}^{R} \frac{n_i}{S} \log\left(\frac{n_i}{S}\right), \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})} \rangle,$$

where R – radiometric image resolution; n_i – number of atomic territorial units of spectral brightness i in this neighborhood; S – an area of the analyzed territory; N – the number of pixels in the analyzed territorial system; x_i – average value of the spectral brightness of an atomic region; $\bar{x}$ – the average value of the spectral brightness of the analyzed territory.

2. Intensity metrics based on calculating reliable and stable image parameters invariant to noise and unwanted distortions. Thus, the color moment characterizes the distribution of the spectral brightness of a territorial area and is defined as a set of mathematical expectations ($\bar{I}$), dispersion (D) and asymmetry (A) of the brightness of an atomic area of a territorial system in a certain spectral range:

$$M = \langle \bar{I}, D, A \rangle = \langle \frac{1}{N}\sum_{j=1}^{N} c_j, \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(c_j - \bar{I}\right)^2}, \sqrt[3]{\frac{1}{N}\sum_{j=1}^{N}\left(c_j - \bar{I}\right)^3} \rangle,$$

where c_j is the brightness of the j-th pixel in a certain spectrum.

3. Hue histogram, which is an indicator built on the principle of histograms and characterizing the distribution of the number of image pixels of certain shades in A groups. To calculate, it is necessary to define the vector of possible shades of pixels through τ. In this case, the number of groups A can be selected manually to increase the information content of the metric or determined based on a rule.

These metrics are different ways of numerically assessing the landscape diversity of territory and can be calculated for areas of different scales (geosystems of different hierarchical levels) based on different data (for example, the image of territory in different spectral bands).

The most important source of information about the invariant properties of the territory is digital elevation models (DEMs), which can be consolidated from various sources, such as SRTM (Shuttle Radar Topography Mission), 3DEP (3D Elevation Program), GMTED (Global Multi-resolution Terrain Elevation Data). Regional geoinformation systems often become an important source of information about the relief of a territory. For the experiment, an algorithm was developed that aggregates data on the DEM of a classified area based on the use of third-party software interfaces (APIs). The original data set was expanded using the attribute parameters of the territory images containing information about the latitude and longitude of the analyzed area.

Based on the digital terrain model analysis, the following descriptors of the D_{DEM} group were calculated.

1. Metrics of the steepness of the territory, which objectively characterize such properties of the territorial system as surface runoff, erosion, and the amount of solar energy received, were calculated based on determining the parameters of the maximum, minimum, mathematical expectation, and deviation of the height difference between neighboring DEM points. The slope value A can then be calculated based on the values of the height map.

$$\begin{cases} A = \tan^{-1}\left(\sqrt{\alpha_e^2 + \alpha_n^2}\right) \\ \alpha_e = \sigma(h_E - h_W) \\ \alpha_n = \sigma(h_N - h_S) \end{cases}$$

where α_e – the steepness of the slope in the direction from east to west; α_n – the steepness of the slope in the direction from north to south; $h_{E,W,N,S}$ – east, west, north, south elevation; σ – DEM scale characteristic (distance between height map points).

2. Slope exposure metrics is a morphometric parameter that objectively reflects the orientation of the study area to the flow of sunlight, as well as the probable direction of water runoff, calculated based on the parameters of the maximum, minimum, mathematical expectation and deviation of the DEM gradient map and characterizing the azimuth of the slope of the earth's surface:

$$F = -\tan^{-1}\left(\frac{\alpha_e}{\alpha_n}\right)$$

3. Statistical characteristics (minimum, maximum, mathematical expectation, standard deviation) of the absolute and relative values of the heights of the territory's DEM also carry information that ultimately reduces the dimensionality of the analyzed data.

Finally, digital landscape maps (D_{LM}) generated in geographic information systems are also a significant source of information that makes it possible to improve the accuracy of metageosystem classification. Despite the fact that the scale of such maps is often smaller than the scale of the classified areas, they integrate a significant amount of information about the enclosing geosystems. Obtaining data on a landscape map area can be carried out by integrating the data preparation module with the GIS system APIs by using attribute information about the coordinates of the analyzed area. During the experiment, the descriptors of the D_{LM} group were calculated based on the regional GIS "Mordovia" data by calculating metrics similar to those used to determine the descriptors of the D_{ERS} group.

Additional sifting of non-informative features to reduce the dimensionality of the analyzed data can be performed based on an algorithm that assumes a random iterative selection of samples from a labeled training data set, followed by updating the significance parameter of each feature based on the difference between the selected sample and the two objects closest to it of the same or alternative class. If there is a sufficient difference in the values of a feature for a certain number of nearest neighbors of the same class, its importance decreases, and vice versa; if there is a difference between the values of a feature for objects of different classes, its importance increases. The feature weight decreases if its value differs more for the nearest objects of the same class than for the nearest objects from different classes; otherwise, the weight increases.

Any reduction in the dimensionality of the analyzed data leads to the loss of a certain amount of information. However, suppose the resulting vector of parameters allows the identification of the territory with acceptable accuracy. In that case, the reduction in dimensionality makes it possible to approach the use of less deep and more resistance to the problem of overfitting machine learning models.

3. Research Results

The study was carried out on a system of test sites deployed in the Republic of Mordovia in the zone of interaction between forest types of geosystems of the Oka-Don lowland and the forest-steppe of the layer-tier Volga Upland. Remote sensing data from the Sentinel-2 satellite were chosen as the initial data for systematization, allowing for the exploration of the territories of individual classes and categories of land. To test the methodology for increasing the efficiency of metageosystem classification, a set of labeled data was formed based on a system of test polygons. To expand the training data set, affine transformations were used, which, however, were applied, taking into account the preservation of key properties of spatial objects, such as slope exposure.

At the first stage of the experiment, a neural network model was designed with two convolutional blocks (the number of filters is 32 and 16, the filter size is 3 by 3 pixels) and a decision-making module based on two densely connected layers, with a capacity of 30 and 10 neurons. The output data of convolutional and densely connected layers is fed to the input of the batch normalization block. The linear rectification operation (ReLU) was chosen as the activation function. To increase the stability of the model to overfitting, a subsampling block was introduced based on taking the maximum value with the dimension of the thinning areas 4 by 4 pixels. Before the output, the Bayesian probability of belonging of the territorial system to a certain class is defined as the output of the layer of the generalized logistic function for the multivariate Softmax case. The decision on whether a territory belongs to a certain class is based on the “winner takes all” principle by selecting a hypothesis for which the estimated probability is maximum. Convolutional neural network models have widely established themselves as a tool

for classifying multidimensional spatially distributed data, the information in the cells distributed based on non-linear patterns.

Training the model on 80% of the labeled data made it possible to obtain a high classification accuracy of spatial objects of heterogeneous classes, equal to 88%. When modeling a severe deficit of labeled data, the relative power of the training sample was reduced to 20%. In comparison, the classification accuracy on the validation data fell to 77%, while the accuracy of identifying some classes of metageosystems decreased by more than 20%. It is a natural result when using convolutional models - for sustainable learning, they require a large power of the training set. It should also be noted that deeper neural network architectures will improve classification accuracy due to the possibility of extracting complex hierarchical features. However, this will increase the requirements for the hardware on which the calculations are performed, or it will take a lot of time to conduct experimental studies.

Fig. 2 shows the curves that characterize the dependence of the mathematical expectation of the classification accuracy on the validation data depending on the training epoch of the fully connected (FC) model. It can be seen that the combined analysis of features gives a significant increase in the accuracy of the classification of metageosystems. Accounting for descriptors calculated based on satellite imagery data of the territory (D_{ERS} group) made it possible to achieve an accuracy of 76%. Involvement of relief descriptors (D_{DEM} group) increases the accuracy by 3%, and metrics calculated based on landscape maps (D_{LM} group) by 11%. Finally, the simultaneous analysis of descriptors of all categories leads to an increase in classification accuracy by almost 12%.

As a result, the low-capacity FC model showed classification accuracy characteristics higher than that of a more cumbersome convolutional model trained on multidimensional spatial data. In particular, higher accuracy was achieved when detecting territories of the following classes: annual and perennial crops, herbaceous vegetation, highways and roads, industrial buildings and rivers.

At the same time, a low-capacity model can be trained without involving a powerful GPU, making it convenient to fine-tune and optimise hyperparameters without access to heavy and expensive hardware. Thus, wide shallow machine learning models trained on the basis of a set of informative territorial descriptors can function and be further trained on relatively thin devices, such as unmanned aerial vehicles.

The next important advantage of the FC model, trained on territorial descriptors, is a stable operation in the face of a shortage of labeled data. Reducing the proportion of the training sample to 20% did not lead to a significant decrease in the accuracy of the classification of metageosystems (while the accuracy of the convolutional neural network model fell by more than 10%).

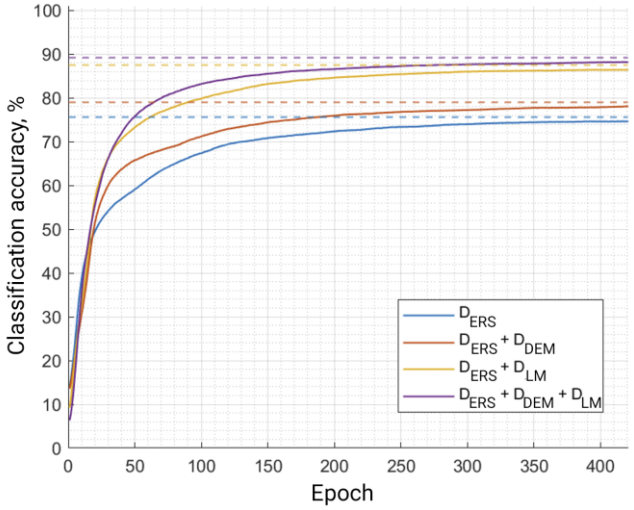


Fig. 2 Change in expectation of classification accuracy on validation dataset depending on the training epoch of the FC-model.

This advantage is due to the sufficient informative load of the allocated territorial descriptors, their invariance to changes in the studied polygon, as well as the good ability of the model to generalize the analyzed data. Thus, the FC model can be reused in the framework of studying new territorial systems and retrained and fine-tuned based on a new labeled data set through transfer learning.

During the experiment, three artificial neural networks (ANN) were designed and combined into an ensemble according to the described method. The first model is based on one densely connected layer of 10 neurons, and the second and third are based on two, with a capacity of 10-10 and 10-20 neurons, respectively. The results of calculating the F1 metric based on the error matrix are shown in Fig. 3.

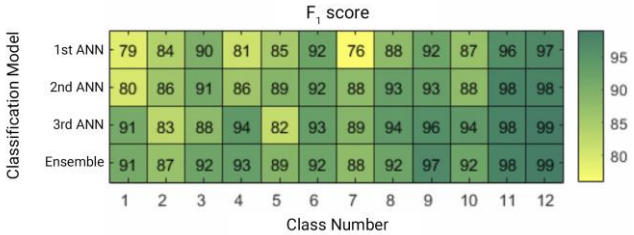


Fig. 3. The F_1 score for classifiers in the classification of geosystems

It can be seen that an increase in the capacity of models does not lead to a clear improvement in the result since more powerful models can be more unstable to overfitting and also require more labeled data for training. When combined into an ensemble, the resulting hypothesis began to be applied on the basis of weighted voting based on the measure of efficiency, which made it possible to avoid gross errors in the classification inherent in each classifier separately. At the

same time, the ensemble only slightly loses accuracy to individual classifiers of the system while maintaining overall resistance to errors of the first and second kind, which are characteristic of an individual classifier when determining objects of a particular class of territory.

To solve the problem of classifying the geosystems of the Inerka test site, spatial data from three different sources were used, which provided information on the dynamic and invariant properties of the study area. The data of remote space monitoring obtained from the Sentinel-2 satellite were chosen as the source of time-varying spectral characteristics of the territory. The analyzed image was obtained on August 30, 2021 (according to the MGRS system, calculated based on the universal transverse Mercator projection, the shooting square is determined by the position of 38UNF), and the spatial resolution of shooting in the visible zones of the spectrum and the near-infrared range is 10 meters per 1 atomic area. Data on the morphometric properties of the territory, characterizing its invariant (slow but irrevocably changing) state, was obtained on the basis of SRTM materials, corrected and distributed by Mapzen based on Amazon cloud application programming interfaces. The mean square error relative to the fixed height for the studied test area is measured within 1.4 meters, and the spatial resolution is 30 meters. The third source of information about the analyzed territory was the regional GIS "Mordovia" data, used to build a digital landscape map containing information about regional land use systems. Not characterized by high spatial resolution, they have a significant information capacity, storing synthetic indicators of the analyzed territory: classes of geosystems and land-use systems.

The level 2A Sentinel-2 image preconditioning was based on the Sen2Cor processor. It included normalization of spectral brightness values, data correction based on atmospheric parameters, corrected for the reflectivity of the terrain and cirrus clouds. To improve the accuracy of the analysis, a set of territorial descriptors was calculated that characterize the properties of the surrounding neighborhood for an atomic site. The synthesized features included the local entropy of the enclosing geosystem (neighbourhood), the spread of spectral brightness relative to the average value, hue histograms, statistical data on elevation maps, exposure, and slope steepness. The calculation of territorial descriptors makes it possible to reduce the dimensionality of the analyzed data (in comparison with the analysis of a satellite imagery fragment) with the inevitable loss of a certain amount of information about the analyzed territory. At the same time, a balance was achieved between the maximum possible facilitation of the allowable capacity of the machine learning model, increasing its resilience to overfitting, and preventing a significant decrease in classification accuracy in the framework of the problem of classifying the geosystems of the test site. The calculated descriptors have the property of visibility: cartograms built on their basis are informative for specialists in the field of data analysis and geosciences. Before machine analysis, territory metrics were normalized by scaling the data vector by its standard deviation value. About a hundred labeled samples were prepared for each territorial class.

Based on the spectral and morphometric properties of the territory, as well as synthetic descriptors of the enclosing geosystems, a parameter vector is formed that is suitable for training by artificial neural networks of low depth based on the use of densely connected layers. Unlike widely used convolutional networks, such models can be trained on a lower amount of labeled data, are characterized by resistance to the overfitting problem, and are less demanding on hardware. The operation of linear rectification (ReLU) was used as the activation function of the layers of the neural network. To increase the stability of the model to overfitting, a subsampling block based on taking the maximum value is introduced. To solve the problem of reducing the accuracy of classification and retraining, normalization layers are introduced into the structure of the neural network. The decision on belonging the territory to a certain class is based on the “winner takes all” principle, by selecting a hypothesis with which the estimated probability is maximum. The total number of fully connected layers is limited to two.

The proposed methodology was tested in the course of design work on the analysis of the geosystems of the polygon «Inerka» (center coordinates: 54°03′ N, 45°53′ E), conducted to analyze the interaction of paragenetic systems of forest-steppe geosystems of the erosion-denudation plain and intrazonal forest landscapes of the valley of the Sura River. The priority geoecological problem is the optimization of tourist and recreational development of the natural monument of republican significance, "Lake Inerka". The results of the classification of spatial data made it possible to identify the following types of geosystems on the territory of the Inerka polygon (Fig. 4).

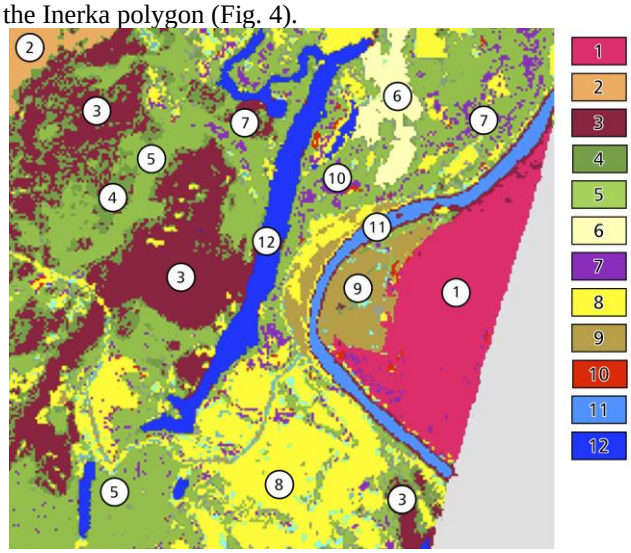


Fig. 4 Map of geosystems of the test polygon “Inerka”

Geosystems class 1 is the coast of the Sura valley with absolute elevations up to 265 m, composed of marls, flasks, and writing chalk with outcrops of bedrock on steep slopes to the day's surface.

The following territorial classes represent the floodplain terraces: 2 – leveled surfaces composed of sands and sandy loams, with soddy-weakly podzolic soils under pine forests; 3 – dune-like remnants of intra-floodplain terrace complexes with hollows with underdeveloped podzolic and soddy-weakly podzolic soils under lichen, green moss, pine forests; 4 – the lower parts of the floodplain terraces with low sandy ridges, the structure of the soil cover is characterized by a mosaic of sod-podzolic and sod-podzolic gley soils under pine, damp birch and aspen forests.

The selected floodplains are decomposed into the following classes of territories: 5 – slightly undulating surfaces composed of loams with interlayers of sands, natural vegetation dominated by floodplain oak forests, and alder forests; 6 – complexes of large and small ridges with underdeveloped soddy sandy and soddy-meadow thin, light loamy soils, with bluegrass-meadow-fescue meadows; 7 – depressions between crests and old rivers with meadow-marsh heavy loamy soils under sedge-mannic-canary meadows; 8 – riverine sandy, wet floodplains with meadows and willows; 9 – riverbed sandbanks (beaches); 10 – alder swamps with peaty-gley and peaty silt-loamy soils.

Natural aquatic complexes are represented by the following classes of geosystems: 11 – natural aquatic complexes of Sura River; 12 – natural aquatic complexes of oxbow lakes.

The process of transformation of the Sura channel is in the development stage. In this regard, the organization of monitoring of ecosystems in the Inerka region is relevant. Studies show that the lake is fed by melt, rain and groundwater. High floods of the Sura that could replenish and clean the lake, due to the interception of meltwater by numerous reservoirs, become very rare. Mapping of the geosystems of the Inerka test site shows their weak resistance to recreational development. The main limiting factors are the composition of Quaternary deposits, the Nature of the relief, the mechanical composition and soil moisture, the thickness of the humus horizon, and the genesis and vegetation composition.

4. Conclusion

Analysis of the effectiveness of the methodology for constructing classifiers ensembles to solve the problem of studying the structure of metageosystems of test sites allows us to draw the following conclusions.

Using ensembles built according to the proposed method allows for rapid automated spatial data analysis to solve the problem of the thematic mapping of metageosystems and natural processes. Ensembles make it possible to approach the problem of preparing data for training models by integrating into a single system of models trained on various combinations of training and validation samples to reduce the impact of errors that occur during the formation of data sets.

Combining models into an ensemble based on the proposed architecture of the metaclassifier makes it possible to increase the stability of the analyzing system: the accuracy of decisions made by the ensemble tends to tend to the accuracy of the most efficient monoclassifier of the system. The system's error in most cases does not exceed the error of the most efficient classifier while avoiding gross systematic errors made by individual monoclassifiers.

The formation of a metaclassifier according to the proposed algorithm is an opportunity to add an element of predictability and control to using neural networks, which are traditionally a "black box". The integration of individual classifiers into ensembles makes it possible to approach the solution to the scientific problem of finding classifier hyperparameters through the combined use of models of the same type with different configurations. The construction of efficient ensembles can be based on models of relatively small width and depth, which makes it possible to design high-precision classifiers, the training of which is less demanding on computing power compared to classical deep models.

The use of classical convolutional neural network models, which have proven themselves well in solving the problem of classifying territorial systems, is associated with some serious limitations: such models can be effectively trained on very large sets of labeled spatial data, but they are subject to the problem of overfitting, are characterized by a poor generalization of information and poor interpretability, and also the processes of their use, training and fine-tuning place high demands on the hardware. Suppose it is impossible to overcome the indicated limitations. In that case, it is advisable to switch to shallow wide, tightly coupled models trained on a set of informational territorial descriptors according to the methodology presented in the article.

The calculation and consolidation of the territorial descriptors proposed by the authors simultaneously lead to a decrease in the dimension of the analyzed data (positive effect) and the inevitable loss of some information about the analyzed territory (negative effect). It is important to find a balance between the two indicated positions to maximise the allowable capacity of the machine learning model, increase its resistance to overfitting, and prevent a significant

decrease in classification accuracy within the framework of a specific problem being solved.

Aggregate analysis of territory descriptors, integrated on the basis of data from different sources, significantly increases the accuracy of metageosystems classification. In the framework of the experiment presented in the article, taking into account the proposed system of descriptors calculated on the basis of satellite imagery data, a digital elevation model and an electronic landscape map made it possible to achieve an accuracy of 89%, which is much more than this parameter for a convolutional neural network model. At the same time, the analysis of relief descriptors increases the accuracy by 3%, and the metrics calculated based on landscape maps - by 11%. Specialists must well interpret the cartograms of the presented descriptors in the field of data analysis in geosciences.

The developed technique for spatial data analysis made it possible to identify types of geosystems on the territory of the Inerka polygon. The priority geoecological problem is the optimization of tourist and recreational development of the natural monument of republican significance, "Lake Inerka". The process of transformation of the Sura channel is in the development stage, and in this regard, the organization of monitoring of ecosystems in the Inerka region is relevant. Studies show that the lake is fed by melt, rain and groundwater. High floods of the Sura that could replenish and clean the lake, due to the interception of meltwater by numerous reservoirs, become very rare. Mapping of the geosystems of the Inerka test site shows their weak resistance to recreational development. The main limiting factors are the composition of Quaternary deposits, the Nature of the relief, the mechanical composition and soil moisture, the thickness of the humus horizon, and the genesis and vegetation composition.

The study carried out is a development of the experience gained earlier in the course of work on developing the regional water balance regulation concept based on the geosystem approach [29]. In particular, a methodology for calculating territorial descriptors and their joint analysis based on neural network algorithms is proposed. In addition to the previously developed land classification method based on the geosystem approach [27], various types of data on host geosystems have been proposed and systematized, the analysis of which makes it possible to improve the accuracy of neural network algorithms.

Some advantages of the approach proposed in the article to improve the efficiency of machine learning models in solving the problem of classifying metageosystems include the stability of the developed solution in the face of a shortage of labeled data, as well as the possibility of reuse in the study of new territorial systems, subject to additional training and fine-tuning.

Funding Statement

The study was supported by the Russian Science Foundation (grant № 22-27-00651), https://rscf.ru/en/project/22-27-00651/.

References

[1] M. F. Goodchild, “Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0,” *International Journal of Spatial Data Infrastructures Research* , vol. 2, no. 2, pp. 24-32, 2004.

[2] X. X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8-36, 2017.

[3] L. Zhang, L. Zhang, and B. Du, “Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 22-40, 2016.

[4] C. Tao, H. Pan, Y. Li, and Z. Zou, “Unsupervised Spectral–Spatial Feature Learning With Stacked Sparse Autoencoder for Hyperspectral Imagery Classification,” *IEEE Geoscience and Remote Sensing Let*, vol. 12, no. 12, pp. 2438-2442, 2015.

[5] Y. Lecun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436, 2015.

[6] S. A. Yamashkin, A. A. Yamashkin, V. V. Zanozin, M. M. Radovanovic, and A. N. Barmin, “Improving the Efficiency of Deep Learning Methods in Remote Sensing Data Analysis: Geosystem Approach*,” IEEE Access*, vol. 8 , pp. 179516-179529, 2020.

[7] Shiji S K, Dr. S.H Krishnaveni, "An Analytical Approach for Reconstruction of Cosmetic Surgery Images Using Euclbp and Sift," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 9, no. 8, pp. 60-71, 2022. Crossref, https://doi.org/10.14445/23488379/IJEEE-V9I8P107.

[8] W. Li, H. Liu, Y. Wang, Z. Li, Y. Jia, and G. Gui, “Deep Learning-Based Classification Methods for Remote Sensing Images in Urban Built-Up Areas,” *IEEE Access*, vol. 7, pp. 36274-36284, 2019.

[9] R. A. Schowengerdt, “Remote Sensing: Models and Methods for Image Processing,” *3rd Ed. Orlando, Fl, Usa: Academic Press* , pp. 387-456, 2006.

[10] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, “When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification Via Learning Discriminative Cnns,” *IEEE Transactions on Geoscience and Remote Sensing* , vol. 56, no. 5, pp. 2811-2821, 2018.

[11] J. G. Lee, and M. Kang, “Geospatial Big Data: Challenges and Opportunities,” *Big Data Research* , vol. 2, no. 2, pp. 74-81, 2015.

[12] D. D. Cham, N. T. Son, N. Q. Minh, N. T. Thanh, and T. T. Dung, “An Analysis of Shoreline Changes Using Combined Multitemporal Remote Sensing and Digital Evaluation Model,” *Civil Engineering Journal* , vol. 6, no. 1, pp. 1-10, 2020.

[13] S. Damuluri, K. Islam, P. Ahmadi, and N. Qureshi, "Analyzing Navigational Data and Predicting Student Grades Using Support Vector Machine," *Emerging Science Journal*, vol. 4, no. 4, pp. 243-252, 2020.
[14] S. Hammal, N. Bourahla, and N. Laouami, "Neural-Network Based Prediction of Inelastic Response Spectra," *Civil Engineering Journal*, vol. 6, no. 6, pp. 1124-1135, 2020.
[15] S. A. Yamashkin, A. A. Kamaeva, A. A. Yamashkin, and E. O. Yamashkina, "Matters of Neural Network Repository Designing for Analyzing and Predicting of Spatial Processes," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, pp. 17-22, 2021.
[16] Y. Bengio, and Y. Lecun, "Scaling Learning Algorithms Towards Ai," *Large-Scale Kernel Machines*, vol. 34, no. 5, pp. 1-41, 2007.
[17] W. Zhao, and S. Du, "Spectral–Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach," *IEEE Trans. Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4544-4554, 2016.
[18] X. Hu, and Q. Weng, "Estimating Impervious Surfaces From Medium Spatial Resolution Imagery Using the Self-Organizing Map and Multi-Layer Perceptron Neural Networks," *Remote Sensing of Environment*, vol. 113, no. 10, pp. 2089-2102, 2009.
[19] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data," *IEEE Geoscience and Remote Sensing Let*., vol. 14, no. 5, pp. 778-782, 2017.
[20] C. Zhao, X. Li, and H. Zhu, "Hyperspectral Anomaly Detection Based on Stacked Denoising Autoencoders," *Journal of Applied Remote Sensing*, vol. 11, no. 4, pp. 042605, 2017.
[21] R. Dong, X. Pan, and F. Li, "Dense U-Net-Based Semantic Segmentation of Small Objects in Urban Remote Sensing Images," *IEEE Access*, vol. 7, pp. 65347-65356, 2019.
[22] W. Li, H. Fu, L. Yu, P. Gong, D. Feng, C. Li, and N. Clinton, "Stacked Autoencoder-Based Deep Learning for Remote-Sensing Image Classification: A Case Study of African Land-Cover Mapping," *International Journal of Remote Sensing*, vol. 37, no. 23, pp. 5632-5646, 2016.
[23] H. Wu, and S. Prasad, "Convolutional Recurrent Neural Networks for Hyperspectral Data Classification," *Remote Sensing*, vol. 9, no. 3, pp. 298.
[24] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land Cover Classification Via Multitemporal Spatial Data By Deep Recurrent Neural Networks," *IEEE Geoscience and Remote Sensing Let*., vol. 14, no. 10, pp. 1685-1689, 2017.
[25] D. E. Kim, P. Gourbesville, and S. Y. Liong, "Overcoming Data Scarcity in Flood Hazard Assessment Using Remote Sensing and Artificial Neural Network, Smart Water," vol. 4, no. 1, pp. 2, 2019.
[26] M. C. Shanker, M. Vadivel, "Hybrid Transfer Learning of Mammogram Images for Screening of Micro-Calcifications," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 9, no. 8, pp. 40-47, 2022.
Crossref, https://doi.org/10.14445/23488379/IJEEE-V9I8P105.
[27] S. A. Yamashkin, A. A. Yamashkin, V. V. Zanozin, M. M. Radovanovic, and A. N. Barmin, "Improving the Efficiency of Deep Learning Methods in Remote Sensing Data Analysis: Geosystem Approach," *IEEE Access*, vol. 8, pp. 179516-179529, 2020.
[28] L. Miklós, E. Kočická, Z. Izakovičová, D. Kočický, A. Špinerová, A. Diviaková, and V. Miklósová, "Landscape as a Geosystem. Cham," Switzerland: Springer, pp. 11-42, 2019.
[29] A. A. Yamashkin, S. A. Yamashkin, M. M. Radovanovic, N. S. Muchkaeva, and I. S. Lyamzina, "Development of the Regional Water Balance Regulation Concept Based on the Geosystem Approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 3, pp. 1672-1683, 2022.