DOI: 10.12737/stp-112202503

Received December 11, 2024
Accepted April 28, 2025

# DATA-DRIVEN APPROACH TO MID-LATITUDE COHERENT SCATTER RADAR DATA CLASSIFICATION

**O.I. Berngardt**
*Institute of Solar-Terrestrial Physics SB RAS,*
*Irkutsk, Russia, berng@iszf.irk.ru*

**Abstract.** A self-consistent, data-driven approach to classifying data obtained at the ISTP SB RAS mid-latitude coherent scatter radars has been developed. Based on 2021 data, a solution of the problem of automatic data classification is presented without their labeling by an expert and without postulating the number of classes. The algorithm automatically labels the data, determines the optimal number of signal classes observed by the radars, and trains a two-layer classifying neural network of an extremely simple structure. The trajectory calculations use the wave optics method and international reference models of the ionosphere and the geomagnetic field. The model is trained on signals coming from the main lobe of the antenna pattern. During training, to adapt part of the data obtained with improved spectral resolution, it is artificially coarsened to the standard resolution. Each signal class determined by the neural network is interpreted from a physical point of view, using statistical characteristics of the signals belonging to it. The number of classes in the data is demonstrated to range from 23 to 35. The significance of various parameters of the input data is assessed. It is shown that the most important parameters for the classification are the calculated scattering height and the elevation of the trajectory at the scattering point, and the least important are the spectral width of the received signal and the calculated number of reflections from the underlying surface.

**Keywords:** decameter radar, SECIRA, ionosphere, automatic classification.

## INTRODUCTION

Often, the problem of interpreting data has several possible scenarios for their explanation, the choice of which may be subjective and depend on an interpreter. Thus, it is important for interpretation of the results to be independent of the interpreter. The problem can be formulated as a data-driven approach — building models based on objective information contained in the data. The paper presents a self-consistent data-driven approach to solving the problem of classifying processed data from ISTP SB RAS coherent scatter radars in terms of the radiophysical mechanisms of formation and propagation of these signals.

The Russian coherent radar network SECIRA [Berngardt et al., 2020b] consists of radars similar to those of the international network SuperDARN [Greenwald et al., 1995; Chisham et al., 2007; Nishitani et al., 2019] in software and hardware. SECIRA radars are software-modified CUTLASS stereo radars [Lester et al., 2004]. Interpretation of received signals usually begins with the classification of data into different types (classes). The main classes of scattered signals are ionospheric scatter from magnetically oriented irregularities, scattering from the underlying surface (Earth and sea), scattering from meteor trails, near-range echo in the ionospheric E layer, etc. [Nishitani et al., 2019].

The problem of classifying coherent scatter radar data by machine learning methods into two classes (ionospheric scatter and ground scatter) has been addressed, for example, in [Ponomarenko et al., 2007; Blanchard et al., 2009], where it is shown that a very simple model is sufficient to divide the data into two classes — just several free parameters. Ribeiro et al. [2011] use an intuitive algorithm for clustering into two clusters, which is basically similar to the DBSCAN algorithm [Ester et al., 1996]. Kunduri et al. [2022] employ the DBSCAN analog to divide data into 9 clusters, with the data preconverted into probabilities of various classes by a neural network model trained on a synthetic dataset generating signals of these 9 classes (0.5, 1, 1.5 and 2 hop ionospheric scatter, as well as scattering from the Earth/sea surface). Kong et al. [2024] deal with the problem of clustering latent representations of signals extracted from them by an autoencoder [Rumelhart et al., 1986; Goodfellow et al., 2016]. A solution to the more complex problem of clustering into 20 classes for the self-learning network when clustering trains a classifier has been proposed in [Berngardt et al., 2022; Berngardt, 2022].

From comparison of the algorithms [Ponomarenko et al., 2007; Blanchard et al., 2009] and [Berngardt et al., 2022; Berngardt, 2022], it might have been expected that the problem of dividing into 20 classes can be solved by a model with a relatively small number (several hundred) of unknown parameters. However, the solution given in [Berngardt, 2022] requires ~30000 free parameters and the use of a polynomial embedding space at the input, which makes the model excessively complex. The model is currently employed in ISTP SB RAS radars, but its improvement requires answering the following questions.

1. How to take into account the type of radar and its mode of operation when training and using the model?

2. How to determine the number of classes of scattered signals in data without involving an expert?

3. What characteristics of the received signal have the greatest effect on the quality of its classification?

The proposed work aims at improving this method. It is shown that the application of the data-driven approach makes it possible to answer the above questions.

## SIGNAL PROPAGATION MODEL AND INPUT MODEL PARAMETERS

The approach put forward in [Berngardt et al., 2022; Berngardt, 2022] involves using unlabeled data to create their classifier. The resulting neural network is a scheme similar to an autoencoder [Rumelhart et al., 1986; Goodfellow et al., 2016], but with multiple decoder heads, where each head (decoder) is trained separately with labels created by a clusterer for a separate experiment. The decoder head is a fully connected layer $y_j = SoftMax\left(\sum_i A_{ij} x_i\right)$ with nonnegative coefficients $A_{ij} \geq 0$ — a projection of a set of hidden classes $x_i$ onto a set of clusters $y_j$, which is implicitly close to defining the total probability through conditional ones. The encoder for all autoencoder heads is the same and is the required data classifier for hidden classes $x_i$. The scheme of the neural network and its training is presented in Figure 1, *a*. It is believed that the autoencoder works effectively when the number of hidden classes is not smaller than the actual number of variables needed to accurately solve the problem [Goodfellow et al., 2016].

The approach proposed in [Berngardt et al., 2022] consists of two stages. The first stage (clustering) is the division of data into slightly overlapping classes. At this stage, the Gaussian mixture approach is applied, in which it is assumed that the data in each cluster has multidimensional Gaussian distribution with unknown parameters, their number is 20. Limitations of this approach include the complexity of its justification: there are many different clustering methods, and their choice will lead to different data clusterings. Ribeiro et al. [2011] employs an algorithm similar to the DBSCAN algorithm for clustering into two clusters. Kunduri et al. [ 2022] employ a DBSCAN analog to divide data into 9 clusters (0.5, 1, 1.5 and 2 hop ionospheric scatter, as well as scattering from the Earth/sea surface), with data pre-converted into probabilities of various classes by the neural network model. Kong et al. [2024] utilizes the AE-K-means algorithm for clustering into two clusters (ground scatter and ionospheric scatter), which is a clustering by the K-means method of features extracted from data using an autoencoder neural network. Therefore, the choice of clustering method is subjective and depends on the researcher.

The second stage (classification) of the algorithm [Berngardt et al., 2022] involves training the classifier on the data labeled at the first stage. Limitations of this algorithm include an unreasonably large neural network that is difficult to interpret and an intuitively selected number of

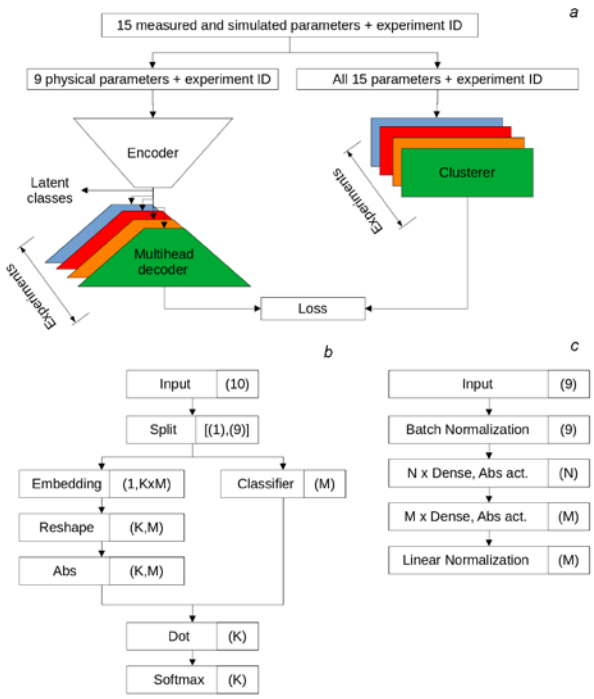


*Figure 1*. Neural networks and their training method: *a* is a training scheme for a wrapped classifier (encoder). Each clusterer and decoder correspond to a single experiment (fixed beam, frequency channel, and day). The number of decoder and clusterer heads is equal to the number of experiments, ~15000. Colors indicate different experiments. Detailed architecture (*b*) of the wrap used to train a classifier. Detailed architecture (*c*) of the neural classifier network (encoder). *K* is the maximum number of clusters after stage 1, *M* is the number of hidden classes in the signal, *N* is the dimension of the hidden layer of the classifier.

classes equal to 20. Note that in other works using neural networks, for example, in [Kunduri et al., 2022; Kong et al., 2024], the choice of neural network architecture is often not justified.

To improve the model within the data-driven approach, the neural network architecture and parameters should be selected optimal from characteristics of the dataset in use.

The need for interpretability of the classes into which we divide signals requires that the number of hidden classes be as small as possible and nevertheless sufficient to confidently describe our data and predict the results. From a mathematical standpoint, the problem reduces to finding a neural network of minimum width that ensures the highest possible quality of the solution.

## STAGE 1. DATA CLUSTERING

### Data used for clustering

When constructing a classifier as in [Berngardt et al., 2022; Berngardt, 2022], the following measured and simulated data was applied. Parameters measured by radar:

1. Time, distance to a scatterer.
2. The Doppler velocity *V* and the measured spectral width *W* are determined from the signal by the FITACF algorithm [Ribeiro et al., 2013]; the work uses the spectral width in the exponential correlation function model (*W*l, velocity units).

3. The elevation angle is found by the algorithm [Berngardt et al., 2021] with meteor radar calibration.

The parameters were obtained by simulating radio wave propagation, using the geometric optics approach in the model ionosphere described by IRI and IGRF.

4. The effective scattering height is calculated as a result of rectilinear (refraction-free) radio wave propagation.

5. The slope angle (sine of the elevation angle) of the trajectory relative to the horizontal at four points in range (1/4, 2/4, 3/4, 4/4 of measured range).

6. Angle between radio wave propagation direction and Earth's magnetic field (angle cosine).

7. Propagation mode is the number of reflections from the underlying (ionospheric) layer or from Earth's surface during propagation to a scatterer.

8. Scattering height.

Only 10 parameters (2, 4–8) were utilized as input data of the classifier in [Berngardt et al., 2022; Berngardt, 2022]. All 15 were used for clustering. In this work, the effective scattering height is excluded from the classifier parameters (item 4).

## Radar type in model training

Only 7- and 8-pulse sounding sequences (most commonly used in SECIRA and SuperDARN radars) have been applied to model training before. Preliminary analysis has shown that when using the full dataset from the ISTP SB RAS radars, the prediction result is sensitive to types of radars and their operating modes. Therefore, in order to build a single model, independent of radar characteristics, when preparing data it is necessary to compensate for differences in their characteristics and modes. Three main differences include the waveform of sounding signals (the type of sounding sequence in use affects the spectral resolution [Berngardt et al., 2020a]), the distance between interferometer and main arrays (affects the uncertainty in calculating the elevation angle of incoming signal [Milan et al., 1997]), and the type of antennas employed (the antenna pattern has an effect on the azimuth and elevation angle dependence of signal power).

The type of antennas in use is the most difficult to account for, so it is ignored in this paper.

Taking into account the sounding signal waveform is compensated by augmentation of data and making all data statistically similar, regardless of the signal type, and considering the distance between the arrays is compensated by eliminating “bad” signals that do not come from the main lobe of the antenna pattern [Milan et al., 1997].

Examine augmentation of radar data. Sounding signals of two main types are most often used in coherent SuperDARN radars: the standard 7-pulse signal [Barthes et al., 1998] and the 8-pulse katscan [Ribeiro et al., 2013], and sometimes the 13-pulse tauscan [Greenwald et al., 2008]. In SECIRA radars, 10-pulse and 16-pulse signals are added to them [Berngardt et al., 2020]. All of them differ in duration and spectral resolution. This especially affects the measurement of the spectral width $W$ of received signal: the shortest 7-pulse signal gives maximum errors; the longest 16-pulse signal, minimum errors.

There are three main approaches to making the data similar regardless of the sounding sequence type: solving the inverse problem, eliminating non-standard data, and intentionally distorting non-standard data.

The first approach, solving the inverse problem, reduces mathematically to the problem of inverting convolution and requires remaking the existing data processing algorithm FITACF; therefore, it is omitted.

The second approach is to exclude the spectral width from consideration, which is obviously ineffective: in all existing SuperDARN/SECIRA signal separation algorithms, spectral width plays an important role.

The approach employed in this work involves deliberate distortion of data obtained with high spectral resolution to a state in which it is difficult to distinguish the data from that acquired with low spectral resolution. In machine learning, such deliberate distortion is called augmentation and is widely applied [Shorten, Khoshgoftaar, 2019].

Within this approach, all data was reduced to the lowest accuracy: the spectral width in the data obtained by longer sequences was increased so that the resulting spectral width distributions were close to the 7-pulse sequence distributions. This approach makes it possible to use data obtained with different spectral resolutions for training.

Since 16-pulse and 7-pulse sequences are most often exploited in SECIRA radars, an assessment was made of necessary additional augmentations of sounding data with 16-pulse signals as compared to 7-pulse signals. The FITACF spectral width estimation algorithm is quite complex, so the necessary spectral width augmentations were determined experimentally according to the formula

$$W_{16,\mathrm{augm}} = W_{16} + \delta W \approx W_7. \quad (1)$$

Here, $W_{16}$ and $W_7$ are the spectral width obtained by the FITACF algorithm from the measurement data for the 16-pulse and 7-pulse sequences respectively; $W_{16,\,\mathrm{augm}}$ is its augmented value; $\delta W$ is the desired augmentation.

Let $\delta W$ be a random variable with unknown probability density $\mathbb{P}_{\delta W}$. Then the probability density of the augmented spectral width $\mathbb{P}_{W_{16,\mathrm{augm}}}$ is a convolution of the probability density $\mathbb{P}_{W_{16}}$ of the spectral width measured by the 16-pulse sequence with the probability density of augmentation $\mathbb{P}_{\delta W}$ and should be approximately equal to the probability density $\mathbb{P}_{W_7}$ of the spectral width measured by the 7-pulse sequence:

$$\mathbb{P}_{W_{16,\mathrm{augm}}}(W) = \mathbb{P}_{W_{16}}(W) * \mathbb{P}_{\delta W}(W) \approx \mathbb{P}_{W_7}(W). \quad (2)$$

Augmentation distribution (convolution kernel $\mathbb{P}_{\delta W}$) was found from spectral width distributions of 16-pulse sequences and 7-pulse sequences by analyzing signals with a low Doppler shift, usually specific for ground scatter signals [Blanchard et al., 2009]. For this purpose, signals with a Doppler shift of no more than 30 m/s were selected from ranges 500–1500 km in experiments where sounding was carried out with alternating types of pulse sequence. This mode was regular

in the EKB radar from April to December 2021. The distributions are exemplified in Figure 2, *a*, *b*.

The problem of finding the distribution of $\mathbb{P}_{\delta W}$ was solved by training a neural network consisting of a single convolution with a width [–100, 100] m/s without activation functions. The coefficients of the found convolution kernel $\mathbb{P}_{\delta W}$ are shown in Figure 2, *d* in orange. Therefore, a random variable was chosen as an augmentation model that approximates this distribution well.

$$\delta W = \tan(\eta)19 - 5\,[\text{м/с}]. \quad (3)$$

Here, η is a random variable having a uniform distribution in the range [0, atan(6)],

$$\eta \sim \mathbb{U}\left(\left[0, \operatorname{atan}(6)\right]\right). \quad (4)$$

Parameters of this model were selected manually to ensure a satisfactory match between the curves in Figure 2, *d*. The use of $\delta W$ resulted in a close distribution of the spectral widths measured by the 7-pulse sequence and the augmented data obtained by the 16-pulse sequence (see Figure 2, *f*).

Figure 2 illustrates the distribution of spectral widths of signals received during regular measurements of 16-pulse and 7-pulse sequences before ($\mathbb{P}_{W_{16}}$, Figure 2, *c*) and after ($\mathbb{P}_{W_{16,\,\text{augm}}}$, Figure 2, *f*) spectral width compensation by the method (1, 3, 4) as compared to the distribution of spectral widths measured by 7-pulse sequence $\mathbb{P}_{W_7}$.

Figure 2, *a*, *b*, and *e* illustrates probability density distributions in velocity—spectral width coordinates for measurements by the 7-pulse, 16-pulse sequences, and for the 16-pulse sequence after its augmentation.

The *V*, *W* distributions shown in Figure 2 *a* for the 7-pulse sequence are known and were used, in particular, to determine the conditions for separating ground scatter signals [Ponomarenko et al., 2007; Blanchard et al., 2009]. A small proportion of negative spectral widths is a known error related to the features of the signal processing algorithm FITACF.

Narrowing (see Figure 2, *b*) of the *W* distribution when measured by a 16-pulse signal as compared to a 7-pulse signal is associated with a higher spectral resolution of the 16-pulse sequence.

Figure 2 indicates that before augmentation the spectral widths obtained by the 16-pulse sequence are much narrower than those obtained by the 7-pulse one, and augmentation provides distributions for the spectral width close to those observed with the 7-pulse sequence.

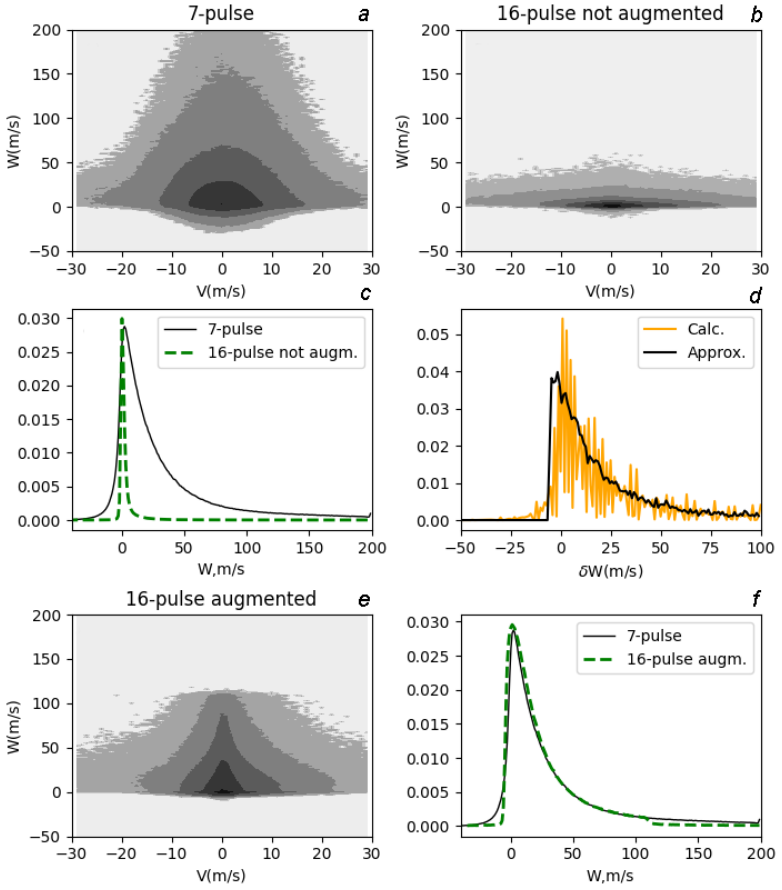


*Figure 2*. Augmentation of spectral broadening of 16-pulse sequences data derived from EKB radar for 2021. Distribution of velocity—spectral width pairs for sequences of two main types — 7-pulse (*a*) and 16-pulse (*b*); spectral width distribution for data (*c*); calculated and simulated distributions of the augmenting additive $\mathbb{P}_{\delta W}$ (*d*); distribution of velocity—spectral width pairs for the 16-pulse sequence after augmentation (*e*); spectral width distribution for data after augmentation (*f*)

### Peculiarities of data clustering and analysis of results

Clustering is the search for splitting data into clusters that are well interpreted by an observer. Within the approach [Berngardt et al., 2022; Berngardt, 2022], it is required not only to split data into clusters, but also to select their number so that the resulting clustering is explicable from the physical standpoint. There are various clustering methods [Saxena et al., 2017], each corresponds to its own data model and usually requires the selection of a hyperparameter on the value of which both the splitting and the total number of clusters depend significantly. To correctly solve the problem of choosing a clustering method and its hyperparameters, two main approaches are generally exploited: using an external expert who evaluates the quality of each specific clustering, or applying a clustering quality metric based on a criterion. In both cases, the assessment is subjective: each expert can interpret the data in their own way, and each quality metric gives its own estimates. The Silhouette coefficient (Rousseeuw, 1987) and a family of information criteria, such as the Bayesian information criterion (BIC) [Schwarz, 1978], are the most commonly used metrics. The choice of criteria often determines the expected form and number of clusters. For example, Kong et al. [2024] compare several clustering methods according to several criteria including Silhouette.

For clustering, by analogy with [Berngardt et al., 2022], we will use 15-dimensional data consisting of parameters measured by radars (velocities, spectral widths, elevation angles, sounding frequencies, time, azimuth/beam number, frequency channel, etc.) and the results of simulation of signal propagation trajectory (angle relative to the geomagnetic field, angle relative to the horizon in different parts of the path, number of signal propagation hops).

Frequency channels and beam numbers are categorical variables that require the choice of a vector representation, which is not obvious in this problem. In this work, unlike [Berngardt et al., 2022; Berngardt, 2022], clustering is used separately on each beam and on each of the two frequency channels, which eliminates the need to search for a vector representation, but significantly increases the number of analyzable independent experiments.

To test the applicability of the GM algorithm to this problem and determine the number of optimal clusters, the clustering problem has been solved in two ways. The first is the GM algorithm with determination of the number of clusters according to the Bayesian information criterion (BIC), hereinafter referred to as GMBIC; it searches for clusters of mostly elliptical form. The second is clustering by the GMsDB algorithm [Berngardt, 2023], which takes GMBIC clusters and combines substantially overlapping elliptical clusters into larger complex clusters. The DBSCAN-GM clustering method, which is similar to GMsDB, has been put forward, for example, in [Smiti et al., 2016], but with a slightly different principle of clustering. Comparing the number of clusters obtained by both algorithms (GMsDB and GMBIC) allows us to figure out how many elliptical clusters do not intersect each other, and if their number is large, indirectly prove that GMBIC may be applied to this problem.

Figure 3 demonstrates distributions of the number of clusters found in radar data by the two methods: GMBIC and GMsDB algorithms with a statistical significance level $\alpha$=0.1 when clusters are combined [Berngardt, 2023]. Panels *a1* and *a2* show the dependences of the number of clusters determined by the two algorithms. The proportionality is seen in the number of clusters, which indicates an approximately constant proportion of intersecting elliptical clusters. Panels *b1* and *b2* illustrate the distribution of the number of GMBIC clusters combined by the GMsDB algorithm. We can see that more complex clusters, which are identified by the GMsDB algorithm, are composed of no more than 3–4 elliptical clusters. The low proportion of complex clusters suggests that clusters generally have a simple elliptical shape. Panels *c1* and *c2* exhibit distributions of the proportion of isolated GMBIC clusters that have no close neighbors. It is apparent that, on average, 80–83 % of clusters determined by GMBIC are isolated. Their high proportion suggests that it is acceptable to apply GMBIC to initial clustering. This also explains the acceptable quality of approximation of coherent radar data clusters achieved by GM-based models [Berngardt et al., 2022; Berngardt, 2022]. Panels *d1*, *d2* show distributions of the number of GMBIC clusters in the data; *e1*, *e2*, distributions of the number of GMsDB clusters in the data. It can be seen that the number of clusters is seen not to exceed 52, and there are, on average, more clusters in MAGW data than in EKB data. Panels *f1* and *f2* display proportions of data in isolated GMBIC clusters. It can be observed that a significant proportion of radar data (from 40 to 80 %) is in isolated elliptical clusters. Their high proportion also indicates that it is acceptable to use the GMBIC method for clustering. For the MAGW radar, the proportion of data in complex clusters exceeds the proportion of such data from the EKB radar.

The GMsDB algorithm tends to combine intersecting clusters into one [Berngardt, 2023], which affects the clustering quality in the case of intersecting clusters. Later on, therefore, the GMBIC clustering was utilized as reference; approximately 80 % of its clusters are isolated and coincide with GMSDB clusters, and only 20 % of clusters in the data are complex (non-elliptical) clusters (see panels *c1*, *c2*). This implies that with GMBIC we can correctly cluster from 50 to 80 % of all data into simple elliptical clusters, and divide complex clusters into several simple ones.

## STAGE 2.
## DATA CLASSIFICATION

### Building an optimal data classifier and its justification

This stage is aimed mainly at finding the minimum fully connected classifier network that repeats clustering

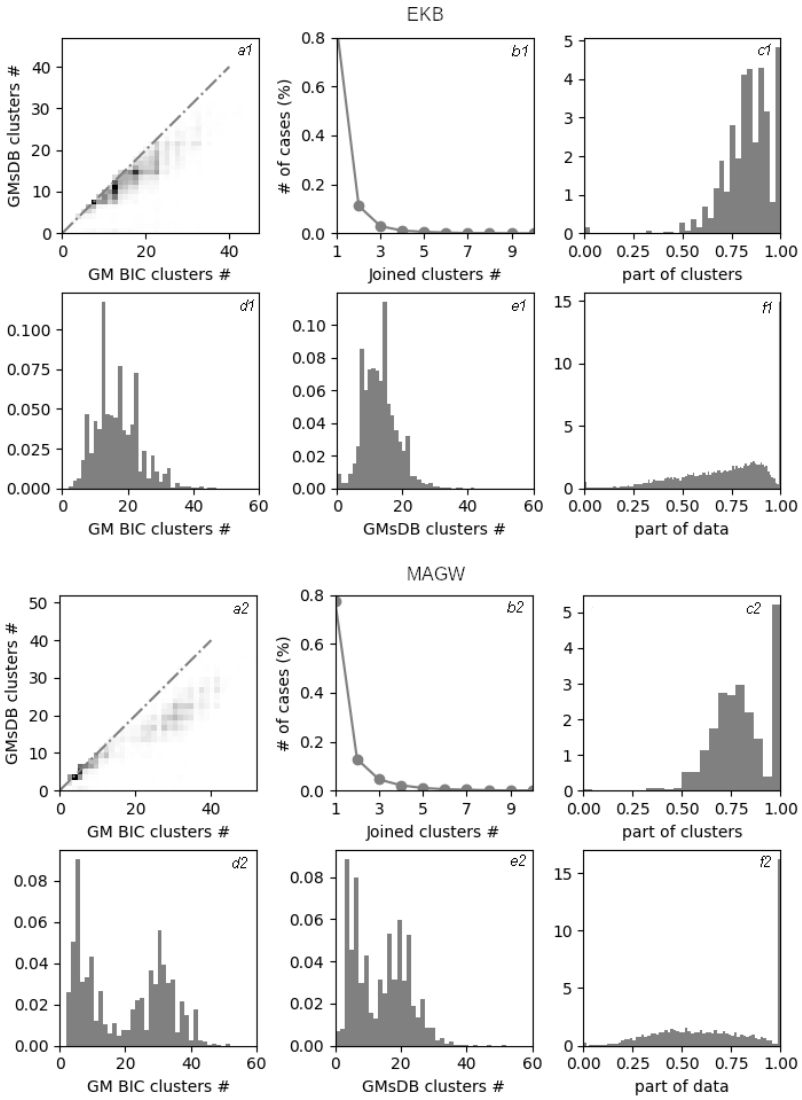


*Figure 3*. Distribution of the number of clusters determined by the GMBIC and GMsDB algorithms in the EKB and MAGW radars: distribution of the number of clusters determined by the two algorithms (*a1*, *a2*); distribution of the number of GMBIC clusters combined by the GMsDB algorithm (*b1*, *b2*); distribution of the proportion of isolated GMBIC clusters (*c1*, *c2*); distribution of the number of GMBIC clusters in the data (*d*1, *d*2); distribution of the number of GMsDB clusters in the data (*e1*, *e2*); proportion of data in isolated GMBIC clusters (*f1*, *f2*)

with high accuracy. The minimum network is important to facilitate the interpretation of its results from a physical standpoint: the number of neurons at the output of the classifier network corresponds to the minimum number of independent signal types in observable data. Let us build a fully connected network consisting of a small number of layers (two layers) with the minimum possible number of neurons in each layer.

To build the network, two datasets have been created: full and shortened. The shortened dataset was used to obtain a good initial approximation for all coefficients of the neural network and to determine its hyperparameters — the minimum number of neurons in each layer. The full dataset was employed for the final solution of the problem and the final network training.

The full dataset was created from ~15000 experiments totaling ~42 million records: (20 million records from the EKB radar and 22 million records from the MAGW radar) and was split at a ratio 4:1 into training and test parts.

The shortened dataset was formed from 1000 experiments (~2.8 million records) randomly selected from the full dataset, and split into training and test parts at the ratio 4:1. The validation parts of the dataset were missing in both cases since cross-validation of the training dataset (in three folds) was applied to training, and three versions of the model were always trained, which was necessary for subsequent analysis.

## Justification of the classifier network architecture

The architecture of neural networks (wrap and classifier) is depicted in Figure 1, *b*, *c*. Here, *K* is the maximum number of clusters after stage 1; *M* is the number of hidden (latent) classes in the data; *N* is the dimension of the hidden layer of the classifier. The network is a significant simplification of the version proposed in [Berngardt et al., 2022; Berngardt, 2022], but provides a better prediction quality.

The choice of the architecture of the new classifier network (see Figure 1, *c*) is based on three principles: a wide two-layer network suffices to approximate continuous functions [Kolmogorov, 1957; Arnold, 1963]; as activation functions of the neural network, it is desirable to use absolute values to maintain the relation with algorithms that have proven themselves well before [Ponomarenko et al., 2007]; as a conversion of network outputs to probabilities, instead of the *Softmax* function, we can normalize nonnegative quantities to their sum.

Let us validate the choice of the absolute value activation function. The model proposed in this paper has been developed from standard approaches to signal separation at coherent scatter radars. The well-known and widely used condition at SuperDARN radars for separating signals into two types (ground scatter and scattering from ionospheric irregularities) according to their spectral characteristics has the form [Ponomarenko et al., 2007]

$$A|V| + B|W| + C > 0, \tag{5}$$

where *A*, *B*, *C* are some constants; *V*, *W* are the measured Doppler shift of a signal and its spectral width. We can assume that in the case of other classes the boundaries can also be described by a superposition of module functions, so it is advantageous to utilize the absolute value function as an activation function.

Let us justify the use of the linear normalization layer. The *Softmax* function is traditionally used at the output of most classifiers, which allows us to normalize neural network outputs so that the values are nonnegative and their sum is 1. However, it is often necessary to perform additional calibration of the obtained values [Guo et al., 2017] or modify the *Softmax* function [Sutton et al., 2018]. Therefore, the choice of the activation function at the classifier output is arbitrary. In this paper, we will utilize the linear normalization layer as activation at the classifier network output:

$$LinearNormalization\left(\vec{x}\right)_i = \frac{x_i}{\sum_j x_j}. \tag{6}$$

When the condition $x_i \geq 0 \; \forall i$ holds (it holds automatically due to the use of absolute activation in the previous layer, see Figure 1, *c*), the output values of the layer, as with *Softmax*, satisfy Kolmogorov's axiomatics in the probability theory [Kolmogoroff, 1933]: they are nonnegative, their sum is 1, and the probability of several mutually exclusive events is the sum of their probabilities. Therefore, the outputs of such a layer can be interpreted as probabilities of the corresponding classes, and this does not require changing the standard loss functions during network training (cross-entropy).

Unlike the previously developed network [Berngardt et al., 2022; Berngardt, 2022], the use of physically adequate (absolute value) activation functions allowed us to solve the problem:

- with fewer input parameters: the new classifier model does not require knowledge of the effective scattering height;
- without methods of increasing data dimension: the classifier does not preliminarily increase data dimension with Polynomial Features;
- significantly reducing the network depth from six to two fully connected layers.

At the input of the classifier (see Figure 1, *c*), there is a batch normalization layer [Ioffe, Szegedy, 2015], which is an adaptive linear scaling of inputs and is employed to speed up the search for optimal network coefficients. If necessary, its coefficients can be added into the coefficients of the first network layer. Subsequent analysis has shown that the constructed neural network provides much higher cluster prediction quality than models [Berngardt et al., 2022; Berngardt, 2022].

## Determining the number of signal classes for classification

The proposed model reduces the problem of signal classification to analyzing encoder outputs (see Figure 1, *a*, *c*) as to the probability that the data belongs to one of several classes. In constructing an interpretable network, it is important to choose the optimal number of neurons in this layer (and the other layers of the neural network) with fixed activation functions.

The initial wide network for the classifier was a network of widths N=300, M=140 in the first and second layers respectively. This can be justified as follows: the number of found clusters in the radar data does not exceed 52 (see Figure 3, *d1*, *d2*), this is the maximum expected number of hidden classes M in the data and the minimum number of neurons in the output layer of the classifier. According to [Berngardt, 2024]), it is advisable to choose an initial number of neurons at least twice the expected minimum number of neurons. The number of neurons in the last layer has therefore been chosen to be about three times the maximum number of clusters; and the number of neurons in the first layer, about 6 times. As it turned out, this architecture is sufficient to search for the minimum number of neurons, and such a neural network can be trained for a reasonable amount of time on an ordinary personal computer. To speed up the search for the minimum number of neurons, the network was trained at the reduced dataset (1000 experiments) described above. Cross-validation according to the algorithm [Berngardt, 2024] was carried out at three folds. As a result, three versions of the model were trained.

According to the algorithm [Berngardt, 2024], finding the minimum number of neurons when evaluating the quality of the network requires quality metrics *Q* that satisfy the condition:

$$\begin{aligned} &Q\left(X_1 \vee X_2\right) = \\ &= \frac{Dim\left(X_1\right)Q\left(X_1\right) + Dim\left(X_2\right)Q\left(X_2\right)}{Dim\left(X_1\right) + Dim\left(X_2\right)}; \\ &X_1 \wedge X_2 = \varnothing, \end{aligned} \tag{7}$$

where $X_1$, $X_2$ are disjoint datasets, and *Dim*(*X*) is the number of elements (samples) in *X*. The Accuracy metric has therefore been taken as a basic metric when searching for the minimum number of neurons.

The number of independent classes was estimated by two methods. The first method provides an upper bound on the minimum sufficient number of classes in the data.

Using the dataset (filtered by the elevation angles and augmented), clustering of each experiment was performed using GMBIC, a neural network (300×140 neurons) was trained, the minimum number of neurons in layers was determined (49 and 35 respectively, Figure 4, *a1*), the minimum network (49×35 neurons) was trained. In this case, a sample was assumed to belong to a given class if all three trained classifiers predicted the same class for it (the ensemble voting classification method). The number of samples in the ranked series of classes is shown in Figure 4, *b1*. From the observed inflection point (a sharp decrease in the number of observations of samples in the class) we can estimate the number of frequently observable classes. It is 27 for the EKB radar and 28 for the MAGW radar (vertical dash-dotted lines in Figure 4, *b*1). Note that reducing the number of classes to <35 (and re-training the classifier with a new number of classes) does not make the inflection position stable: for example, when choosing the number of hidden classes equal to 31, it becomes 23 and 24 for the EKB and MAGW radars respectively (see Figure 4, *c1*, vertical dash-dotted lines). Therefore, decreasing the number of classes to <35 is probably unjustified. Using this method corresponds to the fact that all the clusters we found have a shape close to elliptical, but can significantly intersect. Obviously, if real clusters have a more complex shape, this method can overestimate the number of independent classes. Therefore, such an estimate is in line with the upper bound on the number of classes.

The second method gives a lower bound on the minimum sufficient number of classes. For the original dataset (filtered by elevation angles and augmented), clustering was performed by GMsDB, a wide neural network was trained as by the first method (300×140), and the minimum number of neurons per layer was determined (36 and 23 respectively, Figure 4, *a2*). After that, the final network with a minimum number of

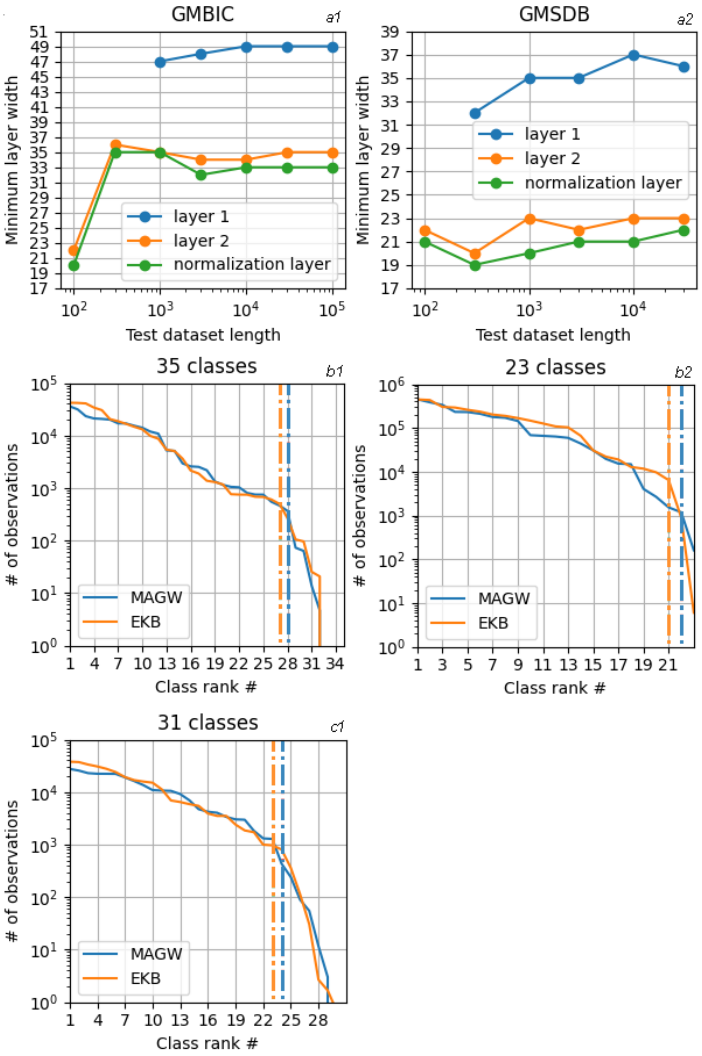


*Figure 4.* Estimated minimum number of neurons in the network and distribution of classes by the number of elements: on the left is the GMBIC clusterer; on the right, the GMsDB clusterer; *a1*, *a2* — the dependence of the minimum number of neurons in classifier layers (first hidden, second hidden, and output normalization ones) on the dataset amount used for the search; *b1*, *b2* — the number of samples in a class as a function of the rank of this class for an optimal network (35 classes and 23 classes respectively); *c1* is the number of samples in a class as a function of the rank of this class for an inoptimal network (31 classes). Vertical lines of respective colors in *b1*, *b2*, *c*1 indicate the boundary between frequently and rarely occurring classes in corresponding radars.

neurons (36×23) was trained using the dataset clustered by GMBIC. This corresponds to the finding of the minimum possible number of disjoint classes. It should be noted that the number of classes confidently defined as the inflection position in the plot in Figure 4, *b2* is only by 1–2 smaller than this value: 21 classes for EKB and 22 for MAGW, which indicates the stability of this separation. Application of this method suggests that the found clusters can have a complex shape and do not intersect. Obviously, if the clusters that actually exist in the data intersect with each other (signal types are poorly distinguishable), this method underestimates the number of independent classes. Therefore, such an estimate is in line with the lower bound on the number of classes. Note that such a lower bound on the number of classes is close to the number of clusters and hidden classes used empirically in [Berngardt et al., 2022; Berngardt, 2022] (20 classes).

Thus, the expected number of classes in the data ranges from 23 to 35, and an error in classification based on the upper bound (35 classes) can lead to separation of complex classes into several parts, and an error in classification based on the lower bound (23 classes) can result in joining of several different, but similar classes into one. Obviously, the former is preferable, and therefore we will use it in the future.

To increase accuracy and generality, the model was retrained on the entire training dataset (more than 15000 experiments, or ~25 million different samples). The model is trained in three versions, using three-fold cross-validation; splitting into folds is random.

### The final algorithm for finding the optimal classifier

The final algorithm for finding the optimal classifier consists of the following stages.

1. Extracting a data part that fits elevation angles below the threshold one [Milan et al., 1997] (28° for the EKB radar and 38° for the MAGW radar) corresponding to the signals coming from the main lobe of antenna pattern.

2. Calculation of trajectory parameters of radio wave propagation, trajectory shape, angle with the magnetic field, and scattering height from radar data, and their aided expansion of the set of parameters measured by radars.

3. Augmentation of the calculated spectral width of received signal for experiments conducted with long pulse sequences (16-pulse) according to (3), (4).

4. Clustering of each experiment (experiments differ in dates, azimuths, and frequency channels) by GMBIC in the previously described 15-dimensional parameter space.

5. Selection of a small dataset (1000 experiments) from experiments used for stages 6–7.

6. Training of a sufficiently wide neural network (see architecture in Figure 1, *c*) on 9 physical parameters, using a network with 300 hidden neurons in the first hidden layer and 140 neurons (latent classes) in the second.

7. Determination of the minimum number of neurons in layers of the resulting network by the algorithm [Berngardt, 2024] with Accuracy as a metric.

8. In all available experiments, the neural network is trained (see architecture in Figure 1, *a*–*c*) with the optimum number of neurons found for the classifier network.

At stages 1–3, we prepare the data; at stages 4–7, for a small part of the dataset, we determine the number of hidden classes in the data and the optimal number of neurons in the network; at stage 8, we train the final optimal classifier at the entire available dataset.

The proposed algorithm automatically determines the number of classes in the data, is fully data-driven, and does not require an expert at any stage. The algorithm is self-consistent and self-learning: it finds all algorithm parameters automatically, except for the list of parameters, used for clustering and classification, and the general network architecture, the reasons for which have been outlined above.

The resulting neural network has 49 neurons in the first layer and 35 neurons in the second, which means there are 35 different classes in the data. The model achieves a clustering repetition quality of 0.92 according to the AUC-PR metric and significantly exceeds the previous networks' quality of 0.68 [Berngardt, 2022; Berngardt et al., 2022].

The resulting neural classifier network (see Figure 1, *c*) provides a minimum number of neural network parameters with high quality of its operation, which subsequently simplifies its interpretation. On the other hand, this number of neurons can be interpreted as the optimum number of radiophysically distinguishable classes in EKB and MAGW radar data.

The final model (see Figure 1, *c*) for determining signal class from its parameters has an analytical form

$$y_k = \left| \sum_{j=1}^{49} C_{kj} \left| \sum_{i=1}^{9} A_{ij} x_i + B_j \right| + D_k \right|, \tag{8}$$

$$k_{\text{detected}} = argmax\left(y_k\right), \tag{9}$$

where $A_{ij}$, $B_j$, $C_{kj}$, $D_k$ are the coefficients that are searched for as a result of network training; $x_i$ denotes input parameters; $k_{\text{detected}}$ is the number of hidden class to which the measured signal with parameters $x_i$ will belong. The number of model parameters can be easily calculated from the above formula and is equal to 2240. The model can be easily implemented for fast calculations and without frameworks of neural networks. The obtained formula for optimal signal classification is structurally close to the results of the Kolmogorov—Arnold theorem [Kolmogorov, 1957; Arnold, 1963]. Its structural relationship is also seen with the standard algorithm for separating signals, scattered from the ionosphere and Earth's surface [Ponomarenko et al., 2007], given in (5) and traditionally used at SuperDARN radars.

## DISCUSSION

### Interpretation of the resulting classes

Cross-validation is used in model training and research, which makes it possible to employ three network variants to construct an ensemble model. As the

analysis of the data for 2021 showed, the classes defined by three versions of the model coincide with a high degree of quality: the adjusted Rand index [Hubert, Arabie, 1985] in pairs between the results of the three models lies in the range 0.936–0.967 for the EKB radar and 0.897–0.937 for the MAGW radar, which indicates a close similarity between the classifications obtained by these models and allows us to apply any of the models separately or three models together (ensemble) to interpretation.

With the ensemble use of the three models, it is convenient to employ a voting mechanism and decide on the signal class when predictions of all models match. If network predictions do not match, the result is placed in a separate class (data that cannot be unambiguously interpreted).

The statistics of parameters of various classes (95 % confidential interval) determined by such an ensemble method in 2021 is presented in Figure 5 separately for EKB and MAGW radars. The classes are divided into three groups, highlighted in color: ground scatter, ionospheric scatter, and signals that are difficult to interpret.

The last class included signals with incredibly high velocities or spectral widths (>1000 m/s). Signals with a low average scattering height (<100 km) were interpreted as ground scatter; the remaining part of the classes, as ionospheric scatter of different types.

Figure 5 presents the statistics of scattering heights (Hiri), the number of propagation hops (Mode), the radar range (Range), the Doppler velocity $V_d$, the spectral width *Wl*, the cosine of the angle of the radio wave trajectory with the geomagnetic field cos(*k*, *B*), the elevation angle of the trajectory with the horizon at the scattering point sin(*k*, *xy*), and the number of observations of this class (# of cases).

The need to use a large number of classes for signals scattered from the ionosphere in automatic data classification has already been suggested and justified [Burrell et al., 2015]. Multiple types of ground scatter signals have already been proposed and substantiated, for example, in [Kunduri et al., 2022]. Due to the complexity and dynamics of the processes occurring in the ionosphere, it is expected that the number of types of signals scattered from the ionosphere will exceed the number of types of signals scattered from the Earth surface.

Three of the found classes have a negligible amount of data (class 1, 22, 27).

Analysis of the behavior of the main signal features shown in Figure 5 allows us to pre-interpret the classes as follows.

### Scattering of ionospheric types

It includes 13 classes: 0, 2, 3, 5–7, 10, 11, 19–21, 27, 32. The total proportion of such signals (from the main lobe of the antenna pattern) is 56 % for EKB and 48 % for MAGW. They can be interpreted as follows (see Figure 5).

1. Class 0 is 1.5th or 2.5th hop aspect scattering in the E/F layer. Heights 100–200 km, the 2nd hop, distances 1500–3000 km, high mostly positive velocities up to 800 m/s, high spectral widths up to 600 m/s, proximity to orthogonality to the magnetic field $\cos\left(\vec{k}, \vec{B}\right) \in [-0.25..0]$, scattering in the ascending branch of the trajectory $\sin\left(\vec{k}, \overrightarrow{xy}\right)[R] > 0$.

2. Class 2 is 0.5th hop aspect scattering in the F layer. Heights 300–450 km, the 1st hop, distances 800–2500 km, high velocities up to 250 m/s, high spectral widths up to 250 m/s, proximity to orthogonality to the magnetic field, scattering in the ascending branch of the trajectory.

3. Class 3 is 1.0st hop non-aspect scattering in the E/F layer or magnetically oriented quasi-bistatic scattering [Kravtsov, Namazov, 1980; Berngardt et al., 2016], when trajectories of incident and scattered waves differ significantly. Heights 80–200 km, the 1st hop, distances 1000–2500 km, high velocities up to 300 m/s, average spectral widths up to 200 m/s, lack of orthogonality to the magnetic field, scattering in the descending part of the trajectory.

4. Class 5 is 1.5th hop aspect scattering in the E/F layer. Heights 50–200 km, the 2nd hop, distances 1500–3000 km, high mostly negative velocities up to 500 m/s, high spectral widths up to 800 m/s, pronounced orthogonality to the magnetic field, scattering mainly in the horizontal or ascending part of the trajectory.

5. Class 6 is 0.5th hop aspect scattering in the E layer. Heights 100–200 km, the 1st hop, distances 350–700 km, low velocities, spectral widths up to 200 m/s, pronounced orthogonality to the magnetic field, scattering mainly in the horizontal or ascending part of the trajectory.

6. Class 7 is presumably Pedersen ray scattering [Ponomarenko et al., 2011] or magnetically oriented quasi-bistatic scattering [Kravtsov, Namazov, 1980; Berngardt et al., 2016]. Heights 200–250 km, the 1st hop, distances 1000–1500 km, low velocities up to 100 m/s, positive for EKB, negative for MAGW, average spectral widths up to 250 m/s, lack of orthogonality to the magnetic field, scattering mainly in the horizontal or descending part of the trajectory.

7. Class 10 is 0.5th hop scatter by Pedersen ray [Ponomarenko et al., 2011] or in the E/F layer. Heights 100–300 km, the 1st hop, distances 500–1500 km, low velocities, average spectral widths up to 250 m/s, weak orthogonality to the magnetic field, scattering near the horizontal part of the trajectory.

8. Class 11 is 0.5th hop ionospheric scatter in the E/F layer. Heights 170–300 km, the 1st hop, distances 500–1000 km, low velocities, average spectral widths up to 200 m/s, weak orthogonality to the magnetic field, scattering in the horizontal or ascending part of the trajectory.

9. Class 19 is 0.5th hop scattering in the E/F layer. Heights 70–300 km, the 1st hop, distances 600–2000 km, average velocities up to 200 m/s, average spectral widths up to 250 m/s, weak orthogonality to the magnetic field, scattering mainly in the horizontal part of the trajectory.

10. Class 20 is a possible analog of near-range echo (scattering at E-layer heights at short radar ranges <300 km [Ponomarenko et al., 2016]), but for F-layer heights (hereinafter F-layer near-range echo). Heights 150–300 km, the 1st hop, distances 250–700 km, average

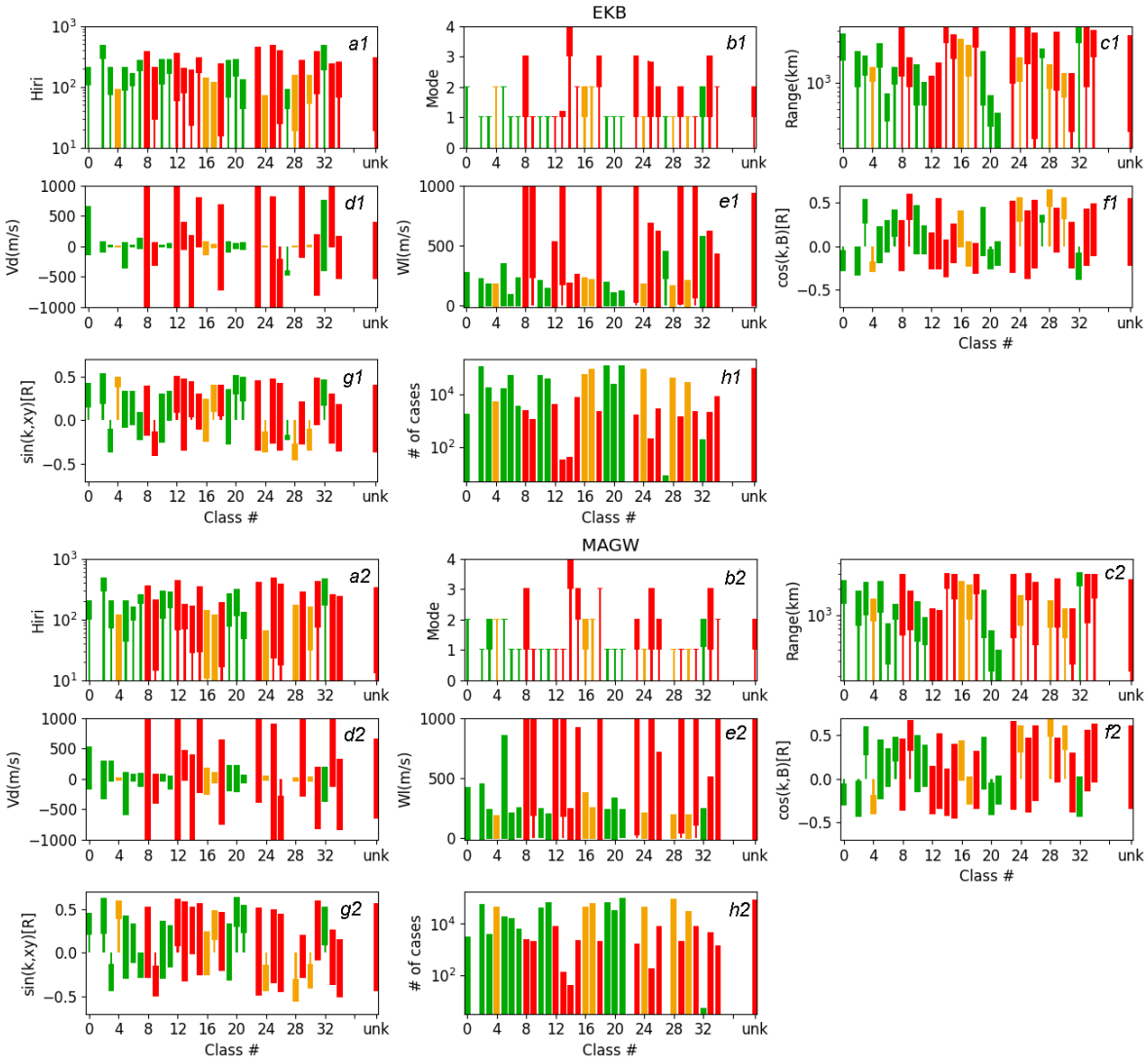


*Figure 5*. Parameters of various classes for the EKB and MAGW radars according to statistics on 2021 (95 % confidential interval) obtained by the ensemble estimation method. Red color indicates presumably noise classes; green, ionospheric scatter; orange, ground scatter. In the last column (unk) are signals interpreted in various ways by different network versions; *a1*, *a2* — scattering height; *b1*, *b2* — number of reflections from the underlying layer or the Earth surface; *c1*, *c2* — radar range; *d1*, *d2* — Doppler velocity; *e1*, *e2* — spectral width; *f1*, *f2* — the cosine of the angle with the magnetic field at the scattering point; *g1*, *g2* — the sine of the elevation angle at the scattering point; *h1*, *h2* — the number of signal observations

velocities up to 200 m/s, high spectral widths up to 300 m/s, weak orthogonality to the magnetic field, scattering in the ascending part of the trajectory.

11. Class 21 is meteor echo [Chisham, Freeman, 2013; Berngardt, 2022] and near-range echo [Ponomarenko et al., 2016]. Heights 60–100 km, the 1st hop, distances 220–400 km, low velocities up to 100 m/s, average spectral widths up to 200 m/s, weak orthogonality to the magnetic field, scattering in the ascending part of the trajectory.

12. Class 27 is 1st hop scattering in the E layer, heights 30–80 km, the 1st hop, distances 2000 km, high velocity (~–400 m/s), high spectral widths up to 500 m/s, orthogonality to the magnetic field is not pronounced, scattering in the descending part of the trajectory.

13. Class 32 is possible 1.5–2.5th hop F-scattering. Heights 200–450 km, 1–3 hops, distances 2000–4500 km, high velocities up to 800 m/s, high spectral widths up to 300 m/s, orthogonality to the magnetic field is not pronounced, scattering in the ascending part of the trajectory.

## Ground scatter

It includes 6 classes: 4, 16, 17, 24, 28, 30. The total proportion of such signals (from the main lobe of the antenna pattern) is 31 % for EKB and 37 % for MAGW. They can be interpreted as follows.

1. Class 4 is 1st hop ground scatter. Heights below 100 km, the 2nd hop, distances 900–1500 km, low velocities, average spectral widths up to 200 m/s, lack of orthogonality to the magnetic field, scattering in the ascending branch of the trajectory.

2. Class 16 is scattering with high spectral widths and velocities, possibly 1st hop ground scatter with strong refraction by short-living ionospheric irregularities. Heights 0–100 km, 1–2 hops, distances 180–3000 km, average velocities up to 200 m/s, high spectral widths

up to 400 m/s, weak orthogonality to the magnetic field, scattering in the horizontal part of the trajectory.

3. Class 17 is 2nd hop ground scatter or 1.5th hop E-scattering. Heights 0–100 km, the 2nd hop, distances 1000–2500 km, low velocities, average spectral widths up to 200 m/s, weak orthogonality to the magnetic field, scattering in the ascending part of the trajectory.

4. Class 24 is 1st hop ground scatter. Heights 0–70 km, the 1st hop, distances 700–2000 km, low velocities, average spectral widths up to 200 m/s, orthogonality to the magnetic field is not pronounced, scattering in the descending part of the trajectory.

5. Class 28 is 1st hop ground scatter, height 20–150 km, the 1st hop, range 800–1500 km, low velocities, average spectral widths up to 200 m/s, lack of orthogonality to the magnetic field, scattering in the descending part of the trajectory.

6. Class 30 is 1st hop ground scatter, the 1st hop, heights 40–150 km, range 600–1200 km, low velocities, average spectral widths up to 200 m/s, lack of orthogonality to the magnetic field, scattering in the descending part of the trajectory.

Proportion of signals in the remaining (uninterpreted) classes is low, 4 % for the EKB radar and 5 % for the MAGW radar. The proportion of signals detected differently by various models is ~10 % in each of the radars. Thus, the proposed method allows us to automatically classify ~85 % of all data received in the main lobe of the antenna pattern.

### Range-time dynamics of different classes

Figures 6, 7 illustrate diurnal variations in signals of different classes according to ISTP SB RAS EKB and MAGW radar data for 2021, using a test dataset that did not participate in training. The advantage of this data representation is that neither range nor time are directly involved in the data classification, and the appearance of grouped areas of points in these coordinates serves as a subjective confirmation of the good quality of the classification. The plots obtained make it possible in some cases to confirm the above interpretation of these classes.

An additional confirmation of the correctness of the classification is the height distribution of signals of several classes (Figure 8): meteor echo/near-range echo (scattering by the E layer at close distances), F-layer near-range echo (scattering by the F layer at close distances), other ionospheric signals and ground scatter signals. It can be seen that the height distribution of meteors determined by the algorithm is well matched to what is expected with a maximum at ~80–100 km [Chisham, Freeman, 2013], the F-layer near-range echo distribution corresponds to heights of ~180 km, ground scatter signals are concentrated at 0–100 km, and ionospheric scatter of other types has a maximum distribution at 180–200 km. The height distributions of signals of different types for the EKB and MAGW radars are similar.

### Degree of importance of various input parameters of the model

One of the urgent issues in identifying the types of scattered signals is the choice of necessary parameters [Burrell et al., 2015; Ponomarenko, McWilliams, 2023]. Within the data-driven approach, we can formulate this problem in terms of the feature importance: which parameters most strongly affect the quality of detection of each specific class by the model we have built. A similar approach has been adopted in [Kong et al., 2024]. In machine learning, there are a large number of different methods for such estimate [Huang et al., 2020]. One of the universal methods to do this is the permutation feature importance [Breiman, 2001], in which the importance of an input parameter for prediction is estimated from the change in prediction quality when values of this parameter are randomly permutated in the dataset.

Since it is desirable not only to arrange the input parameters in order of importance, but also to find the optimal combination of such parameters for classification, it is advisable to use greedy modifications of the algorithm (the modification adopted in this work is given in Appendix 1).

Figure 9 shows the degree of importance of various input parameters for determining different classes ($\Delta Q_{\text{lopt}}$), a higher value corresponds to a more important feature. The degree of importance for the total classification is also given (column “total”). The results are presented for each network variant obtained in cross-validation. The cells for which the value is missing represent insignificant parameters.

Important for the classification is most often seen to be the height at which the signal is scattered, elevation of the radio wave propagation trajectory at the scattering point, as well as approximately equally the angle with the geomagnetic field at the scattering point and the elevation angle in the middle of the signal propagation path. The least important parameters are the signal propagation mode and the spectral width of the received signal. This is consistent with qualitative expectations: knowing the scattering height and the propagation trajectory of signals really makes it easy to distinguish between their different types. For ground scatter, the scattering height should be ~0; for meteor echoes, ~90 km; for ionospheric scatter, from 100 to 400 km. Elevation of the trajectory and the angle with the magnetic field will allow us to distinguish ordinary scattering from the aspect one characteristic of plasma instabilities of the ionospheric E and F layers.

Thus, the radio propagation trajectory shape and the scattering height are the most important for classification of scattered signals. These parameters cannot be measured directly by a radar and require simulation of radio wave propagation. To determine them, we should know the sounding frequency, the three-dimensional antenna pattern, the measured elevation angle and azimuth of the incoming radio wave, as well as the three-dimensional structure of the refractive index of the ionosphere. Obviously, in complex situations when the propagation trajectory is difficult to predict (there are no reliable measurements of the elevation angle of the incoming radio wave or there is no sufficiently accurate model of the ionosphere), this method will give significant errors, which explains the extensive usage of simpler methods based on measuring velocity and spectral width [Ponomarenko et al., 2007; Blanchard et al. al., 2009] at high-latitude radars. As the work shows, the method we have developed can be applied to calibrated mid-latitude radars using IRI model.

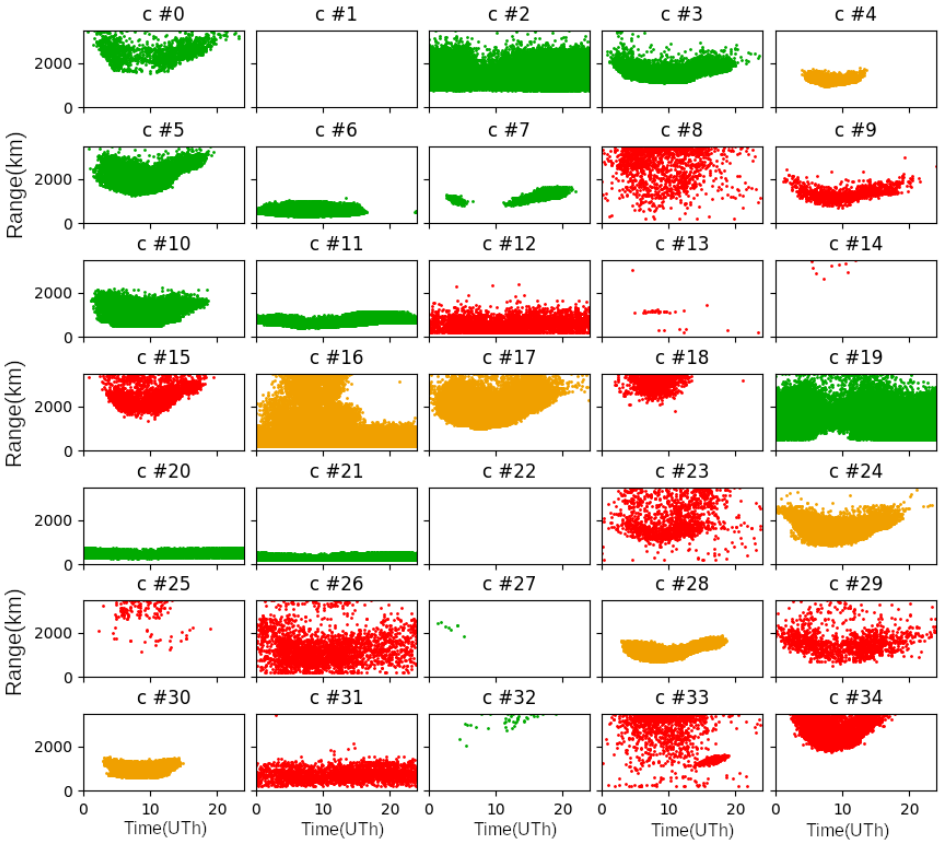


*Figure 6.* Range-time distributions of signals at the EKB radar: the result of the distribution of data by class at the test dataset by voting at the ensemble of three networks. Indefinite data is excluded. Colors indicate different signal types as in Figure 5

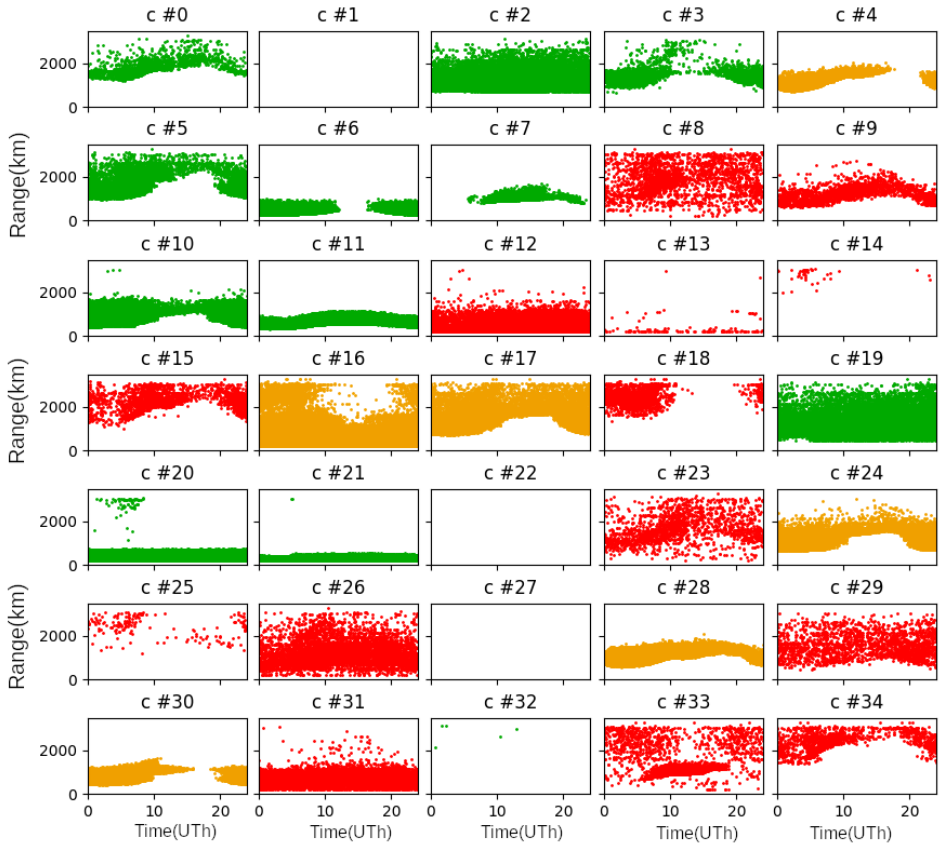


*Figure 7.* Range-time distributions of signals in the MAGW radar: the result of the distribution of data by class at the test dataset by voting at an ensemble of three networks. Indefinite data is excluded. Colors indicate different signal types as in Figure 5

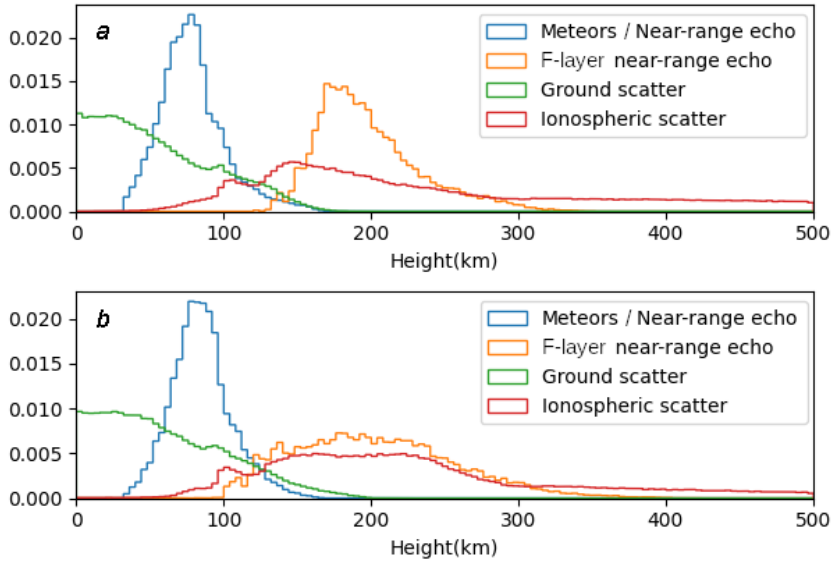


*Figure 8.* Distribution of scattering of various types by heights obtained through ray tracing, using radar data and IRI at EKB (*a*) and MAGW (*b*) for 2021. Distributions of meteors/near-range echo (class 21), F-layer near-range echo (scattering by F layer at close ranges) (class 20), ionospheric scatter of other types (classes 0, 2, 3, 5–7, 10, 11, 19, 27, 32), and ground scatter (classes 4, 16, 17, 24, 28, 30)

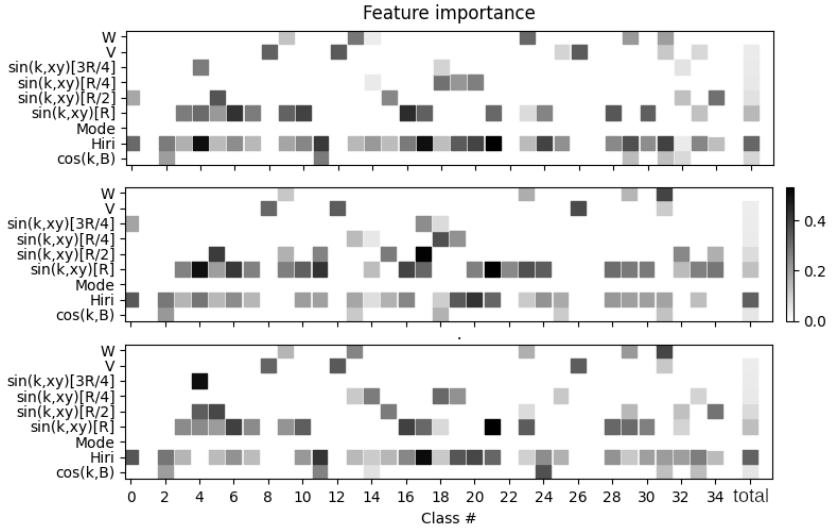


*Figure 9.* The degree of importance of various parameters for determining classes by three different network variants obtained from cross-validation training, as well as the total degree of importance of each parameter for the classification result (column "total")

## Testing the algorithm based on observations in 2023

To test the model's performance, the data that had not previously been employed in training and under different geophysical conditions was processed using MAGW data for the first half of 2023.

The processing results (range-time distributions of signals of different classes and distributions of different classes by characteristics) are presented in Figures 10, 11. There is a good qualitative agreement with the results of processing of initial (training) data for 2021 (see Figures 6, 7). IRI-2020 was used to simulate radio wave propagation [Bilitza et al., 2022].

The main feature of the 2023 data is a significant difference in the level of ionospheric disturbance. According to the Royal Observatory of Belgium, in 2021 the annual average number of sunspots was 30, while for the first half of 2023 it was 129. This leads both to more active scattering of various ionospheric types and to degradation of accuracy of prediction of radio wave propagation by IRI under disturbed conditions (the error in trajectory calculations usually increases with increasing range). Another feature of the data is likely to be a less accurate calibration of the radar by the elevation angle (see, e.g., the change in meteor distribution in Figure 11, *b*).

Note that the proportion of ground scatter signals decreased to 24 % compared to 2021, the proportion of signals scattered by the ionosphere increased to 51 %, the proportion of uninterpreted signals doubled to 10 %, and the proportion of signals differently determined by different networks increased 1.5 times, to 15 %. Thus, in the first half of 2023, the algorithm made it possible

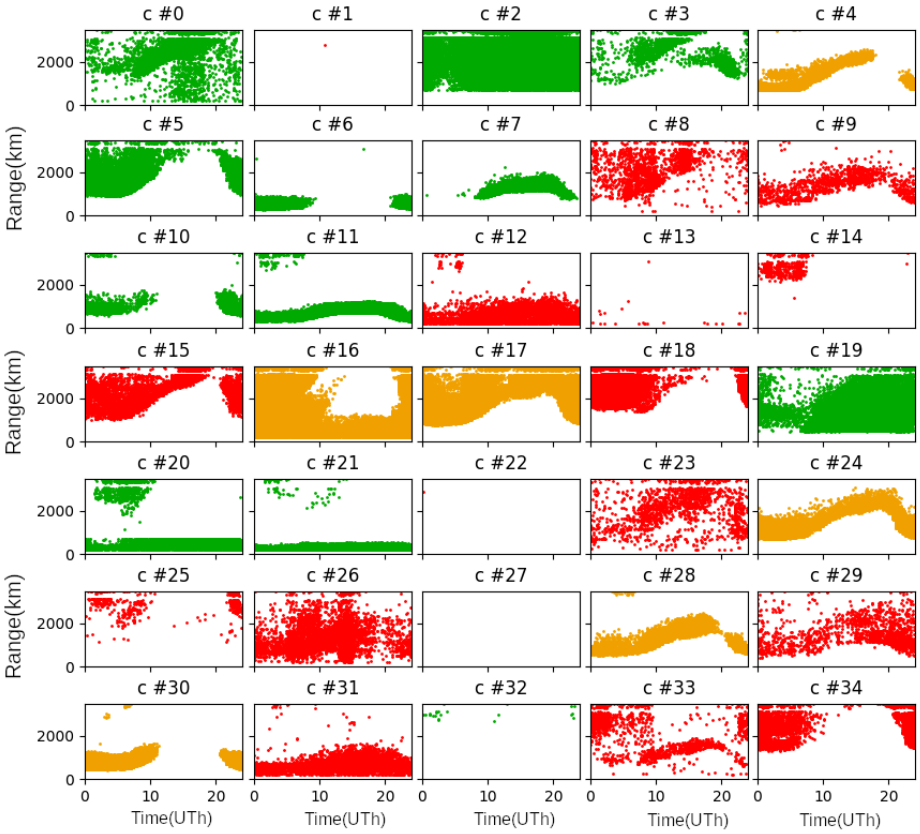


*Figure 10*. Results of the MAGW radar data classification for January–May 2023: range-time distribution of observations of each class. Colors indicate different signal types as in Figure 5

to automatically interpret 75 % of all signals, which is 10 % less than in 2021.

Figure 10 shows advantages and limitations of the proposed method. The advantages include the qualitative agreement of range-time distributions of signals of various classes with the data for 2021, which suggests that the method can be applied to new data.

Among limitations is the observable degradation of accuracy of determining the classification quality at ranges above 2000 km. The most significant is class 20 (F-layer near-range echo) in Figures 6, 7, 10. Comparing the Figures indicates that in 2023 the method more often makes mistakes at ranges above 2000 km, where a significant error in trajectory calculations is expected to accumulate. A similar effect is observed in some other classes: ionospheric scatter (classes 10, 11) and meteor/near-range echo (class 21). An indirect sign of a decrease in the quality of calculations is also changes in the mode composition of signals (see Figures 5, 11): almost all classes began to include higher modes than in 2021, which indicates difficulties in trajectory calculations, and may also be due to an increase in the level of background ionospheric disturbance in 2023 as compared to 2021. High velocities in ground scatter classes also indicate an increase in the ionospheric disturbance level and the accompanying large-scale wave activity in the background ionosphere.

The limitations of the model include the inability to separate very close classes, for example, the E-layer near-range echo and meteor echo, combined by this model into one class (class 21). This limitation is caused by two of its features: locality (it ignores the temporal behavior of irregularities over long lifetimes since it uses the equivalent standard spectral resolution of the 7-pulse sequence limiting lifetimes to ~50 ms), and the accuracy of height determination (parameters of these irregularities cannot be separated with required accuracy due to insufficient accuracy in determining the elevation angle).

## Seasonal and diurnal features of observations of various classes

Seasonal and diurnal features of observation of various classes of signals during 2021 are shown in Figure 12.

Panels *a*–*d* plot the daily dependence of the occurrence of various classes (demonstrated in local solar time at the calculated scattering point) and the seasonal dependences of observation of various classes of signals. We can see that most EKB radar signals are observed during the day; at the MAGW radar, the diurnal effect in signals is less pronounced. The seasonal dependence is more pronounced at MAGW and less pronounced at EKB. A similar effect may be due to the fact that MAGW is located more to the pole than EKB. Therefore, the illumination for the MAGW radar has a more pronounced seasonal dynamics, and ionospheric dynamics is controlled by the magnetosphere to a greater extent than in EKB.

Panels *e* and *f* present the statistics on 2nd and 1st hop ground scatter (classes 17 and 24). As expected, this scattering is observed mainly during the daytime when

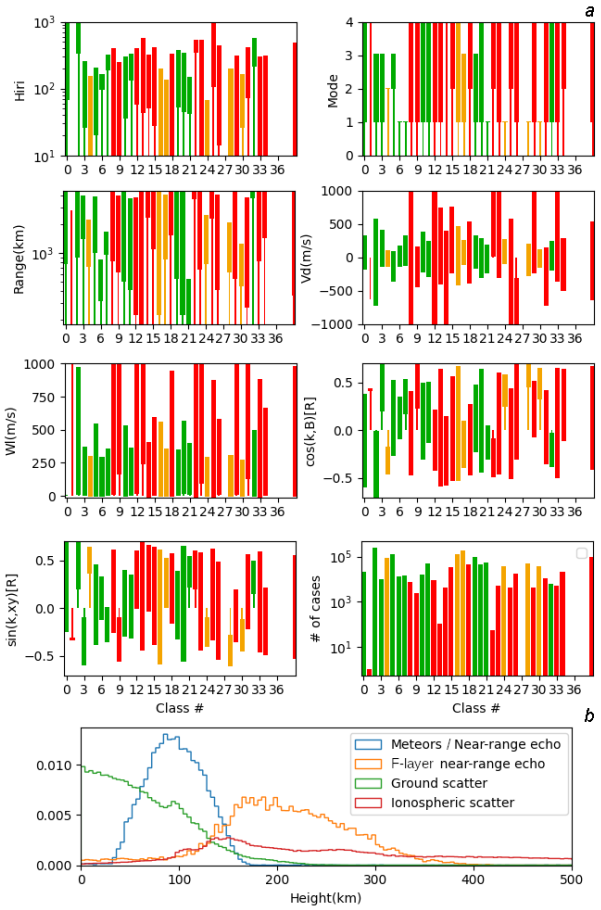


*Figure 11*. Results of the MAGW radar data classification for January–May 2023: 95 % interval of changes in the main parameters in each class, colors and types of parameters are similar to those in Figure 5 (*a*); height distribution of signals of various types; signal types are similar to those in Figure 8 (*b*)

the electron density in the ionosphere is high enough to reflect the radio signal from the ionosphere, and in summer signals are shielded by the near-range echo of the E and F layers.

Panel *g* displays the F-layer near-range echo (class 20) most intense in summer, which is one of the causes for the shielding of ground scatter signals. It is apparent that at the MAGW radar this type of scattering can also be observed in the solar terminator region.

Panel *h* shows a mixed class of signals — meteor echo/near-range echo (class 21), which is difficult to separate by scattering characteristics — close heights and low velocities [Ponomarenko et al., 2016]. Signals of this class are demonstrated to be most often observed at night (meteor observations) and in summer (near-range echo observations).

Panel *i* exhibits possible scattering by Pedersen ray (class 10) [Ponomarenko et al., 2011]. This type of scattering is most often observed in winter during the daytime.

Panel *j* presents a possible candidate for quasi-bistatic scattering by magnetically oriented irregularities (class 7), the possibility of which was predicted in [Kravtsov, Namazov, 1980; Berngardt et al., 2016] and is related to the fact that signal paths in forward and reverse directions may not coincide; therefore, the condition of orthogonality of scattering to the geomagnetic field for the trajectory of the received scattered radio wave is not met [Berngardt et al., 2016].

Panels *k*, *l* give two examples of ionospheric scatter: 0.5 hop aspect scattering in the F layer (class 2), and 0.5 hop scattering in the E/F layer with weak aspect sensitivity (class 19). Scattering of these types is seen to be generally observed during the unlit time periods, which qualitatively corresponds to empirical patterns.

Panels *e*–*l* indicate that many irregularities intensify near the solar terminator, which is associated with its high spatio-temporal dynamics. Figure 12 also shows that scattered signals of many types can be divided into mainly daytime (mostly ground scatter) and mainly nighttime (mostly scattering by ionospheric irregularities).

## CONCLUSION

The paper has attempted to solve the problem of automatic classification of coherent scatter radar data by minimizing the influence of subjective human opinion on preparation and interpretation of data, as well as to analyze the resulting solution.

Within the self-consistent data-driven approach, we have developed a method for automatically constructing such a classifier. This method allowed us to construct, train, and study compact mathematical model (8)–(9), which makes it possible to automatically classify EKB and MAGW radar data by using radiophysical features of their propagation and scattering. The number of free model parameters is 2240, and the number of found signal classes is 35. The work is a generalization, improvement, and mathematically more rigorous development of the approach outlined in previous papers [Berngardt et al., 2022; Berngardt, 2022].

The following results have been obtained.

An empirical model has been developed for augmenting the results of sounding with 16-pulse sequence up to 7-pulse sequence results (3)–(4). Such an augmentation is necessary to train a single classifying model independent of the type of sounding signal in use.

We have created and trained a neural network (classifier), using absolute value activation functions [Vallés-Pérez et al., 2023] and an output layer of linear normalization (see Figure 1, *c*). The method of searching for the minimum number of neurons in a fully connected layer of the neural network [Berngardt, 2024] allowed us to find the minimum number of neurons in this network: 49 and 35 neurons in the first and second layers respectively. The number of neurons in the second layer corresponds to the number of hidden signal classes in the data. The trained neural network provides a high quality of repetition of the results of the clustering in use (0.92 from the AUC-PR metric), significantly exceeding the previous model's quality metric of 0.68 [Berngardt et al., 2022; Berngardt, 2022]. The resulting network is much smaller than the previous version [Berngardt et al., 2022; Berngardt, 2022] and has 2240 parameters.

By comparing the results of network training for two clusterings (Gaussian mixture+BIC criterion for estimating the number of clusters (GMBIC) and GMsDB)

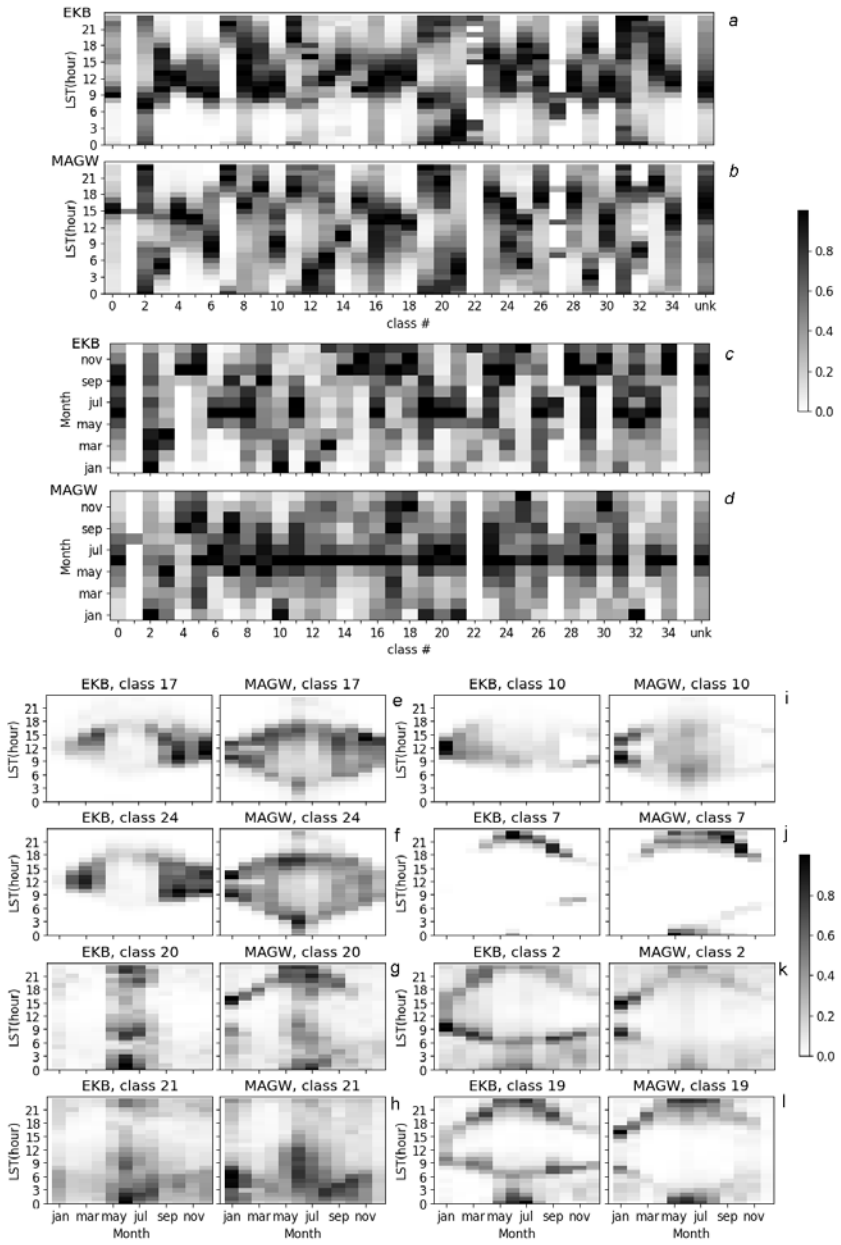


*Figure 12.* Time dependence of the occurrence of various classes, demonstrated in the local solar time at the calculated scattering point (*a*, *b*); seasonal dependence of the occurrence of various classes (*c*, *d*); seasonal-time dependences of observations of various classes: *e*, *f* — 2nd and 1st hop ground scatter (classes 17 and 24); *g* — F-layer near-range echo (class 20); *h* — meteor echo/near-range echo (class 21); *i* — Pedersen ray scattering (class 10); *j* — quasi-bistatic scattering (class 7); *k* — 0.5 hop aspect scattering in the F layer (class 2); *l* — 0.5 hop scattering in the E/F layer with weak aspect sensitivity (class 19)

[Berngardt, 2023], it has been shown that the number of separable classes in radar data varies from 23 to 35 (see Figure 4); 35 was selected for convenience of further interpretation.

Depending on the dataset for network training (three-fold cross-validation), the shape of the found classes may vary slightly: The adjusted Rand index between the model predictions is 0.936–0.967 for the EKB radar, and 0.897–0.937 for the MAGW radar. Using three versions of the trained model, we have constructed and analyzed the ensemble voting classification model. The cases when all the three models did not predict an identical class for the data were combined into a separate class, interpreted as an uncertain prediction result. The proportion of such data is relatively small (10 %), which indicates a high continuity of the results of each network and the possibility of their separate use.

We have analyzed the identified 35 observable signal classes and preliminarily interpreted each of these classes. It has been shown that 19 classes can be interpreted from the physical standpoint (13 types of ionospheric scatter and six types of ground scatter), the rest include high velocities and spectral widths (~1000 m/s). We have found the parameters that most strongly affect the determination of each class, and it has also been shown that the most important parameters are the scat-

tering height and the elevation angle of signal propagation at the scattering point. The spectral width of the signal and its propagation mode are among the least important parameters (see Figure 9).

We have illustrated range-time distributions of these signals at the EKB and MAGW radars (see Figures 6, 7), which demonstrate that many interpreted classes have the expected diurnal variation.

The results of the analysis of MAGW radar data for 2023, not used in model training, have shown the successful performance of the model under more disturbed ionospheric conditions and have revealed limitations of this model: the expected dependence on the quality of calculations of the radio wave propagation trajectory leading to a decrease in classification quality at ranges above 2000 km (see Figures 10, 11).

The proposed method allowed us to automatically classify ~85 % of all data received in the main lobe of the antenna pattern of the EKB and MAGW radars in quiet 2021 and ~75 % of MAGW data for the first half of disturbed 2023. The parameters of the three versions of the trained model are available at [https://github.com/berng/WrappedClassifier/tree/master/v.3.0]. The results of real-time signal processing by the proposed algorithm are available at [http://sdrus.iszf.irk.ru/node/107].

The research was financially supported the Russian Science Foundation (Grant No. 24-22-00436) [https://rscf.ru/project/24-22-00 436/].

## Appendix 1

### ALGORITHM OF ESTIMATING PARAMETER SIGNIFICANCE

To determine the set of the most important parameters for classification, we use the following (greedy) modification of the permutation algorithm:

1. Select the hidden class $C$ that we want to analyze.

2. For the entire dataset $x_{ij}$, predict hidden classes for all available data.

3. Select a part from the dataset $x_{ijC}$ whose predictive class corresponds to the selected class ($x_{ijC}$: $Prediction(x_{ijC})=C$), then work only with it;

4. Create an empty set of important features $F=\varnothing$;

5. If some features were previously identified as important, permutate each corresponding column independently, and add unimportant features to the dataset without permutation:

$$x'_{i,jC} = \begin{cases} i \notin F : x_{i\,jC} \\ i \in F : permutate\left(x_{i\,jC}\right) \end{cases}.$$

6. Calculate the classification quality for the dataset $x'_{i,jC}$, corresponding to this permutation as the proportion of correct predictions of this class in the dataset $x_{ijC}$ of step 3 (with respect to the original full dataset $x_{ij}$ of step 2, we actually calculate the Precision metric with respect to the selected class $C$):

$$Q_F = \frac{Dim\left(Prediction\left(x'_{i\,jC}\right)=C\right)}{Dim\left(x_{i\,jC}\right)}.$$

7. Choose the $l$th column in the dataset $x'_{i\,jC}$ and calculate the permutation quality from it.

$$x^{(l)}_{i,\,jC} = \begin{cases} i \neq l : x'_{i\,jC} \\ i = l : permutate\left(x'_{i\,jC}\right) \end{cases},$$

$$Q_{F,\,l} = \frac{Dim\left(Prediction\left(x^{(l)}_{i\,jC}\right)=C\right)}{Dim\left(x_{i\,jC}\right)}.$$

8. Calculate the reduction of quality:

$$\Delta Q_l = Q_F - Q_{F,\,l}.$$

9. Select the column $l_{opt}$, in which the reduction of quality $\Delta Q_l$ during permutation is the largest,

$$l_{opt} = argmax\left(\Delta Q_l\right).$$

10. If this column matches one of the previously selected columns ($l_{opt} \in F$) or if there is no reduction in quality (the lowest quality found is no worse than the quality obtained in step 6, $\left(\Delta Q_{l_{opt}} \le 0\right)$ — end the search.

11. In all other cases, add the found column to the list of important parameters $\left(F = F \vee l_{opt}\right)$ and return to step 5.

The resulting set of $F$ values is a set of parameters that significantly affect the prediction result of class $C$.

The algorithm is a search algorithm with greedy addition of features and has a stopping criterion that allows us to account only for significant features whose permutation worsens the classification quality.

For greater statistical significance of the method, the quality $Q_F$, $Q_{F,\,l}$ in steps 6, 7 is defined not as the quality $Q$ determined from all elements of the datasets $x'_{i,\,jC}$, $x^{(l)}_{i,\,jC}$, but as the 95 % interval of values of this quality [min($Q$), max($Q$)], using a set of random samples (with repetitions) from these datasets of the same length (bootstrap method). The reduction in quality $\triangle Q_{l_{opt}}$ in this case (step 8) is estimated as the difference between the upper bound of the confidential interval on the parameter being tested (step 7) and the lower bound of the confidential interval on the dataset of the previous step (step 6): $\triangle Q_l = \min\left(Q_F\right) - \max\left(Q_{F,l}\right)$. This algorithm allows us to identify more important parameters with an approximate significance level of 0.05, and reject statistically less significant parameters.

## Appendix 2

### EXPLANATION OF SOME TERMS

- *argmax*($f(x)$) is a function that returns the value of the argument $x$ such that the function $f(x)$ is maximum; for vectors, the number of the coordinate whose value is maximum;

- $\overrightarrow{SoftMax}(\vec{x})$ is a vector-valued function common in machine learning that returns a vector with nonnegative features according to the formula

$$SoftMax_i(\vec{x}) = \frac{\exp(x_i)}{\sum_j \exp(x_j)};$$

- $\overrightarrow{BatchNormalization}(\vec{x})$ — superposition of two linear transformations $\vec{x}$: the first returns a value with zero mean and unit variance at the training dataset, and the second returns coefficients that are fitted during the training to improve the quality of model training;
- The width of the neural network layer is the number of neurons in this layer;
- The fold is a part of a dataset when learning with the cross-validation method. The entire dataset is generally divided into equal parts (folds);
- The scattering regions at propagation hops 0.5, 1, 1.5, etc. are explained in Figure 13. Boundaries of separation are the region of reflection from the ionosphere and the ground scatter region — the places where the vertical radio wave vector component changes sign.

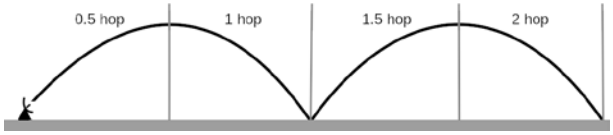


*Figure 13*. Explanation of the hop scattering regions

## REFERENCES


Arnold V. On the function of three variables. *American Mathematical Society Translations*. 1963, pp. 51–54.

Barthes L., Andrè R., Cerisier J.-C., Villain J.-P. Separation of multiple echoes using a high-resolution spectral analysis for SuperDARN HF radars. *Radio Sci*. 1998, vol. 33, no. 4, pp. 1005–1017. DOI: 10.1029/98RS00714.

Berngardt O.I. The first comparative analysis of meteor echo and sporadic scattering identified by a self-learned neural network in EKB and MAGW ISTP SB RAS radar data. *Solar-Terrestrial Physics*. 2022, vol. 8, no. 4, pp. 63–72. DOI: 10.12737/stp-84202206.

Berngardt O.I. Superclustering by finding statistically significant separable groups of optimal gaussian clusters. 2023. DOI: 10.48550/arXiv.2309.02623.

Berngardt, O.I. Minimum number of neurons in fully connected layers of a given neural network (the first approximation). 2024. DOI: 10.48550/arXiv.2405.14147.

Berngardt O.I., Kutelev K.A., Potekhin A.P. SuperDARN scalar radar equations. *Radio Sci*. 2016, vol. 51, no. 10, pp. 1703–1724. DOI: 10.1002/2016RS006081.

Berngardt O.I., Grkovich K.V., Fedorov R.R. Synthesis of Symmetric Sounding Sequences for Ekaterinburg Coherent Decameter Radar. *Radiophysics and Quantum Electronics*. 2020a, vol. 62, no. 11, pp. 721–733. DOI: 10.1007/s11141-020-10018-y.

Berngardt O., Kurkin V., Kushnarev D., Grkovich K., Fedorov R., Orlov A., Harchenko V. ISTP SB RAS decameter radars. *Solar-Terrestrial Physics*. 2020b, vol. 6, no. 2, pp. 63–73. DOI: 10.12737/stp-62202006.

Berngardt O.I., Fedorov R.R., Ponomarenko P., Grkovich K.V. Interferometric calibration and the first elevation observations at EKB ISTP SB RAS radar at 10–12 MHz. *Polar Science*. 2021, vol. 28, p. 100628. DOI: 10.1016/j.polar.2020.100628.

Berngardt O.I., Kusonsky O.A., Poddelsky A.I., Oinats A.V. Self-trained artificial neural network for physical classification of ionospheric radar data. *Adv. Space Res.* 2022, vol. 70, no. 10, pp. 2905–2919. DOI: 10.1016/j.asr.2022.07.054.

Bilitza D., Pezzopane M., Truhlik V., Altadill D., Reinisch B.W., Pignalberi A. The International Reference Ionosphere Model: A review and description of an ionospheric benchmark. *Rev. Geophys*. 2022, vol. 60, no. 4, e2022RG000792. DOI: 10.1029/2022RG000792.

Blanchard G.T., Sundeen S., Baker K.B. Probabilistic identification of high-frequency radar backscatter from the ground and ionosphere based on spectral characteristics. *Radio Sci*. 2009, vol. 44, no. 5. DOI: 10.1029/2009RS004141.

Breiman Leo. Random forests. *Machine Learning*. 2001, vol. 45, no. 1, pp. 5–32. DOI: 10.1023/a:1010933404324.

Burrell A.G., Milan S.E., Perry G.W., Automatically determining the origin direction and propagation mode of high-frequency radar backscatter. *Radio Sci.* 2015. vol. 50, no. 12, pp. 1225–1245. DOI: 10.1002/2015RS005808.

Chisham G., Freeman M.P. A reassessment of SuperDARN meteor echoes from the upper mesosphere and lower thermosphere. *J. Atmos. Solar-Terr. Phys*. 2013, vol. 102, pp. 207–221. DOI: 10.1016/j.jastp.2013.05.018.

Chisham G., Lester M., Milan S.E., Freeman M.P., Bristow W.A., McWilliams K.A., et al. A decade of the Super Dual Auroral Radar Network (SuperDARN): scientific achievements, new techniques and future directions. *Surveys in Geophysics*. 2007, vol. 28, pp. 33–109. DOI: 10.1007/s10712-007-9017-8.

Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. B: *Proc. Second International Conference on Knowledge Discovery and Data Mining*. KDD96. Portland, Oregon: AAAI Press, 1996, pp. 226–231.

Goodfellow I., Bengio Y., Courville A. Deep learning. *Adaptive Computation and Machine Learning*. MIT Press. 2016.

Greenwald R.A., Baker K.B., Dudeney J.R., Pinnock M., Jones T.B., Thomas E.C., et al. Darn/Superdarn: A global view of the dynamics of high-latitude convection. *Space Sci. Rev*. 1995, vol. 71, pp. 761–796. DOI: 10.1007/BF00751350.

Greenwald R.A., Oksavik K., Barnes R., Ruohoniemi M.J., Baker J., Talaat E.R., et al. First radar measurements of ionospheric electric fields at sub-second temporal resolution. *Geophys. Res. Lett.* 2008, vol. 35, no. 3. DOI: 10.1029/ 2007GL032164.

Guo Chuan, Geoff Pleiss, Yu Sun, Weinberger K.Q. On calibration of modern neural networks. 2017. DOI: 10.48550/arXiv.1706.04599.

Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Sci. Rev*. 2020, vol. 37, p. 100270. DOI: 10.1016/j.cosrev. 2020.100270.

Hubert L., Arabie P. Comparing partitions. *J. Classification.* 1985, vol. 2, no. 1, pp. 193–218. DOI: 10.1007/bf01908075.

Ioffe S., Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. DOI: 10.48550/arXiv.1502.03167.

Kolmogoroff A. Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer Berlin Heidelberg. 1933. DOI: 10.1007/978-3-642-49888-6.

Kolmogorov A.N. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR*. 1957, pp. 953–956.

Kong Xing, Liu E., Shi S., Chen F. The implementation of deep clustering for SuperDARN backscatter echoes. *Adv. Space Res.* 2024, vol. 74, no. 1, pp. 243–254. DOI: 10.1016/j.asr.2024.03.039.

Kravtsov Y.A., Namazov S.A. Characteristics of scattering of radio waves from magnetically oriented inhomogeneities of the ionosphere near critical frequency. *Radiotekhnika i elektronika* [J. Communications Technology and Electronics]. 1980, pp. 459–466. [In Russian].

Kunduri B.S.R., Baker J.B.H., Ruohoniemi J.M., Thomas E.G., Shepherd S.G. An examination of SuperDARN backscatter modes using machine learning guided by ray-tracing. *Space Weather*. 2022, vol. 20, no. 9, e2022SW003130. DOI: 10.1029/2022SW 003130.

Lester M., Chapman P.J., Cowley S.W.H., Crooks S.J., Davies J.A., Hamadyk P., et al. Stereo CUTLASS — A new capability for the SuperDARN HF radars. *Ann. Geophys*. 2004, vol. 22, no. 2, pp. 459–473. DOI: 10.5194/angeo-22-459-2004.

Milan S.E., Jones T.B., Robinson T.R., Thomas E.C., Yeoman T.K. Interferometric evidence for the observation of ground backscatter originating behind the CUTLASS coherent HF radars. *Ann. Geophys*. 1997, vol. 15, no. 1, pp. 29–39. DOI: 10.1007/s00585-997-0029-y.

Nishitani N., Ruohoniemi J.M., Lester M., Baker J.B.H., Koustov A.V., Shepherd S.G., et al. Review of the accomplishments of mid-latitude Super Dual Auroral Radar Network (SuperDARN) HF radars. *Progress in Earth and Planetary Science*. 2019, vol. 6, no. 1. DOI: 10.1186/s40645-019-0270-5.

Ponomarenko P.V. Blessing Iserhienrhien и Jean-Pierre St.-Maurice. Morphology and possible origins of near-range oblique HF backscatter at high and midlatitudes. *Radio Sci*. 2016, vol. 51, no. 6, pp. 718–730. DOI: 10.1002/2016 RS006088.

Ponomarenko P., McWilliams K.A. Climatology of HF Propagation Characteristics at Very High Latitudes From SuperDARN Observations. *Radio Sci*. 2023, vol. 58, no. 5, e2023RS007657. DOI: 10.1029/2023RS007657.

Ponomarenko P.V., Waters C.L., Menk F.W. Factors determining spectral width of HF echoes from high latitudes. *Ann. Geophys*. 2007, vol. 25, no. 3, pp. 675–687. DOI: 10.5194/ angeo-25-675-2007.

Ponomarenko P.V., Koustov A.V., St.-Maurice J.-P., Wiid J. Monitoring the F-region peak electron density using HF backscatter interferometry. *Geophys. Res. Lett*. 2011, vol. 38, no. 21. DOI: 10.1029/2011GL049675.

Ribeiro A.J., Ruohoniemi J.M., Baker J.B.H., Clausen L.B.N., de Larquier S., Greenwald R.A. A new approach for identifying ionospheric backscatter in midlatitude SuperDARN HF radar observations. *Radio Sci*. 2011, vol. 46, no. 4. DOI: 10.1029/2011RS004676.

Ribeiro A.J., Ruohoniemi J.M., Ponomarenko P.V., Clausen L.B.N., Baker J.B.H., Greenwald R.A., Oksavik K., de Larquier S. A comparison of SuperDARN ACF fitting methods. *Radio Sci*. 2013, vol. 48, no. 3, pp. 274–282. DOI: 10.1002/rds.20031.

Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Computational and Applied Mathematics*. 1987, vol. 20, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

Rumelhart D.E., Hinton G.E., Williams R.J. *Learning Internal Representations by Error Propagation. B: Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press. 1986. DOI: 10.7551/mitpress/5236.003.0012.

Saxena Amit, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Tiwari A., et al. A review of clustering techniques and developments. *Neurocomputing*. 2017, vol. 267, pp. 664–681. DOI: 10.1016/ j.neucom.2017.06.053.

Schwarz Gideon. Estimating the dimension of a model. *The Annals of Statistics*. 1978, vol. 6, no. 2, pp. 461–464. DOI: 10.1214/aos/1176344136.

Shorten C., Khoshgoftaar T.M. A survey on image data augmentation for deep learning. *J. Big Data*. 2019, vol. 6, no. 1. DOI: 10.1186/s40537-019-0197-0.

Smiti A., Elouedi Z. Fuzzy density based clustering method: Soft DBSCAN-GM. *B: 2016 IEEE 8th International Conference on Intelligent Systems*, 2016, pp. 443–448. DOI: 10.1109/IS.2016.7737459.

Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction. Second*. The MIT Press. 2018.

Vallés-Pérez I., Soria-Olivas E., Martıénez-Sober M., Serrano-López A.J., Vila-Francés J., Gómez-Sanchıés J. Empirical study of the modulus as activation function in computer vision applications. *Engineering Applications of Artificial Intelligence*. 2023, vol. 20, p. 105863. DOI: 10.1016/ j.engappai.2023.105863.

URL: https://github.com/berng/WrappedClassifier/tree/master/v.3.0 (accessed April 15, 2025).

URL: http://sdrus.iszf.irk.ru/node/107 (accessed April 15, 2025).

URL: https://rscf.ru/project/24-22-00436/ (accessed April 15, 2025).

Original Russian version: Berngardt O.I., published in Solnechno-zemnaya fizika. 2025, vol. 11, no. 2, pp. 22–44. DOI: 10.12737/szf-112202503. 

*How to cite this article*

Berngardt O.I. Data-driven approach to mid-latitude coherent scatter radar data classification. *Sol.-Terr. Phys*. 2025, vol. 11, iss. 2, pp. 19–38. DOI: 10.12737/stp-112202503.