Solar-Terrestrial Physics. 2022. Vol. 8. Iss. 4. P. 63–72. DOI: 10.12737/stp-84202206. © 2022 O.I. Berngardt. Published by INFRA-M Academic Publishing House

UDC 004.852, 551.59
DOI: 10.12737/stp-84202206

Received May 04, 2022
Accepted September 19, 2022

THE FIRST COMPARATIVE ANALYSIS OF METEOR ECHO AND SPORADIC SCATTERING IDENTIFIED BY A SELF-LEARNING NEURAL NETWORK IN EKB AND MAGW ISTP SB RAS RADAR DATA

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

Abstract. The paper describes the current version (v.1.1) of the algorithm for automatic classification of signals received by ISTP SB RAS decameter coherent scatter radars. The algorithm is a self-learning neural network that determines the type of scattered signals from the results of physical modeling of radio wave propagation, using radar data and international reference models of the ionosphere and geomagnetic field. According to MAGW and EKB ISTP SB RAS radar data for 2021, the algorithm self-learns to classify scattered signals into initially unknown classes based on physically interpreted parameters of radio wave propagation and data measured by the radar, with 15 frequently observed out of 20 possible hidden classes identified, 14 of which can be interpreted from a physical point of view. To demonstrate the operation of the algorithm, we present the first statistical analysis of observations of signals assigned by the algorithm to classes which we interpret as scattering by meteor trails and scattering with the sporadic E layer respectively. Through a statistical analysis of EKB and MAGW radar data during 2021–2022, we demonstrate the range-altitude characteristics of signals of these types. A correlation is shown between the hourly average numbers of observations of both classes, as well as between the hourly average line-of-sight velocities obtained for both classes. The results obtained make it possible to interpret these classes as a meteor echo and sporadic scattering respectively, and to use radar data to study the interaction between the neutral atmosphere (studied from meteor scattering data) and the lower ionosphere (studied from observations of sporadic scattering). Currently, this classification algorithm works in ISTP SB RAS radars in automatic mode.

Keywords: machine learning, signal classification, coherent scatter radars, meteor echo, sporadic scattering.

INTRODUCTION

The problem of classifying multidimensional experimental data is complex, it is studied intensively in geophysics [Siwei, Ma, 2021]. One of the tools for diagnostics and monitoring of the magnetosphere, ionosphere, and upper atmosphere is the Super Dual Auroral Radar Network (SuperDARN) and similar pulsed decameter coherent scatter radars. Today there are more than 35 such instruments in the world [Nishitani et al., 2019]. A large amount of data provided by these radars is difficult to interpret automatically. Each radar transmits sequences of sounding pulses and receives scattered signals used to study scattering irregularities in the atmosphere and ionosphere. The received radar signals are a mixture of signals formed by various physical scattering mechanisms [Nishitani et al., 2019]. An important problem in interpreting data from these radars is therefore to identify scattered signals of different types. Currently, various methods are used to address the problem [Blanchard et al., 2009; Ribeiro et al., 2011; Lavygin et al., 2020], including statistical and machine learning methods.

Since 2012, ISTP SB RAS has been operating the EKB coherent scatter radar in the Sverdlovsk Region; and since 2020, the MAGW radar in the Magadan Region [Berngardt et al., 2020]. The field of view of the EKB radar is −4°–+48°; that of the MAGW radar is −66°– –16°. The radars operate in a frequency band 8–20 MHz, which increases the number of various scattered signals they receive and extends the radar range to 3500–4500 km due to hop radio wave propagation. On the other hand, this complicates the interpretation of the received signals due to the difficulty in considering paths of these radio waves in the ionosphere. These radars have hardware and software similar to those of SuperDARN radars. At the end of 2020, both radars started regular elevation measurements and were well calibrated. Using the elevation angle of a received signal allows us to estimate its propagation path and to formulate the problem of automatic classification of received signals in terms of their propagation and scattering. SuperDARN radars generally do not use elevation information to automatically classify data.

Such a problem is usually solved by initially classifying data into classes: scattering by the ionosphere, scattering by the Earth surface, scattering by meteors, etc., generally without taking into account the propagation path. The proposed approach addresses the problem from a different angle as construction and training of a scheme that will allow the algorithm to independently divide the data into appropriate classes according to propagation of these signals, and then will enable a researcher to interpret each class from a physical point of view.

The approach combines the learning process (classification) and the process of automatic clustering into a single scheme for determining previously unknown classes in data only from physically interpreted parameters, both measured by a radar and obtained from numerical simulation. This ensures, on the one hand, the automation of the learning process based on a huge amount of available data, and on the other hand, the physical interpretability of the results.

The main idea of the proposed two-stage method is to use clustering at the first stage for all available data — both experimental and numerically simulated. This clustering is mathematical, it is substantiated only by the requirement relatively understandable from physical standpoint that the signals under study are divided into some bounded domains, close in shape to multidimensional ellipsoids, in a multidimensional parameter space. This corresponds to the assumption that the signals having different physical mechanisms of formation should somehow differ from each other in the entire multidimensional space of the parameters measured and obtained from numerical simulation.

After such data clustering, at the second stage, the classifier (artificial neural network) is trained, on the one hand, so that it can use only the parameters well interpreted from physical standpoint, we have selected from all available parameters, and, on the other hand, so that the classes it receives ("hidden classes") will be similar to those from clustering. In terms of machine learning, the approach is close to finding the optimal vector representation of data (optimal embedding), the coordinates of which are the probabilities of signal belonging to each of the hidden classes. The learning method is called Wrapped Classifier with Naive Teacher.

It is generally easiest for researchers to classify radar data visually by analyzing diurnal variations of a signal, taking into account the continuity of signal observation regions and their dynamics in time—distance coordinates. For example, meteor echoes are usually observed as time-fragmented signals at short (up to 450 km) distances. Scattering by the Earth surface in these coordinates has the form of a horseshoe-shaped region, where long distances (ends of the horseshoe) correspond to sunrise and sunset; and close ones (the middle of the horseshoe), to noon [Nishitani et al., 2019]. Scattering by field-aligned irregularities is often observed after sunset in the form of a region extended in distance and time. Clustering into some bounded domains has therefore be chosen as the clustering algorithm, with each measurement day at each radar (experiments) clustered independently of the others. The peculiarity of the algorithm is that the classifier has trained to divide classes in the entire set of available experiments (radars, days) so that the classification it generates is close to splitting of each of these experiments into clusters individually up to permutation of class numbers.

MATHEMATICAL DESCRIPTION OF THE PROBLEM AND CLASSIFIER TRAINING

The proposed algorithm improves on the model [Berngardt et al., 2022] and is its simpler and better interpreted version from physical and mathematical points of view.

The initial dataset used to train the algorithm includes both measured and simulated parameters. It is a set of 15-dimensional vectors $\vec{X}_{e,m}$, where the index e numbers experiments (days and radars), and m is the ordinal number of the vector of observations within one experiment. Coordinates of this vector correspond to different parameters.

Some of the parameters are directly measured by the radar: time, azimuth, radar range, sounding signal frequency, elevation angle of a received signal, Doppler frequency shift, and received signal spectral bandwidth.

The other part is obtained from numerical simulation of radio wave propagation based on estimated signal parameters and experimental set up — time, geographic coordinates of the radar, radar range, elevation angle, azimuth of sounding, and sounding frequency. As a result of the simulation, eight parameters are calculated which characterize the signal propagation path to a scattering point: scattering height, effective scattering height, the number of propagation hops, angle between the propagation path and the geomagnetic field at the scattering point, and angles between the path and the horizon at the scattering point and at each quarter of the radar range to the scattering point. The propagation path is calculated by the geometrical optics method (ray tracing) [Ginzburg, 1970] in the nonmagnetized ionosphere approximation. The output parameters of the radio signal propagation model, which are used for training the neural network, are shown in Figure 1.

To describe the path, the following eight parameters have been chosen: sine of the angle between the ray trajectory and the horizontal direction at a point at 1/4, 2/4, 3/4, and 4/4 of the radar range to a scattering point

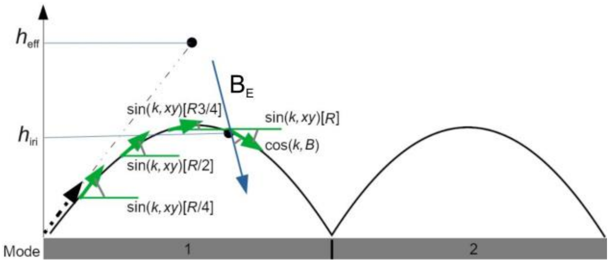

Figure 1. Scheme of calculation of physical parameters determined by numerical simulation. The solid black line is the radio signal propagation path calculated using ray tracing. The blue arrow is the geomagnetic field direction; horizontal green lines mark the horizon direction; green solid arrows, the radio signal propagation direction

(group delay) — the parameters sin(*k*, *xy*)[*R*/4], sin(*k*, *xy*)[*R*/2], sin(*k*, *xy*)[*R*3/4], sin(*k*, *xy*)[*R*] respectively; cosine of the angle between the ray trajectory and the geomagnetic field at a point at the radar range of the signal — parameter cos(*k*, *B*); Mode is the number of reflections from the Earth surface +1; scattering height (the height of the trajectory point at a fixed radar range) with regard for refraction (h_{iri}) and without regard for refraction in the ionosphere (h_{eff}). Additional two parameters that are quite often used for separating signals of different types are the recorded Doppler frequency shift and spectral signal broadening (m/s). The trajectory parameters are selected such that they are close to zero in the following cases: 1) scattering by field-aligned irregularities, cos(*k*, *B*) is close to zero; 2) scattering by the Earth surface (first-hop groundscatter), sin(*k*, *xy*)[*R*/2] is close to zero since near this point reflection occurs from the ionosphere and the radio wave vector at the point is almost horizontal; 3) double scattering by the Earth surface (second-hop groundscatter), sin(*k*, *xy*)[*R*/4], sin(*k*, *xy*)[*R* 3/4] are close to zero since near these points reflection occurs from the ionosphere and the radio wave vector at them is almost horizontal. The sign of sin(*k*, *xy*)[*R*] allows us to determine the direction of the radio wave vector: toward Earth (negative) or from Earth (positive), which also simplifies subsequent interpretation of signals. The Mode parameter also allows us to separate one-hop and two-hop signals, which is important for interpretation. Heights, velocities, and spectral widths provide a more correct interpretation of a scattered signal and are often used in various methods of classifying data from SuperDARN and analogous radars [Ribeiro et al., 2013]. The ten parameters described above are the same as in [Berngardt et al., 2022]. Thus, the classifier model takes into account the regional peculiarities of the formation of scattered signals of different classes only through the international reference models of the ionosphere and geomagnetic field (IRI, IGRF), and observation of signals of a particular class is manifested only in the relative rate of occurrence of such signals, which is one of the physical assumptions of this model. As will be shown below, differences do exist in the rate of occurrence of signals of different classes in different radars. Obviously, this assumption is rather rough, but in the first approximation, as shown in [Berngardt et al., 2022] and in this paper, it yields interpretable results.

Ionospheric refraction is calculated using the International Reference Ionosphere IRI [Bilitza et al., 2014] with parameters recommended by the developers. The geomagnetic field is calculated with the aid of the International Geomagnetic Reference Field IGRF [Thébault, 2015]. As input parameters for calculating the radio wave trajectory we utilized the elevation angle of a received signal (it was assumed to coincide with the elevation angle of radiation); radar beam azimuth; operating frequency of the radar; geographic location of the radar; date; time; radar range (group delay of a signal) from the radar to the scattering point. The required smoothness of the ionosphere in the calculations was provided by its approximation by second-order local B splines. The ionosphere is assumed to be two-dimensional inhomogeneous (in the plane of signal propagation in distance and height). Features of the simulation are discussed in more detail in [Berngardt et al., 2022].

The first stage of training the model is to split all available 15-dimensional data on $\vec{X}_{e,m}$ into clusters — compact areas that resemble ellipsoids in a 15-dimensional space. To do this, the distribution $P_e\left(\vec{X}_{e,m}\right)$ of data values is approximated by a linear combination of twenty 15-dimensional Gaussian distributions $p_{y_{e,n}}$ with unique parameters for each experiment

$$P_e\left(\vec{X}_{e,m}\right) \approx \sum_{y_{e,n}=1}^{N} A_{y_{e,n}} p_{y_{e,n}}\left(\vec{X}_{e,m} \middle| \vec{\Theta}_{y_{e,n}}\right),$$
$$\sum_{y_{e,n}=1}^{N} A_{y_{e,n}} = 1. \tag{1}$$

Here *N*=20, the weights $A_{y_{e,n}}$ and parameters $\vec{\Theta}_{y_{e,n}}$ of the 15-dimensional Gaussian distributions are calculated by fitting experimental data on $\vec{X}_{e,m}$ with the aid of the maximum likelihood method [Dempster et al., 1977]. This problem is solved by searching for unknown parameters $A_{y_{e,n}}$, $\vec{\Theta}_{y_{e,n}}$, which define the division of the $\vec{X}_{e,m}$ dataset, obtained in each of the experiments *e*, into clusters with numbers $y_{e,n}$.

The cluster numbers $y_{e,n}$, to which the corresponding points $\vec{X}_{e,m}$ belong, are the numbers from 1 to *N*, which are obtained automatically by the clustering algorithm.

At the second stage, the final data classifier is trained through the constructed data clustering $y_{e,n}$. The problem of constructing an optimal classifier is to find a function $\vec{g}()$ that receives a vector $\vec{x}_{e,m}$ at the input as a projection of the vector $\vec{X}_{e,m}$ onto a 10-dimensional subspace of the selected physical parameters convenient for subsequent interpretation.

Thus, the classification that we want to obtain should, on the one hand, repeat the clustering, derived at the previous stage, quite well; on the other, use only those parameters for this purpose that we will later be able to confidently interpret from a physical point of view. This solution is found as an approximate solution of the problem

$$\vec{f}_e\left(\vec{g}\left(\overrightarrow{PF}\left(\vec{x}_{e,m}\right)\right)\right) \approx \overrightarrow{OHE}\left(y_{e,n}\right). \tag{2}$$

Here, the wrap function $\vec{f}_e\left(\vec{z}\right)$ is unique for each experiment *e* and has the form

$$\vec{f}_e\left(\vec{g}\right) = \overrightarrow{soft\max}\left(\sum_{l=1}^{K} C_l \left(\sum_{k=1}^{K} b_{k,l,e} g_k\right)\right),$$
$$b_{k,l,e}, C_{l,e} \geq 0,$$
$$soft\max_j\left(\vec{z}\right) = \frac{e^{z_j}}{\sum_i e^{z_j}}. \tag{3}$$

All unknown parameters (C_l, $b_{k,l,e}$ and the function parameters $\vec{g}()$) are found by selecting experimental data as solutions of (2). Dimensions of the vectors $\vec{f}_e, \vec{g}$ are the number of clusters (N=20) and hidden classes (K=20) taken to be equal. Summation in the formula is carried out using hidden classes k to ensure the correct permutation of class numbers within each individual experiment e (this process is discussed in more detail in [Berngardt et al., 2022]). This permutation is unique for each experiment e, and the summation by hidden classes l is carried out for a possible combination of similar hidden classes and actually tries to make a prediction of the probability of a class predicted by Naive Teacher from a linear combination of probabilities of hidden classes yielded by the wrapped classifier.

The problem was solved by one of the widely used upgrades of the gradient descent method — the moment method ADAM [Goodfellow et al., 2016]. As an optimality condition, the classical approach to solving classification problems has been utilized — minimizing the Weighted Cross Entropy *WCE* [Goodfellow et al., 2016]

$$WCE = -\sum_{j,k} W_k Y_{right,\,j,k} \log\left(Y_{left,\,j,k}\right), \tag{4}$$

where W_k are balancing weights inversely proportional to the number of data in class k; the index j numbers objects in the learning dataset: pairs (experiment number, object number in the experiment), $Y_{right,\ j,\ k}$, $Y_{leftt,\ j,\ k}$ are the right and left sides of Equation (2) respectively.

There are two criteria of optimality of the resulting classifier. The first, mathematical, is that the classification of each experiment performed by the wrapped classifier up to permutation of class numbers most closely corresponds to the division of this experiment made by Naive Teacher. Fulfillment of this criterion is the result of training the neural network, and deviations from such an optimal classification are used in training as a function that we want to minimize by training the neural network (also known as loss function, Expression (4)). The quality of the result is evaluated numerically by the quality metric (the so-called internal quality assessment AURPC), which is described below.

The second criterion, physical, is that the data on each individual class obtained from the classification can be interpreted from a physical point of view as signals received due to a specific scattering mechanism. Fulfillment of this criterion is checked by statistically analyzing the data by an expert after being classified by a trained classifier (the so-called external quality assessment). Optimization of this criterion consists in selecting the structure (architecture) of the neural network and input parameters to achieve the required quality of interpretation. An example of such an expert analysis in a similar problem is given in [Berngardt et al., 2022].

As Naive Teacher we have chosen the probabilistic Gaussian Mixture method [Vander Plas, 2016] presenting data as random, having a probability distribution equal to a linear combination of multidimensional normal distributions whose parameters (weight, mean, and covariance matrix) are determined automatically during data analysis, and the number of these normal distributions is fixed and set by the researcher. The choice of the number of classes is governed by the following. The Gaussian Mixture model is characterized by the fact that it divides the data exactly into the number of classes given by N. The scheme of constructing the wrapped classifier places only upper limit on the number of hidden classes K. If possible, the real number of hidden classes with a non-zero number of elements may become smaller after training, i.e. the algorithm can combine close classes if this does not degrade the accuracy of the approximation. Thus, for N and K it is convenient to choose a sufficiently large number that exceeds the expected number of different types of signals in advance. Radar studies usually identify about a ten of such types qualitatively distinguishable by the range or spectral characteristics of signal: scattering of one-hop and two-hop signals by the Earth surface, meteor scattering, scattering by field-aligned irregularities of the E layer (two-stream and gradient-drift), scattering by field-aligned irregularities of the F layer, scattering by sporadic layers, mesospheric echo, etc. The number of classes N, K was therefore chosen to be twice the expected number of different classes such that the wrapped classifier could correspondingly reduce the number of detected classes if necessary. This is also related to the simplicity of the clusterer employed (Gaussian Mixture), which aims to roughen and simplify the clustering in the hope that the statistical nature of its errors in each experiment will allow the formation of hidden classes more resistant to such errors. As further analysis has shown, the actual number of interpreted hidden classes with a significant number of observations, in fact, ranges from 15 to 18.

The function $\overrightarrow{OHE}\left(y_{e,n}\right)$ has only one non-zero coordinate corresponding to the cluster number $y_{e,n}$,

$$OHE_m\left(y_{e,\,n}\right) = \delta_{y_{e,\,n},m}. \tag{5}$$

The classifier function $\vec{g}()$ is approximated by a three-layer fully connected neural network with about 30 thousand free parameters, 133 neurons, and *ReLU* activation functions in each hidden layer, and the *SoftMax* activation function in the output layer. Thus, the function is normalized so that

$$\begin{gathered}\sum_{i=1}^{K} g_i = 1, \\ g_i \geq 0.\end{gathered} \tag{6}$$

This allows us to interpret the output of the function $\vec{g}()$ as probabilities that $\vec{x}_{e,\,m}$ belongs to the hidden classes $\{1..K\}$ defined by the coordinates g_i. To improve the quality of training, the layers are separated from each other by batch normalization layers. The network architecture is illustrated in Figure 2, *b*.

As testing has shown, the three-layer network $\vec{g}()$ is enough for qualitatively solving the problem, more or less deep networks do not improve the quality.

To improve the quality of the classifier, the dimension

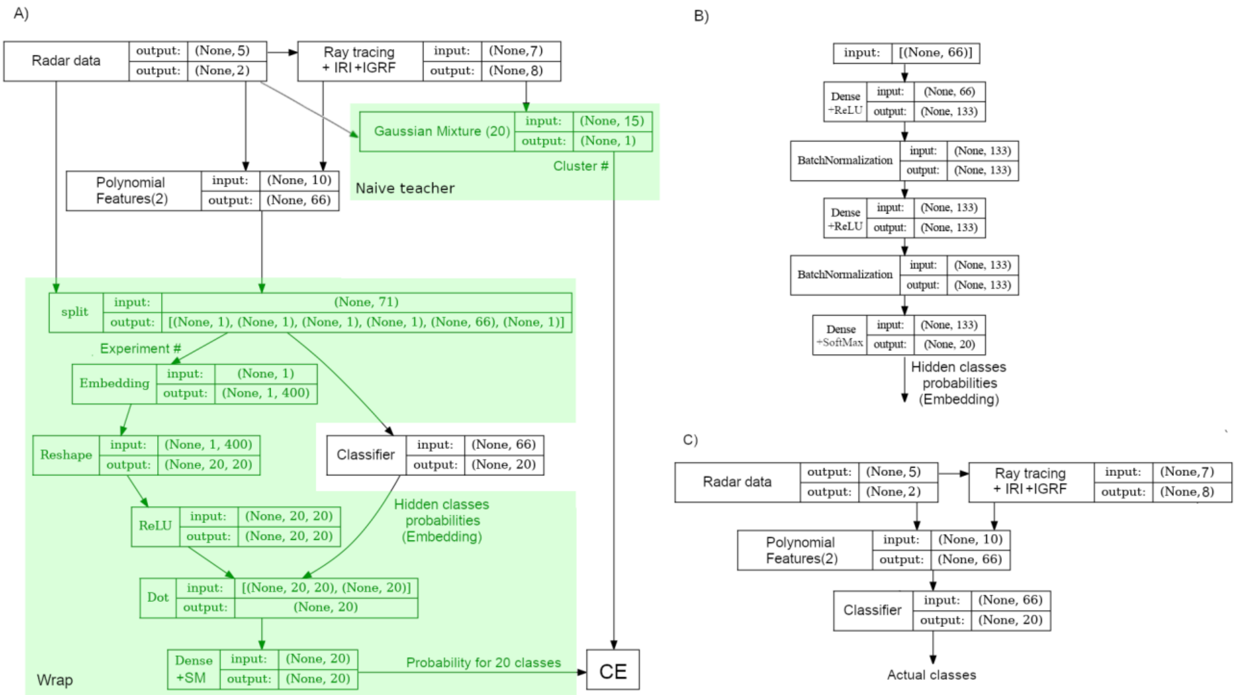

Figure 2. Architecture of the Wrapped Classifier with Naive Teacher and vector dimensions: network architecture in learn mode (*a*); classifier architecture (*b*); network architecture in processing mode (*c*). "None" stands for the number of records in the dataset

of the input data was increased by the method of feature space. The efficiency of its use in a similar problem is shown in [Berngardt et al., 2022]. It is assumed that in this variant its use will also be effective. It was chosen in the form of a polynomial transform $\overrightarrow{PF}\left(\vec{x}\right)$

$$\overrightarrow{PF}\left(\vec{x}\right)=\left(1,\, x_0,\, x_1,\, ...,\, x_{N-1},\, x_0^2,\, x_0x_1,\right.$$
$$\left. x_0x_2,\, ...,\, x_{N-2}x_{N-1},\, x_{N-1}^2\right). \tag{7}$$

This transform increases the initial dimension N of the input vector $\vec{x}_{e,m}$ to $N+N(N+1)/2+1$ and thereby simplifies the solution of the data classification problem. The use of this transform was facilitated by the fact that the weighted sum of squared Doppler shift and signal bandwidth is already widely utilized as a good criterion for identifying groundscatter signals [Blanchard et al., 2009]. Thus, taking into account squares of the input features and cross products of the input features can increase the classification efficiency. A characteristic benefit of using the feature space *PF* at the input of the neural network is that the neural network during learning eliminates the coordinates that are not essential for the optimal solution, leaving only the essential ones. Increasing the space dimension will allow us to employ simpler networks for the classification, but by increasing the number of free coefficients of the neural network and by increasing the size of the dataset required for its training. The efficiency of applying the polynomial feature space to this problem has been demonstrated in [Berngardt et al., 2022] when considering the previous version of the wrapped classifier. Omitting *PF* in the clusterer was intended to roughen the clusterer and introduce more errors into it in each individual experiment in order to make the hidden classes of the classifier more resistant to errors due to the statistical nature of learning.

A large number of unknown values of $C_{l,e}$, $b_{k,\,l,e}$ in (3) makes it possible to optimally approximate an arbitrary permutation of $y_{e,n}$ by the coordinates g_i of the function $\vec{g}()$ for each experiment e separately, so that $\vec{g}$ does not depend on e and, as expected, is determined solely by the shape of signal clusters in the N-dimensional space of their parameters.

The main difference between this model and the previous version described in [Berngardt et al., 2022] is the model shape $\vec{f}_e\left(\vec{g}\right)$ and the requirement $C_{l,e}\geq 0$ in (3). This makes it possible to interpret the function $\vec{g}$ more confidently as the probability of hidden classes from which we can obtain cluster probabilities $\vec{f}_e$ through a simple linear transform with nonnegative coefficients. Moreover, this modification made it possible to significantly simplify the function model $\vec{g}$, to obtain better separation results, and to conduct its better training due to a smaller number of free coefficients in the model (30 thousand instead of 80 thousand free parameters).

Thus, the proposed classification scheme consists of two consecutive neural networks (Figure 2, *a*) one of which (classifier) provides an optimal representation of data in the form of a 20-dimensional vector, which further we interpret as probabilities of hidden classes. The second network (Wrap) converts the probabilities of hidden classes into cluster numbers of a dataset clustered by Naive Teacher. The classifier network and the wrap network are jointly trained to best match their output to clustering of this dataset by Naive Teacher.

The clusterer independently clusters data from experiment to experiment (from day to day, from radar to radar): the same types of scattering in different experiments may have different cluster numbers. Therefore, when interpreting the clustering results, we should renumber them. We cannot completely trust this teacher and be sure that its clustering is correct. That's why we call this teacher naive. Relying on the results of its clustering, we build and train an optimal classifier that performs optimal classification based on physics of radio wave propagation and has greater capabilities for generalization and interpretation than Naive Teacher.

The number of neurons in each layer of the classifier (133) was chosen in accordance with the Kolmogorov—Arnold theorem [Arnold, 1963] for the most optimal representation of the input data. The classifier has about 30 000 trained parameters, and the wrap has about 153 000 ones, which is significantly less than in the previous version of the classifier [Berngardt et al., 2022].

To speed up network training, each stage was performed sequentially, with the obtained datasets stored in the repository. To interpret new data points, only a trained classifier and a radio wave propagation model are required. The wrap network and the clusterer are not used in the forecast (Figure 2, *c*).

The classifier model receives ten input parameters at the input, only two of which (Doppler shift and spectrum width) were directly measured by the radar, and the remaining eight were obtained from numerical simulation based on the measured received signal parameters. These parameters are shown in Figure 1; they have been discussed in more detail in [Berngardt et al., 2022].

This architecture (classifier+wrap) allows us to automatically renumber the cluster numbers for each experiment independently and increases the accuracy of reconstructing the classifier function $\vec{g}$. During training, as a loss function the weighted cross-entropy is utilized, where the weights are the inverse element number in the cluster encountered in the training dataset. This enables us to automatically balance the dataset and thereby improve the quality of fitting.

AURPC (AUC-RP) is used as a prediction quality metric since it works quite correctly in case of possible class imbalance. To detect overtraining, we use the early stopping method of training after more than twenty unsuccessful training epochs based on the AURPC metric in the validation dataset. The gradient descent method is used with the ADAM optimizer with a packet size of 32. For the training, the dataset from the two radars for January–September 2021 (~3 million records) was divided into training, validation, and test datasets in the ratio 64:16:20 %. The learning process corresponds to that described in [Berngardt et al., 2022].

Figure 3 illustrates the division of experimental data into classes through the example of the data for April 2022. The neural network assigned the class numbers independently during training, but in what follows we interpret only classes 2 and 13.

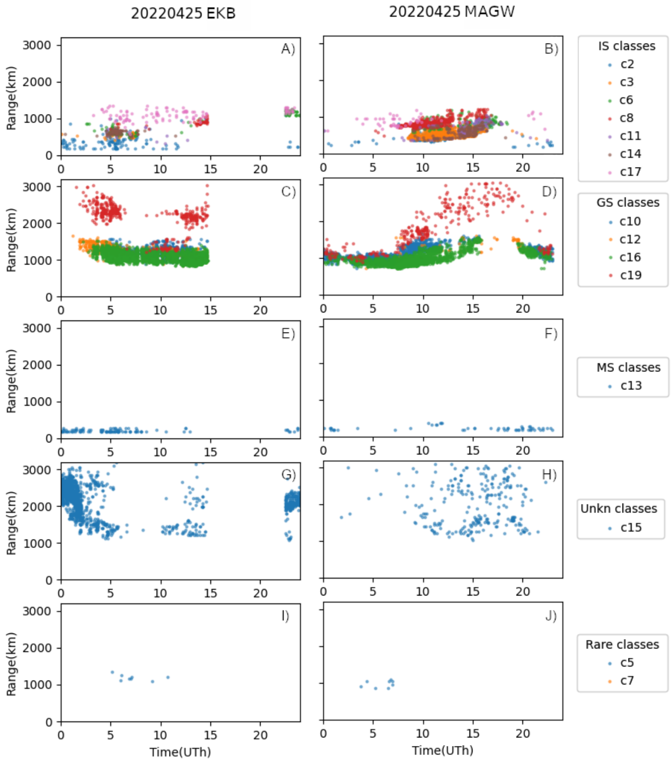

Figure 3. Example of operation of the algorithm using ISTP SB RAS data from EKB (left) and MAGW (right) radars on April 25, 2022. From top to bottom: ionospheric scattering classes (*a*, *b*); groundscattering classes (*c*, *d*); meteor scattering classes (*e*, *f*); unidentified classes (*g*, *h*); rarely observed classes (*i*, *j*)

Figure 4 shows the distribution of the number of detected signals in different classes. The actual number of hidden classes with a non-zero number of objects is seen to be 18, two of which, 9 and 18, are poorly defined data (data distributed approximately evenly in the range-time diagram, we usually do not analyze later on), and several classes that are very rare (1, 5, and 7, with classes 1, 5 found only for the EKB radar). Class 15 is uninterpreted in terms of propagation — the average scattering heights h_{iri} exceed the height of the F2-layer maximum, which suggests that in this case the radio signal propagation path is calculated incorrectly: either the ionosphere in these cases does not correspond to the model (this is possible given the existing accuracy of the IRI model) or the angle of signal arrival is estimated incorrectly (this is also possible given the difficulties in calculating the elevation angle of signal arrival). Thus, the total number of interpreted classes is 14, and their total percentage in the data from various radars is from 50 to 60 %, which suggests that about half of the recorded radar data can be automatically interpreted in terms of model propagation of radio signals by the proposed method. This is in good agreement with the qualitative expectations and results of the previous version of the algorithm [Berngardt et al., 2022]. Figure 4 indicates that there are radar-dependent features in the data. In particular, classes 1, 5 are observed only on EKB. There is also a clear imbalance in the observation of the same classes on different radars (for example, 8, 10, 16), which might have been caused by both the technical characteristics of the radars (different noise levels and slightly different antenna gain coefficients) and regional features of the ionosphere (the MAGW field of view is more deviated from the north than the EKB one).

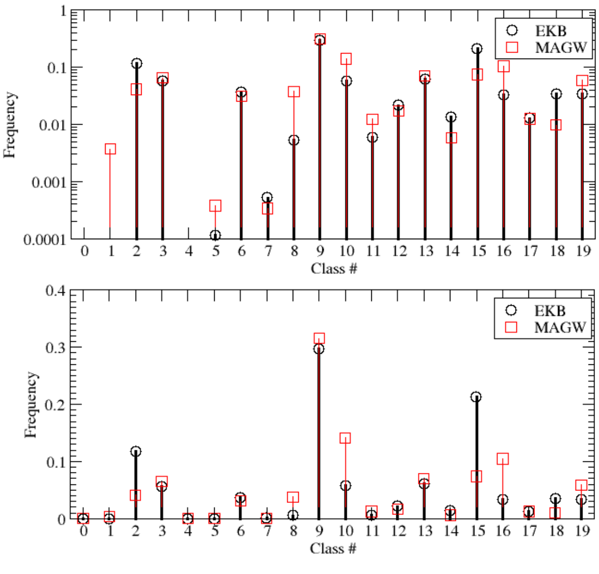

Figure 4. Frequency distribution of observed signal classes in the EKB and MAGW ISTP SB RAS data during training on logarithmic (top) and linear (bottom) scales

FIRST PRELIMINARY RESULTS AND THEIR INTERPRETATION

Further study is devoted to class 13, interpreted as meteor scattering, and to class 2, interpreted as a result of sporadic scattering in the ionosphere. Detailed analysis of the remaining classes defined by the neural network is beyond the scope of this work. To substantiate this interpretation of the classes identified independently by the proposed neural network, Figure 5 shows the altitude-range distributions of EKB and MAGW radar data belonging to these classes for 2021–2022.

The distribution of the heights of the appearance of class 13 on both radars is seen to be within 50–150 km with a maximum near 80–90 km, which corresponds closely to the scattering by meteor trails [Chisham, Freeman, 2013; Fedorov, Berngardt, 2021]. The range is from the minimum radar range of 180 km to ~400 km with a strong decrease in the occurrence rate with distance, which also corresponds closely to the statistics of observations of meteor scattering found from the data by another method [Fedorov, Berngardt, 2021].

The altitude-range distribution of class 2 differs from that of class 13. The range of distribution heights has expanded from 0 to 250 km, near ranges and low altitudes most likely corresponding to groundscatter with high elevation angles (it is not yet possible to verify owing to the problem of phase uncertainty in interferometric observations of radars of this type [Chisham, Freeman, 2013]), which at such short ranges corresponds to scattering by lower ionospheric layers with altitudes below 200 km. This can be interpreted as scattering by the sporadic layer and subsequent scattering by the Earth surface.

This mechanism is supported by an increase in the effective scattering height below 500 km, which may be due to underestimation of sporadic layers by the IRI model. The second portion of the scattered signals is concentrated near 100 km and is observed mainly at 300–500 km. This suggests the possibility of interpreting the class 2 signals as a mixture of direct scattering by the sporadic layer and scattering by the Earth surface after scattering by the sporadic layer.

The interpretation of class 2 as suitable for the diagnosis of the sporadic layer is also supported by line-of-sight velocities close to neutral wind velocities, diagnosed from the Doppler shift of the frequency of the received signal. Figure 3, *a*, *b*, *e*, *f* shows that classes 13 and 2 are very fragmentary (sporadic) in time and space (marked with blue colors in Figure 3, *a*, *b*, *e*, *f*); therefore, for further comparative statistical analysis it is convenient to use their 1-hour averaged parameters.

Figure 6 exemplifies the behavior of the line-of-sight velocity (top) and the number of scattered signals (bottom) averaged over 1 h. Black color marks class 13 (meteor echo); red color, class 2 (sporadic scattering). These two classes are seen to correlate well in terms of the rates and periods of occurrence of signals. In terms of formation mechanisms, class 2 can be interpreted as scattering by sporadic layers, whose formation mechanism in the lower ionosphere is one of the controversial issues in its research, may be associated with meteors [Malhotra et al., 2008], and requires high temporal and spatial resolution for a reliable study. Such a high temporal resolution (from units of seconds to minutes) is

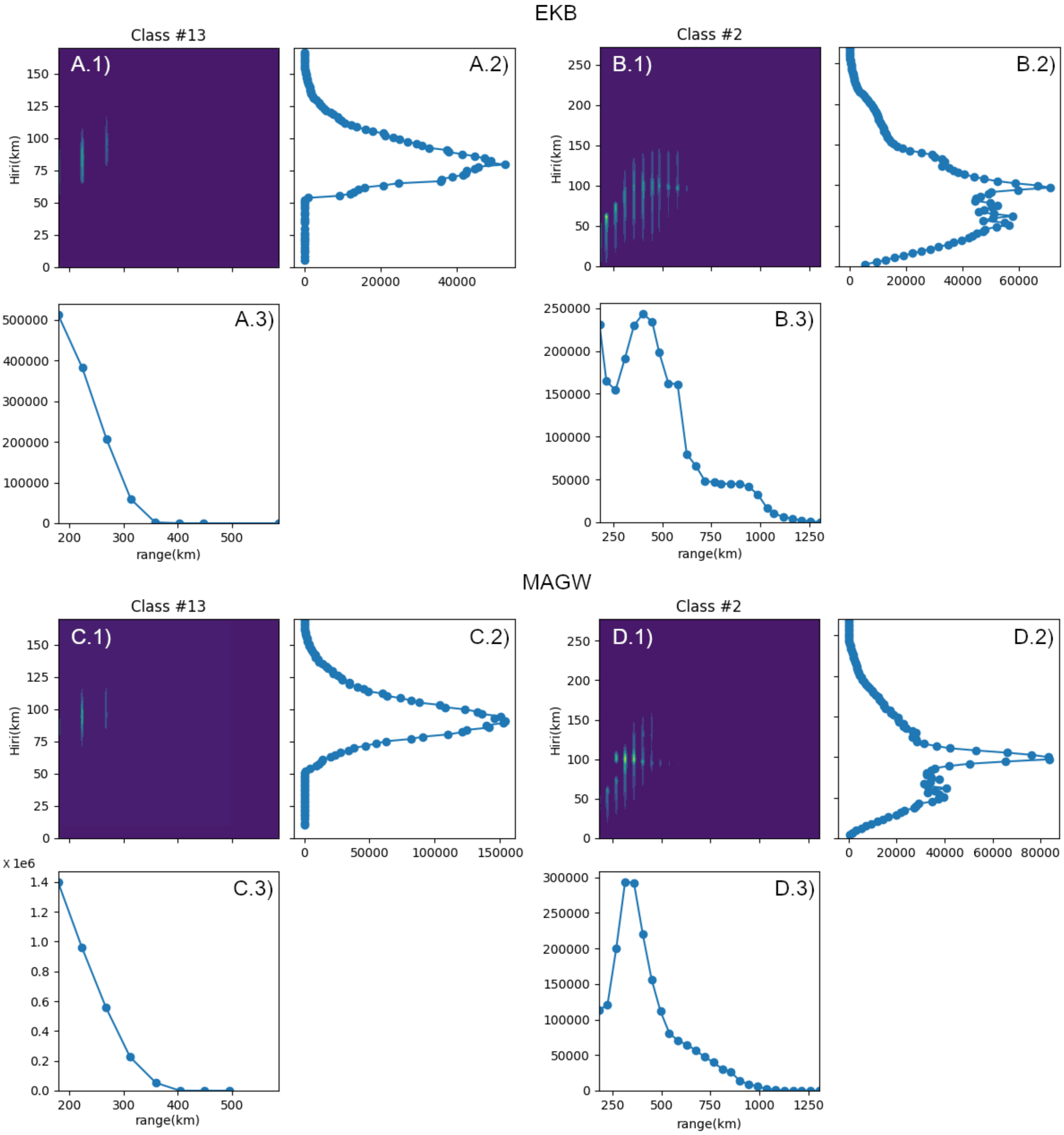

Figure 5. Comparison of altitude-range distributions of classes 13 and 2 for January 2021 – March 2022. Altitude-range distribution of the rate of observation of traces (*a1–d1*); rate of occurrence of scatterers depending on the height calculated from signal propagation path (*a2–d2*); signal distribution as a function of radar range (*a3–d3*)

provided by SuperDARN radars and similar EKB and MAGW radars of ISTP SB RAS. Statistical confirmation of this correlation for the full dataset on 2021–2022 is illustrated in Figure 7. Linear (Pearson) and rank (Spearman) correlations are shown in Table.

The correlations are seen to range from 0.52 to 0.76, which confirms the significant positive correlation shown in Figure 7 between hourly average signal parameters in the meteor and sporadic echo classes. In all cases, the calculated significance level of the absence of correlation (p-value) is lower than 10^{-6} (omitted in Table), which indicates the significance of the correlation. A slight excess of the Spearman correlation over the Pearson correlation (~10 %) indicates the presence of not only a strong linear correlation, but possibly a weak nonlinear correlation. The different slope in panels of Figure 7, A, C for the number of observations of sporadic echo on the radars can be associated with both the features of the sounding geometry (the MAGW radar field of view is more inclined to the west than the EKB radar one) and regional features of the ionosphere, and requires additional analysis. The method of classifying signals received by radars described in the work and the preliminary analysis carried out in the work make it possible in the future to use a comparative analysis of these two classes for diagnosing the neutral ionosphere coupling in the lower ionosphere.

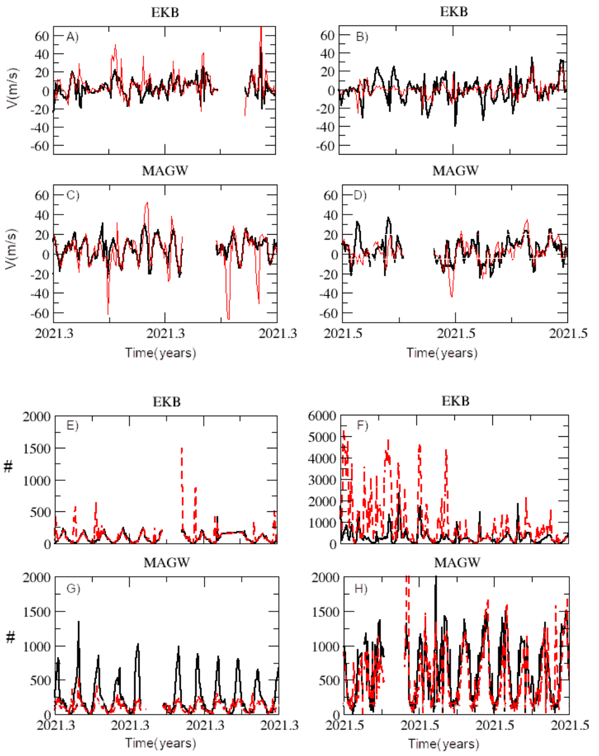

Figure 6. Line-of-sight velocities (*a*–*d*) measured with the EKB and MAGW radars and the number of detected signals (*e*–*h*) in spring (*a*, *b*, *e*, *g*) and summer (*b*, *d*, *f*, *h*) in 2021 for classes 13 (black line) and 2 (red line)

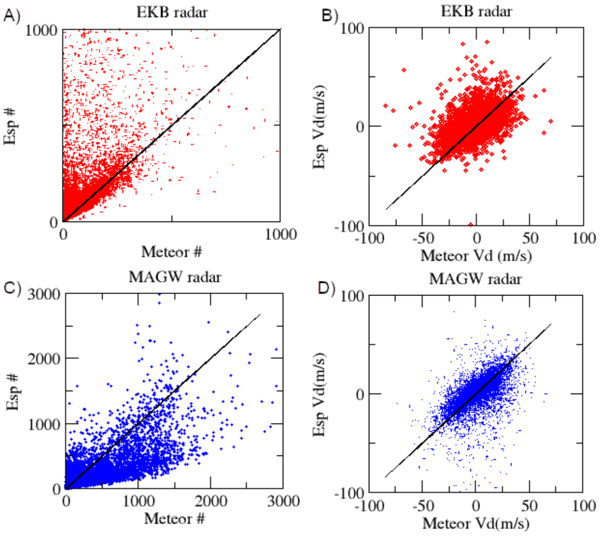

Figure 7. Relationship between the number of scattered signals (*a*, *c*) and hourly average velocities (*b*, *d*) in classes 13 (meteor echo) and 2 (sporadic scattering) as recorded by the EKB (*a*, *b*) and MAGW (*c*, *d*) radars

Pearson and Spearman correlations between meteor and sporadic echoes according to EKB and MAGW data

Correlation coefficient	EKB	MAGW
Pearson by the number of observations	0.636	0.718
Spearman by the number of observations	0.716	0.761
Pearson by the Doppler velocity	0.527	0.564
Spearman by the Doppler velocity	0.591	0.652

CONCLUSIONS

The paper describes the current version of the algorithm for automatic classification of signals (v.1.1) received by ISTP SB RAS decameter coherent scatter radars. The algorithm, which we called the Wrapped Classifier with Naive Teacher, is a self-learning neural network that determines the type of scattered signals from the results of sounding and numerical simulation of radio wave propagation. Radio signal propagation is calculated by the geometrical optics (ray tracing) method using radar data and the international reference models of the ionosphere (IRI) and geomagnetic field (IGRF). The model self-learns using the results of preliminary classification of data by Naive Teacher. The naive teacher algorithm is clustering of data by a statistical model of a mixture of multidimensional normal distributions. It splits the data into clusters, which are then used to search for hidden classes in the data according to those physical parameters (obtained from radar data, as well as as from physical simulation of radio wave propagation) that can then be effectively interpreted. The neural network self-learned using MAGW and EKB ISTP SB RAS data for 2021.

The resulting hidden classes found by the classifier can be physically interpreted by statistically analyzing the distribution of physically interpreted parameters of signals belonging to each class. This classifier is an upgrade of the classifier proposed earlier in [Berngardt et al., 2022], and has a better quality of division into classes, as well as a simpler architecture with fewer free parameters found as a result of self-learning.

To demonstrate the operation of the classification algorithm, the first statistical analysis of observations of signals assigned by the algorithm to classes 13 and 2, interpreted by us as scattering by meteor trails and scattering with the sporadic E layer respectively, has been carried out. The comprehensive analysis of EKB and MAGW statistical data for 2021–2022 has revealed the range-altitude characteristics of signals of these types. The correlation has been shown between the hourly average numbers of observations of both classes, as well as between the hourly average line-of-sight velocities obtained in the data of both classes. We believe that the results will allow us to use radar data for studying the interaction between the neutral atmosphere (studied from meteor scattering data) and the lower ionosphere (studied from observations of the sporadic E layer) at spatially close points with high temporal resolution. Currently, the data classification algorithm works in the ISTP SB RAS radars in automatic mode [http://sdrus.iszf.irk.ru/node/95]; the code of the trained neural network of the classifier is available at [https://github.com/berng/WrappedClassifier].

The ISTP SB RAS EKB radar is included in the Shared Equipment Center “Angara” [http://ckp-rf.ru/ckp/3056]. The radars were operated with the financial support of the Ministry of Science and Higher Education of the Russian Federation (subsidy No. 075-GZ/S3569/278). The EKB and MAGW ISTP

SB RAS data are available at [http://sdrus.iszf.irk.ru/ekb/page_example/simple]. The model was trained partially using the equipment of the Public Access Center Bioinformatics of the Federal Research Center Institute of Cytology and Genetics SB RAS (ICG SB RAS). I am grateful to I.S. Petrushin (Irkutsk State University) for useful discussions. The work was financially supported by RFBR-CNRS Grant No. 21-55-15012.

REFERENCES

Arnold V.I. On functions of three variables. *American Mathematical Society Translations. Ser. 2*. 1963, vol. 28, pp. 51–54. (*Translation of Dokl. Akad. Nauk SSSR*. 1957, vol. 114, iss. 4, pp. 679–681).

Berngardt O.I., Kurkin V.I., Kushnarev D.S., Grkovich K.V., Fedorov R.R., Orlov A.I., Kharchenko V.V. ISTP SB RAS decameter radars. *Solar-Terr. Phys*. 2020, vol. 6, iss. 2, pp. 63–73. DOI: 10.12737/stp-62202006.

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, iss. 10, pp. 2905–2919. DOI: 10.1016/j.asr.2022.07.054. (In print).

Bilitza D., Altadill D., Zhang Y., Mertens C., Truhlik V., Richards Ph., et al. The International Reference Ionosphere 2012 — a model of international collaboration. *J. Space Weather Space Climate*. 2014, vol. 4, id. A07, 12 p. DOI: 10.1051/swsc/2014004.

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, iss. 5, RS5012. DOI: 10.1029/2009rs004141.

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.

Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society: Ser. B (Methodological)*. 1977, vol. 39, no. 1, pp. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

Fedorov R.R., Berngardt O.I. Monitoring observations of meteor echo at the EKB ISTP SB RAS radar: algorithms, validation, statistics. *Solar-Terr. Phys.* 2021, vol. 7, no. 1, pp. 47–58. DOI: 10.12737/stp-71202107.

Ginzburg V.L. *The Propagation of Electromagnetic Waves in Plasmas*. Pergamon Press, 1970, 615 p.

Goodfellow I., Bengio Y., Courville A. *Deep Learning (Adaptive Computation and Machine Learning Ser)*. MIT Press, 2016, 800 p.

Lavygin I.A., Berngardt O.I., Lebedev V.P., Grkovich K.V. Identifying ground scatter and ionospheric scatter signals by using their fine structure at Ekaterinburg decametre coherent radar. *IET Radar, Sonar and Navigation*. 2020, vol. 14, iss. 1, pp. 167–176. DOI: 10.1049/iet-rsn.2019.0192.

Malhotra A., Mathews J.D., Urbina J. Effect of meteor ionization on sporadic-E observed at Jicamarca. *Geophys. Res. Lett.* 2008, vol. 35, iss. 15. DOI: 10.1029/2008GL034661.

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 Sci*. 2019, vol. 6, iss. 1, p. 27. DOI: 10.1186/s40645-019-0270-5.

Ribeiro A.J., Ruohoniemi J.M., Baker J.B.H., Clausen S., de Larquier S., Greenwald R.A. A new approach for identifying ionospheric backscatter in midlatitude SuperDARN HF radar observations. *Radio Sci*. 2011, vol. 46, iss. 4, RS4011. DOI: 10.1029/2011RS004676.

Ribeiro A.J., Ruohoniemi J.M. Ponomarenko P.V., Clausen L.B.N., Baker J.B.H., Greenwald R.A., et al. A comparison of SuperDARN ACF fitting methods. *Radio Sci*. 2013, vol. 48, iss. 3, pp. 274–282. DOI: 1002/rds.20031.

Siwei Yu., Ma J. Deep learning for geophysics: current and future trends. *Rev. Geophys*. 2021, vol. 59, iss. 3, e2021RG000742. DOI: 10.1029/2021rg000742.

Thébault E., Finlay C.C., Beggan C.D., Alken P., Aubert J., Barrois O., et al. International Geomagnetic Reference Field: the 12th generation. *Earth, Planets and Space*. 2015, vol. 67, iss. 1, p. 79. DOI: 10.1186/s40623-015-0228-9.

Vander Plas J. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc., 2016, 548 p.

URL: http://sdrus.iszf.irk.ru/node/95 (accessed October 12, 2022).

URL: https://github.com/berng/WrappedClassifier (accessed October 12, 2022).

URL: http://ckp-rf.ru/ckp/3056 (accessed October 12, 2022).

URL: http://sdrus.iszf.irk.ru/ekb/page_example/simple (accessed October 12, 2022).

Original Russian version: O.I. Berngardt, published in Solnechno-zemnaya fizika. 2022. Vol. 8. Iss. 4. P. 66–76. DOI: 10.12737/szf-84202206. © 2022 INFRA-M Academic Publishing House (Nauchno-Izdatelskii Tsentr INFRA-M)

How to cite this article

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. Iss. 4. P. 63–72. DOI: 10.12737/stp-84202206.