Solar-Terrestrial Physics. 2026, vol. 12, iss.1, pp. 11–17. DOI: 10.12737/stp-121202602. © 2026 E.D. Tulnikov, V.S. Makhmutov, M.V. Philippov. Published by INFRA-M Academic Publishing House

DOI: 10.12737/stp-121202602
EDN: RJXVBM

Received August 04, 2025
Accepted November 11, 2025

APPLICATION OF MACHINE LEARNING METHODS TO DETERMINE SPECTRAL CHARACTERISTICS OF RADIATION IN THE "SOLNTSE-TERAHERTZ" EXPERIMENT

E.D. Tulnikov
P.N. Lebedev Physical Institute RAS, Moscow, Russia, tulnikov.ed@yandex.ru

V.S. Makhmutov
P.N. Lebedev Physical Institute RAS, Moscow, Russia, mahmutovvs@lebedev.ru
Moscow Institute of Physics and Technology, Moscow, Russia

M.V. Philippov
P.N. Lebedev Physical Institute RAS, Moscow, Russia, filippovmv@lebedev.ru

Abstract. This paper explores the possibility of using machine learning methods for analyzing observations from the "Solntse-Terahertz" scientific equipment, developed at the Lebedev Physical Institute for installation on the Russian segment of the ISS. The scientific equipment consists of eight detectors with target frequencies ranging from 0.4 to 12.0 THz. One of the primary goals of the experiment is to study solar flares whose spectra in this range often have a U-shaped form. The primary focus in determining the spectral parameters is on identifying spectral indices of the decaying and rising parts of the spectrum, as well as the position of the turnover point. The algorithms were trained using model data on the intensity of radiation passing through optical paths of the instrument. The data was obtained by numerical integration methods. The analysis has shown that the Stacking algorithm demonstrates the highest accuracy in determining the spectral parameters and can be integrated into the data processing system for future experiment on the ISS, enabling the automatic preliminary restoration of solar flare spectrum parameters.

Keywords: Sun, flare, submillimeter radiation, machine learning.

INTRODUCTION

The Laboratory of Solar and Cosmic Ray Physics of the Lebedev Physical Institute (Dolgoprudny Scientific Station (DSS) of LPI) has created scientific equipment (SE) "Solntse-Terahertz" [Kalinin et al., 2021], designed to implement a space experiment at the Russian segment of the ISS. The purpose of the experiment is to measure solar radiation and flares in the terahertz range (0.4–12 THz) in order to study the nature of solar activity and the mechanisms of charged particle acceleration on the Sun [Wedemeyer et al., 2016; Kaufmann et al., 2004; Krucker et al., 2013] and other astrophysical objects (such as ultraluminous infrared galaxies and blazars) [Tulnikov et al., 2025].

The equipment includes eight detectors intended to detect radiation at frequencies of 0.4, 0.8, 1.0, 3.0, 5.0, 7.0, 10.0, and 12.0 THz. The optical system of each channel [Kvashnin et al., 2021] contains

- mirrors with a scattering coating for attenuation of visible and infrared radiation;
- low-pass and band-pass filters that highlight the target frequency range;
- an optical chopper modulating a 10 Hz signal for correct operation of a detector;
- an optoacoustic converter (Golay cell) with an amplifier as a radiation detector [Kalinin et al., 2021].

Detailed characteristics of the optical system and the filters are given in [Tulnikov et al., 2024]. The choice of the Golay cell [Philippov et al., 2024a] is due to its uniform response in the target spectral range. The electronic part [Philippov et al., 2024b] consists of amplifiers, optical chopper drivers, a power supply unit, and an automatic thermal management system.

The possibility of detecting terahertz flare emission with such equipment was confirmed during the Antarctic experiment GRIPS (January 19–30, 2016) [Kaufmann et al., 2016; Duncan et al., 2016].

Tulnikov et al. [2024] have shown that the transmission coefficients of optical paths for 3, 5, 7, and 12 THz channels in a low-frequency region are commensurate with the transmission in the vicinity of the target frequency. This fact does not complicate the analysis of monotonously increasing spectra in the entire range (0.4–12 THz) since the intensity of radio emission in the vicinity of the target frequency exceeds that in the low-frequency region 10–1000 times, depending on the spectral index. However, when analyzing frequency spectra in which the intensity in the low-frequency region is comparable to the intensity in the vicinity of the target frequency, this leads to an ambiguous relationship between the intensity measured by SE and the actual radiation intensity in the vicinity of the target frequency, which impedes the recovery of the spectrum.

To identify spectrum parameters, it is necessary to simulate the response of SE to a spectrum with given parameters and select a set of parameters for which the model response of SE would coincide with the experimental one. Since it is impossible to search through all possible combinations of spectrum parameters due to their continuity, we must develop an algorithm capable of determining the emission spectrum parameters at the input of optical paths from the emission intensity recorded by various SE channels. This naturally leads to the formulation of a regression problem in machine learning, where the input variables are signals from eight SE channels, and the output variables are radio emission frequency spectrum parameters.

In [Kaufmann et al., 2004; Tsap et al., 2016; De Castro et al., 2005; Cristiani et al., 2007, 2009], an interesting feature of the radio emission spectrum of a number of solar flares has been found: the expected decrease in radiation fluxes at frequencies above 70–90 GHz is replaced by a significant increase in radiation at frequencies above ~200–400 GHz, i.e. a U-shaped spectrum is observed in the region of submillimeter radiation from solar flares. The purpose of this work is to develop an algorithm capable of identifying the U-shaped spectrum parameters (the spectral index of the decaying part of the spectrum, the position of the turnover point, and the spectral index of the rising part of the spectrum) from the radiation intensity recorded by the device.

1. DATA ACQUISITION AND PROCESSING

As noted above, existing observations of solar flares in the subterahertz range indicate that there may be a point in the terahertz range, after which the decaying gyrosynchrotron spectrum is replaced by a rising one [Kaufmann et al., 2004; Tsap et al., 2016; De Castro et al., 2005; Cristiani et al., 2007, 2009], i.e. a turnover point. We have therefore decided to examine the U-shaped spectrum decreasing with the spectral index γ to the turnover frequency ν_0 in the range 0.4–12.0 THz and then increasing with the spectral index α (Figure 1). We have analyzed spectra in the range 0.016–15 THz. The spectral energy density was normalized to the value of the flux at a frequency of 0.405 THz, which was obtained in [Kaufmann et al., 2004].

To acquire model data on the intensity of radiation passing through the optical system of the device at the radiation spectrum described above, we have used the technique described in [Tulnikov et al., 2024]. It involves the numerical integration of the product of the radiation spectrum entering the optical path of each SE channel and the transmission function of the corresponding optical path. Calculations were performed for γ=–1.0÷–3.0 in 0.5 increments and for the position of the turnover point in the range 0.4–12.0 THz in increments of 0.2 THz to 1 THz and then in 1 THz increments, and for α=1.0÷4.0 in 0.5 increments. Thus, the total size of the array of the spectra considered was 9765 records. Division into train and test samples was carried out at a 4:1 ratio, i.e. the train sample contains 7812 records; and the test sample, 1953 records.

Since the intensity is determined from the shape of the spectrum up to a multiplier, we decided to utilize all possible ratios of the intensity of radiation passing through optical paths of all channels to the intensity of radiation passing through the optical paths of channels with a lower target frequency as features for training the model. Thus, we have obtained 28 features. A large number of correlated features have been found which, when training the model, can cause problems related to the ambiguity of selecting coefficients for correlated features (see Figure 2). To decrease the number of correlated features, the principal component analysis (PCA) method was employed [Jolliffe, Cadima, 2016]; therefore, averages of all features were reduced to 0; and standard deviations, to 1. We took 11 features describing 99.9 % of data spread. Averages of all features were also reduced to 0; and standard deviations, to 1.

2. RESULTS AND THEIR DISCUSSION

Several algorithms have been used to solve the regression problem: linear regression (LR) without regularization, the method of the *K*-nearest neighbors (KNN)

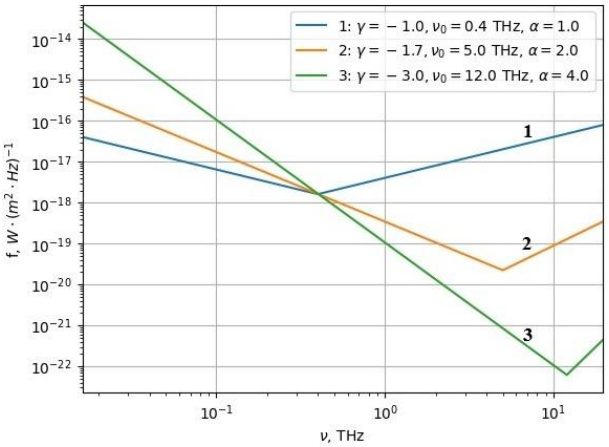

Figure 1. Examples of the U-shaped radio emission spectra considered: a kink in the spectrum at frequencies of 0.4, 5, and 12 THz

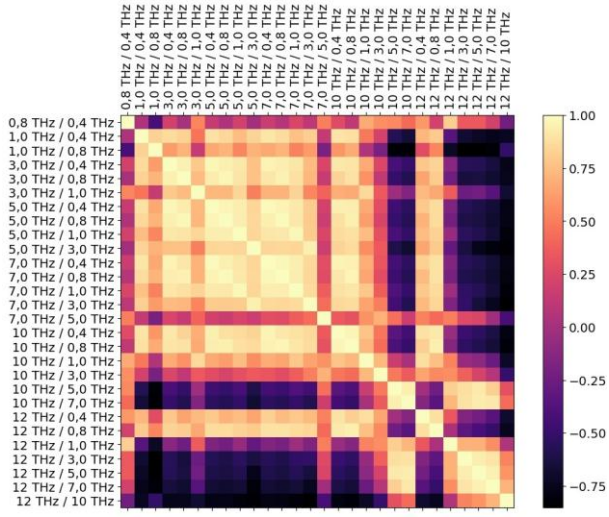

Figure 2. Heat map of the correlation coefficient matrix between features. All features in which the absolute value of the correlation coefficient exceeds 0.5 (light and dark areas) are considered correlated. When training a model, this can cause problems related to the ambiguity of the selection of coefficients for the correlated features

[Altman, 1992], the support vector machine (SVM) method [Cortes, Vapnik, 1995], the decision tree (DT) [Quinlan, 1986], random forest (RF) [Breiman, 2001], CatBoost [Prokhorenkova et al., 2018], XGBoost [Chen, Guestrin, 2016], LightGBM [Ke et al., 2017], and Stacking [Wolpert, 1992] with architecture in which the random forest (RF) is trained on results of the KNN, RF, CatBoost, XGBoost, and LightGBM methods. Hyperparameters of the models that may differ from the standard values presented in the open source Scikit-learn library [Pedregosa et al., 2011] for the LR, KNN, SVM, DT, RF, and Stacking models and in corresponding libraries for the CatBoost, XGBoost, and LightGBM models are listed in Table 1. The principle of the Stacking method training is that the basic models (in our case, KNN, RF, CatBoost, XGBoost, LightGBM) are trained on a part of the train sample. Their predictions based on a previously unused part of the train sample are taken as meta-features for training a meta-model (in our case, RF) that provides the final prediction. When generating predictions, they are first generated by basic models, and then a final prediction is made from them by the meta-model. The obtained values of the RMS deviation *RMSE* and the mean absolute relative error *MAPE* in the train and test samples for γ, ν_0, α are presented in Figures 3 and 4 respectively.

According to the results from the test sample, it is clear that the KNN and Stacking algorithms best tackle the problem. The deviation of zero for the KNN method in the train sample suggests, nonetheless, that the model is overfitted. To further improve the accuracy, we analyzed the data with which the models produced results with the greatest deviations. It has been found that the models worst distinguish between spectral data with a turnover frequency above 8 THz. This fact is consistent with the data for training models since above 8 THz there is information from only two channels, which makes it difficult to determine the position of the turnover point near the right boundary of the spectrum under study from the response of the instrument.

To identify the data whose processing is a priori difficult for the algorithms, the binary classification problem was solved (class 1 — the position of the turnover point is not higher than 8 THz, class 2 — the position of the turnover point is above 8 THz). Several algorithms have been employed to solve it: logistic regression (LogR) [Hosmer et al., 2013] without regularization, KNN, DT, and RF. Hyperparameters of the models, which may differ from the standard values provided in the open source Scikit-learn library for the LogR, KNN, DT, and RF models, are listed in Table 2. The accuracy and F1 metrics in the train and test samples are presented in Table 3.

The results show that all the selected algorithms, except for LogR, almost accurately determine the spectra for which ν_0<8 THz.

After extracting data with ν_0<8 THz, the size of the train sample was 5729 records; and the test sample, 1432 records. The principal component method was also applied to this dataset [Jolliffe, Cadima, 2016]; therefore, averages of all features were reduced to 0; and standard deviations, to 1. We left 11 features describing 99.9 % of data spread. Then, averages of all the features were also reduced to 0; and the standard deviations, to 1.

Table 1

Hyperparameters of models for solving the regression problem for different spectrum parameters

Model	γ	ν_0	α
LR	–	–	–
KNN	n_neighbors = 2 weights = ‘distance’ metric = ‘minkowski’	n_neighbors = 2 weights = ‘distance’ metric = ‘minkowski’	n_neighbors = 2 weights = ‘distance’ metric = ‘minkowski’
SVM	kernel = ‘rbf’	kernel = ‘rbf’	kernel = ‘rbf’
DT	max_depth = 18 criterion = ‘squared_error’	max_depth = 14 criterion = ‘squared_error’	max_depth = 19 criterion = ‘squared_error’
RF	max_depth = 36 n_estimators = 300 criterion = ‘squared_error’	max_depth = 21 n_estimators = 250 criterion = ‘squared_error’	max_depth = 46 n_estimators = 250 criterion = ‘squared_error’
CatBoost	loss_function = ‘*RMSE*’	loss_function = ‘*RMSE*’	loss_function = ‘*RMSE*’
XGBoost	objective = ‘reg:squarederror’	objective = ‘reg:squarederror’	objective = ‘reg:squarederror’
LightGBM	boosting_type = ‘gbdt’ n_estimators = 100 num_leaves = 31	boosting_type = ‘gbdt’ n_estimators = 100 num_leaves = 31	boosting_type = ‘gbdt’ n_estimators = 100 num_leaves = 31
Stacking	hyperparameters of KNN, RF, CatBoost, XGBoost, LightGBM are given above; hyperparameters of the RF meta-model: max_depth = 35 n_estimators = 100	hyperparameters of KNN, RF, CatBoost, XGBoost, LightGBM are given above; hyperparameters of the RF meta-model: max_depth = 20 n_estimators = 100 criterion = ‘squared_error’	hyperparameters of KNN, RF, CatBoost, XGBoost, LightGBM указаны выше; hyperparameters of the RF meta-model: max_depth = 36 n_estimators = 100

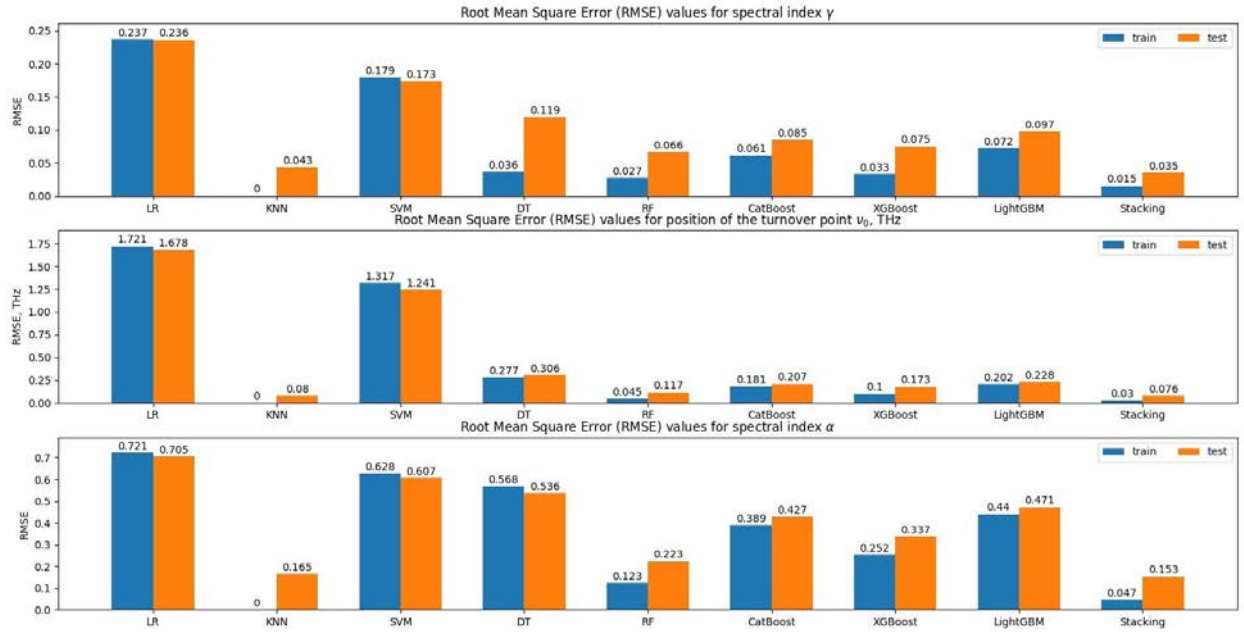

Figure 3. Root-mean-square deviation *RMSE* in the train and test samples for spectral index γ (*a*), position of the turnover point (frequency) ν_0 (*b*), and spectral index α (*c*)

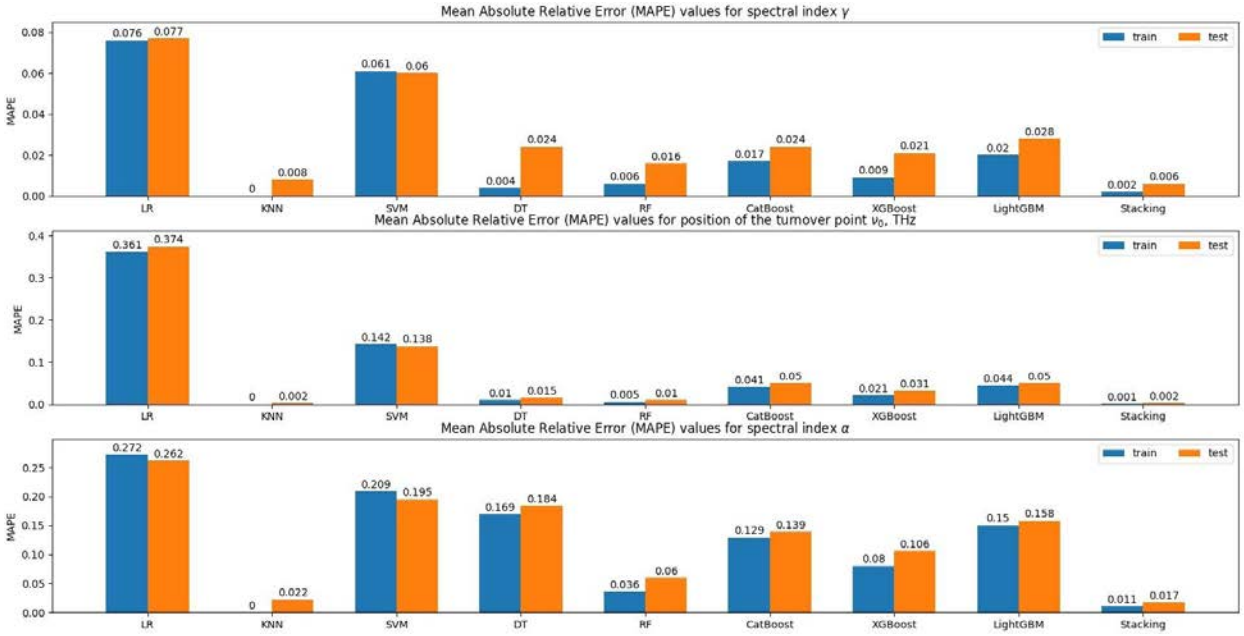

Figure 4. Mean absolute relative error *MAPE* in the train and test samples for spectral index γ (*a*), position of the turnover point (frequency) ν_0 (*b*), and spectral index α (*c*)

Table 2

Hyperparameters of models for solving the problem of classification from the position of a turnover point

Model	Hyperparameters
LogR	–
KNN	n_neighbors = 2 weights = 'distance' metric = 'minkowski'
DT	max_depth = 17 criterion = 'gini'
RF	max_depth = 31 n_estimators = 100 criterion = 'gini'

Table 3

Accuracy and F1 metrics in the train and test samples

Model	Metrix	train	test
LogR	accuracy	0.809	0.830
	F1	0.589	0.626
KNN	accuracy	0.999	0.998
	F1	0.997	0.994
DT	accuracy	1	0.994
	F1	1	0.989
RF	accuracy	1	0.997
	F1	1	0.995

We have also used several algorithms to solve the regression problem with the new dataset: LR without regularization, KNN, SVM, DT, RF, CatBoost, XGBoost, LightGBM, and Stacking. Hyperparameters of the models that may differ from the standard values provided in the open source Scikit-learn library for the LR, KNN, SVM, DT, RF, and Stacking models and in the corresponding libraries for the CatBoost, XGBoost, and LightGBM models are presented in Table 1. The obtained *RMSE* and *MAPE* values in the train and test samples for γ, ν_0, α are displayed in Figures 5 and 6 respectively.

According to the results from the test sample, it is clear that the KNN and Stacking algorithms best tackle the problem. The deviation of zero for the KNN method in the train sample suggests, nonetheless, that the model is overfitted. The separation of data on spectra with ν_0>8 THz improved the results of the LR, SVM, DT, CatBoost, XGBoost, and LightGBM methods for determining ν_0 and α and the results of RF for identifying α. In other cases, such separation did not yield improved results. Thus, the U-shaped flare data processing algorithm may consist of the following steps.

1. Applying the transformation, which for all model data reduces averages of all features to 0 and standard deviations to 1, to experimental data.
2. Applying the transformation, obtained by the PCA method for all model data, to the resulting data.
3. Applying the transformation, which reduces averages of all features obtained after PCA for all model data to 0 and the standard deviations to 1, to the acquired data.
4. Processing of the received data by pre-trained Stacking algorithms to derive γ (*RMSE*=0.035), ν_0 (0.076 THz), and α (0.153).

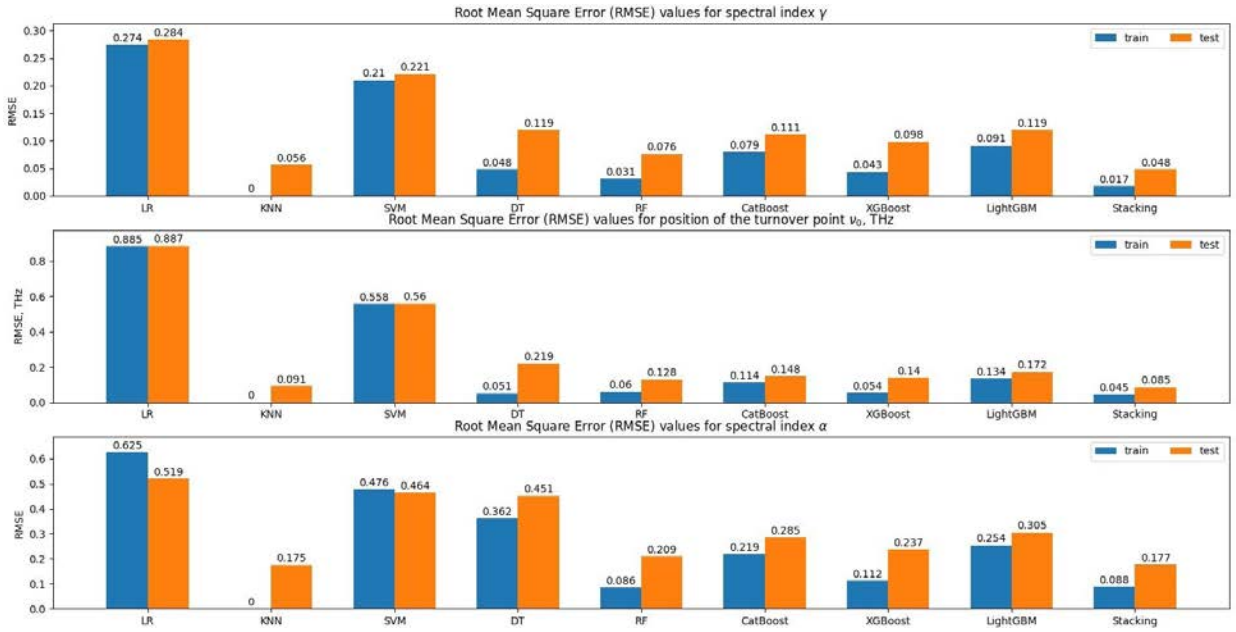

Figure 5. The same as in Figure 3. Data with $\nu_0 \leq 8$ THz

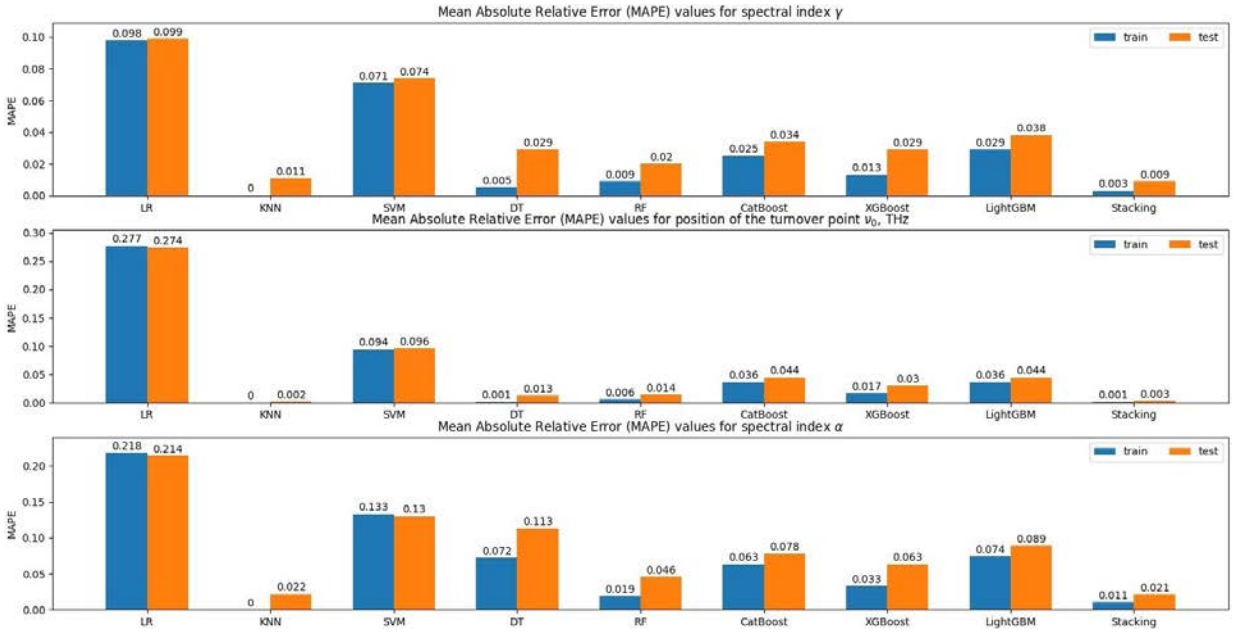

Figure 6. The same as in Figure 3. Data with $\nu_0 \leq 8$ THz

3. CONCLUSIONS

In this work, we have developed a technique based on machine learning algorithms to restore the parameters of U-shaped solar flare spectra in the terahertz range (0.4–12 THz) from data obtained by simulating the operation of the optical system of the "Solntse-Terahertz" scientific equipment. The emphasis was on determination of the spectral indices of the decaying part γ and of the rising part α of the spectrum, as well as the position of the turnover point (frequency) ν_0.

The analysis has revealed that the Stacking algorithm demonstrates the best accuracy in identifying the spectrum parameters. It shows the smallest deviations in the identified parameters when using spectra for which the turnover point is both above and below 8 THz. This suggests that the algorithm is able to determine the parameters of even slightly different spectra such that the turnover point is above 8 THz.

The developed algorithm can be integrated into the data processing system of a future experiment on the ISS, providing automatic preliminary restoration of the parameters of solar flare spectra. Note that for weak flares the noise of the measured signal has a significant effect on the shape of the recorded spectrum. Going forward, in order to improve the interpretation of data, it is necessary to examine the effect of noise in the experimental data on the accuracy of the algorithm results. It is also important to pay attention to the determination of the parameters of spectra of other shapes: monotonously rising and monotonously decaying with increasing emission frequency.

REFERENCES

Altman N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*. 1992, vol. 46, no. 3, pp. 175–185.

Breiman L. Random forests. *Machine Learning*. 2001, vol. 45, pp. 5–32.

Chen T., Guestrin C. Xgboost: A scalable tree boosting system. *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

Cortes C., Vapnik V. Support-vector networks. *Machine Learning*. 1995, vol. 20, pp. 273–297.

Cristiani G., De Castro C.G.G., Luoni M.L., et al. Observed flux density enhancement at submillimeter wavelengths during an X-class flare. *Adv. Space Res*. 2007, vol. 39, iss. 9, pp. 1445–1450. DOI: 10.1016/j.asr.2007.02.039.

Cristiani G., De Castro C.G.G., Mandrini C.H., et al. Asymmetric precipitation in a coronal loop as explanation of a singular observed spectrum. *Adv. Space Res*. 2009, vol. 44, iss. 11, pp. 1314–1320. DOI: 10.1016/j.asr.2009.06.012.

De Castro C.G.G., Kaufmann P., Raulin J.P. Recent results on solar activity at submillimeter wavelengths. *Adv. Space Res*. 2005, vol. 35, iss. 10, pp. 1769–1773. DOI: 10.1016/j.asr.2005.03.083.

Duncan N., Saint-Hilaire P., Shih A.Y., et al. First flight of the Gamma-Ray Imager/Polarimeter for Solar flares (GRIPS) instrument. *Space Telescopes and Instrumentation 2016: Ultraviolet to Gamma Ray.* Edinburgh, 2016, vol. 9905, p. 876. DOI: 10.1117/12.2233859.

Hosmer Jr D.W., Lemeshow S., Sturdivant R.X. *Applied Logistic Regression*. John Wiley & Sons, Inc., 2013, 510 p.

Jolliffe I.T., Cadima J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016, vol. 374:20150202, iss. 2065. DOI: 10.1098/rsta.2015.0202.

Kalinin E.V., Philippov M.V., Makhmutov V.S., et al. A study of the characteristics of a terahertz radiation detector for the Solntse-Terahertz scientific apparatus. *Cosmic Res*. 2021, vol. 59, iss. 1, pp. 1–5. DOI: 10.1134/S0010952521010032.

Kaufmann P., Raulin J.P, De Castro C.G.G., et al. A new solar burst spectral component emitting only in the terahertz range. *ApJ. Lett*. 2004, vol. 603, iss. 2, pp. L121–L124. DOI: 10.1086/383186.

Kaufmann P., Abrantes A., Bortolucci E.C., et al. THz solar observations on board of a trans-Antarctic stratospheric balloon flight. *41st International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz)*, Copenhagen, 2016, p. 1. DOI: 10.1109/IRMMW-THz.2016.7758395.

Ke G., Meng Q., Finley T., et al. Lightgbm: A highly efficient gradient boosting decision tree. *31st Conference on Neural Information Processing Systems (NIPS–2017),* Long Beach, 2017, vol. 30.

Krucker S., De Castro C.G.G., Hudson H.S., et al. Solar flares at submillimeter wavelengths. *Astron. Astrophys. Rev*. 2013, vol. 21, iss. 1, pp. 1–45. DOI: 10.1007/s00159-013-0058-3.

Kvashnin A.A., Logachev V.I., Philippov M.V., et al. Optical system design of the detector for solar terahertz emission measurements. *Space Engineering and Technology*. 2021, vol. 35, iss. 4, pp. 22–30. DOI: 10.33950/spacetech-2308-7625-2021-4-22-30.

Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine learning in Python. *J. Machine Learning Res*. 2011, vol. 12, pp. 2825–2830. DOI: 10.5555/1953048.2078195.

Philippov M.V., Makhmutov V.S., Razumeyko M.V. Scientific equipment for the Sun-Terahertz space experiment: study of the temperature effect in the Golay cell. *Measurement Techniques*. 2024a, vol. 67, iss. 3, pp. 195–202. DOI: 10.1007/s11018-024-02335-9.

Philippov M.V., Makhmutov V.S., Maksumov O.S., et al. Electronics Unit for "Sun-Terahertz" Scientific Equipment. *Instrum. Exp Tech*. 2024b, vol. 67, iss. 3, pp. 545–553. DOI: 10.1134/S0020441224700829.

Prokhorenkova L., Gusev G., Vorobev A., et al. CatBoost: unbiased boosting with categorical features. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018).* Montréal, 2018, vol. 31.

Quinlan J.R. Induction of decision trees. *Machine Learning*. 1986, vol. 1, pp. 81–106.

Tsap Y.T., Smirnova V.V., Morgachev, A.S., et al. On the origin of 140 GHz emission from the 4 July 2012 solar flare. *Adv. Space Res.* 2016, vol. 57, iss. 7, pp. 1449–1455. DOI: 10.1016/j.asr.2015.12.037.

Tulnikov E.D., Logachev V.I., Makhmutov V.S., et al. Characteristics of the optical system of the Solntse-Terahertz scientific equipment. *Cosmic Res*. 2024, vol. 62, iss. 6, pp. 551–557. DOI: 10.1134/S0010952524600434.

Tulnikov E.D., Makhmutov V.S., Philippov M.V. Review of studying submillimeter radiation from the Sun and astrophysical sources. *Bull. Russian Academy of Sciences: Physics*. 2025, vol. 89, iss. 6, pp. 854-857. DOI: 10.1134/S1062873825711298.

Wedemeyer S., Bastian T., Brajša R., et al. Solar science with the Atacama Large Millimeter/Submillimeter Array — a new view of our Sun. *Space Sci. Rev*. 2016, vol. 200, pp. 1–73. DOI: 10.1007/s11214-015-0229-9.

Wolpert D.H. Stacked generalization. *Neural Networks*. 1992, vol. 5, iss. 2, pp. 241–259.

Original Russian version: Tulnikov E.D., Makhmutov V.S., Philippov M.V., published in Solnechno-zemnaya fizika. 2026, vol. 12, no. 1, pp. 14–20. DOI: 10.12737/szf-121202602.

How to cite this article

Tulnikov E.D., Makhmutov V.S., Philippov M.V. Application of machine learning methods to determine spectral characteristics of radiation in the “Solntse-Terahertz” experiment. *Sol.-Terr. Phys*. 2026, vol. 12, iss. 1, pp. 11–17. DOI: 10.12737/stp-121202602.