Multiclass Classification of Solar Flares in Imbalanced Data Using Ensemble Learning and Sampling Methods

Ryoma Matsuura
University of California, Los Angeles
Department of Atmospheric and Oceanic Sciences
Los Angeles, CA 90095, USA

Haodi Jiang
Sam Houston State University
Department of Computer Science
Huntsville, TX 77340, USA
Correspondence: haodi.jiang@shsu.edu

Jason T. L. Wang
New Jersey Institute of Technology
Department of Computer Science
Newark, NJ 07102, USA

Abstract

Solar flares are intense bursts of radiation across the electromagnetic spectrum on the surface of the Sun. They are categorized into four classes: B, C, M, and X, depending on their intensity, with X-class flares being the strongest. Being able to predict a flare's class before its occurrence is critical for anticipating the severity of its impact on Earth. We used the Space-weather HMI Active Region Patches (SHARP) parameters available from Stanford's Joint Science Operations Center (JSOC) to train machine learning models to classify these flares. However, predicting the flare class is a challenging task, as it is a multiclass classification problem involving imbalanced data due to the small number of X-class flares in a solar cycle. We propose a new method that uses a combination of random undersampling and the synthetic minority oversampling technique (SMOTE) to combat the imbalanced data problem. Furthermore, we develop an ensemble algorithm that uses nine classifiers as base learners and logistic regression as meta-learner. Experimental results show that the proposed method is effective in predicting solar flares, especially the most intense X-class flares, within the next 24 hours.

Introduction

A solar flare is a major eruptive event in the solar system. This event converts the magnetic energy stored in the Sun's magnetic field into radiation that covers a wide range of wavelengths. The radiation is a huge threat to astronauts on missions and orbiting satellites. Flares can be categorized into different classes on the basis of their intensity. Intense flares are often accompanied by coronal mass ejections (CMEs), which are large-scale magnetic structures that contain coronal material. CMEs travel at very high speeds and can cause large geomagnetic storms when they are directed toward Earth. Geomagnetic storms can have a negative impact on various technological infrastructures, such as the Global Positioning System (GPS) and electrical power grids. The negative impact on these systems, upon which our society is highly dependent, would critically damage the economy of a nation (Moldwin 2023).

Due to the critical impact flares can have, prediction of flare events is very important for protecting our technological infrastructure. Many studies have been performed to understand the causes and precursors of flares. They have traditionally been done by building physics-based models of the flares. However, physics-based models are far from satisfactory for accurately predicting flares ahead of time (Jiao et al. 2020). In recent years, the use of machine learning and statistical models in the prediction of solar flares has become common. Many of the studies that used machine learning for prediction attempted to predict whether an active region would produce a flare belonging to a particular class in a certain time frame. Because the class labels are given on the basis of the peak soft X-ray flux, predicting the classes of flares allows scientists to estimate the intensity of the flares and their impact on Earth. Many machine learning models have been tested on flare data, including long short-term memory (LSTM) networks and convolutional neural networks (CNNs) (Liu et al. 2019; Zheng, Li, and Wang 2019; Wang et al. 2020; Sun et al. 2022; Datla, Jiang, and Wang 2023).

However, predicting flares is a challenging task, in part due to the uneven class distribution in the data set. Machine learning models perform poorly with imbalanced data, as reported in the literature (Sun et al. 2007; Wang, Tian, and Liu 2019). An earlier study (Liu et al. 2017) attempted to predict flares of classes B, C, M, and X (ordered from the smallest to the largest flares, with flares of class X being the most intense) within the next 24 hours using SHARP parameters (Bobra et al. 2014). Those authors resampled the data with random undersampling to combat the class imbalance problem. Later, researchers (Abduallah et al. 2021) extended that earlier study by using an ensemble algorithm that makes predictions based on majority voting among random forests (Breiman 2001), multilayer perceptrons (Rosenblatt 1958), and extreme learning machines (Huang, Zhu, and Siew 2006). Our study aims to further extend the above work to improve the predictive capability on solar flare data. We achieve this goal by combining established machine learning models in an ensemble algorithm. Furthermore, we employ sampling methods to address the challenge of imbalanced classification. The major contributions of our work are:

1. integrating the synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002) into our method on top of

Table 1: SHARP Parameters Used in This Study

Parameter	Description
TOTUSJH	Total unsigned current helicity
TOTBSQ	Total magnitude of Lorentz force
TOTPOT	Total photospheric magnetic free energy density
TOTUSJZ	Total unsigned vertical current
ABSNJZH	Absolute value of the net current helicity
SAVNCPP	Sum of the modulus of the net current per polarity
USFLUX	Total unsigned flux
AREA_ACR	Area of strong field pixels in the active region
TOTFZ	Sum of z-component of Lorentz force
MEANPOT	Mean photospheric magnetic free energy
R_VALUE	Sum of flux near polarity inversion line
EPSZ	Sum of z-component of normalized Lorentz force
SHRGT45	Fraction of Area with shear >45°

random undersampling to handle the heavily imbalanced data;

2. using nine base models in our ensemble algorithm to increase the prediction accuracy;
3. using logistic regression, instead of simple majority voting, as the meta-learner in the ensemble algorithm to make the final decision based on the base models' outcomes.

In contrast to existing methods (Liu et al. 2019; Zheng, Li, and Wang 2019; Wang et al. 2020; Sun et al. 2022), our work focuses on the multiclass classification of solar flares with imbalanced data.

Data

We use the Space-weather HMI Active Region Patches (SHARP) related data products downloaded from the hmi.sharp data series at the Joint Science Operations Center (`http://jsoc.stanford.edu/`). SHARP data products are derived from NASA's Solar Dynamics Observatory's (SDO) Helioseismic and Magnetic Imager (HMI) observations. These data products, released in 2010, are commonly used to predict eruptive events such as flares from solar active regions (Bobra et al. 2014). In 2014, a separate data series, cgem.Lorentz, was produced based on SHARP data to include estimates of the Lorentz force. Using these data series, the researchers (Bobra and Couvidat 2015) calculated 25 parameters that characterize active regions and used a univariate feature selection algorithm to rank the parameters. On the basis of the feature selection, they suggested that only the top 13 parameters are important for predicting solar flares. These parameters are listed in Table 1. For more details on these parameters, see (Bobra and Couvidat 2015; Liu et al. 2017).

Using SHARP parameters and the Geostationary Operational Environmental Satellite (GOES) X-ray flare catalog prepared by the National Centers for Environment Information (NCEI), researchers (Liu et al. 2017) constructed a set of 845 flare samples. This data set contains flares that occurred between May 2010 and December 2016. On the basis of flare intensity, the flare samples are classified into four categories with the number of samples in each category enclosed in parentheses: B (128), C (552), M (142), and X (23). Our work attempts to predict the class/category of a flare that would occur within the next 24 hours using this data set and the 13 SHARP physical parameters listed in Table 1.

Methodology

Class Imbalance

As indicated above, the data set at hand contains significantly less X-class flare samples (23 out of 845 samples) and significantly more C-class flare samples (552 out of 845 samples) than the flares of the other classes. Flares of different classes do not occur at the same frequency. Therefore, this is an imbalanced data set.

Having an imbalanced data set is common in many areas of study. However, machine learning models are known to perform poorly when trained on a data set with a heavy class imbalance. Various studies have shown that the prediction performance for the class to which fewer samples belong (minority class) is particularly poor with imbalanced data sets (Sun et al. 2007; Wang, Tian, and Liu 2019). This is because these machine learning models attempt to maximize the overall prediction accuracy across all classes, and therefore pay more attention to classes to which more samples belong (majority class) and pay less attention to samples from the minority class (Wang, Tian, and Liu 2019). However, in many real-world problems that involve imbalanced data sets, such as rare disease detection, the performance of the minority class prediction is more important (Sun et al. 2007). This is also true for flare prediction, because X-class flares are the most impactful flares while being rare compared to smaller flares such as C-class flares.

Undersampling

One possible way to improve the prediction performance when we have an imbalanced data set is to reduce the number of samples in the majority class before training a machine learning model using undersampling techniques. In our work, we undersample the C-class flares to match the number of M-class flares in the training data set, where the M-class has the second-largest number of samples, through random undersampling as in previous work (Liu et al. 2017; Abduallah et al. 2021). In doing so, we used the RandomUnderSampler function provided by the imbalanced-learn package in Python. Random undersampling is a simple undersampling method that removes randomly selected samples from the majority class to increase the proportion of other classes. Although random undersampling makes the training data set more balanced, using it may lose important information in the majority class.

Oversampling

Additionally, because the number of the X-class flares is very small, we chose to oversample the data set using the synthetic minority oversampling technique (SMOTE). SMOTE is an oversampling method that oversamples the minority class by generating synthetic data samples of that class (Chawla et al. 2002). Specifically, SMOTE generates synthetic data samples by taking each sample S from the minority class and generating a sample at a random point between S itself and one of its K-nearest minority class neighbors in the feature space (Chawla et al. 2002; He and Garcia 2009).

Using the SMOTE method, we doubled the number of X-class flares in the training data set prior to training. As a result of our random undersampling and SMOTE oversampling, we significantly reduced the proportion of C-class flares while increasing the proportion of X-class flares in the training data set. Together, the sampling methods improve the accuracy of flare prediction.

The Ensemble Algorithm

Using an ensemble algorithm or using multiple "weak" models to build one strong model is a way to improve prediction performance in solving machine learning problems. Various forms of ensemble algorithms have been developed, among which stacked generalization is widely used (Wolpert 1992). In stacked generalization, different types of model (such as random forests, support vector machines, etc.) are trained separately to make predictions individually. Then a meta-learner uses the predictions made by the individual models and attempts to make the final decision based on the individual predictions. Stacked generalization has been used for solar flare prediction in several studies. For example, researchers employed stacking of CNN and LSTM networks to achieve better prediction performance in certain settings (Sun et al. 2022).

Our stacking method is based on the work in (Abduallah et al. 2021), but takes it a step further by utilizing more base models. Additionally, we hypothesized that with more base models, it is unlikely that all the models perform just as well as each other, in which case there may be a way to make the final prediction better than a simple majority-voting strategy. Therefore, we used multinomial logistic regression as the meta-learner to make the final prediction of the flare classes based on the classifications done by the individual base models. Note that when using multinomial logistic regression, the classes are weighed by the weights adjusted to be inversely proportional to the class frequencies. Later in the experiments section, we will compare this meta-learner with the simple majority-voting strategy.

Base Models

In the related study (Abduallah et al. 2021), random forests (RF) (Breiman 2001), multilayer perceptrons (MLP) (Rosenblatt 1958), and extreme learning machines (ELM) (Huang, Zhu, and Siew 2006) were used as base learners to form the ensemble model with the majority voting strategy for multiclass classification of solar flares. RF is a bagging ensemble algorithm in which multiple decision trees are fit on sub-samples before a majority vote is taken among the trees. MLP and ELM are both variations of a feedforward artificial neural network with an input layer, an output layer, and hidden layers.

In addition to these three models, we adopted the following six commonly used classifiers as base learners: extremely randomized trees (ERT) (Geurts, Ernst, and Wehenkel 2006), support vector machines (SVM) (Cortes and Vapnik 1995), K nearest neighbors (KNN) (Fix and Hodges 1989), radius-based nearest neighbors (R-NN) (Bentley 1975), adaptive boosting based on decisions trees (ADA) (Freund and Schapire 1997), and gradient boosting (GB) (Friedman 2001). ERT, also known as extra trees, is an ensemble classifier similar to RF with more randomness included. SVM is a popular classification model that searches for a hyperplane that separates the classes while maximizing the margin between the classes. KNN is a simple classifier that determines the classification of a data point based on the majority vote of its K nearest neighbors in the feature space. R-NN is similar to KNN, except that the majority vote is taken based on all points within a certain radius in the feature space rather than the K nearest neighbors. ADA is an ensemble algorithm based on a series of "weak" models that are trained iteratively while weights are assigned to each sample to place emphasis on misclassified samples in the next iteration. GB is another ensemble algorithm similar to adaptive boosting but uses gradient descent to optimize the loss function when training its "weak" models.

Hyperparameter Settings

We used the Python scikit-learn package to implement the base models. For all models, any parameter not specified below is set to its default value. Our RF is made up of 500 trees, and 4 randomly chosen features are considered when looking for the best split of a node. Our MLP includes 3 hidden layers, each consisting of 100 neurons. The ELM has one hidden layer that contains 200 neurons, and the hyperbolic tangent function is used as the activation function. The number of trees in the ERT is set to 500. For the SVM, the radial basis function (RBF) kernel was used where Hyperparameters C and γ were set to 50 and 0.01, respectively. For the KNN, K was chosen to be 6 where a weighing function that is inversely proportional to the distance was used to place a higher value on the nearest neighbors. A radius of 4 was chosen for the R-NN, and the points were weighed by the same weighing function as the KNN. Our ADA is based on 500 decision trees with a maximum depth of 8 and the learning rate is set to 1. The GB utilizes 500 trees with a maximum depth of 3 and the learning rate is set to 0.01. All of these parameter values are chosen to optimize prediction performance.

Experiments and Results

Experimental Setup

We incorporate random undersampling and SMOTE oversampling into our 10-fold cross-validation scheme to under-

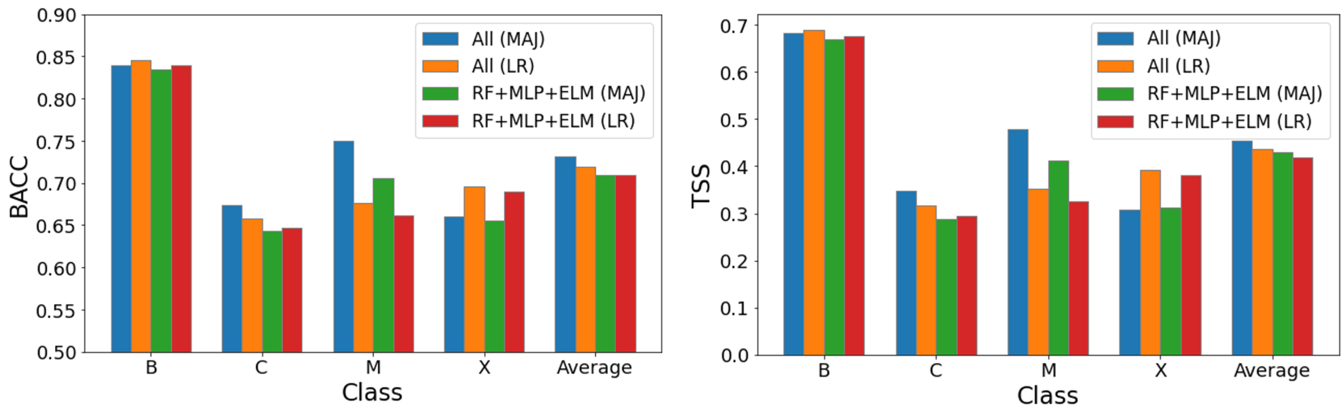

Figure 1: Flare prediction results obtained by using different sets of base models with MAJ or LR being the meta-learner in the ensemble algorithms (stacked generalization).

sample the C class and oversample the X class in the training set as described in the Methodology section. In performing the 10-fold cross-validation, we randomly partition the data set into 10 subsets or folds of equal size while preserving the ratio of the classes in each fold. Then we use nine folds to train the models while leaving the remaining one fold for testing. This partitioning/training/testing process is repeated 10 times. In each time, we denote the training and test sets as A and B, respectively.

Since training the meta-learner (multinomial logistic regression) requires predictions from every base model and the corresponding label to be fed to the meta-learner, all base models are trained and tested within A. Here, we further perform another 10-fold cross-validation on the set A. Random undersampling and SMOTE oversampling are again incorporated into the training set in this 10-fold cross-validation process. For each fold, the nine base models in our stacking ensemble are first trained to perform multiclass classification. These base models produce probability estimates for each flare class. The predictions made by the base models serve as the training data for the meta-learner, where each training sample for the meta-learner contains the predicted class probabilities from each base model and the actual class label for the corresponding input. After the meta-learner is trained, each base model will be retrained using the set A before testing on the set B. During testing, we first obtain the class probabilities from each base model for a test sample in B. Then, we feed these probabilities to the trained meta-learner to produce a set of class probabilities. We then make the final decision based on the class probabilities produced by the meta-learner by selecting the class with the highest probability as the predicted class for the test sample. There are two 10-fold cross-validations, and therefore 100 experiments are performed.

Evaluation Metrics

We adopt two metrics to evaluate the performance of the machine learning models studied here: balanced accuracy (BACC) and true skill statistics (TSS), both of which are well suited for imbalanced classification problems (Chawla 2005; Abduallah et al. 2021; Georgoulis et al. 2024). These metrics are computed considering our task as four separate binary classification problems, one for each flare class (Abduallah et al. 2021), instead of a four-class classification problem. For each class i, the true positive (TP_i) is defined as the number of samples that are correctly predicted to be in class i. The true negative (TN_i) is defined as the number of samples that are not in class i and are predicted not to be in class i. The false positive (FP_i) is defined as the number of samples that are mistakenly classified in class i. The false negative (FN_i) is defined as the number of samples that are in class i, but are not predicted to be in class i.

BACC and TSS for each class can be calculated through the following formulas:

$$\mathrm{BACC}_i = \frac{1}{2}\left(\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} + \frac{\mathrm{TN}_i}{\mathrm{TN}_i + \mathrm{FP}_i}\right), \tag{1}$$

$$\mathrm{TSS}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} - \frac{\mathrm{FP}_i}{\mathrm{TN}_i + \mathrm{FP}_i}. \tag{2}$$

Computing these metrics on our results allows us to easily compare the results obtained using different machine learning methods, both class by class and by the average score across all four flare classes.

Results of Stacked Generalization

Figure 1 compares the BACC and TSS obtained using the nine base models described in the Methodology section against those obtained using the 3 base models RF, MLP and ELM in (Abduallah et al. 2021). The word “All” in the figure indicates that all nine base models including RF, MLP, ELM, ERT, SVM, KNN, R-NN, ADA, GB, are used in the ensemble algorithms (stacked generalization). The word “MAJ” denotes that the simple majority voting strategy is used as the meta-learner, while the word “LR” indicates that multinomial logistic regression is used as the meta-learner in the ensemble algorithms. Both random undersampling and SMOTE oversampling were used to sample training data during the training phase. Figure 1 shows that adopting the nine base models proposed here is better than adopting the

Table 2: Flare Prediction Results with Stacked Generalization

BACC	Class B	Class C	Class M	Class X	Average
All (MAJ)	0.840±0.057	0.674±0.072	**0.750±0.076**	0.661±0.125	**0.731±0.055**
All (LR)	**0.845±0.062**	0.658±0.077	0.676±0.075	**0.696±0.092**	0.719±0.049
ERT	0.838±0.054	**0.679±0.077**	0.744±0.080	0.642±0.126	0.726±0.054
SVM	0.820±0.066	0.661±0.074	0.735±0.075	0.660±0.115	0.719±0.052
RF	0.830±0.059	0.666±0.074	0.732±0.078	0.641±0.125	0.717±0.055
ADA	0.833±0.057	0.670±0.070	0.741±0.083	0.613±0.122	0.714±0.051
GB	0.825±0.051	0.650±0.078	0.710±0.077	0.627±0.127	0.703±0.055
KNN	0.814±0.061	0.637±0.070	0.706±0.072	0.651±0.117	0.702±0.045
R-NN	0.776±0.070	0.649±0.073	0.725±0.078	0.641±0.116	0.698±0.049
ELM	0.798±0.061	0.633±0.078	0.666±0.081	0.664±0.112	0.690±0.053
MLP	0.776±0.075	0.621±0.073	0.655±0.084	0.668±0.105	0.680±0.050
TSS	**Class B**	**Class C**	**Class M**	**Class X**	**Average**
All (MAJ)	**0.698±0.106**	**0.361±0.145**	**0.490±0.134**	0.287±0.249	**0.459±0.094**
All (LR)	0.689±0.124	0.316±0.155	0.352±0.151	**0.392±0.184**	0.437±0.098
ERT	0.677±0.108	0.358±0.155	0.488±0.159	0.284±0.252	0.452±0.107
SVM	0.639±0.131	0.322±0.148	0.471±0.151	0.321±0.229	0.438±0.103
RF	0.660±0.118	0.332±0.147	0.465±0.157	0.282±0.250	0.435±0.111
ADA	0.666±0.115	0.340±0.139	0.481±0.166	0.227±0.245	0.429±0.103
GB	0.628±0.127	0.299±0.152	0.435±0.161	0.237±0.243	0.399±0.113
KNN	0.627±0.122	0.274±0.141	0.412±0.145	0.302±0.235	0.404±0.090
R-NN	0.552±0.140	0.298±0.145	0.451±0.157	0.282±0.231	0.396±0.099
ELM	0.596±0.121	0.267±0.156	0.333±0.163	0.329±0.223	0.381±0.105
MLP	0.552±0.151	0.242±0.146	0.311±0.169	0.337±0.210	0.360±0.099

three base models, whether MAJ or LR is used as the meta-learner, in terms of the average scores of both BACC and TSS. We note that the method in (Abduallah et al. 2021) involves RF, MLP and ELM combined with MAJ, denoted as RF + MLP + ELM (MAJ) in Figure 1. The highest average scores for BACC and TSS are achieved when MAJ is used with all base models. On the other hand, LR performs much better than MAJ in predicting the X-class flares. Thus, if one wants to predict all four flare classes, one would need to use MAJ combined with all nine base models. If one wants to focus on predicting the X-class flares, LR combined with the nine base models is recommended.

Table 2 presents the means and standard deviations of BACC and TSS over the 100 experiments in the two 10-fold cross-validations used in our study for all machine learning methods, individual or combined. Both random undersampling and SMOTE oversampling were used to sample training data during the training phase. Table 2 shows that the average scores of both BACC and TSS are maximized when the nine base models are used altogether and the majority vote (MAJ) is taken. ERT (extremely randomized trees) is the best individual model that achieves the highest average scores, both in BACC and in TSS, among the nine base models. When looking at the performance of predicting the rare, strongest X-class flares, the BACC and TSS obtained by using all nine base models combined with multinomial logistic regression (LR) as the meta-learner are higher than those obtained by all other methods. This finding is consistent with what we see in Figure 1, showing that this form of stacked generalization (LR) may not produce the highest average scores, but does very well in predicting the rare, strongest X-class flares.

Results of Sampling

Figure 2 compares the performance metric values obtained by applying different sampling techniques to training data where RUS denotes random undersampling. All nine base models were used in the ensemble algorithms. Figure 2 shows that for both majority voting (MAJ) and multinomial logistic regression meta-learners, the average BACC and TSS scores are maximized when random undersampling (RUS) and SMOTE oversampling are used together. The worst average BACC and TSS scores are obtained when MAJ is used with neither RUS nor SMOTE implemented. Between the two sampling techniques, RUS is better than SMOTE in producing the average BACC and TSS scores. On the other hand, SMOTE is more effective than RUS in predicting X-class flares.

Conclusion

In this paper, we present a new method for predicting solar flares using SHARP physical parameters. This method consists of a stacking ensemble algorithm that uses nine well-known machine learning models as base learners and multinomial logistic regression (LR) as the meta-learner. Furthermore, the method adopts random undersampling (RUS) and SMOTE oversampling to overcome the imbalanced classification issue. The proposal method outperforms the method in earlier study (Abduallah et al. 2021), Specifically, our experimental results show that higher average BACC and TSS scores can be achieved using RUS and SMOTE than when

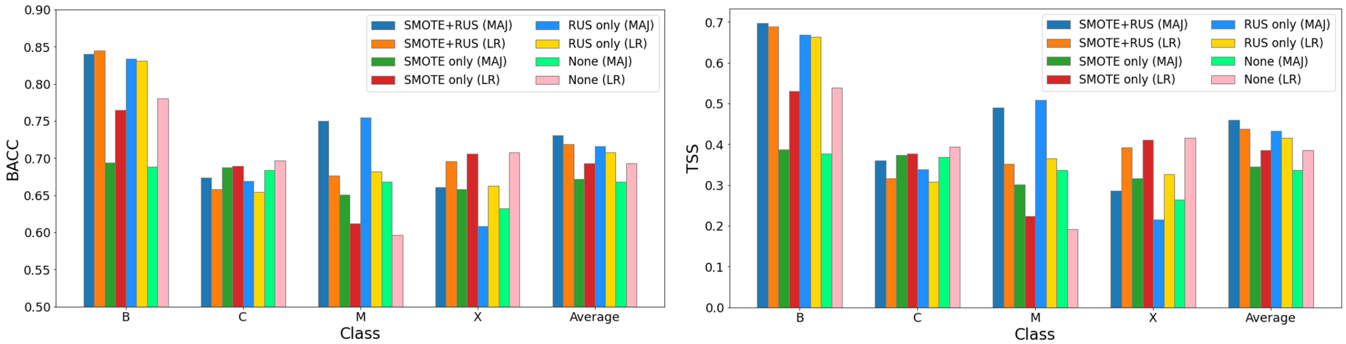

Figure 2: Flare prediction results obtained by different sampling techniques.

using only one or none of the sampling methods. The highest average BACC and TSS scores are obtained using RUS, SMOTE, and the majority voting (MAJ) meta-learner. However, for the prediction of the rare, strongest X-class flares, LR is better than MAJ. On the basis of these results, we conclude that the proposed method is viable for solar flare prediction in imbalanced data.

References

Abduallah, Y.; Wang, J. T. L.; Nie, Y.; Liu, C.; and Wang, H. 2021. DeepSun: Machine-learning-as-a-service for solar flare prediction. *Research in Astronomy and Astrophysics* 21(7):160.

Bentley, J. L. 1975. Survey of techniques for fixed radius near neighbor searching. Technical report, Stanford Linear Accelerator Center, USA.

Bobra, M. G., and Couvidat, S. 2015. Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal* 798(2):135.

Bobra, M. G.; Sun, X.; Hoeksema, J. T.; Turmon, M.; Liu, Y.; Hayashi, K.; Barnes, G.; and Leka, K. 2014. The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: SHARPs–space-weather HMI active region patches. *Solar Physics* 289(9):3549–3578.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321–357.

Chawla, N. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20(3):273–297.

Datla, G. V.; Jiang, H.; and Wang, J. T. L. 2023. An interpretable lstm network for solar flare prediction. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, 526–531.

Fix, E., and Hodges, J. L. 1989. Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique* 57(3):238–247.

Freund, Y., and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1):119–139.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 1189–1232.

Georgoulis, M. K.; Yardley, S. L.; Guerra, J. A.; Murray, S. A.; Ahmadzadeh, A.; Anastasiadis, A.; Angryk, R.; Aydin, B.; Banerjee, D.; Barnes, G.; Bemporad, A.; Benvenuto, F.; Bloomfield, D. S.; Bobra, M.; Campi, C.; Camporeale, E.; DeForest, C. E.; Emslie, A. G.; Falconer, D.; Feng, L.; Gan, W.; Green, L. M.; Guastavino, S.; Hapgood, M.; Kempton, D.; Kitiashvili, I.; Kontogiannis, I.; Korsos, M. B.; Leka, K.; Massa, P.; Massone, A. M.; Nandy, D.; Nindos, A.; Papaioannou, A.; Park, S.-H.; Patsourakos, S.; Piana, M.; Raouafi, N. E.; Sadykov, V. M.; Toriumi, S.; Vourlidas, A.; Wang, H.; Wang, J. T. L.; Whitman, K.; Yan, Y.; and Zhukov, A. N. 2024. Prediction of solar energetic events impacting space weather conditions. *Advances in Space Research*.

Geurts, P.; Ernst, D.; and Wehenkel, L. 2006. Extremely randomized trees. *Machine Learning* 63(1):3–42.

He, H., and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9):1263–1284.

Huang, G.-B.; Zhu, Q.-Y.; and Siew, C.-K. 2006. Extreme learning machine: theory and applications. *Neurocomputing* 70(1-3):489–501.

Jiao, Z.; Sun, H.; Wang, X.; Manchester, W.; Gombosi, T.; Hero, A.; and Chen, Y. 2020. Solar flare intensity prediction with machine learning models. *Space Weather* 18(7):e2020SW002440.

Liu, C.; Deng, N.; Wang, J. T. L.; and Wang, H. 2017. Predicting solar flares using SDO/HMI vector magnetic data products and the random forest algorithm. *The Astrophysical Journal* 843(2):104.

Liu, H.; Liu, C.; Wang, J. T. L.; and Wang, H. 2019. Predicting solar flares using a long short-term memory network. *The Astrophysical Journal* 877(2):121.

Moldwin, M. 2023. *An Introduction to Space Weather*. Cambridge University Press.

Rosenblatt, F. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6):386.

Sun, Y.; Kamel, M. S.; Wong, A. K.; and Wang, Y. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40(12):3358–3378.

Sun, Z.; Bobra, M. G.; Wang, X.; Wang, Y.; Sun, H.; Gombosi, T.; Chen, Y.; and Hero, A. 2022. Predicting solar flares using CNN and LSTM on two solar cycles of active region data. *The Astrophysical Journal* 931(2):163.

Wang, X.; Chen, Y.; Toth, G.; Manchester, W. B.; Gombosi, T. I.; Hero, A. O.; Jiao, Z.; Sun, H.; Jin, M.; and Liu, Y. 2020. Predicting solar flares with machine learning: Investigating solar cycle dependence. *The Astrophysical Journal* 895(1):3.

Wang, Q.; Tian, Y.; and Liu, D. 2019. Adaptive FH-SVM for imbalanced classification. *IEEE Access* 7:130410–130422.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5(2):241–259.

Zheng, Y.; Li, X.; and Wang, X. 2019. Solar flare prediction with the hybrid deep convolutional neural network. *The Astrophysical Journal* 885(1):73.