Rare-Event Time Series Prediction: A Case Study of Solar Flare Forecasting

Azim Ahmadzadeh*, Berkay Aydin*, Dustin J. Kempton*, Manolis K. Georgoulis†‡,
Sushant S. Mahajan†§, Maxwell Hostetter*, and Rafal A. Angryk*
*Department of Compute Science, Georgia State University, Atlanta, GA, USA
†Department of Physics & Astronomy, Georgia State University, Atlanta, GA, USA
†RCAAM of the Academy of Athens, Athens, Greece
Email: *{aahmadzadeh1,baydin2,dkempton1,mhostetter1,angryk}@cs.gsu.edu,
‡manolis.georgoulis@phy-astr.gsu.edu, §mahajan@astro.gsu.edu

***Abstract*—We present a case study for time series prediction models in extreme class-imbalance problems. We have extracted multiple properties from the Space Weather ANalytics for Solar Flares (SWAN-SF) benchmark dataset which comprises of magnetic features from over 4075 active regions over a period of 9 years to create the forecasting dataset used in this study. In the extracted dataset, the class-imbalance ratio is 1:60, where the minority class is formed by instances of strong solar flares (GOES M- and X-class). This ratio reaches to 1:800 if we only consider the strongest class of flares (GOES X-class). This case of extreme imbalance, along with the temporal coherence of the sliced time series, provides us with an interesting set of challenges in the forecasting of scarce real-life phenomena. We have explored remedies to tackle the class-imbalance issue such as undersampling, oversampling and misclassification weights. In the process, we elaborate on common mistakes and pitfalls caused by ignoring the side effects of these remedies, including how and why they weaken the robustness of the trained models while seemingly improving the performance.**

***Index Terms*—class imbalance, sampling, time series, flare forecast**

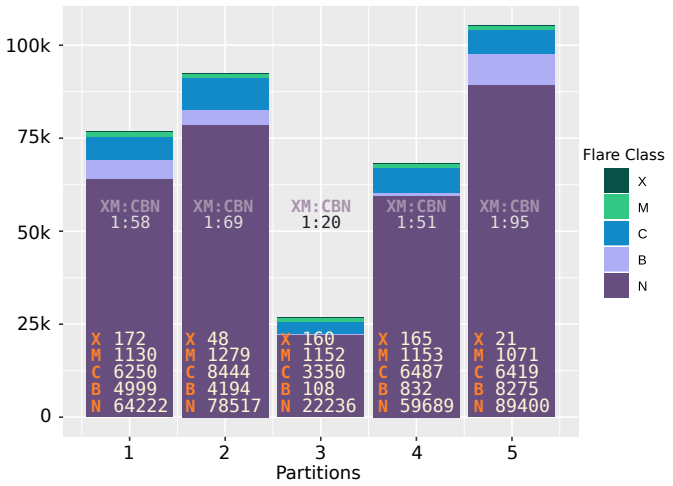

Fig. 1. Frequency and imbalance ratio of all five flare classes across different partitions of SWAN-SF benchmark dataset.

I. Introduction

Any collection of data must be accompanied by a rigorous data cleaning process. It requires a thorough investigation by the experts of the domain and data scientists to produce a reliable dataset. Nonetheless, there are some challenges which are inherited from the subject under study due to unique characteristics of the data which should be identified, understood and dealt with appropriately. Class-imbalance issue is one of the main problems of this kind, which is present in many natural or other nonlinear dynamical systems. This is often due to the nature of the events, not the data collection process.

Class-imbalance in spite of being a well known issue is often not treated properly. This is particularly true when the primary objective is not machine learning per se but the testing and scrutiny of domain-specific theories. The complexity of the problem at hand and the absence of data experts very often underestimate the needed level of care, resulting in unrealistic and unreliable analyses.

In this study we present an example of an extremely imbalanced dataset, namely the time series features of solar magnetic data, and examine flare prediction with the goal of showing the footprint of extreme class-imbalance on real-world problems. In addition, we show the impact of disregarding an interesting phenomenon called "temporal coherence" in spatiotemporal datasets. We demonstrate the impact and biases of different approaches and discuss how they should be interpreted from the perspective of the subject under study by involving domain experts. We hope that this work raises awareness to interdisciplinary researchers and enables them to spot and tackle similar problems in their respective areas.

II. Common Challenges in Solar Flare Prediction

In spite of more than 20 years of research and meaningful advances, solar flare prediction remains an outstanding problem. In the following sections, we present a few challenges commonly faced while tackling it. Predicting (or forecasting) the occurrence of solar flares is a typical 21st century rare-event task. Flares are sudden and substantial enhancements of high energy electromagnetic radiation (like extreme ultra violet and X-rays) at local solar scales which pose a threat to humans and equipment in space. They are automatically detected and classified by the National Oceanic and Atmospheric Admin-

istration's (NOAA) constellation of GOES satellites based on their peak flux in soft X-ray wavelengths on a logarithmic scale as A-, B-, C-, M- and X-class solar flares. A and B-class flares are difficult to distinguish from the random variations in the Sun's background X-ray level, but C-class flares and above are detected reliably by GOES satellites. The most intense of these classes, namely M and X, are most often targeted for prediction due to their potentially adverse space-weather ramifications.

A. Extreme Class-Imbalance

The frequency distribution of the peak X-ray fluxes of flares is nearly a perfect power law with a dynamical range spanning several orders of magnitude. A statistical analysis of NOAA's flare reports during solar cycle 23 (1995 to 2008) shows that around 50% of active regions produce C-class flares, while 10% produce M-class flares and less than 2% produce X-class flares. Solar cycle 24 (2009 to present), from which our data discussed in Section 3 are taken [1], exhibit a much weaker major flare crop, making class-imbalance a conspicuous problem to deal with (for a review, see [2]).

B. Point-in-time vs Time Series Forecasting

Solar flare forecasting has been humanity's first attempt toward space weather forecasting. As such, numerous magnetic properties and forecast methods have been proposed since the early 1990s [3], [4]. A quick perusal of the voluminous literature, however, will show that the vast majority of these methods correspond to point-in-time forecasting, namely, to using the instantaneous value of one or more parameters in order to produce a binary or probabilistic flare forecast over a preset forecast horizon. However, flares are an inherently dynamical phenomenon, with clear pre-flare and post-flare phases, characterized by certain evolutionary trends [5], [6]. Because of this, time series of aspiring flare forecasting parameters should be used, rather than isolated points in time. We believe that the sheer level of difficulty of this task was the key factor for the (over-)simplifying point-in-time assumption. However, it may be precisely this compromise that may have hampered non-incremental progress toward flare prediction. Therefore, our main goal in this study is to explore the difficulties which arise when we take into account the temporal evolution of magnetic parameters of active regions rather than looking at a single snapshot in time.

C. Non-representative Datasets

While there is no shortage of satellites and instruments which map the magnetic field of the Sun's photosphere over the past 25 years [7]–[11] we should not forget that (i) the temporal span of high-quality solar data is still limited and (ii) training forecast methods on certain parts of a solar cycle is not necessarily optimal for forecasting other parts of the same and/or different cycles, due to the continuously modulating background of magnetic activity. Therefore, there exists a problem of heterogeneous and/or non-representative data. Coupled with simpler, yet unjustified, selections of random undersampling for majority class events, the sampled subsets of data fail to become representative of the overall flare population.

III. SWAN-SF Dataset: A Multivariate Time Series Data

Multiple flare prediction studies [3], [12], [13] and the European Union FLARECAST project [14], [15] emphasize machine learning for flare prediction, but they use point-in-time measurements. Up until now, we are unable to determine whether the current imperfect accuracy measurements or skill scores are a result of dataset specifics (i.e., point-in-time use) or of the quality of the machine learning models themselves. Here we will use a benchmark dataset, named as Space Weather ANalytics for Solar Flares (SWAN-SF), released recently by [1] and made entirely of multivariate time series, aiming to carry out an unbiased flare forecasting and hopefully set the above question to rest.

The five partitions of SWAN-SF dataset (see Fig. 1) are temporally separated so that the partitions contain approximately an equal number of X- and M-class flares. The data points in this dataset are time series slices of physical (magnetic field) parameters extracted from the flaring and flare-quiet regions, in a sliding fashion. That is, for a particular flare with a unique id, k equal-length multivariate time series are collected from a fixed period of time in the history of that flare. This period is called an *observation window*, denoted by T_{obs}, and spans over 24 hours. Given that t_i indicates the starting point of the i-th slice of the multivariate time series, the $(i+1)$-th slice starts at $t_i + \tau$, where $T_{obs} = 8\tau$. This kind of sliding observation inherits the fact that very often the k slices behave very similar due to their temporal and spatial closeness. In other words, the physical parameters describing the behavior of the region corresponding to a particular flare are not significantly different from one slice to the next. These similar slices, if described in the multi-dimensional feature space of our data, are located too close to each other to be considered distinct instances. And their closeness does not reflect any characteristics of those data points, except our slicing methodology. We refer to this phenomenon as *temporal coherence*[1] of data. This is a key concept to understand some of the challenges we would like to address in this study.

A. Final Forecast Dataset

To use the benchmark SWAN-SF, two main approaches might be taken: One is to preprocess the time series and feed them directly into supervised models. The other is to extract a set of statistical features from the time series and then train the models based on the derived descriptors. Our interest in the analysis of different sampling methods guides us to use the extracted features.

Statistical Features To the best of our knowledge, SWAN-SF data is the first flare data benchmark that focuses on time series, rather than point-in-time data points for a period just

[1] We introduce this concept in the context of data manipulation, and it should not be confused with *temporal coherence* in Optics or any other topic.

short of a decade. Therefore, perhaps except in a few instances, there are no established theories as to which characteristics of the time series may show a significant flare-predictive capability. We hereby build a prediction dataset relying on the set of first four statistical moments of the time series, namely, their *mean*, *variance*, *skewness*, and *kurtosis*. To allow comparison with previous studies, we also consider a point-in-time feature, namely the *last value* of each time series. Moreover, we also add *median* to the list to compensate for the effect of outliers on *mean*.

The obtained dataset of the extracted features has a dimensionality of 144, resulting from the computation of the 6 above-mentioned statistics on the 24 physical parameters of the SWAN-SF dataset. Data points of this dataset are labeled by 5 different classes of flares, namely GOES X, M, C, B, and N. The latter represents flare-quite instances or GOES A-class events.

Throughout this study, we only use the *last value* feature to keep the number of variables low. At the end, however, we present the contribution of other statistical features as well to show the benefit of using time series instead of point-in-time data instances. We also conduct our experiments on a binary class data by merging X and M classes into a super-class called XM, and C, B, and N classes into another super-class denoted by CBN. This simplification allows us to only focus on the challenges we mentioned before, which is the primary objective of this study.

Preprocessing After computing the above-mentioned features, the dataset requires a minimal preprocessing due to the presence of some missing values. Since this accounts for a very small fraction of the data (i.e., $< 0.01\%$), we simply utilize linear interpolation to reproduce those values. In addition, we use zero-one data transformation to normalize the data for our experiments, since otherwise the optimal hyperplanes found by SVM will be meaningless.

IV. Class-imbalance and Temporal coherence

In this section, we discuss different challenges for machine learning algorithms, caused directly or indirectly by two important characteristics of our dataset, namely, class-imbalance and temporal coherence. Without lose of generality, we use the SVM classifier which, like many other learners, is sensitive to these issues. Therefore, instead of analyzing the specific characteristics of this particular classifier, we focus on the common denominators of the well-known classifiers in view of imbalanced datasets.

A. Class-imbalance Problem

In class-imbalanced data, the population of one or more data classes is far less than that of the majority class(es). In situations of significantly less dense data classes, special treatment is required, knowing that machine learning models generally perform best when classes are roughly equal in size. Here we use the terms "minority", or "positive", class to refer to the less frequent group and "majority", or "negative", class, conversely. Stronger flares (GOES X- and M- class) form our minority class, and weaker events (GOES C-, B- and A-class) belong to our majority class. Fig. 1 illustrates the distribution of all classes in each partition.

Classification models, in general, aim to reduce the cost of their objective function by minimizing the total number of misclassifications. In an imbalanced dataset, since the density of the majority class is significantly higher than that of the minority class, many instances of the majority class should be sacrificed (i.e., misclassified) for a correct classification of an instance from the minority class. The SVM classifier in particular, searches for optimal hyper-planes to make such separations. An imbalanced dataset most likely preserves the imbalanced density of the classes even close to the decision boundaries (where the ideal class regions overlap or meet). In such a situation, a hyper-plane that is supposed to pass through the boundaries will be shifted into the region of the minority class to reduce the total number of incorrect classifications/predictions by getting all the positive classes right. This leads to higher true-negatives (i.e., correct predictions of CBN-class flares) and lower true-positives (i.e., correct predictions of XM-class flares). In other words, a model in a class-imbalance space always favors the majority class. This is of particular concern because virtually all class-imbalance problems aim to predict minority, rather than majority, events.

Another angle to this problem is to decide on the right choice of a performance measure. Many well-known performance metrics are significantly impacted by class-imbalance, including accuracy, precision (but not recall), and the f1-score. This is mainly because these measures ignore the number of misclassifications. For instance, a model that classifies all instances as the negative (majority) class may result in a very high (often asymptotic to 1.0) accuracy, while learning little or nothing about the minority class. For the particular case of class-imbalance there are defined less susceptible measures such as TSS (True Skill Statistic[2] [16]) or the HSS (Heidke Skill Score [17], [18]), with TSS being reportedly more robust for solar flare prediction [13]. In this study, we use only TSS since our main objective is to show the changes in the models' performance and not to find an operational-ready model.

Undersampling and Oversampling A simple approach tackling the class-imbalance issue is to enforce a balance between classes by *undersampling* (that is, taking out instances from the majority class) or *oversampling* (that is, providing more instances to the minority class by replication). This results in using roughly only as many negative instances (majority) as there are positive instances (minority) in the training phase, thus achieving a perfect $1:1$ balance ratio. This solution, however, comes at some cost. When undersampling, for instance, we leave out a great portion of the data during training, therefore not learning from the entire collection. To avoid the enormous data waste, a very large dataset should be available overall. When oversampling, we add replicates of existing instances. This may cause a model to memorize data structures

[2]This is also known as Hanssen-Kuipers Discriminant.

TABLE I
DIFFERENT UNDERSAMPLING AND OVERSAMPLING APPROACHES APPLIED TO *Partition* 3 OF THE SWAN-SF DATASET, SHOWCASING THE VARYING EXPANSION/SHRINKAGE FACTORS FOR DIFFERENT CLASSES.

Method	Expansion/Shrinkage Factor					Description
	X	M	C	B	N	
Undersampling 1	1.00	1.00	0.05	0.05	0.05	preserves climatology in sub-class level
Undersampling 2	1.00	0.14	0.03	0.98	0.00	X-base undersampling; enforces a sub-class balance
Undersampling 3	7.20	1.00	0.23	7.11	0.03	M-base undersampling; enforces a sub-class balance
Oversampling 1	19.58	19.58	1.00	1.00	1.00	preserves climatology in sub-class level
Oversampling 2	7.66	7.66	1.00	1.00	0.30	same as oversampling 1 but it suppresses N
Oversampling 3	31.41	4.36	1.00	31.02	0.15	C-base oversampling; enforces a sub-class balance
Oversampling 4	208.46	28.95	6.64	205.89	1.00	N-base oversampling; enforces a sub-class balance

instead of generalizing and learning about them which is, as expected, very prone to overfitting.

While both of these techniques seem fairly straightforward and easy to implement, one should be extra careful when applying them to a multi-class data, such as the flare dataset. This is true despite the fact that we converted it to a binary class problem. Alternative avenues exist depending on whether the balance in the sub-class level (i.e., $|\mathrm{X}| = |\mathrm{M}|$ and $|\mathrm{C}| = |\mathrm{B}| = |\mathrm{N}|$) is also required or not. Notice that this is an addition to the primary goal of our sampling which aims to achieve a balance between the super-classes. When undersampling, for instance, if this additional balance is decided, we first need to decide which class in the minority group is considered the "base" class. Letting GOES X be the base class, we must undersample from M-class flares first to balance X and M classes and then undersample from the majority (CBN) class. This yields a balanced dataset in both super-class and sub-class levels. For convenience, we call this undersampling method an X-based undersampling. To show a few different sampling methodologies, we list some results on *Partition* 3 of the SWAN-SF dataset, in Table. I.

A quick look at Table. I shows that the choice of the sampling method plays a critical role. Knowing that GOES C class represents the strongest flares in the weak-flare class (majority), it is expected that a higher fraction of C-class instances in this group results in a harder prediction problem for the model. In other words, a sampling methodology has to contort the climatology of flares to achieve the desired balance. This change affects the distribution of samples in the feature space by making the decision boundary (where GOES C and M classes overlap) denser or sparser. As an example, Oversampling 3 and 4 may result in a slightly easier data for a predictive model since it replicates more X-class flares than M-class flares relative to what the flare climatology suggests.

The above suggests that performance of different models on the same dataset is only comparable if they all employ identical sampling methodologies.

Misclassification Weights Another classical remedy for the class-imbalance problem is penalizing misclassification of different classes differently. SVM, like some other machine learning algorithms, can incorporate different weights in its objective function. For details on how this is mathematically implemented, we refer the interested readers to [19]. For the experiments in this study, inspired by the class-imbalance ratio, we adjust the weights using $w_j = \frac{n}{k \cdot n_j}$, where n is the total population, n_j is the population in class j, and k represents the number of classes.

B. Cross Validation

Cross validation is a family of statistical techniques, typically used to determine the generalization power of a model. Regardless of the employed cross validation technique (k-fold, leave-p-out, stratified or purely random), it is very often assumed that a random selection is allowed. However, in many real-world data collections, random sampling must take into account the spatiotemporal characteristics of the data, which is rooted in the temporal (or spatial) coherence of data. The temporal coherence in SWAN-SF dataset, as we discussed in Sec. III, prohibits this practice, since it yields an overly optimistic performance of the model's generalization, a phenomenon known as 'overfitting'.

To avoid this mistake, we select training and testing instances from different partitions of dataset. To showcase the impact of randomly splitting the data from one partition, we conduct an experiment that is presented in Sec. V-C.

Validation Set Disregarding the temporal coherence of the data and random sub-sampling for obtaining the training and testing sets has an additional and perhaps a more important impact: when the test set is obtained by randomly splitting the data, the tested model obscures the overfitting sign, i.e., a significant difference between the training and testing performance. Therefore, the use of any sorts of sampling methodology on the test set (either for reducing the imbalance ratio, as we discussed before, or in cross validation) must be avoided at all costs if it distorts the actual distribution of the events. Cross validation is also used for optimization of models' hyper-parameters or the data-driven parameters. We wish to reiterate that for all such tasks, another subset of data should be used which is known as the validation set, and the test set must never be exposed to the model except for reporting the final performance of the model. For instance, to tune SVM's hyper-parameters, namely c and γ for achieving an optimal hyper-plane by the model, a validation set must be defined in order to reflect the changes that the model takes in. Based on our expectation of the acceptable performance, we may let the model improve upon the validation set's feedback. Only when we believe that the model has reached its highest performance we can use the test set to measure its robustness.

Similar to the capital impact of different sampling methodologies on performance, as we discussed previously, sampling modification of the test set is another way of creating non-robust models with seemingly high performance.

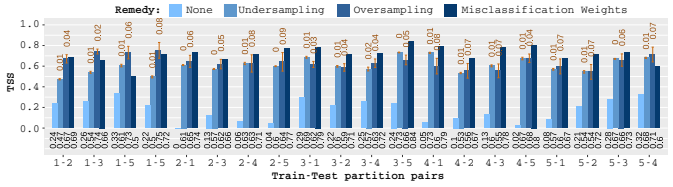

Fig. 2. Experiments Z, A, B, and C. TSS of SVM trained and tested on all possible permutations of partition pairs, using three different remedies for the class-imbalance issue (oversampling, undersampling and misclassification weights) and compared to a baseline where no remedy was employed. Each set of columns includes the partition pairs (gray font) the mean TSS values (black font) and their respective variance (orange font).

C. Hyper-parameter Tuning

Temporal coherence of the dataset also affects the way we tune the hyper-parameters of our models. Any supervised learning model requires optimization of its hyper-parameters in a data-driven manner. Since the discovered hyper-parameters should remain optimal over the entire dataset (including the upcoming data points for a predictive model) the training and validation sets, as well as the test set, must each be representative of the entire dataset. In our dataset, because of the temporal coherence random sampling does not produce such subsets. Hence, tuning process remains confined to the partitions. Although the solution we proposed for cross validation using NOAA AR Number, may also be used to tackle this problem, it is very likely that a model highly optimized on one partition simply does not perform well globally. Therefore, we believe that this is a problem yet to be investigated more thoroughly since at this point it is very clear to us that flare forecast problem has a dynamic and periodic behavior, for which ensemble models may be more appropriate.

V. Experiments, Results, and Interpretations

In this section, we present all experiments conducted to showcase the challenges discussed previously. We also elaborate on the interpretation of these experiments in regards to our overarching flare forecasting task.

Notice that the objective of this study is not to achieve a robust model with high performance, but to compare models trained differently. Therefore, although any changes on the data (e.g., using different normalization, data split, or sampling techniques) require re-tuning of the hyper-parameters, without loss of generality we rely on our pre-tuned hyper-parameters for SVM: $c = 1000$, $\gamma = 0.01$, with a Radial Based Function (RBF) kernel.

We would like to stress that we choose to use TSS as our performance measure only to show the changes that different treatments cause. A higher TSS is not necessarily evidence of a better forecast model, as it may well be coupled with a very low HSS.

A. Baseline

To establish a baseline for the experiments, the model first needs to learn from the available data without any special treatment at the data input process or on the model configuration.
Experiment Z: Baseline This experiment is as simple as training SVM on all instances of one partition and testing the model on another partition. We try this on all possible partition pairs, resulting in 20 different trials, to illustrate how the difficulty of the prediction task varies as the partitions are chosen from different parts of the solar cycle. The results are visualized in Fig. 2, along with the impact of discussed class-imbalance remedies that we further discuss in the following sections.

B. Tackling Class-imbalance Issue

In Sec. IV-A, we discussed three different approaches towards tackling the class-imbalance problem. To show the impact of each solution, we carry out some experiments and discuss the results below.

The following experiments share a common structure: SVM is trained and tested on all 20 permutations of partitions pairs independently. In each round, the model learns from instances in the training partition and is then tested against the (different) testing partition. To measure the confidence of a model's performance when a certain sampling method is employed, we repeat the experiment 10 times and report the variance and mean value of TSS.
Experiment A: Undersampling In the fitting phase, the model takes in a subset of the training partition generated by a X-based undersampling method (Table. I; Undersampling 2). This enforces a $1:1$ balance not only in the super-class level (i.e., $|\text{XM}| = |\text{CBN}|$) but also in the sub-class level (i.e., $|\text{X}| = |\text{M}|$ and $|\text{C}| = |\text{B}| = |\text{N}|$). The trained model is then tested against all other partitions one by one to examine the robustness of the model. The undersampling step is only taken in the training partition, as undersampling of the test partition distorts reality and would not reflect the true model performance. The consistent and significant impact of this remedy is evident in Fig. 2.
Experiment B: Oversampling Similar to Experiment A, but using Oversampling 3 of Table. I instead. Again, no over- or undersampling takes place in the testing set. Comparing the results of oversampling with undersampling in Fig. 2, shows a close correspondence between the two models in terms of their mean TSS values; typically, differences are within applicable uncertainties.
Experiment C: Mis-classification Weights We use the imbalance ratio of the super-classes as the weights. For instance when working with *Partition* 3, since the minority-to-majority ratio is $1:20$, we set $w_{\text{XM}} = 20$ and $w_{\text{CBN}} = 1$. As shown in Fig. 2, this solution outperforms both undersampling and oversampling approaches in terms of their TSS. It is worth pointing out that employing misclassification weights has the advantage of a data-driven tunability that may be better suited than over- and undersampling to achieve more robust forecast models.

C. Impact of Cross Validation

In Sec. IV-B, we discussed the theoretical impact of random sampling, embedded in many cross validation methods, on a temporally coherent dataset. The following experiment is designed to put the validity of this discussion to test.

Experiment D: Data Splits This time, SVM is both trained and tested on randomly chosen instances of the same partition. Technically, this is a k-fold cross validation using a random sub-sampling method with $k = 10$. The results are then juxtaposed with those obtained by training SVM on one partition and testing it on another. We equipped SVM in both scenarios with misclassification weights, to eliminate the need for an additional sampling layer. Therefore, the only determining factor is whether the instances are sampled from the same partition or not. Let it be clear that sampling from a single partition does not mean any overlapping between the training and testing sets.

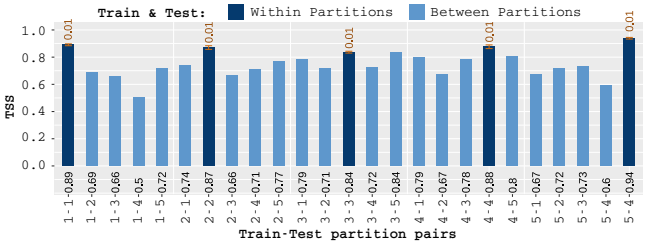

Fig. 3. Experiment D. TSS-values of SVM performance when trained and tested either using a 10-fold cross validation sub-sampling on a single partition (dark blue) or assigning different partitions for training and testing (light blue). Labels and font colors have the same meaning as in Figure 2. In all cases, the SVM has been equipped with misclassification weights.

Fig. 3 presents this comparison. When SVM is trained and tested on a single partition, performance is boosted very significantly with TSS $\in (0.84, 0.94)$ with an average TSS$\simeq 0.88$. Training and testing on different partitions yields TSS $\in (0.50, 0.84)$ with an average TSS$\simeq 0.71$. This remarkable difference should not be viewed as evidence of the robustness of the model but rather points to memorization and hence overfitting, caused when a forecast model is trained and tested on a temporally coherent dataset. It is the lower performance when the model is trained and tested on different partitions that better encapsulates its true robustness.

D. Oversampling Impact

In Sec. IV-A, we showed that there are multiple variants of oversampling and undersampling approaches. We also presented how this affects flare distributions in *Partition* 3 as an example. Below we test how different oversampling impacts TSS values across different partitions.

Experiment F: Oversampling With or Without Sub-Class Balance We use Oversampling 1 and 3 (from Table. I) in the training phase to remedy the class-imbalance problem and then we test the trained model against all other partitions. Our results are shown in Fig.4. For them, one sees a relatively similar, consistent performance, although the C-based Oversampling 3 seems to give a statistically higher performance.

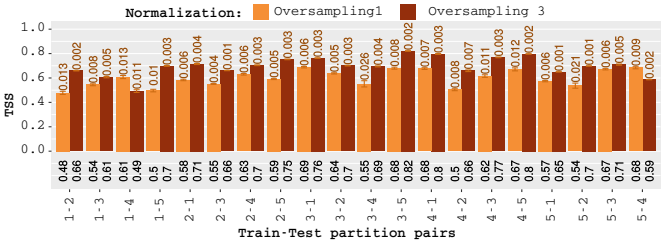

Fig. 4. Experiment F. TSS-values of SVM performance impacted by two different oversampling methods; Oversampling 1 (orange columns), where the climatology of sub-classes are preserved, and Oversampling 3 (burgundy columns), where the sub-classes are forced to reach a $1:1$ balance ratio by considering C-class to be the base. Labels and font colors have the same meaning as in Figure 2.

This said, it becomes clear that different oversampling methods give non-identical performances. Therefore, comparison of any two forecasting models on similar datasets will be fair only if the employed sampling methodologies are identical.

E. Using Other Time Series Features

We reserve the last experiment for presenting the benefit of using time series of SWAN-SF dataset, as opposed to other point-in-time datasets that we were trying to mimic by using *last value* as the statistical feature extracted from the time series in SWAN-SF dataset. Below we compare some other basic statistics.

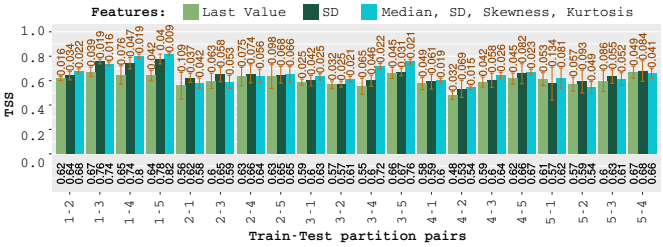

Fig. 5. Experiment G. TSS-values of SVM performance on 3 different feature spaces. Undersampling 2 (from Table. I is used to remedy the class-imbalance issue.

Experiment G: SVM With Other Statistical Features SVM is trained and tested on partitions pairs, using Undersampling 2 from Table. I as a class-imbalance remedy, using (i) *last value*, (ii) *standard deviation*, and (iii) *median*, *standard deviation*, *skewness*, *kurtosis*. As illustrated in Fig. 5, *standard deviation* results in statistically better performance than *last value*, and using the four-number summary seems to outperform *standard deviation*. This is a very good indication that different characteristics of time series carry some important pieces of information that may significantly improve reliability of a forecast model.

It is beyond the scope of this work to find an optimal solution using more statistical features. We therefore leave further investigation in this direction for future studies.

VI. Summary, Conclusions, and Future Work

We used SWAN-SF benchmark dataset as a case study to highlight some of the challenges in working with imbalanced datasets, which are very often overlooked by scientists of the domain. We also addressed an interesting characteristic of some datasets, that we called *temporal coherence*, inherited from the spatial and temporal dimensions of the data. Using several different experiments, we showcased some pitfalls and overlooked consequences of disregarding those peculiarities, and we discussed the impact of different remedies in the context of flare forecast problem.

There are still many interesting cases left to be discussed that we plan to include in our future studies. In several experiments, for instance, we noticed that despite the improvement in models' performances in terms of TSS, other measures such as HSS showed a moderate deterioration in the models. A new measure that reflects both these skill scores appears to be necessary to avoid many misleading interpretations. In spite of many studies, it is still an unsolved problem. Another avenue for further investigation is to incorporate NOAA AR Numbers in the sampling phase so that the temporal coherence of data that confines normalization and hyper-parameter tuning tasks to only subsets of data, can be bridged over.

We hope that this work raises awareness not only to the scientists interested in flare forecast problem, but all interdisciplinary researchers who might be dealing with imbalanced and temporally coherent datasets, and enables them to spot and tackle similar problems in their respective areas.

Acknowledgment

This work was supported in part by two NASA Grant Award [No. NNH14ZDA001N], and one NSF Grant Awards [No. AC1443061 and AC1931555]. The AC1443061 award has been supported by funding from the Division of Advanced Cyber infrastructure within the Directorate for Computer and Information Science and Engineering, the Division of Astronomical Sciences within the Directorate for Mathematical and Physical Sciences, and the Division of Atmospheric and Geospace Sciences within the Directorate for Geosciences.

References

[1] R. A. Angryk *et al.*, "Multivariate time series dataset for space weather data analytics (manuscript submitted for publication)," *Scientific Data*, 2019.

[2] M. J. Aschwanden *et al.*, "25 years of self-organized criticality: solar and astrophysics," *Space Science Reviews*, vol. 198, no. 1-4, pp. 47–166, 2016. [Online]. Available: https://link.springer.com/article/10.1007\%2Fs11214-014-0054-6

[3] G. Barnes *et al.*, "A comparison of flare forecasting methods. i. results from the "all-clear "workshop," *The Astrophysical Journal*, vol. 829, no. 2, p. 89, Sep 2016. [Online]. Available: https://iopscience.iop.org/article/10.3847/0004-637X/829/2/89

[4] M. K. Georgoulis, "On Our Ability to Predict Major Solar Flares," *Astrophysics and Space Science Proceedings*, vol. 30, p. 93, Jan 2012.

[5] A. Benz, "Flare observations," *Living Reviews in Solar Physics*, vol. 5, 02 2008. [Online]. Available: doi.oeg/10.12942/lrsp-2008-1

[6] L. Fletcher, B. R. Dennis, H. S. Hudson, S. Krucker, K. Phillips, A. Veronig, M. Battaglia, L. Bone, A. Caspi, Q. Chen, P. Gallagher, P. T. Grigis, H. Ji, W. Liu, R. O. Milligan, and M. Temmer, "An observational overview of solar flares," *Space Science Reviews*, vol. 159, no. 1, p. 19, Aug 2011. [Online]. Available: https://doi.org/10.1007/s11214-010-9701-8

[7] S. Tsuneta, K. Ichimoto, Y. Katsukawa, S. Nagata, M. Otsubo, T. Shimizu, Y. Suematsu, M. Nakagiri, M. Noguchi, T. Tarbell, A. Title, R. Shine, W. Rosenberg, C. Hoffmann, B. Jurcevich, G. Kushner, M. Levay, B. Lites, D. Elmore, T. Matsushita, N. Kawaguchi, H. Saito, I. Mikami, L. D. Hill, and J. K. Owens, "The solar optical telescope for the hinode mission: An overview," *Solar Physics*, vol. 249, no. 2, pp. 167–196, Jun 2008. [Online]. Available: https://doi.org/10.1007/s11207-008-9174-z

[8] P. H. Scherrer, J. Schou, R. I. Bush, A. G. Kosovichev, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, J. Zhao, A. M. Title, C. J. Schrijver, T. D. Tarbell, and S. Tomczyk, "The helioseismic and magnetic imager (hmi) investigation for the solar dynamics observatory (sdo)," *Solar Physics*, vol. 275, no. 1, pp. 207–227, Jan 2012. [Online]. Available: https://doi.org/10.1007/s11207-011-9834-2

[9] P. H. Scherrer, R. S. Bogart, R. I. Bush, J. T. Hoeksema, A. G. Kosovichev, J. Schou, W. Rosenberg, L. Springer, T. D. Tarbell, A. Title, C. J. Wolfson, and I. Zayer, *The Solar Oscillations Investigation — Michelson Doppler Imager*. Dordrecht: Springer Netherlands, 1995, pp. 129–188. [Online]. Available: https://doi.org/10.1007/978-94-009-0191-9_5

[10] A. Tritschler, T. R. Rimmele, S. Berukoff, R. Casini, J. R. Kuhn, H. Lin, M. P. Rast, J. P. McMullin, W. Schmidt, F. Woger, and D. Team, "Daniel k. inouye solar telescope: High-resolution observing of the dynamic sun," *Astronomische Nachrichten*, vol. 337, no. 10, pp. 1064–1069, 11 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asna.201612434

[11] N. Fox, "Parker Solar Probe: A NASA Mission to Touch the Sun," in *EGU General Assembly Conference Abstracts*, vol. 20, Apr 2018, p. 10345.

[12] M. G. Bobra and S. Couvidat, "Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm," *The Astrophysical Journal*, vol. 798, no. 2, p. 135, 2015. [Online]. Available: doi.org/10.1088/0004-637X/798/2/135

[13] D. S. Bloomfield *et al.*, "Toward reliable benchmarking of solar flare forecasting methods," *The Astrophysical Journal*, vol. 747, no. 2, p. L41, Feb 2012. [Online]. Available: https://iopscience.iop.org/article/10.1088/2041-8205/747/2/L41

[14] F. Benvenuto, M. Piana, C. Campi, and A. M. Massone, "A hybrid supervised/unsupervised machine learning approach to solar flare prediction," *The Astrophysical Journal*, vol. 853, no. 1, p. 90, jan 2018. [Online]. Available: https://doi.org/10.3847\%2F1538-4357\%2Faaa23c

[15] K. Florios, I. Kontogiannis, S.-H. Park, J. A. Guerra, F. Benvenuto, D. S. Bloomfield, and M. K. Georgoulis, "Forecasting solar flares using magnetogram-based predictors and machine learning," *Solar Physics*, vol. 293, no. 2, p. 28, Jan 2018. [Online]. Available: https://doi.org/10.1007/s11207-018-1250-4

[16] F. Woodcock, "The evaluation of yes/no forecasts for scientific and administrative purposes," *Monthly Weather Review*, vol. 104, no. 10, pp. 1209–1214, 1976. [Online]. Available: https://doi.org/10.1175/1520-0493(1976)104⟨1209:TEOYFF⟩2.0.CO;2

[17] G. Barnes and K. Leka, "Evaluating the performance of solar flare forecasting methods," *The Astrophysical Journal Letters*, vol. 688, no. 2, p. L107, 2008. [Online]. Available: https://doi.org/10.1086/595550

[18] J. Mason and J. Hoeksema, "Testing automated solar flare forecasting with 13 years of michelson doppler imager magnetograms," *The Astrophysical Journal*, vol. 723, no. 1, p. 634, 2010. [Online]. Available: https://doi.org/10.1088/0004-637X/723/1/634

[19] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," *Methods in molecular biology (Clifton, N.J.)*, vol. 609, pp. 223–39, 01 2010. [Online]. Available: doi.org/https:10.1007/978-1-60327-241-4_13