Making Stuff Up : Ameliorating Data Scarcity in Flare Forecasting through Synthetic Multivariate Time Series Generation with Deep Learning

Presenter: *Rafal Angryk*

March 2022

Data Mining Lab @ Georgia State University

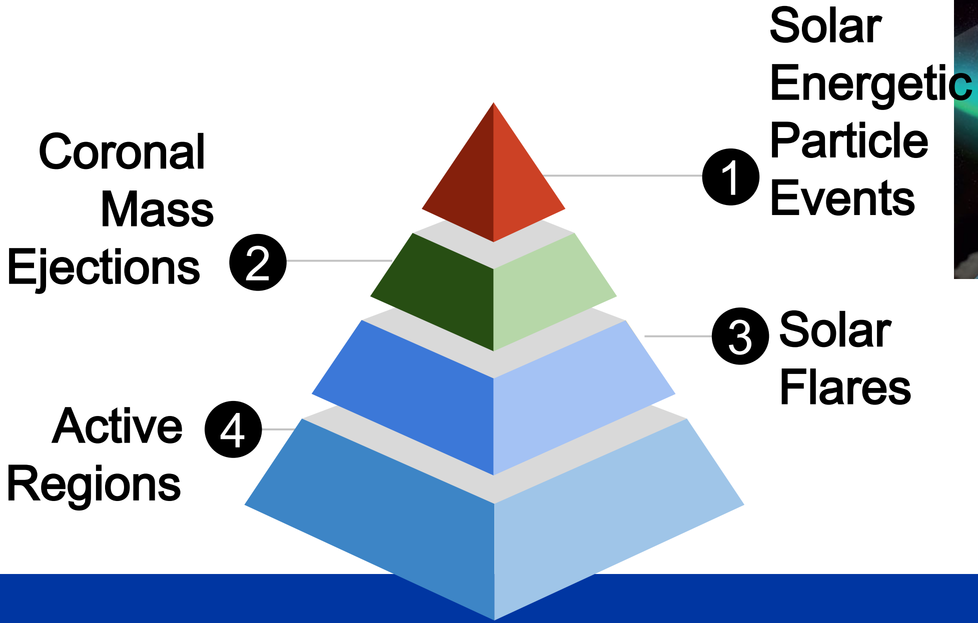
Before we go anywhere, I want to thank to my awesome co -authors from GSU:

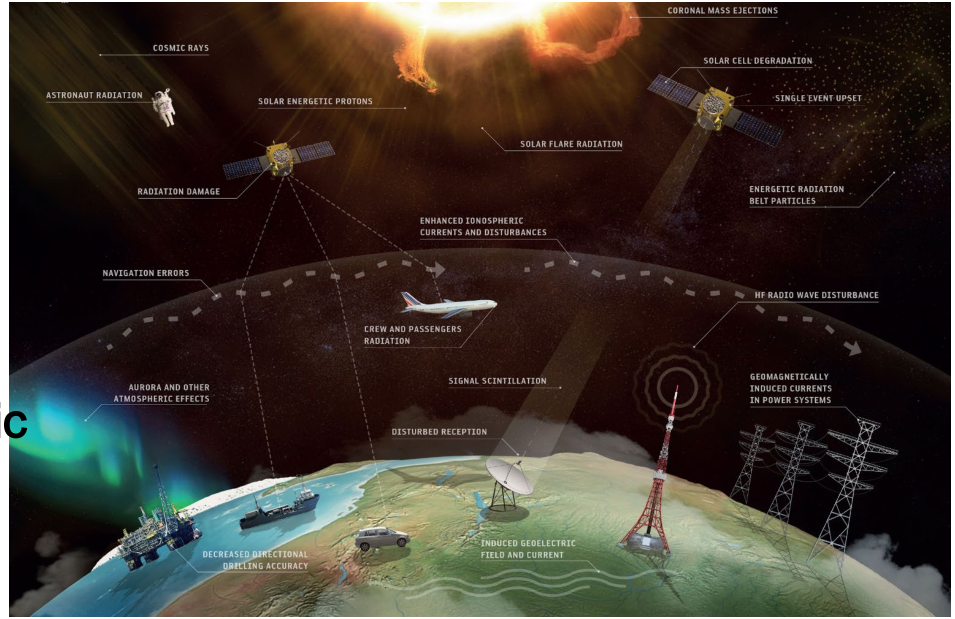
1. Yang Chen, Ph.D. Student
2. Dr. Dustin J. Kempton, Research Assistant Professor
3. Dr. Azim Ahmadzadeh , Postdoctoral Research Associate

MOTIVATION

Motivation (1)

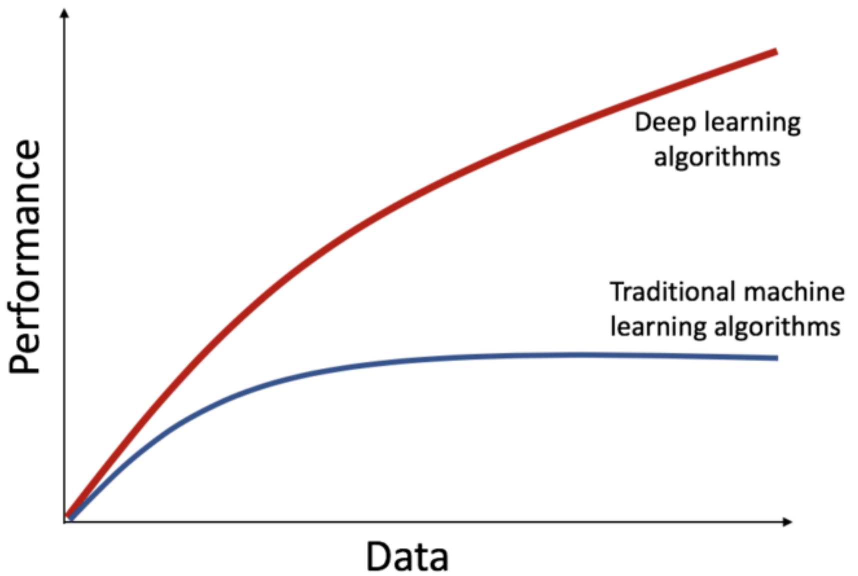
Prediction of Solar Flares is important! ☺





Motivation (2)

ML is so cool, but is data hungry!



Motivation (3)

We have easily accessible big data! ☺

SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

Multivariate time series dataset for space weather data analytics

Rafal A. Angryk[1], Petrus C. Martens[2], Berkay Aydin[1], Dustin Kempton[1], Sushant S. Mahajan[2], Sunitha Basodi[1], Azim Ahmadzadeh[1], Xumin Cai[1], Soukaina Filali Boubrahimi[1], Shah Muhammad Hamdi[1], Michael A. Schuh[1] & Manolis K. Georgoulis[2,3]

HARVARD Dataverse

Add Data ▾ Search ▾ About Us

SWAN-SF

Version 1.2

Angryk, Rafal; Martens, Petrus; Aydin, Berkay; Kempton, Dustin; Mahajan, Sushant; Basodi, Sunitha; Ahmadzadeh, Azim; Xumin Cai; Filali Boubrahimi, Soukaina; Hamdi, Shah Muhammad; Schuh, Micheal; Georgoulis, Manolis, 2020, "SWAN-SF", https://doi.org/10.7910/DVN/EBCFKM, Harvard Dataverse, V1

Cite Dataset ▾ Learn about Data Citation Standards.

Description	We introduce and make openly accessible a comprehensive, multivariate time series (MVTS) data extracted from solar photospheric vector magnetograms in Spaceweather HMI Active Region Patch (SHARP) series. (2019-01-31)
Subject	Astronomy and Astrophysics; Computer and Information Science
Keyword	space weather analytics, flare prediction, multivariate time series

Files | Metadata | Terms | Versions

Search this dataset...

Filter by
File Type: All ▾ Access: All ▾

1 to 7 of 7 Files

addenda.tar.gz
Gzip Archive - 32.5 MB
Published Apr 4, 2020
112 Downloads
MD5: 82c...eae

partition1_instances.tar.gz
Gzip Archive - 1.2 GB
Published Apr 4, 2020
410 Downloads
MD5: b54...e4f

4,098 MVTS data instances extracted from solar photospheric vector magnetograms in Space weather HMI Active Region Patch (SHARP) series, spans **~8 years (05/2010 – 08/2018,** includes **51 parameters**, integrates **over 10,000 flare reports**.

Motivation (4)

Do we though?

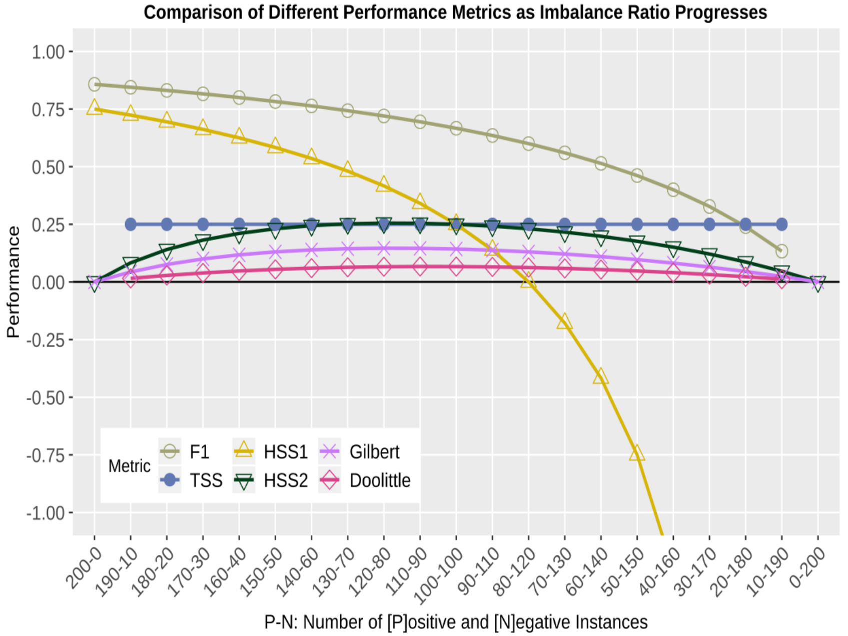
Each partition of SWAN-SF contains approximately an equal number of X- and M-class flares.

As we can see, an extreme imbalance ratio between flare samples (X and M) and non-flare samples (C, B and N) exists in every partition of SWAN-SF

Motivation (5)

Only a few measure out there are truly insensitive to class imbalance.

Truth Table/Confusion Matrix for the model does not change (it correctly predicts 75% of positive instances, and 25% of negative instances), but the class balance does (imbalance transforms from (p = 0,n = 200) to (p = 200,n = 0).



THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES, 254:23 (15pp), 2021 June https://doi.org/10.3847/1538-4365/abec88

CrossMark

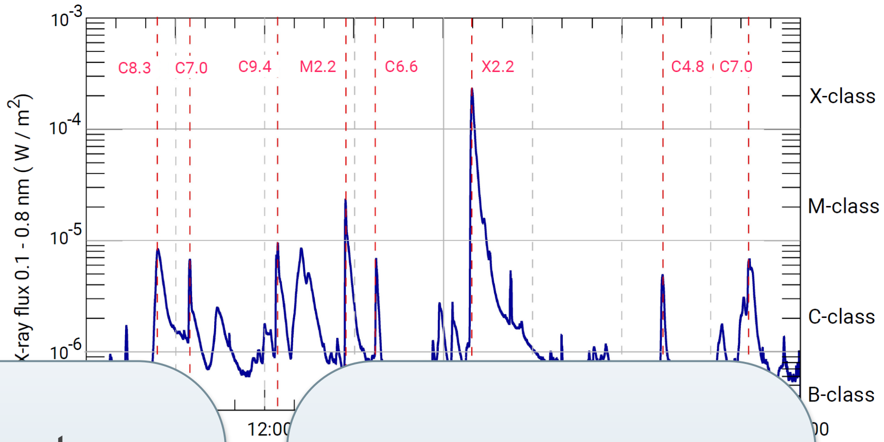
How to Train Your Flare Prediction Model: Revisiting Robust Sampling of Rare Events

Azim Ahmadzadeh[1], Berkay Aydin[1], Manolis K. Georgoulis[2], Dustin J. Kempton[1], Sushant S. Mahajan[3], and Rafal A. Angryk[1]
[1] Georgia State University, Atlanta, GA 30302, USA; aahmadzadeh1@cs.gsu.edu
[2] RCAAM of the Academy of Athens, Soranou Efesiou 4, Athens, GR-11527, Greece
[3] Institute for Astronomy, University of Hawai'i at Mānoa, Honolulu, HI, USA
Received 2021 January 31; revised 2021 March 1; accepted 2021 March 5; published 2021 May 17

Abstract

We present a case study of solar flare forecasting by means of metadata feature time series, by treating it as a prominent class-imbalance and temporally coherent problem. Taking full advantage of pre-flare time series in solar active regions is made possible via the Space Weather Analytics for Solar Flares (SWAN-SF) benchmark data set, a partitioned collection

Problem definition



In majority of real-life data cases, we have to deal with an **extreme class-imbalance issues. SWAN-SF is no exception.**

Significant imbalance may affect any ML classifier by injecting a bias towards the majority classes.

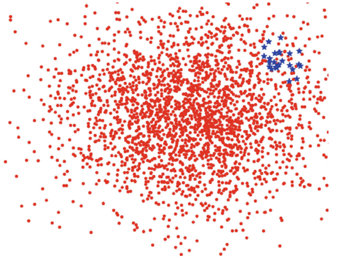
Research Question:

- **Can we figure out a proper treatment for this issue?**

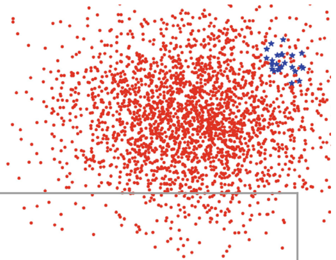
BACKGROUND

Data augmentation for Image Data

No.	Method	Description
1	**Oversampling/ Undersampling**	The simplest remedies for constructing a balanced dataset, but they don't introduce any new information.
2	**Transformation-based techniques**	Performing one or more of data transformations on the existing images. However, these transformations are not applicable to all situations. • For example, the chirality of a solar filament image would be changed if a reflection or affine transformation is performed.
3	**Generative adversarial network (GAN)-based Algorithms**	Providing an alternative way to perform the data augmentation. • To learn an underlying distribution of real samples. • To produce synthetic samples based on the learned distribution.



Data augmentation for TS Data



No.	Method	Description
1	**Oversampling/ Undersampling**	The simplest remedies for constructing a balanced dataset, but they don’t introduce any new information.
2	**SMOTE/RUSO/ RNSO/RNOSO**	They provide statistical interpretations with its generated samples. They only generate point-in-time synthetic samples, not time series.
3	**CGAN - Conditional generative adversarial network**	Generating informative synthetic time series data based on real data. Controlling the category of generated samples. • This allows us to mitigate the class-imbalance issue by generating the samples of minority classes (e.g. flare samples). Providing stable and faster training compared to the vanilla GAN.

- SMOTE: Synthetic Minority Over-sampling Technique.
- RUSO: Random Uniform Synthetic Oversampling.
- RNSO: Random Normal Synthetic Oversampling.
- RNOSO: Random NOise Synthetic Oversampling.

GENERATIVE ADVERSARIAL NETWORKS

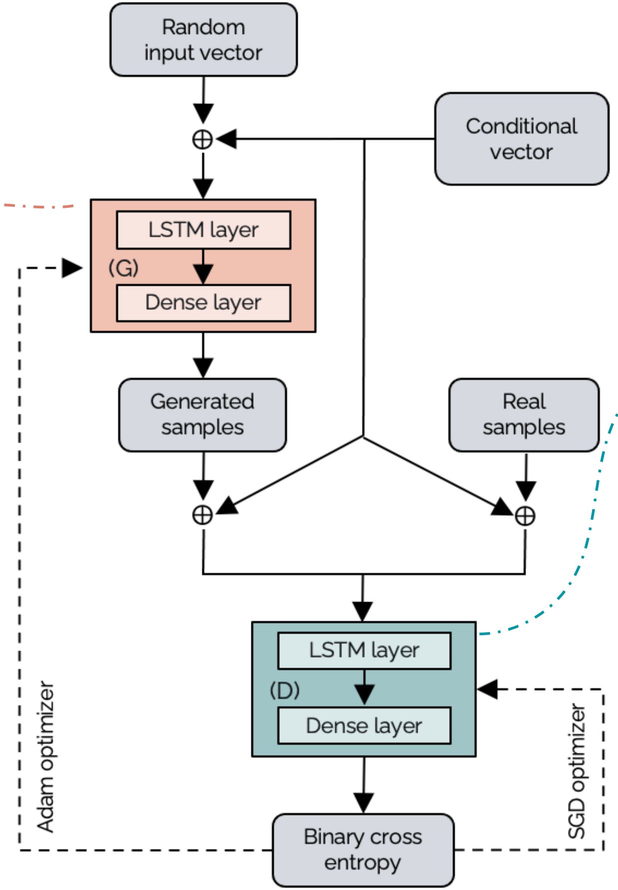
CGAN Algorithm

Generator (G)

- Input-1: Random input vector.
- Input-2: Conditional vector.
- Output: Generated samples
- **Goal:** Generating samples as realistic as possible.

Conditional vector: Labels, encoded into the one -hot representation.

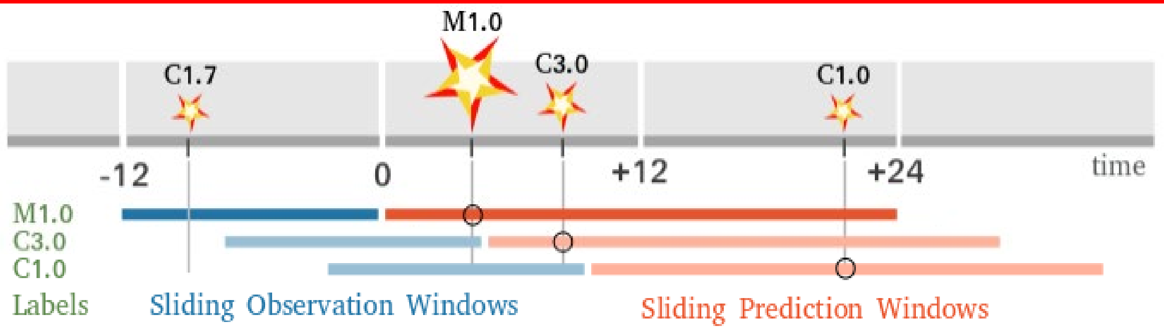
FL	[1, 0]
NF	[0, 1]



Discriminator (D)

- Input-1: Generated and real samples.
- Input-2: Conditional vector.
- Output: Predictions of inputs (e.g. synthetic or real).
- **Goal:** Maximizing its ability of differentiating synthetic and real samples.

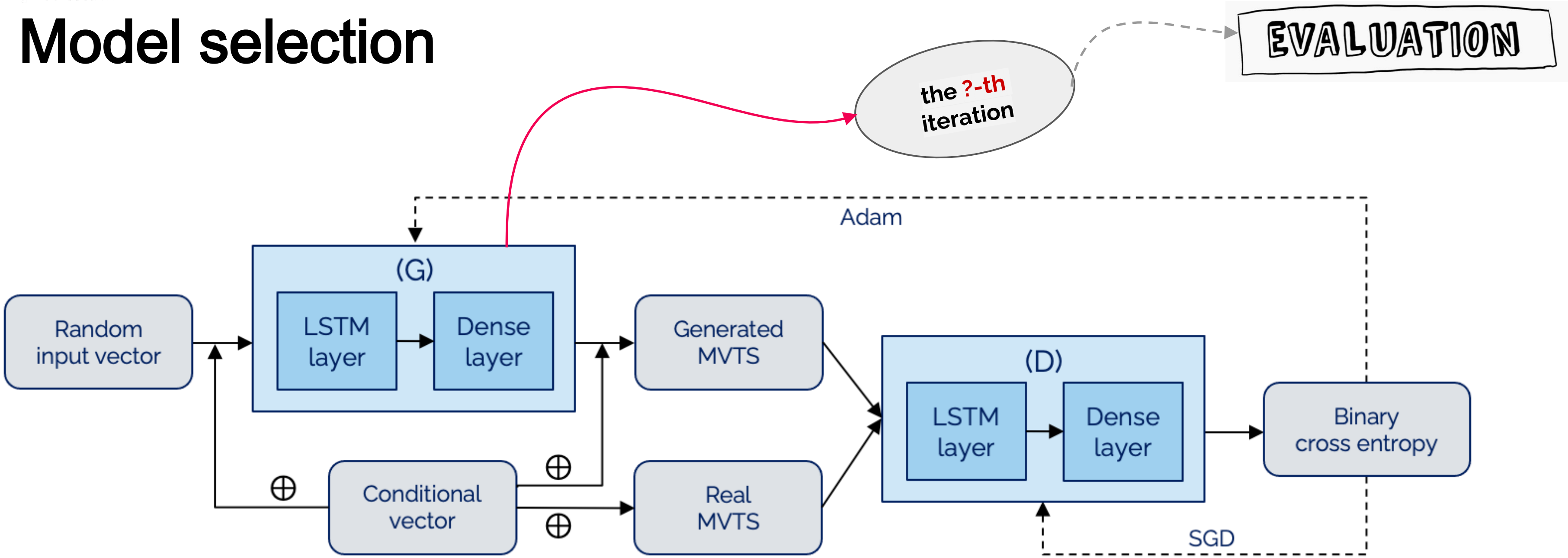
Experimental settings



- All experiments are conducted using MVTS samples from **SWAN-SF** dataset:
  - (1) Training CGAN models (on Partition 1 from SWAN-SF)
  - (2) Evaluation of generated samples.
- Only **four** magnetic field parameters were selected for this study:

No.	Param	Description
1	TOTUSJH	Total unsigned current helicity.
2	ABSNJZH	Absolute value of the net current helicity.
3	SAVNCPP	Sum of the absolute value of the net current perpolarity.
4	TOTBSQ	Total magnitude of Lorentz force.

Model selection



- To determine which DNN model should be used to generat e synthetic MVTS samples.

Evaluation (1)

- **Kullback –Leibler Divergence (DKL)**
- When given sets of real (T) and synthetic (S) samples, with equal number of multivariate time series. For each instance, we extract its mean, median, and standard deviation. We then construct the corresponding probability distributions PT and PS, with setting the bin size to M=20. To quantitatively measure the similarity, we calculate the Kullback –Leibler (KL) divergence between distributions of PT and PS:

$$D_{KL}(P_T || P_S) = \sum_{m \in M} P_T(m) \cdot log\left(\frac{P_T(m)}{P_S(m)}\right)$$

- **The KL divergence is a non -negative measure, which means** DKL(PT ||PS) ⩾ 0. The smaller value indicates the higher similarity **between PT and PS.**

Evaluation (2)

$$AA_{TS} = \frac{1}{2}(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(d_{TS}(i) > d_{TT}(i)) + \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(d_{ST}(i) > d_{SS}(i)))$$

- Adversarial Accuracy (AA)
- Is used for measuring the similarity of two sets of data samples through their nearest neighbors. The distance function d is defined in as the minimum (Euclidean) distance between each real sample and all synthetic samples

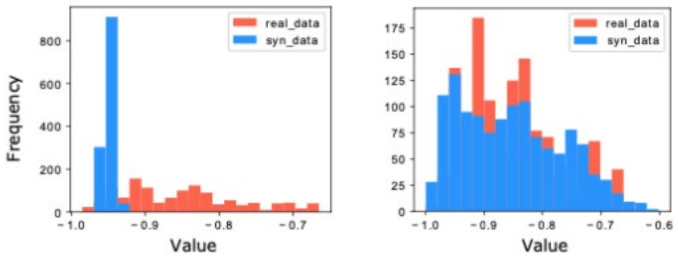
$$\begin{cases} d_{TS}(i) = min_j ||X_T^i - X_S^j||_2 \\ d_{TT}(i) = min_{j,j\neq i} ||X_T^i - X_T^j||_2 \end{cases}$$

- The range of AA is [0, 1]: The outcome 1 indicates that there is no resemblance between the set of real samples and the set of synthetic samples. The outcome 0 indicates that the two sets are exactly the same, yielding no new information.
- The desirable outcome of AA is close to 0.5, implying that the real and synthetic samples generated by the generators are indistinguishable.

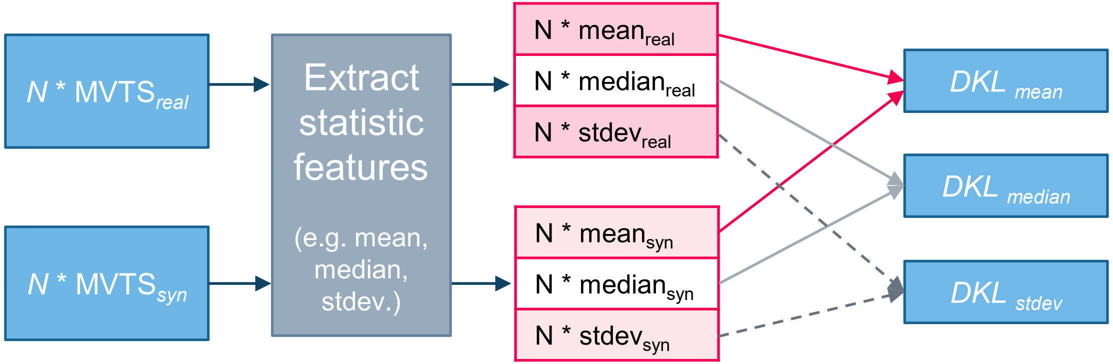
Model selection on DKL

- Based on descriptive statistics of synthetic MVTS samples generated at different iterations.
- TOTUSJH of 1,254 real M/ X flares from Partition1, vs. TOTUSJH of 1,254 artificially generated ones

TOTUSJH



Mean



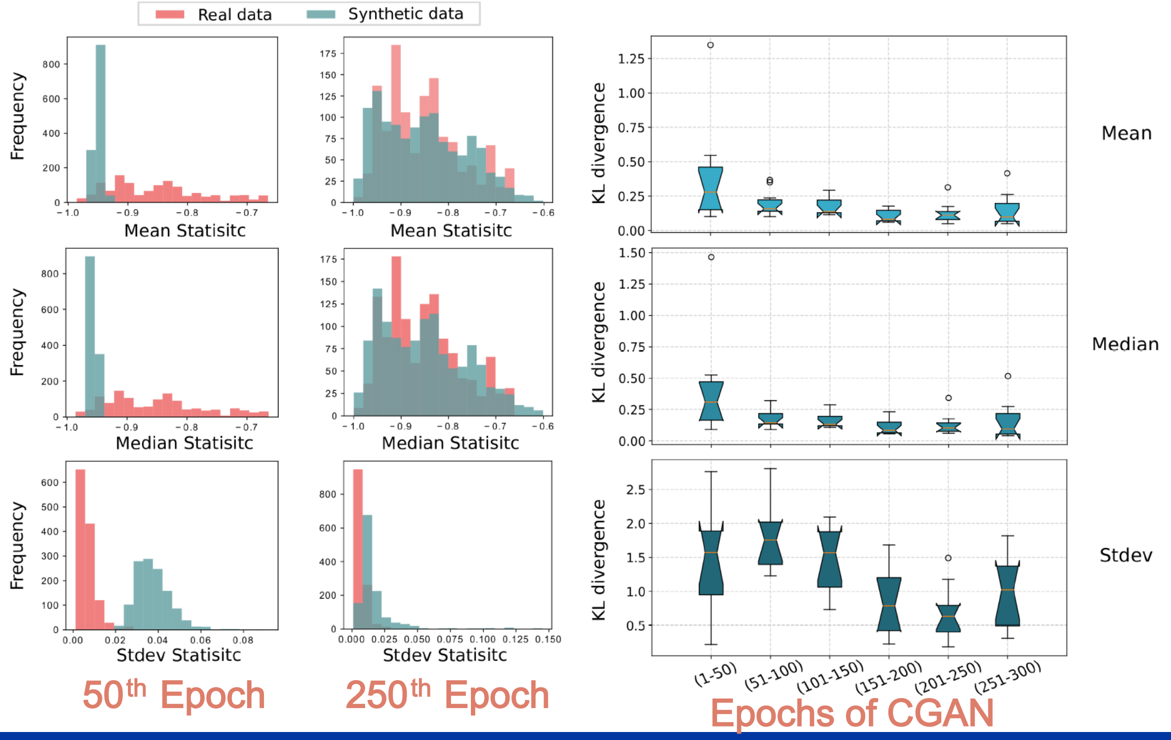
Kullback –Leibler Divergence - The smaller the better.

TOTUSJH	Total unsigned current helicity.

Model selection on DKL

TOTUSJH	Total unsigned current helicity.

- TOTUSJH of 1,254real M/X flares from Partition1, vs. TOTUSJH of 1,254 synthetic ones
- The KL divergence is a non-negative measure, which means DKL(PT ||PS) $\geqslant$ 0.
- The smaller value indicates the higher similarity between PT and PS.
- Last column show the distributions of KL divergence scores across all intermediate models divided into six groups.

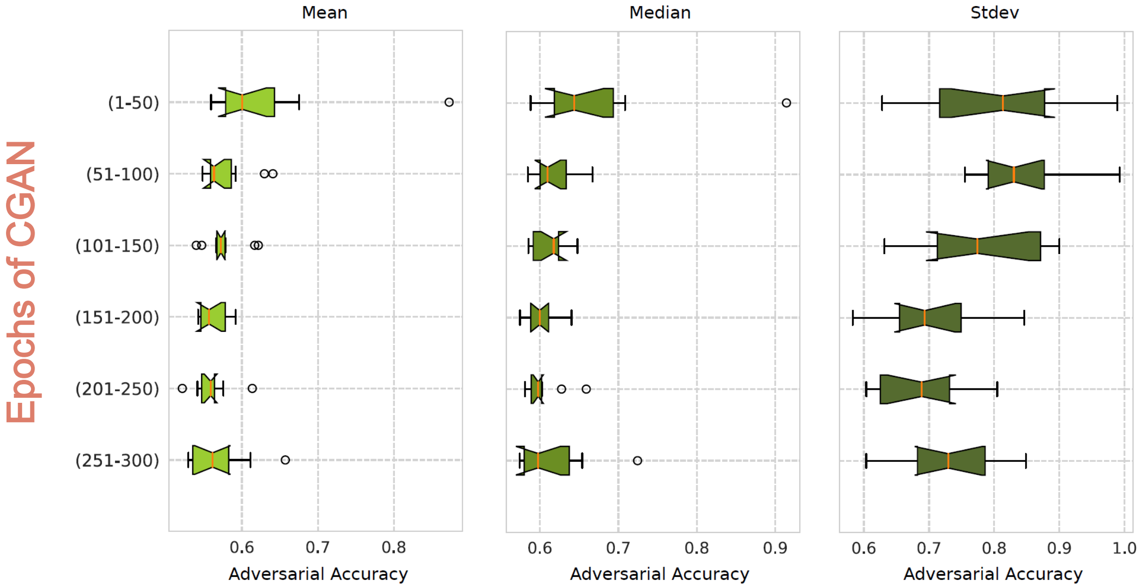


TOTUSJH	Total unsigned current helicity.

Model selection on AA

- Distributions of Adversarial Accuracy of three descriptive statistics for TOTUSJH: mean, median, and standard deviation, presented for all intermediate CGAN models (divided into six groups).

- The desirable outcome of AA is close to 0.5, implying that the real and synthetic samples generated by the generators are indistinguishable.



- AA = 1: there is no resemblance between the sets. AA = 0: two sets are exactly the same, yielding no new information.

AND SO WHAT?

Forecasting Experiment — A

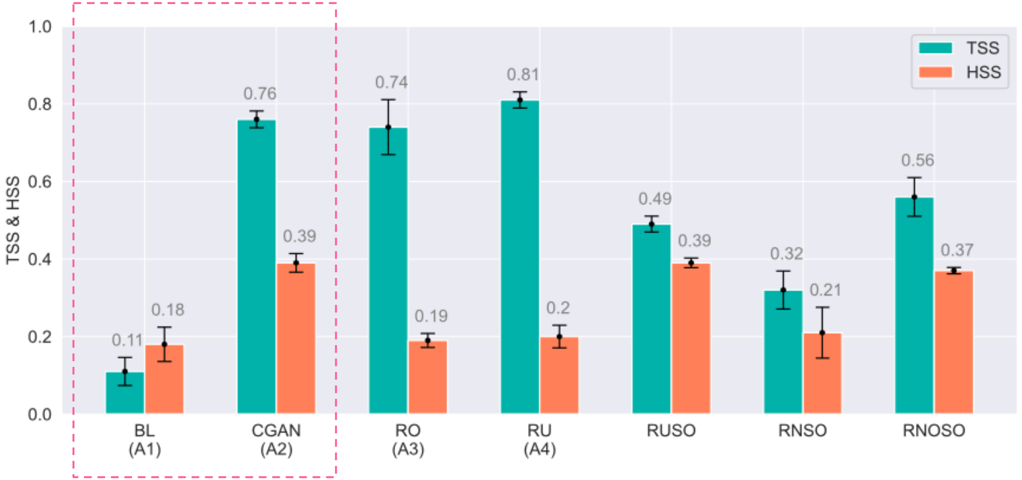
- The first attempt to evaluate synthetic samples by performing flare forecasting. (using SVM).
- Partitions 2, 3 and 5 from SWAN -SF used for testing.

Group	No.	Method	Description	Statistic
A	A1	Baseline (BL)	No data augmentation applied on P1.	last value
	A2	Synthetic Oversampling using CGAN (CGAN)	Adding synthetic flaring samples to the minority class of P1.	
	A3	Random Oversampling (RO)	Randomly oversampling samples of the minority class on P1.	
	A4	Random Undersampling (RU)	Randomly undersampling samples of the majority class on P1.	

Result — A

True skill statistic: $$TSS = \frac{tp}{tp + fn} - \frac{fp}{fp + tn}$$

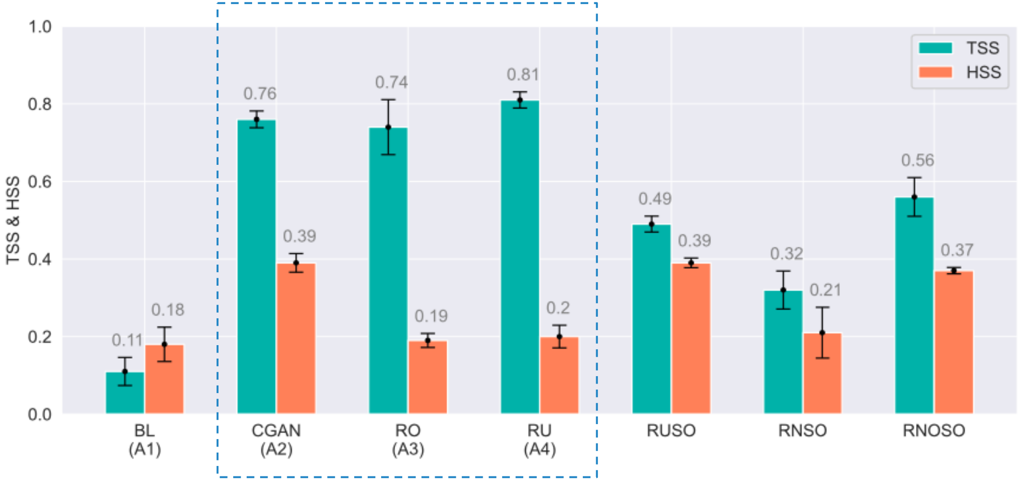
Heidke skill score: $$HSS2 = \frac{2 \cdot ((tp \cdot tn) - (fn \cdot fp))}{(tp + fn) \cdot (fn + tn) + (fp + tn) \cdot (tp + fp)}$$



BL	Baseline
RO	Random Oversampling
RU	Random Undersampling

- Compared to A1, the performance of A2 is improved significantly.
- TSS shows an increase from 0.11 to 0.76, and HSS2 from 0.18 to 0.39.

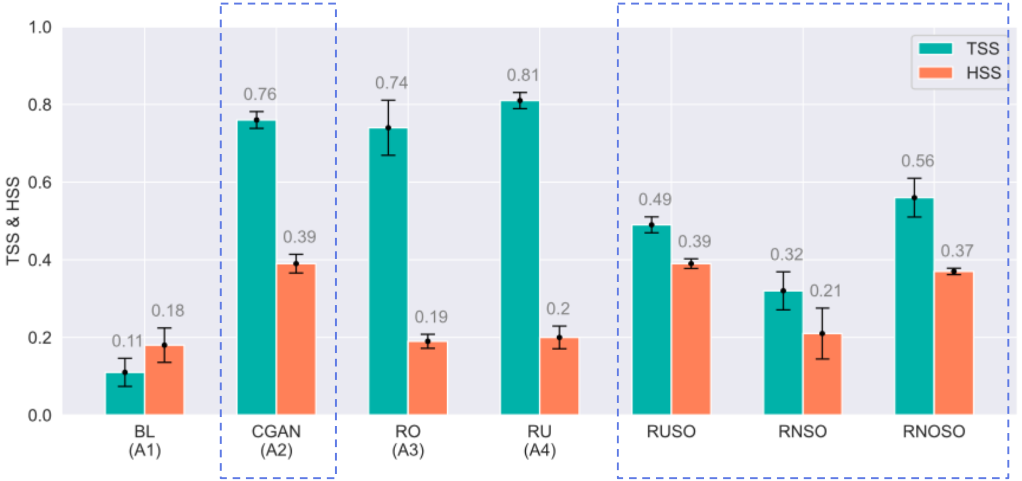
Result — A



BL	Baseline
RO	Random Oversampling
RU	Random Undersampling

- Comparing to A3 and A4, the HSS2 improvement of A2 is evident. (~0.19 → 0.39)
- Comparing to A3, the improvement of A2 might come from balancing the dataset with informative synthetic flaring instances.

Result — A



RUSO	Random Uniform Synthetic Oversampling
RNSO	Random Normal Synthetic Oversampling
RNOSO	Random NOise Synthetic Oversampling

- Compared to the statistic -based oversampling methods [1], the CGAN-based method achieves a significant improvement in terms of TSS while maintaining HSS2 at its highest value, i.e., 0.39.

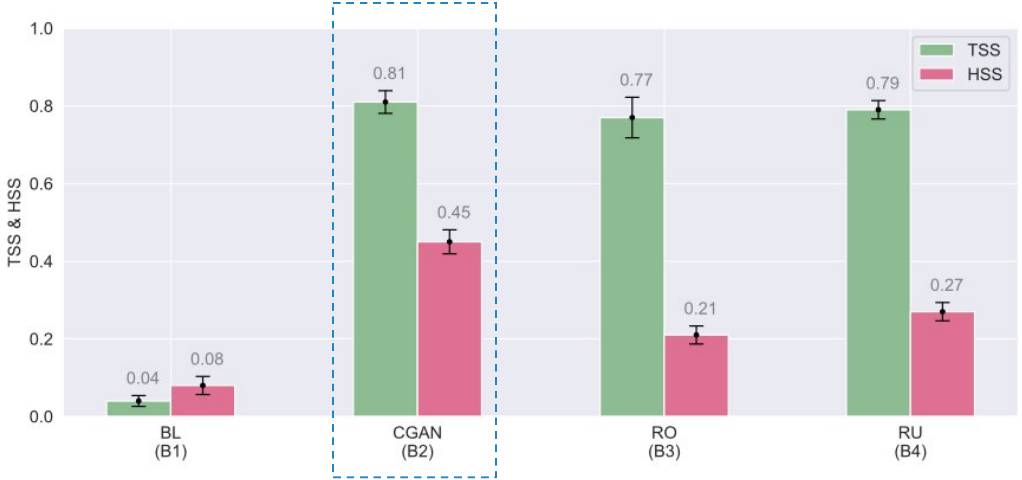
1. Hostetter , M., & Angryk, R. A.: First Steps Toward Synthetic Sample Generation for Machine Learning Based Flare Forecasting. 2020

Forecasting Experiment — B

- The first attempt to evaluate synthetic samples by performing flare forecasting. (using SVM).
- Partitions 2, 3 and 5 from SWAN -SF used for testing.

Group	No.	Method	Description	Statistic
B	B1	Baseline (BL)	No data augmentation applied on P1.	median & standard deviation
	B2	Synthetic Oversampling using CGAN (CGAN)	Adding synthetic flaring samples to the minority class of P1.	
	B3	Random Oversampling (RO)	Randomly oversampling samples of the minority class on P1.	
	B4	Random Undersampling (RU)	Randomly undersampling samples of the majority class on P1.	

Result — B



BL	Baseline
RO	Random Oversampling
RU	Random Undersampling

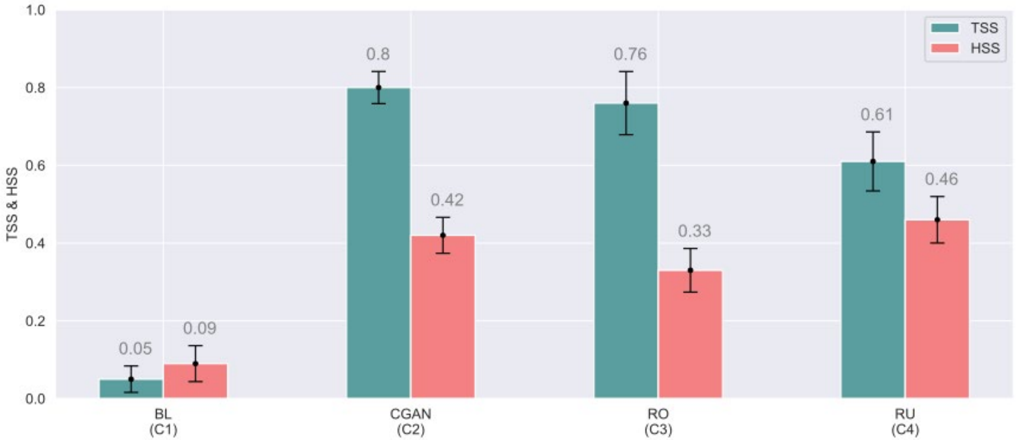
- We observe that B2 achieves the highest TSS and HSS2.
- It shows that the CGAN model can successfully learn the median and standard deviation of real multivariate time series samples.

Forecasting Experiment — C

- To further validate if synthetic samples can learn the temporal characteristics beyond descriptive statistics of median and standard deviation. (using T -SVC classifier)
- Partitions 2, 3 and 5 from SWAN -SF used for testing. Partition 4 used for optimizing T -SVC classifier.

Group	No.	Method	Description	Input
C	C1	Baseline (BL)	No data augmentation applied on P1.	time series
	C2	Synthetic Oversampling using CGAN (CGAN)	Adding synthetic flaring samples to the minority class of P1.	
	C3	Random Oversampling (RO)	Randomly oversampling samples of the minority class on P1.	
	C4	Random Undersampling (RU)	Randomly undersampling samples of the majority class on P1.	

Result — C



BL	Baseline
RO	Random Oversampling
RU	Random Undersampling

- Although C2 does not obtain the highest HSS2 score, it gives a promising TSS and HSS2 pairing .
- It shows synthetic samples share similar temporal characteristics with real samples.

Sum up

- We utilized the conditional generative adversarial network (CGAN) to perform data -informed augmentation of multivariate time series (MVTS) on SWAN-SF.
- We perform the model selection by using two methods: the Kullback - Leibler divergence metric, and the Adversarial Accuracy.
- We use the synthetic MVTS samples to balance the training dataset and compare the classification performance with other class - imbalance remedies.
- Overall, the results show that the CGAN model can indeed generate realistic multivariate time series samples.

Thank you.

Any comments?