Time Series Decomposition Using Wavelet and Fourier Transforms for Enhanced Solar Flare Forecasting

Victor Solomon[1], Omkar Rayala[1], Manya Rampuria[1], Abdul Afrid[1], Junzhi Wen[1], Rafal Angryk[1]
[1]Department of Computer Science, Georgia State University, Atlanta, GA, USA
{vsolomon3, orayala1, mrampuria1, amohammed34, jwen6, rangryk}@student.gsu.edu

Abstract

Time series decomposition enables the extraction of meaningful components by filtering out irrelevant or noisy frequencies. This study investigates the use of four transformation techniques—three wavelet-based (Haar, Symlets, Daubechies) and one Fourier-based (Discrete Fourier Transform)—to reconstruct multivariate time series data for solar flare prediction. The reconstructed data is used to train and evaluate two classifiers: Time Series Forest (TSF) for time series inputs, and Random Forest (RF) for transformed non-time-series representations.

We describe the data preparation pipeline, model training, and comparative performance analysis of TSF and RF. Additionally, we assess the effectiveness of data reduction by evaluating model performance on reduced datasets. Results show that TSF models consistently outperform RF models, and that reduced datasets yield competitive predictive performance. These findings suggest the potential for computational efficiency in near-real-time flare prediction without significant loss in accuracy. Evaluation metrics include the True Skill Statistic (TSS) and Heidke Skill Score (HSS2).

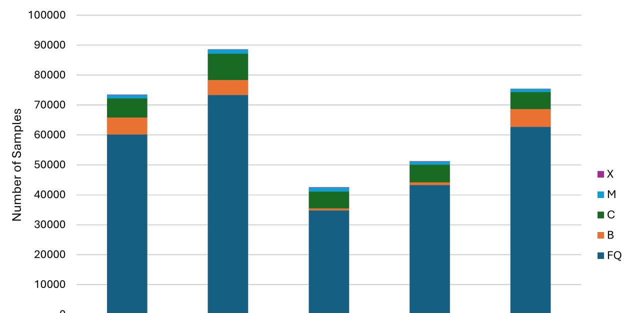
1 Introduction

Solar flares are powerful bursts of radiation from active solar regions linked to sunspots and intense magnetic fields (Longcope and Dana 2020). Classified by NOAA into five levels—A, B, C, M, and X—each step represents a tenfold increase in X-ray intensity. The SWAN-SF dataset used in this study consolidates these into main flare classes.

When directed at Earth, solar flares can disrupt satellites, power systems, and pose radiation risks (Larsen and Erik 2021), making reliable prediction essential. Recent work highlights the value of time series (TS) data in capturing the evolving magnetic behavior of active regions (Ji et al. 2021). Decomposing TS into frequency-based components using wavelet and Fourier transforms helps extract meaningful patterns by filtering out noise (Schlüter and Deuschle 2010).

Although studies such as (Chen, Kempton and Angryk, 2024) use TS data for flare prediction, a few number of studies have assessed the benefits of decomposition. This research fills that gap by applying Haar, Symlet, Daubechies wavelets, and Discrete Fourier Transform (DFT) to improve model performance. We evaluate Time Series Forest (TSF) models on decomposed data and compare them with Random Forest (RF) models on non-time-series formats, including reduced datasets for faster forecasting. The objectives of this study is to investigate the effectiveness of TS decomposition using wavelets and Fourier transforms in improving the performance of Time Series Forest (TSF) models for solar flare prediction. We also compare the impact of TS and Non-Time Series (NTS) data on the performance of TSF and Random Forest (RF) respectively and wether reduced data variations of the data can produce comparable or better results as compared with the initial data. This approach is particularly valuable in operational settings, where accurate predictions are needed in near real-time.

The paper is structured as follows: Section II reviews relevant transforms, Section III describes the dataset, Section IV explains the methodology, Section V presents results, and Section VI concludes with key insights.



	P1	P2	P3	P4	P5
X	165	72	136	153	19
M	1089	1392	1288	1012	971
C	6416	8810	5639	5956	5763
B	5692	4978	685	846	5924
FQ	60130	73368	34762	43294	62688

Figure 1: Data distribution across each partition of the SWAN-SF

2 Review of Literature

Several methods have been used to decompose time series (TS) data. The Fourier transform, one of the earliest tech-

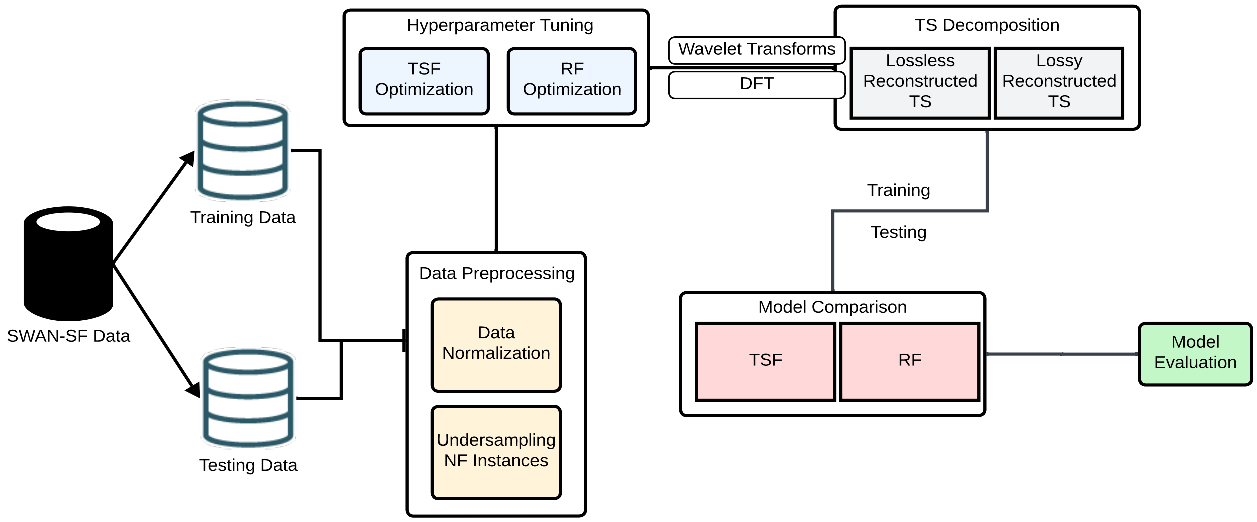


Figure 2: Experimental pipeline for model training and evaluation. The process includes data preprocessing, time series decomposition using DFT and wavelet transforms, model tuning, and comparison.

niques, decomposes a signal into sine and cosine components (Ghaderpour, Pagiatakis, and Hassan 2021). Wavelet transforms, building on the Fourier transform, offer a more comprehensive analysis by considering both scale and location of components (Singh and Singh 2017). Scale governs compression, while location determines the component's position. This allows wavelets to capture both local spatial patterns and temporal dynamics. (Hajiabotorabi 2019) introduced a discrete wavelet transform-recurrent neural network (DWT-RNN) model, using B-spline wavelet decomposition for high-frequency TS prediction. Evaluation on stock indices showed superior performance compared to other models.

(Masoud 2022) applied Discrete Wavelet Transform to satellite-derived ocean color-Chlorophyll-a (Chl-a) TS, finding ”db10” and ”db5” from the Daubechies family optimal for denoising. (Li et al. 2021) integrated Short-Time Fourier Transform (STFT) into neural networks, extracting information from both time and frequency domains, but the model's complexity hindered its performance against non-time-series models.

The Haar wavelet's simplicity makes it computationally efficient. (Zhang, Ho and Huang 2005) used it to decompose TS and extract features with reduced time complexity.

This research explores the use of HWT, SWT, DWT, and DFT for decomposing multivariate TS to improve solar flare prediction.

3 SWAN-SF Dataset

The SWAN-SF dataset (Angryk et al. 2020) is a multivariate time series resource for solar flare research. It is derived from NOAA's SHARP series and HMI vector magnetograms, converting NOAA flare categories into binary labels: flaring (X/M-class) and non-flaring (A/B/C-class).

It contains 4,098 samples, each with 60 time steps at 12-minute intervals over a 12-hour window, using a 24-hour prediction horizon. Each sample includes 51 features, but this study focuses on five key parameters from (Chen et al. 2021): TOTUSJH, TOTBSQ, TOTPOT, TOTUSJZ, and ABSNJZH.

As detailed in (Ahmadzadeh et al. 2021), a sliding window was used, resulting in overlapping sequences. To reduce leakage, the dataset is partitioned into five temporally disjoint subsets.

Due to class imbalance (e.g., FL:NF ratio of 1:29 in Partition 3, Figure 1), flaring events are evenly split, and a sampling strategy is used (Section IV).

4 Methodology

4.1 Time Series Decomposition Algorithms

This study applies three wavelet transforms—Haar (HWT) (Yonus and Hassan 2020), Symlet (SWT) (Akujuobi 2022), and Daubechies (DWT) (Taoufiq et al. 2022) and the Discrete Fourier Transform (DFT) (Wang 1984), to decompose time series (TS) data. These transformations disaggregate the data into component signals, facilitating better interpretation and classification.

(Lilly 2017) demonstrated that wavelet transforms outperform Fourier-based methods in localizing signals in both time and frequency domains—a property that is essential for modeling the non-stationary behavior of solar flares. All selected methods support both lossless and lossy reconstruction, which is leveraged in this study to evaluate predictive performance under varying levels of data fidelity.

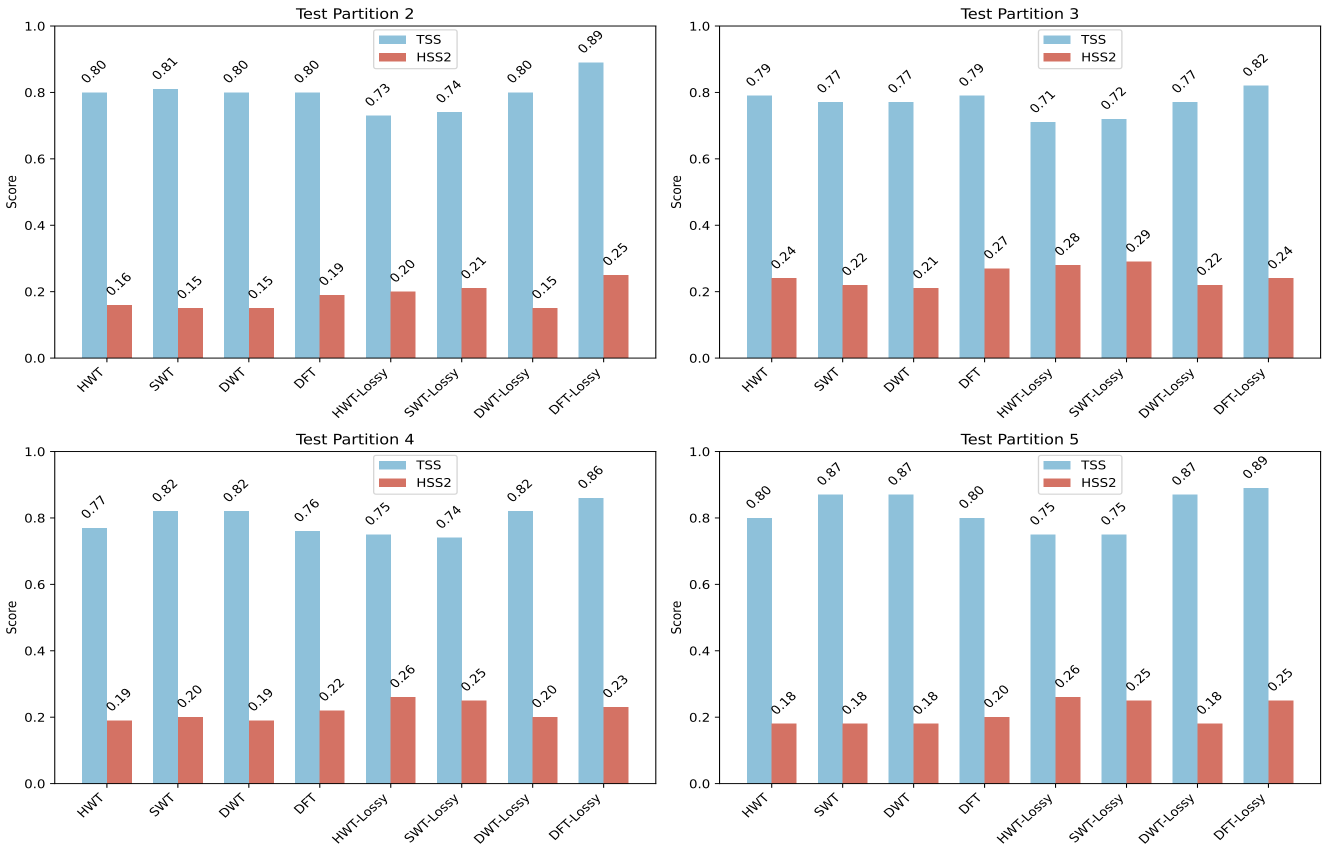


Figure 3: The TSS and HSS2 performance evaluation metrics for the TSF model using the reconstructed TS data from different transforms for training and testing across partitions of the data. HWT refers to the lossless reconstructed TS using HWT. HWT-Lossy refers to the lossy reconstructed TS using HWT. The same convention is maintained for the other transform methods.

Haar Wavelet Transform The Haar wavelet employs a simple step function to separate low- and high-frequency components, as shown below:

$$\psi(t) = \begin{cases} 1, & 0 \leq t < \frac{1}{2}, \\ -1, & \frac{1}{2} \leq t < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Daubechies Wavelet Transform Daubechies wavelets are compactly supported and provide high accuracy for both smooth and discontinuous signals. The order p controls the number of vanishing moments, enabling the capture of polynomial trends up to degree $p - 1$.

Symlet Wavelet Transform SWTs are modified DWTs that improve symmetry and reduce phase distortion. They retain the same functional form as DWT but offer better reconstruction quality for certain signal levels.

Discrete Fourier Transform DFT decomposes discrete signals into a sum of sinusoidal components, each characterized by a specific frequency, amplitude, and phase. Unlike wavelet-based methods, it offers only a frequency domain representation and lacks time localization. However, its effectiveness in capturing regular periodic patterns justifies its inclusion in this study.

4.2 Experimental Setup

The experimental workflow is illustrated in Figure 2. Partition 1 is used exclusively for training, while the remaining partitions serve as test sets. Two classifiers are trained: Time Series Forest (TSF) on TS data, and Random Forest (RF) trianed on NTS data. The NTS representation is obtained through time series decomposition and reconstruction, as described later.

To address class imbalance, the training data is normalized and undersampled. Simulated annealing is used for hyperparameter tuning. Both lossless and lossy reconstructions are evaluated; in the lossy case, low-pass filtering is applied. TSF is trained on the original, lossless, and lossy TS data, while RF is trained on NTS data derived from the reconstructed components.

4.3 Addressing Class Imbalance

All SWAN-SF partitions exhibit substantial class imbalance. To mitigate this, only the training set is balanced by randomly undersampling the non-flare (NF) class to match the number of flare (FL) instances. This strategy is computationally efficient and reduces training time compared to oversampling methods.

4.4 Normalization Strategy

Given the multivariate nature of the dataset, normalization ensures consistent scaling across features. Robust scaling is applied to both training and test sets due to its resilience to outliers.

4.5 Hyperparameter Tuning

Simulated annealing is employed to optimize the hyperparameters of both the TSF and RF classifiers. This probabilistic search algorithm is capable of escaping local minima by occasionally accepting suboptimal solutions, thereby enabling a more effective exploration of the search space. Compared to grid search, simulated annealing offers better convergence in complex, high-dimensional landscapes.

4.6 Model Training

Solar flare prediction is framed as a binary classification task (FL vs. NF). TSF is implemented using the `pyts` library (Faouzi and Janati 2021), and RF using `scikit-learn` (Kramer and Kramer 2016). TSF trains on both lossless and lossy reconstructed TS data, while RF uses NTS data from the decomposed components. In the lossy case, low-pass filters preserve dominant frequencies to retain key signal characteristics.

4.7 Evaluation Metrics

It is essential to use the accurate evaluation metric for flare prediction as it involves rare event prediction and as this needs to be robust to class imbalance. Although training data is balanced, the testing sets used in this study retains class imbalance that is obtainable in the real world. Owing to this, True Skill Statistic (TSS) and Heidke Skill Score (HSS2) are used to evaluate models.

Components of the confusion matrix are used to compute TSS and HSS2 as shown below:

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

TSS is robust to class imbalance but treats FP and FN equally, which may not reflect operational costs. To complement this, HSS2 is calculated as:

$$HSS2 = \frac{2 \times ((TP \times TN) - (FN \times FP))}{P \times (FN + TN) + (TP + FP) \times N} \tag{3}$$

Both metrics range from -1 (worst score) to 1 (perfect score). A score of 0 indicates random performance.

5 Experiments and Results

5.1 Experiment I: Impact of TS Reconstruction on TSF Model

This experiment investigates the effect of lossless and lossy reconstruction of TS on prediction accuracy. The TSF model is trained and tested using the original SWAN-SF dataset. A sample of this data consists of 60 timestamp observations of the 5 magnetic parameters. A lossless reconstruction of the TS is performed and used for training and testing the model. The lossless reconstructed data resulted in similar TSS and HSS2 values as the original TS across all testing partitions. To avoid redundancy we use the lossless reconstructed TS in place of the original TS data. A lossy reconstruction of the TS is performed using the different transforms across each of the five partitions. A low-pass filter approach is used to determine which components of the data are to be kept or removed in the reconstruction. We borrow the concept applied by (Paula et al. 2022) which suggested that flaring data usually has an SNR above a 20dB threshold. This SNR value is used to select a retain ratio for the low-pass filter used by each transform for a lossy reconstruction.

After setting this threshold, the retain ratio that produces an SNR of 20dB for each of the transforms is used to generate the reconstructed TS. The same process is repeated for all the other transforms that have been used.

The detailed TSS and HSS2 for each of the transform is shown in Figure 3 for each of the testing partition. The results demostrate that the DFT-Lossy outperforms the other methods in Partition 2. The DFT-Lossy has the best TSS and the SWT-Lossy has the best HSS2 in Partition 3. The DFT-Lossy outperforms the other methods in the TSS metric while the HWT-Lossy outperforms the other methods in the HSS2 score in Partition 4 and 5. However, it can be noted that for each of the transforms, there is a trade-off between the TSS and HSS2 values across testing partitions.

5.2 Experiment II: The Impact of Time Series vs. Non-Time Series Data on Flare Binary Classification

This research evaluates the impact of TS and NTS data on flare prediction. Specifically, TSF models trained on TS data are compared with RF models trained on NTS data. Variants of the TS dataset are created by reducing the observation window: starting from 720 minutes, it is reduced to 320 minutes, producing samples with 30 observations. Further reductions to 96 and 48 minutes result in samples with 8 and 4 observations, respectively. The data reduction strategy examines whether reduced data can achieve comparable or improved results to the original TS.

The NTS data is generated by decomposing the original TS data with a 720-minute window using each transform, without reconstructing it to the time domain. This results in variants with 30, 8, and 4 observations for accurate comparisons between TSF and RF models. Wavelet transforms allow easy NTS curation by adjusting the detail level for the desired observations. In contrast, DFT requires a more nuanced approach, as it lacks a built-in parameter for reducing the number of points. DFT decomposition initially yields 60 phases and 60 amplitudes per 60-observation sample. To align with TS variants, the last 15 phases are concatenated with the last 15 amplitudes, producing 30 observations per sample. This is repeated for 8- and 4-observation samples.

Results in Figures 4, 5, 6, and 7 show that TS data outperforms NTS data in TSS and HSS2 about 90

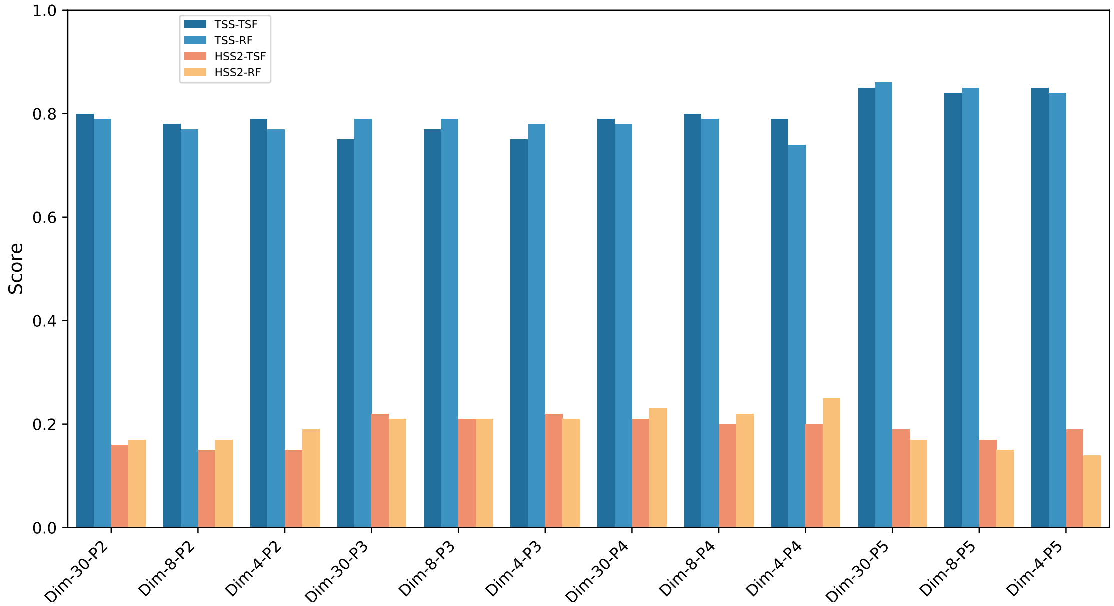


Figure 4: Results of Experiment II using the DFT. The first two bars in each grouped bar display the TSS values of the TSF and RF models respectively, while the last two bars in each grouped bar display the HSS2 values of the TSF and RF models respectively. In this figure, the naming convention ”TSS-Dim-30-P2” refers to the ”TSS value for Partition 2, consisting of 30 observations per sample.” Similarly, ”HSS2-Dim-30-P2” indicates the ”HSS2 value for Partition 2, consisting of 30 observations per sample”.

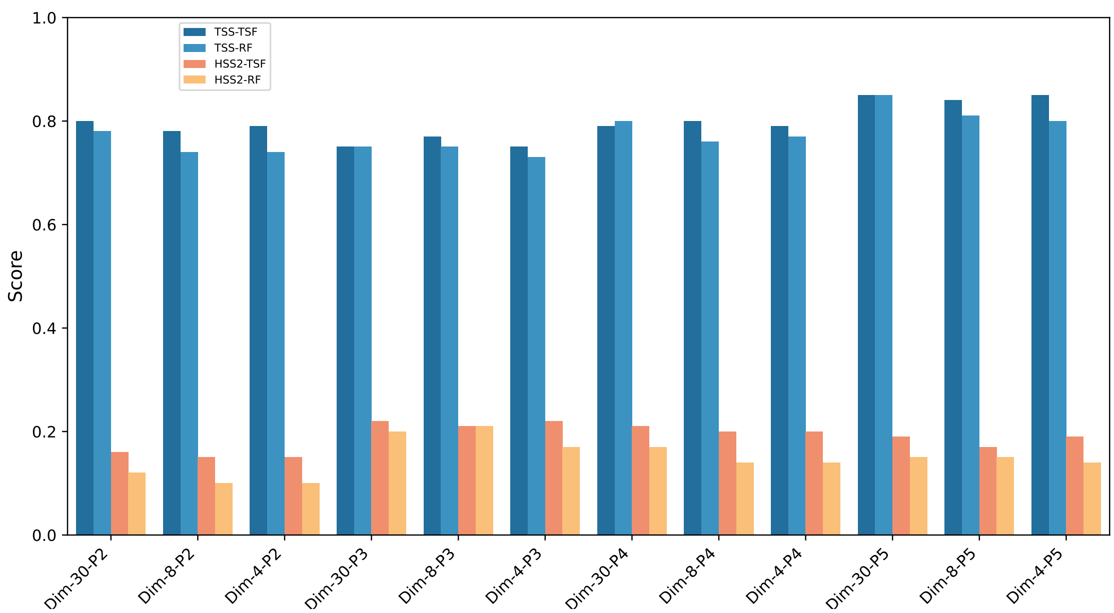


Figure 5: Performance of TSF and RF models in Experiment II using HWT. The same convention is maintained as with Figure 4.

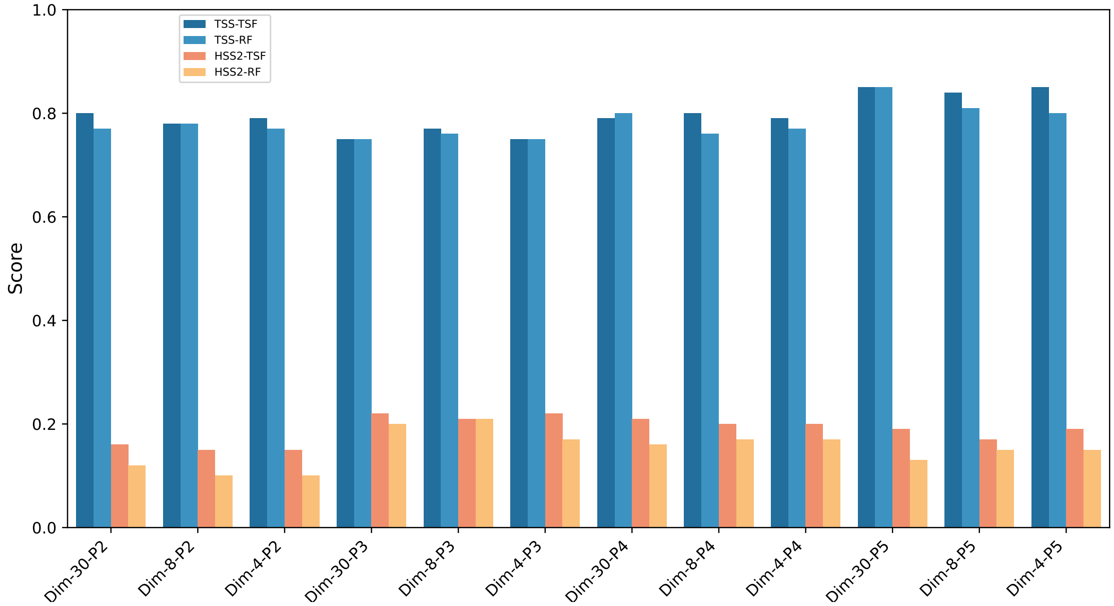


Figure 6: Performance of TSF and RF models in Experiment II using DWT. The same convention is maintained as with Figure 4.

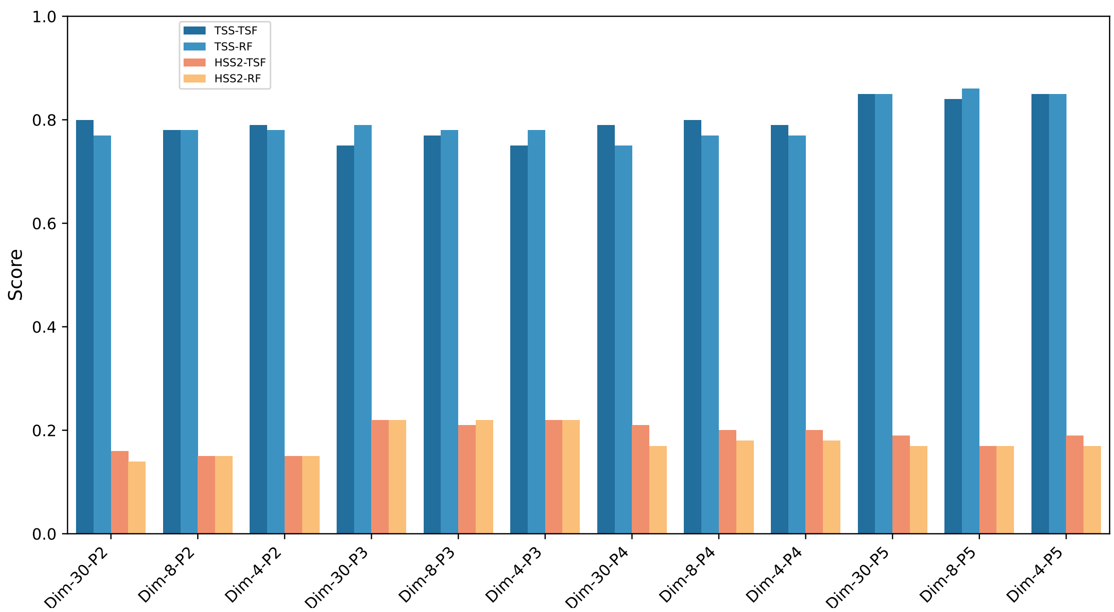


Figure 7: Performance of TSF and RF models in Experiment II using SWT. The same convention is maintained as with Figure 4.

6 Conclusion and Future Work

We proposed a method for comparing TS and NTS models in flare prediction, alongside a data reduction strategy that yields results comparable to the original dataset. Our findings show that TS-based classifiers, such as TSF, generally outperform NTS-based classifiers like RF, achieving higher TSS and HSS2 for our use case. We claim that the improvement in performance is attributed to the temporal structure of the solar surface activity in the data that is preserved by TS models. Additionally, reducing the observation window in the SWAN-SF data, resulting in fewer observations per sample, produces comparable TSS and HSS2 scores. This data reduction approach can enhance training efficiency and faster real-time prediction. We also demonstrated that wavelet and Fourier transforms improved the TSF model's performance, with lossy reconstruction yielding optimal results in certain cases. Future work should explore training on standardized sample dimensions (e.g., 8 observations per sample) and evaluate the robustness of the data reduction strategy across various SWAN-SF dataset variations. Further, ablation studies examining different filtering thresholds for lossy reconstruction of solar data are recommended.

References

Longcope, D. (2020). Using Kepler's laws and Rutherford scattering to chart the seven gravity assists in the epic sunward journey of the Parker Solar Probe. American Journal of Physics, 88(1), 11-19.

Larsen, E. (2021). Predicting solar flares with remote sensing and machine learning. arXiv preprint arXiv:2110.07658.

Ji, A., Aydin, B., Georgoulis, M. K., & Angryk, R. (2020, December). All-clear flare prediction using interval-based time series classifiers. In 2020 IEEE international conference on big data (big data) (pp. 4218-4225). IEEE.

Schlüter, S., & Deuschle, C. (2010). Using wavelets for time series forecasting: Does it pay off? (No. 04/2010). IWQW Discussion Papers.

Chen, Y., Kempton, D. J., & Angryk, R. A. (2024, May). FFAD: A Novel Metric for Assessing Generated Time Series Data Utilizing Fourier Transform and Auto-Encoder. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 129-139). Singapore: Springer Nature Singapore.

Ghaderpour, E., Pagiatakis, S. D., & Hassan, Q. K. (2021). A survey on change detection and time series analysis with applications. Applied Sciences, 11(13), 6141.

Singh, U., & Singh, S. N. (2017). Application of fractional Fourier transform for classification of power quality disturbances. IET Science, Measurement & Technology, 11(1), 67-76.

Hajiabotorabi, Z., Kazemi, A., Samavati, F. F., & Ghaini, F. M. M. (2019). Improving DWT-RNN model via B-spline wavelet multiresolution to forecast a high-frequency time series. Expert Systems with Applications, 138, 112842.

Moradi, M. (2022). Inter-comparison of single-sensor and merged multi-sensor ocean color chlorophyll-a products in the shallow turbid waters-case study: Persian Gulf. International Journal Of Coastal, Offshore And Environmental Engineering (ijcoe), 7(2), 1-10.

Li, S., Chowdhury, R. R., Shang, J., Gupta, R. K., & Hong, D. (2021, November). Units: Short-time fourier inspired neural networks for sensory time series classification. In Proceedings of the 19th ACM conference on embedded networked sensor systems (pp. 234-247).

Zhang, H., Ho, T., & Huang, W. (2005, May). Blind feature extraction for time-series classification using Haar wavelet transform. In International Symposium on Neural Networks (pp. 605-610). Berlin, Heidelberg: Springer Berlin Heidelberg.

Angryk, R. A., Martens, P. C., Aydin, B., Kempton, D., Mahajan, S. S., Basodi, S., ... & Georgoulis, M. K. (2020). Multivariate time series dataset for space weather data analytics. Scientific data, 7(1), 227.

Ahmadzadeh, A., Aydin, B., Georgoulis, M. K., Kempton, D. J., Mahajan, S. S., & Angryk, R. A. (2021). How to train your flare prediction model: Revisiting robust sampling of rare events. The Astrophysical Journal Supplement Series, 254(2), 23.

Younus, M. A., & Hasan, T. M. (2020, April). Effective and fast deepfake detection method based on haar wavelet transform. In 2020 International Conference on Computer Science and Software Engineering (CSASE) (pp. 186-190). IEEE.

Akujuobi, C. M. (2022). Wavelets and wavelet transform systems and their applications. Berlin/Heidelberg, Germany: Springer International Publishing.

Nouhaila, B. O. U. A. L. O. U. L. O. U., Taoufiq, B. D., & Benayad, N. S. I. R. I. (2022). An intelligent approach based on the combination of the discrete wavelet transform, delta delta MFCC for Parkinson's disease diagnosis. International Journal of Advanced Computer Science and Applications, 13(4), 562-571.

Wang, Z. (1984). Fast algorithms for the discrete W transform and for the discrete Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(4), 803-816.

Lilly, J. M. (2017). Element analysis: A wavelet-based method for analysing time-localized events in noisy time series. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 473(2200), 20160776.

Faouzi, J., & Janati, H. (2020). pyts: A python package for time series classification. Journal of Machine Learning Research, 21(46), 1-6.

Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.

Chen, Y., Kempton, D. J., Ahmadzadeh, A., & Angryk, R. A. (2021). Towards synthetic multivariate time series generation for flare forecasting. In Artificial Intelligence and Soft Computing: 20th International Conference, ICAISC 2021, Virtual Event, June 21–23, 2021, Proceedings, Part I 20 (pp. 296-307). Springer International Publishing.