Enhancing SEP Event Prediction through Time Series Data Augmentation

Pouya Hosseinzadeh [1] Soukaina Filali Boubrahimi [1] Shah Muhammad Hamdi [1]

[1]Department of Computer Science, Utah State University, Logan, Utah, USA

Abstract

Solar energetic particles (SEPs), which arise from significant solar events, can cause substantial damage to both space-based and terrestrial infrastructure. Predicting these events is challenging due to their rarity and the limited data available. This study improves the prediction of ∼30, ∼60, and ∼100 MeV SEP events by synthetically augmenting the dataset. By employing machine-learning techniques, particularly time series forest (TSF), alongside data augmentation methods like the synthetic minority oversampling technique (SMOTE), we achieved significant improvements in prediction accuracy and F1-score, reaching approximately 90% accuracy for ∼100 MeV SEP events.

Introduction

- **Solar Energetic Particles (SEPs)**: High-energy particles from the Sun during solar flares and CMEs.
- **Impact of SEPs**: Risks to satellites, communication systems, and astronaut safety.
- **Health Risks**: High-energy SEP events pose severe radiation risks to astronauts.
- **Technological Disruptions**: SEP events can disrupt satellite operations and communications.
- **Prediction Challenges**: Limited historical data due to the rarity of high-energy SEP events.
- **Research Objective**: Enhance prediction accuracy of ∼30, ∼60, and ∼100 MeV SEP events with synthetic data. Also, building a hierarchical prediction framework.

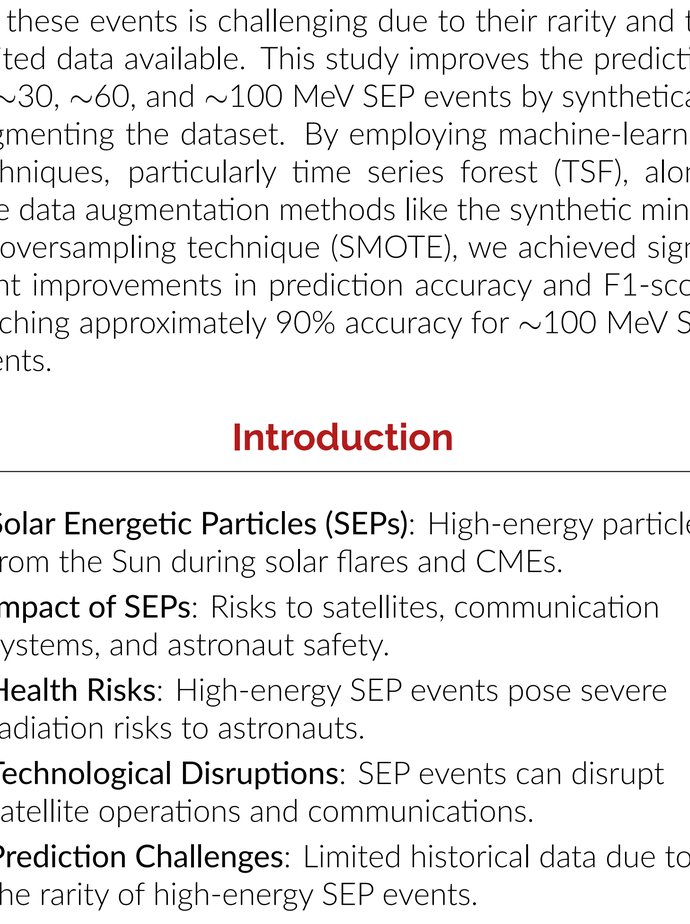

Figure 1. Solar flare image captured by SDO of AIA 304, accessible from `https://helioviewer.org/` and 5 hr proton flux data .

- Figure 2 outlines the entire process from data collection to classification.
- Shows the sources of SEP events and non-SEP events, including GSEP Event Catalog, HEK, and GOES Proton Flux Data.
- Demonstrates the application of data augmentation techniques such as Gaussian noise, SMOTE, and ADASYN.
- Illustrates the steps involved in generating synthetic SEP samples to balance the dataset.
- Highlights the flow from raw data collection, through augmentation, to machine learning classification.
- Emphasizes the importance of each step in improving prediction accuracy and addressing class imbalance.

Figure 2. Data collection, augmentation, and classification pipeline.

Datasets

- **GSEP Event Catalog**: Comprehensive list of SEP events covering three solar cycles.
- **Heliophysics Events Knowledgebase (HEK)**: Source for determining non-SEP events.
- **GOES Proton Flux Data**: Historical proton flux data from GOES satellites, covering different energy bands.

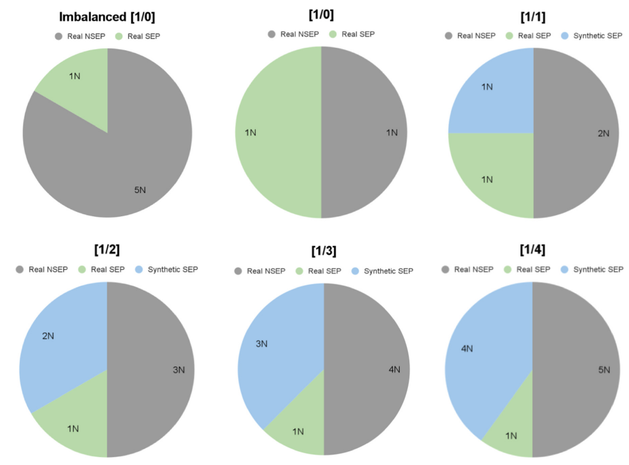

Figure 3. Class distribution with different SEP [Real/Synthetic] ratios.

Methodology

- **Data Augmentation Techniques**: Applied Gaussian noise, SMOTE, and ADASYN to generate synthetic SEP samples and balance the dataset.
- **Gaussian Noise, SMOTE, and ADASYN**: Added random noise to time series data to increase robustness, and generated synthetic samples between minority class samples and their neighbors, focusing on difficult-to-classify examples.
- **Classification Models**: Utilized Time Series Forest (TSF), ROCKET, and Shapelet Transform (SHAPELET) to classify SEP events with improved accuracy.

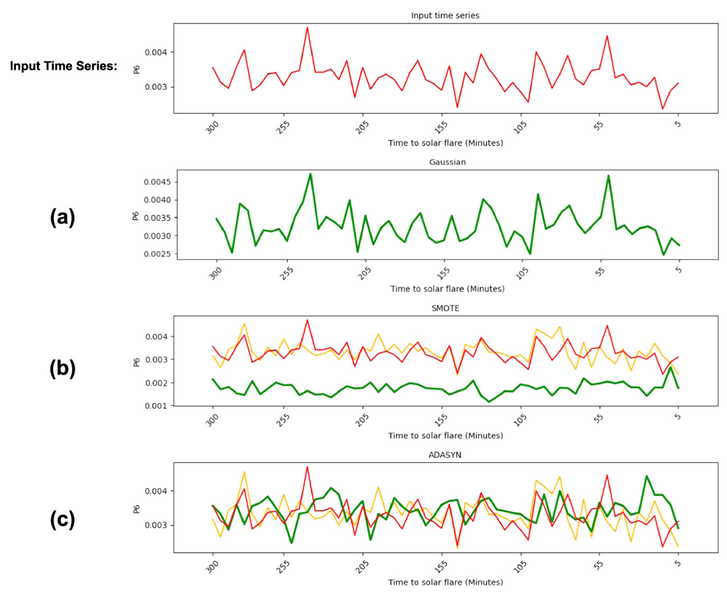

Figure 4. Synthetic time series generation using Gaussian, SMOTE, and ADASYN.

- Figure 5 highlights the similarities between real and SMOTE-generated synthetic time series data.
- Demonstrates the effectiveness of SMOTE in creating realistic synthetic data that closely matches the original time series.
- Shows how data augmentation can help enhance the performance of machine learning models.
- Emphasizes the importance of synthetic data in improving the robustness and accuracy of SEP event prediction models.
- Visual comparison to validate the quality of the synthetic data generated by SMOTE.

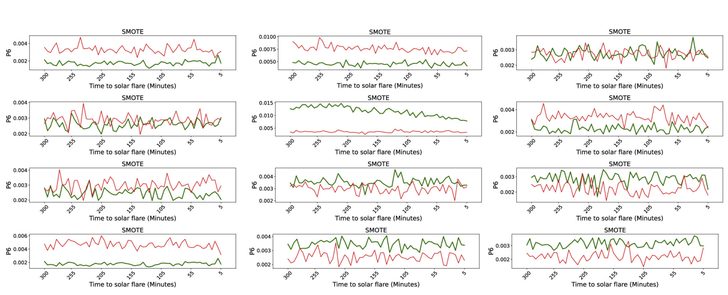

Figure 5. Comparison of input time series (in red) and SMOTE-generated time series (in green) highlighting their similarities.

Results

- Improved accuracy and F1-score of classifiers (SHAPELET, ROCKET, TSF) with SMOTE data augmentation. Comparative performance differences among classifiers for 100 MeV, 60 MeV, and 30 MeV SEP predictions.

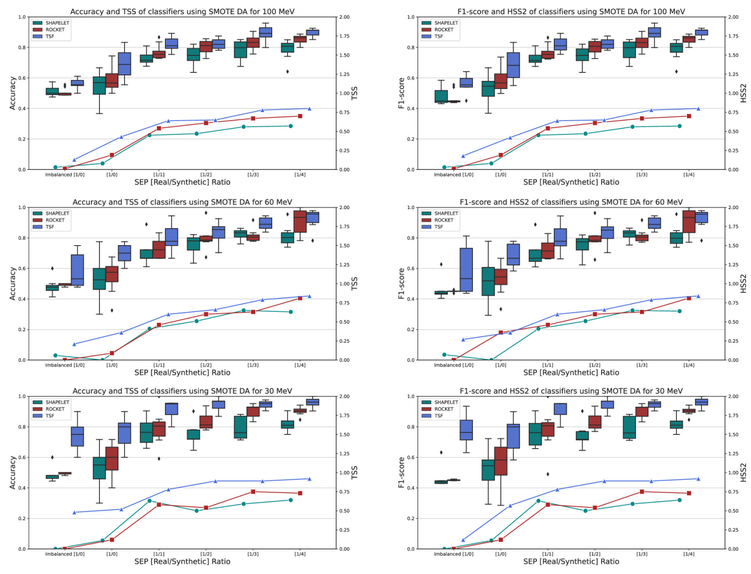

Figure 6. Classifier performance with SMOTE.

Hierarchical Framework:

- Shows the hierarchical prediction framework used for SEP event prediction.
- The model is designed to predict SEP events across different energy bands (∼30, ∼60, and ∼100 MeV).
- Highlights the division of tasks into different stages to improve prediction accuracy.
- Each stage of the hierarchy refines the prediction by focusing on specific energy ranges.
- Shows how the hierarchical model integrates various data sources and preprocessing techniques to enhance prediction performance.

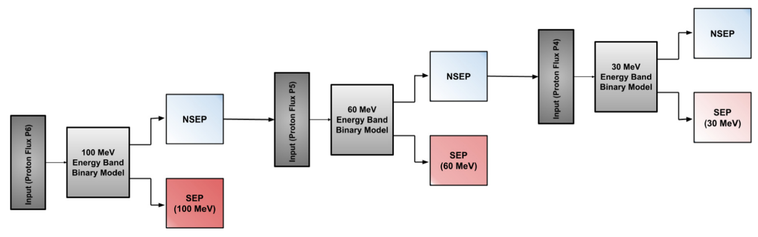

Figure 7. Hierarchical model structure for SEP event prediction across different energy bands.

- Figure 8 compares the accuracy of hierarchical and binary Time Series Forest (TSF) classifiers.
- Displays the impact of applying SMOTE (Synthetic Minority Over-sampling Technique) on the classifiers' performance.
- Indicates that using SMOTE improves the accuracy of both hierarchical and binary TSF classifiers.
- Provides a visual comparison of the effectiveness of data augmentation in enhancing SEP event prediction.
- Highlights the differences in prediction accuracy for various energy bands when using the hierarchical model versus the binary approach.

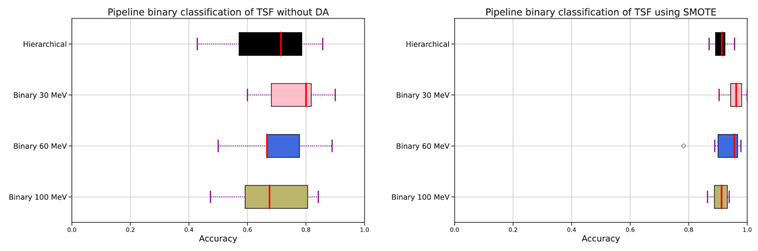

Figure 8. Accuracy of hierarchical and binary TSF classifiers with and without SMOTE.

Future Work

- Explore additional data augmentation techniques to further improve prediction accuracy.
- Implement real-time prediction systems to provide timely warnings for SEP events.
- Analyze the impact of combining multiple data sources, such as solar wind and magnetic field data, on prediction performance.