Solar Flare Prediction Using Magnetic Field and SDO Image Data

Intermediate-Term Report — November 2025

Author: Sattar Saif Shaik (Github)

This project develops a multimodal deep-learning system integrating magnetic-field time-series and solar imagery (SDO) data to predict solar flare events, with mid-term work focused on preprocessing pipelines, dataset management, and baseline classification models.

Why Solar Flare Prediction Matters

The Problem

Solar flares are high-energy eruptions that disrupt satellites, communication networks, and power grids worldwide. Early detection enables critical protective measures.

The **Solar Dynamics Observatory (SDO)** continuously monitors magnetic activity via the **Helioseismic and Magnetic Imager (HMI)**, capturing signatures that precede flare events.

Key Objectives

Enable **early-warning systems** through data-driven prediction.

Integrate **physics-based observations** with deep learning.

Develop a **structured, reproducible workflow** capable of scaling to tens of GB of solar data.

Made with GAMMA

Research Objectives and Challenges

Build a multimodal framework learning from both magnetic-field time-series and SDO imagery to enable data-driven, physics-informed flare forecasting.

Key Challenges and Goals

Independent Dataset Alignment

The magnetic dataset from Big Data 2020 has no timestamps, while SDO imagery is time-indexed, causing data misalignment.

Modality Fusion

Develop separate deep models per modality, then investigate alignment and feature-level fusion techniques (if alignment issue is addressed).

Scalable Pipeline

Execute structured workflow from robust data processing through feature engineering to production modeling.

Made with GAMMA

Magnetic Data Processing Pipeline

Implemented using **Dask** for distributed, memory-aware computation, converting 17 GB of raw JSON data into compact Parquet feature sets.

01. Data Splitting & Memory Planning

Split 17 GB raw data by record_id into train/dev/test/holdout before processing. Dask configured for 18 GB RAM systems with spilling and pausing thresholds.

02. Index Stabilization & Tagging

Reset indices, sort by record_id, add seq_id for temporal sequencing.

03. Feature Transformation

Apply log(1+x) to all skewed magnetic features (USFLUX, TOTPOT, PIL_LEN, etc.) to normalize distributions.

04. Time-Series Engineering

Computed rolling means/std (5-step), differences (1-step), lags (3-step) via partitioned groupby on record_id.

05. Feature Selection & Cleanup

Retain 14 core features covering energy, flux, shear, trend, volatility. Fill NaNs with 0 for consistency.

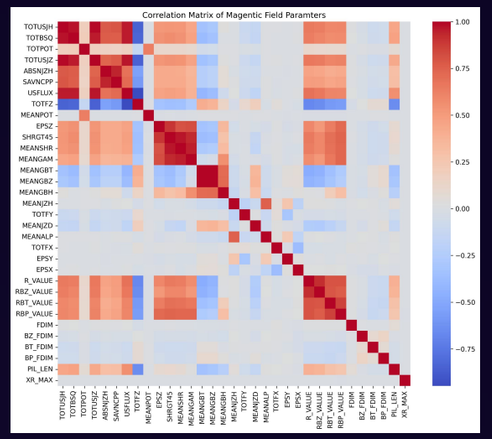
06. Compressed Output Storage

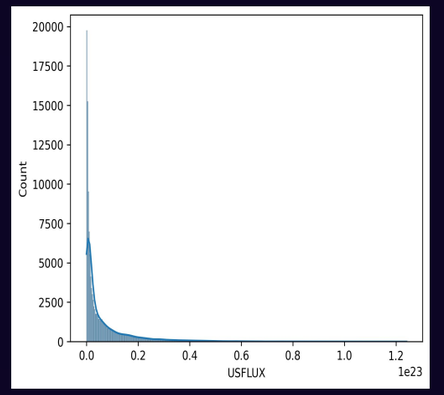
Write splits to Parquet (Snappy compression, PyArrow schema) under data/processed/magnetic_data/.

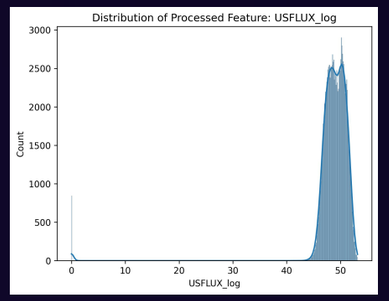
Made with GAMMA

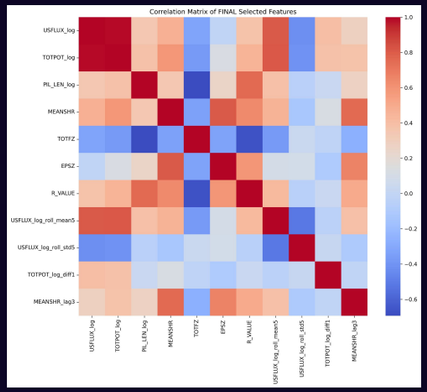
EDA-Driven Feature Engineering

Transform raw, complex solar data into a clean, feature-rich dataset for modeling.









Key Findings from EDA

Data Structure: Grouped processing needed for thousands of short time-series.
Skewness: Critical features like `TOTPOT` and `USFLUX` were right-skewed.
Redundancy: High multicollinearity (r > 0.9) identified among features.
Log-transformation: `log(1+x)` normalized distributions and reduced multicollinearity.

Feature Engineering Pipeline

01. Stabilization

Data was indexed and sorted by `record_id` with a `seq_id` for temporal order.

02. Transformation & Selection

Skewed features were log-transformed (`log(1+x)`) and 7 foundational features selected to reduce multicollinearity.

03. Advanced Time-Series Features

Rolling means/standard deviations, one-step differences, and lag features were engineered to capture trends.

Final Result: A Clean, 14-Feature Dataset

This process yielded a robust, 14-feature dataset ready for predictive modeling.

Made with GAMMA

Baseline Classification Results

Binary Classification (Flare vs No Flare)

Model	Type	Macro F1
Logistic Regression	Linear baseline	77%
Random Forest	Nonlinear	73%
SVM (RBF)	Kernel	76%

Multiclass Classification (C/M/X/Q Flare Types)

Model	Type	Macro F1
Softmax Regression	Linear baseline	77%
Random Forest	Tree ensemble	73%
SVM	Kernel	76%

Engineered features carry significant predictive signal. Further gains require temporal modeling (LSTM/GRU), class balancing, and feature scaling for deep learning.

Made with GAMMA

Technical Challenges & Solutions

Data Volume (17 GB)

Used Dask DataFrame with on-disk compute and sequential splitting to exceed RAM limits.

Memory Constraints

Configured Dask spilling limits, processed splits separately, and employed aggressive garbage collection.

Multicollinearity

Performed correlation-threshold feature selection to reduce redundancy and improve model interpretability.

Storage Efficiency

Saved intermediate outputs as compressed Parquet (Snappy) with PyArrow schemas for rapid I/O.

NaN Handling

Filled missing values from rolling/lag and the intiial data features with 0 to maintain dataset integrity post-processing.

Monitoring & Progress

Integrated dask.diagnostics.ProgressBar for visibility into long-running distributed jobs.

Made with GAMMA

Key Learnings and Achievements

1

Mastered Distributed Data Processing with Dask

Practical experience with Dask, including partitioning, lazy evaluation, and spill-to-disk management for efficient processing of large datasets.

2

Developed a Scalable Magnetic Data Pipeline

Built a modular preprocessing system that efficiently converts raw JSON data into optimized Parquet files, ensuring data integrity and rapid access.

3

Engineered Interpretable Time-Series Features

Created 14 physically meaningful features from raw magnetic field data, enhancing model interpretability and predictive power.

4

Addressed Real-World Data and System Challenges

Successfully tackled practical issues such as data imbalance, multicollinearity, and memory constraints in complex scientific computing environments.

5

Established Reliable Baseline Models and Evaluation

Implemented various classification algorithms and shifted evaluation metrics to F1-score and True Skill Statistic (TSS) for robust model assessment.

6

Prepared the Groundwork for Multimodal Deep Learning

Structured the data workflow and pipeline to facilitate future integration of magnetic field data with SDO image data for advanced deep learning applications.

Artifacts: 14-feature processed datasets, automated reproducible pipeline, baseline classification benchmarks.

Made with GAMMA

Next Steps & Future Directions

1

Temporal Deep Learning

Develop LSTM/GRU networks to capture temporal dependencies in magnetic time-series.

2

CNN Image Pipeline

Build convolutional neural network models for SDO solar imagery feature extraction.

3

Dataset Alignment and Fusion

Explore timestamp interpolation or external matching methods to combine modalities.

4

Multimodal Fusion

Explore different Fusion strategies (if dataset alignment is successful).

5

Production Optimization

Address class imbalance, hyperparameter optimization, and benchmark against existing forecasting systems.

Made with GAMMA

Concluding Reflections

✓ Pre-Learning Foundation Complete

Learning about Dask distributed computing and solar physics domain knowledge established the technical foundation.

✓ Data Processing Pipeline Complete

Complete pipeline processes 17 GB of magnetic data into production-ready feature sets.

✓ Baseline Classification Models Demonstrated Predictive Potential

Baseline models achieve 72–78% accuracy, validating feature engineering and magnetic signal strength.

→ Next Milestone: Deep Learning & Multimodal Integration

Develop LSTM/GRU models for magnetic data, CNN models for SDO imagery, and explore fusion techniques to unlock next-generation solar flare forecasting.