IDENTIFIKASI DAN PREDIKSI BERBASIS SUPERVISED LEARNING PROSES PRA-FLARE KELAS M DAN X MENGGUNAKAN DATA SHARP

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Magister dari
Institut Teknologi Bandung

Oleh
CHANDRA ALIF FERNANDA
NIM 20322003

PROGRAM STUDI PASCASARJANA ASTRONOMI
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI BANDUNG
2023

ABSTRAK

Suar surya atau *solar flare* merupakan salah satu fenomena paling energetik yang terjadi di tata surya. Kemunculan suar surya dapat mempengaruhi dinamika plasma dan energi di tata surya, tak terkecuali Bumi. Namun hingga kini, teori pembentukkan suar surya masih belum diketahui sepenuhnya. Namun demikian, peran magnetisme dalam pembangkitan suar surya telah dikonfirmasi. Di sisi lain, suar surya sebagai suatu proses pelepasan energi tentuntya diawali dengan proses pengumpulan energi. Proses pengumpulan energi pra-*flare* ini diduga tercermin pada parameter magnetiknya. Oleh karena itu, analisis parameter magnetik penting dilakukan untuk memahami kemunculan suar surya itu sendiri.

Tesis ini bertujuan untuk menganalisis peran parameter magnetik yang diperoleh dari database *Spaceweather HMI Active Region Patch* (SHARP) dalam rangka memperdiksi kemunculan suar suya kelas M dan X. Selain itu, tesis ini juga bertujuan untuk memprediksi kemunculan suar surya kelas M dan X dengan meninjau parameter magnetik yang paling berperan dalam kemunculan suar surya.

18 parameter magnetik SHARP dari tanggal 10 Mei 2010 sampai 31 Desember 2021 ditinjau sebagai parameter yang berpotensi untuk digunakan sebagai fitur dalam prediksi menggunakan *support vector machine*. Kemudian ditinjau posisi temporal suar surya kelas M dan X relatif terhadap puncak setiap parameter magnetik pada daerah aktifnya untuk melihat peran parameter tersebut sebagai pendahulu suar surya. Selain itu, perubahan atau fluktuasi signifikan yang terjadi sebelum suar surya diasumsikan sebagai suatu proses pra-*flare*. Digunakan *Short Time Fourier Transform* (STFT) untuk melihat proses ini dengan lebih jelas.

Berdasarkan distribusi jarak temporal suar surya kelas M dan X relatif terhadap puncak setiap parameter magnetik, tidak ditemukan adanya parameter yang konsisten dalam berperan sebagai penanda pendahulu suar surya. Hal ini disimpulkan dari nilai standar deviasi distribusi jarak temporal yang tinggi (>60 jam) dan nilai rata-rata yang mendekati 0 jam. Hal ini mengimplikasikan proses pra-*flare* yang unik pada setiap suar surya.

Selain itu, diperoleh peringkat signifikansi parameter magnetik sebagai fitur dalam mengidentifikasi proses pra-*flare*. Diperoleh peringkat 12 parameter magnetik yang menunjukkan adanya korelasi evolusi dengan proses pra-*flare* berturut-turut yaitu TOTPOT, ABSNJZH, SAVNCPP, TOTUSJH, TOTUSJZ, USFLUX, AREA_ACR, MEANPOT, SHRGT45, MEANSHR, MEANGAM, dan R_VALUE. Dengan memperhatikan signifikansi fitur ini, diperoleh jumlah fitur yang paling optimal untuk dijalankan pada *support vector machine* adalah sebanyak 7 fitur dan teramati peningkatan presisi relatif sebesar 40% terhadap pengurangan fitur.

Kata kunci: suar surya, prediksi suar surya, *supervised learning*, SHARP, Python.

BAB I

PENDAHULUAN

I.1. Latar Belakang

Matahari adalah objek dengan pengaruh terbesar di tata surya. Selain sebagai objek paling massif di tata surya, Matahari juga berperan penting dalam dinamika materi dan energi antarplanet. Energi yang dipancarkan Matahari menjadi penggerak iklim bagi planet-planet di tata surya, tak terkecuali planet Bumi. Studi yang mempelajari efek Matahari dan sumber kosmis lain terhadap magnetosfer, ionosfer, dan termsofer Bumi dan dampaknya pada teknologi dan kehidupan adalah cuaca antariksa.

Cuaca antariksa juga digunakan sebagai istilah yang menggambarkan dinamika lingkungan geoantariksa. Aktivitas permukaan Matahari merupakan faktor utama yang menetukan keadaan cuaca antariksa. Salah satu bentuk aktivitas permukaan Matahari yang paling energetik adalah suar surya atau *solar flare*.

Suar surya adalah fenomena letusan energi dan gelombang elektromagnetik yang terjadi pada atmosfer Matahari. Fenomena ini sering diikuti oleh kenaikan kecerlangan pada daerah aktif di piringan Matahari yang mengalami letusan energi. Sampai saat ini mekanisme terbentuknya suar surya masih belum diketahui dengan baik, namun astronom percaya bahwa pemicu terjadinya flare adalah rekoneksi magnetik yang terjadi di atmosfer Matahari (Hudson, 1995).

Rekoneksi magnetik terjadi ketika dua garis medan magnet yang memiliki polaritas berlawanan bertemu dan membentuk ikalan medan magnet yang lebih kecil. Proses ini melepaskan energi magnetik yang tersimpan dalam medan magnet dan mengubahnya menjadi energi kinetik plasma dan radiasi elektromagnetik. Pelepasan energi yang cepat pada proses inilah efek yang diamati pada suar surya (Priest & Forbes, 2000).

Suatu suar surya biasanya memiliki rentang energi sekitar 10^{20} Joule. Untuk suar surya yang massif, energi yang dilepaskan dapat mencapai orde 10^{25} Joule. Oleh karena itu, fenomena suar surya dapat mempengaruhi cuaca antariksa secara signifikan. Intesitas radiasi yang tinggi pada panjang gelombang pendek yang dipancarkan oleh suar surya dan semburan partikel bermuatan pada lontaran massa korona (LMK) yang mengiringi suar surya dapat mempengaruhi kehidupan di Bumi. Selain itu, fluktuasi medan magnet yang dibawa oleh suar surya juga dapat mempengaruhi medan magnet Bumi dan merusak teknologi berbasis angkasa maupun

landas Bumi yang rentan terhadap fluktuasi medan magnet (Jyothi, 2021). Oleh karena itu, studi suar surya dan upaya prediksi kemunculannya menjadi bidang penelitian yang berkembang pesat di era modern ini.

Studi mengenai suar surya telah dilakukan setidaknya sejak 1859 oleh Carrington yang mengamati adanya bercak putih pada bintik Matahari yang bergerak selama terjadinya suar surya. Dari pengamatan ini, Carrington mengasosiasikan fenomena suar surya dengan bintik Matahari.

Upaya prediksi kemunculan suar surya sering dilakukan melalui analisis bintik Matahari dengan karakteristik tertentu seperti yang dilakukan oleh McIntosh pada tahun 1990. McIntosh membuat suatu model klasifikasi bintik Matahari yang menunjukkan korelasi kuat dengan probabilitas kemuculan suar surya pada kelas bintik tertentu dalam model klasifikasinya. Pada tahun 1997, Kálmán menemukan hubungan antara kemunculan suar surya dengan fluks magnetik dalam pergerakan bintik Matahari. Penelitian yang dilakukan oleh Herdiwijaya dan Imelda pada 2006 menunjukkan adanya kaitan erat antara kemunculan suar surya dengan konfigurasi magnetik dari suatu grup bintik Matahari.

Dewasa ini, pemahaman mengenai keterkaitan rekoneksi magnetik dengan kemunculan suar surya telah diterima secara luas. Teori yang memodelkan kemunculan suar surya seperti 2D CSHKP (Carmichael, 1964; Sturrock, 1966; Hirayama, 1974; Kopp and Pneuman, 1976) dan model 3D terbaru (Janvier dkk., 2014) berhasil mendeskripsikan proses erupsi beserta efek-efek fisisnya. Namun model-model ini masih kurang menjelaskan mengenai mekanisme pemicu ledakan. Hingga saat ini, model pemicu suar surya yang definitif masih belum ditemukan. Oleh karena itu, prediksi suar surya secara teoritik masih menjadi tantangan besar.

Kendati lokasi terjadinya rekoneksi berada pada lapisan atmosfer atas Matahari, data magnetik pada lapisan ini masih belum tersedia secara ekstensif. Di sisi lain, parameter magnetik Matahari pada laipsan fotosfer telah tercatat secara ekstensif semenjak misi SDO pada 2010 (Bobra dkk., 2014). Oleh karena itu, penelitian mengenai suar surya banyak difokuskan untuk menemukan hubungan empirik antara kemuculan suar surya dan konfigurasi parameter magnetik fotosfer. Dengan berkembangnya kekuatan komputasi dan meningkatnya ketersediaan data, penelitian dalam membangun model prediksi suar surya mulai berfokus pada pembelajaran mesin (*machine learning*) dalam menemukan hubungan empirik ini. Beberapa penelitian dalam mengahsilkan model prediksi suar surya menggunakan pembelajaran mesin dan data magnetik fotosfer antara lain penelitian oleh Colak dan Qahwaji (2007), Bobra dan Couvidat (2015), Nishizuka dkk. (2021), dan Zhang dkk. (2022).

Berdasarkan perkembangan penelitian mengenai prediksi kemunculan suar

surya, ketersediaan data, dan kemajuan pembelajaran mesin dalam membuat suatu model prediksi, penulis terispirasi untuk membuat model prediksi berbasis *supervised learning* terhadap kemunculan suar surya, khususnya kelas X dan M dengan memanfaatkan data terbaru dari SHARP. Adapun model *supervised learning* yang akan diterapkan guna memperoleh model prediksi *support vector machine*. Model *supervised learning* ini dipilih berdasarkan hasil penelitian sebelumnya (Fernanda, 2022) yang menunjukkan bahwa *support vector machine* memiliki performa yang paling tinggi di antara model *supervised learning* lainnya.

I.2. Rumusan Masalah

- Rumusan masalah yang diajukan oleh penulis disusun berdasarkan beberapa asumsi. Asumsi pertama didasari oleh fakta bahwa suar surya merupakan proses pelepasan energi. Oleh karena itu, wajar jika kita asumsikan bahwa suar surya didahului oleh suatu proses pra-*flare* yang merupakan proses penumpukan energi magnetik. Proses ini tercermin dalam evolusi parameter magnetiknya. Dengan asumsi demikian, rumusan masalah pertama adalah "**Apakah terdapat suatu proses di parameter magnetik tertentu yang selalu mendahului terjadinya suar surya?**".
- Asumsi yang kedua adalah proses pra-*flare* yang tercerminkan dalam parameter magnetik terlihat sebagai fluktuasi atau perubahan nilai prameter secara tiba-tiba pada daerah aktif sebelum suar surya terjadi. Berdasarkan asumsi tersebut, rumusan masalah yang kedua adalah "**Bagaimana kenampakan proses pra-flare pada setiap parameter magnetik?**".
- Dengan menginvestigasi kedua rumusan masalah di atas, penulis dapat mengajukan rumusan masalah yang ketiga yaitu "**Bagaimana pengaruh pengurangan jumlah fitur berdasarkan proses pra-flare yang tampak pada parameter magnetic terhadap performa model prediksi?**"

I.3. Batasan

Adapun batasan masalah pada penilitian ini antara lain kelas suar surya, posisi daerah aktif Matahari, dan rentang waktu data pengamatan.

- Kelas suar surya: Suar surya yang ditinjau hanya suar surya kuat dengan kelas M atau X. Sedangkan suar surya dengan kelas selain M dan X dianggap tidak terjadi suar surya atau *non-flaring* karena dianggap terlalu lemah.

- Posisi daerah aktif: Data pengamatan daerah aktif yang ditinjau adalah daerah aktif dengan posisi dalam rentang $\pm 69^{o}$ dari bujur meridian piringan Matahari. Batasan ini diambil untuk meminimalkan error pengukuran akibat efek proyeksi terhadap kelengkungan pada tepian piringan Matahari.
- Rentang waktu pengamatan: Batasan berupa rentang data dalam domain waktu, yaitu dari 2 Mei 2010 pukul 00:00 UTC sampai 31 Desember 2021 pukul 23:48 UTC, muncul akibat ketersediaan data dari sumber utama, SHARP, yang hanya menyediakan data mulai dari 1 Mei 2010.

I.4. Tujuan

Berdasarkan rumusan masalah yang telah diuraikan, penulis merumuskan tujuan penelitian sebagai berikut:

1. Menentukan peran masing-masing parameter magnetik terhadap kemunculan suar surya
2. Memprediksi proses pra-*flare* untuk suar surya kelas X dan M dengan menggunakan *supervised learning*

I.5. Metodologi

Metodologi yang digunakan dalam penulisan penelitian ini meliputi studi pustaka serta pengumpulan dan pengolahan data parameter magnetik daerah aktif dari SHARP dan data kemunculan suar surya dari NOAA. Adapun data yang digunakan dalam penelitian ini adalah data dari instrumen HMI (*Helioseismic and Magnetic Imager*) yang merupakan bagian dari misi SDO (*Solar Dynamics Observatory*). Adapun pengolahan dan analisis data dilakukan dengan menggunakan bahasa pemrograman Python.

I.6. Sistematika Penulisan

Penulisan Penelitian ini terbagi ke dalam 6 bab. Bab I berisi tentang pendahuluan yang terdiri dari latar belakang, rumusan dan batasan masalah, tujuan, metodologi, dan sistematika penulisan, Selanjutnya bab II berisi tentang tinjauan pustaka yang menjadi dasar dan pendukung dalam penelitian ini. Kemudian, bab III berisi tentang uraian data dan instrumen penelitian. Setelah itu, bab IV berisi proses pengolahan data yang dilakukan dalam penelitian. Bab V berisi hasil pengolahan data dan diskusi serta pandangan penulis mengenai hasil. Terakhir, bab VI berisi simpulan yang diperoleh dari penelitian ini.

BAB II

KONSEP DASAR DAN STUDI PUSTAKA

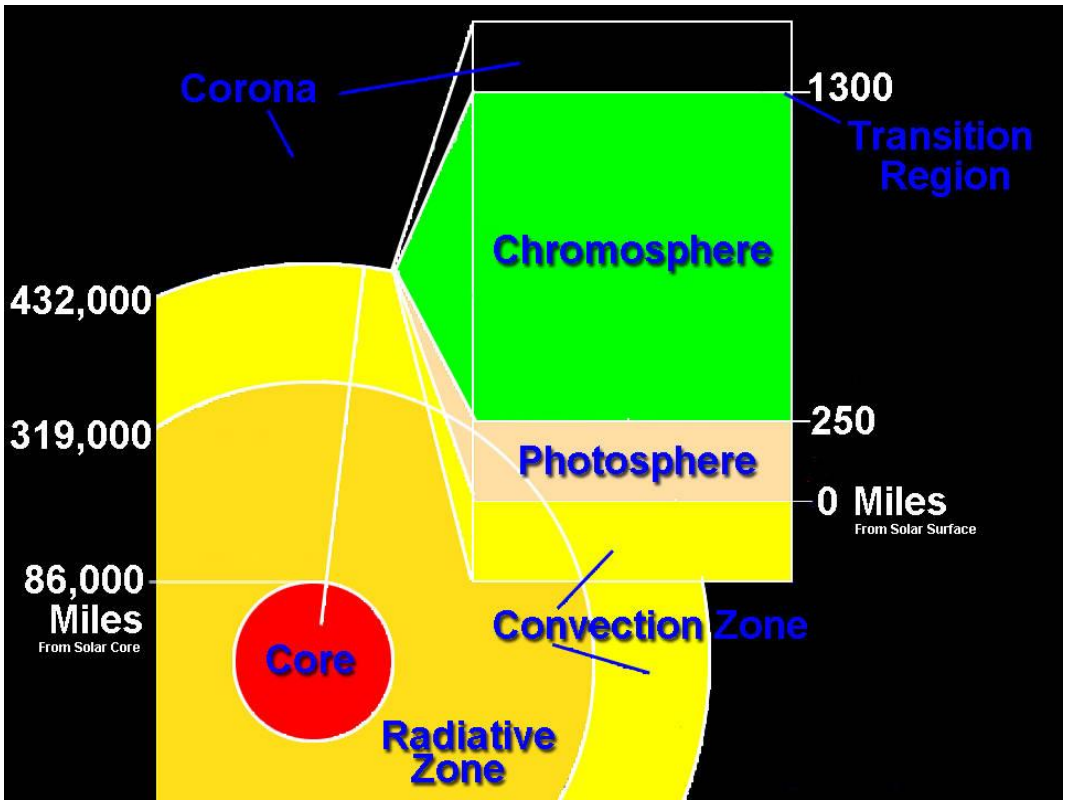
II.1. Aspek Teoritik

II.1.1 Matahari

Matahari adalah bintang utama sistem tata surya. Matahari merupakan bintang deret utama dengan klasifikasi G2V. Dengan massa $1{,}988 \times 10^{30}$ kg, fraksi massa Matahari terhadap keseluruhan massa tata surya mencapai 99%. Selain itu, reaksi fusi yang berlangsung di inti Matahari menghasilkan daya sebesar $3{,}82 \times 10^{26}$ Watt (William, 2018). Energi yang dipancarkan Matahari ini adalah penggerak utama dinamika energi di tata surya. Oleh karena itu, Matahari merupakan objek yang paling berpengaruh baik secara gravitasional maupun dari segi gelombang elektromagnetik dan dinamika plasma.

Dengan produksi energi yang masif, Matahari memiliki temperatur yang sangat tinggi. Oleh karena itu, komposisi materi di Matahari didominasi oleh plasma dan sedikit gas netral. Tanpa adanya materi padat maupun cair di Matahari, permukaan Matahari biasanya didefinisikan sebagai lapisan fotosfer, yaitu lapisan atmosfer dimana gelombang elektromagnetik diradiasikan sebagai radiasi benda hitam ke luar angkasa. Lapisan ini memiliki temperatur efektif sebesar 5772 K (William, 2018).

Seperti yang diilustrasikan oleh gambar II.1, lapisan lain yang merupakan bagian dari atmosfer Matahari berturut-turut berdasarkan ketinggiannya terhadap permukaan Matahari adalah kromosfer dan korona. Kromosfer adalah lapisan dengan ketebalan 2000 km di atas fotosfer. Sedangkan korona merupakan lapisan terluar dari atmosfer Matahari yang menjulang hingga 10.000 km di atas permukaan Matahari. Kedua lapisan ini memiliki karakteristik adanya kenaikan temperatur seiring dengan ketinggian. Material yang sampai di lapisan atas atmosfer Matahari akan dibawa oleh angin Matahari menuju ruang antarplanet. Selain itu, terdapat lapisan transisi yang memisahkan kromosfer dan korona yang memiliki ketebalan sekitar 100 km. Pada lapisan ini terjadi kenaikan temperatur yang signifikan, dari 8000 K ke 500.000 K.



Gambar II.1. Bagan struktur atmosfer Matahari.
Sumber: https://www.nasa.gov/mission_pages/iris/multimedia/layerzoo.html

Atmosfer Matahari adalah lingkungan yang sangat dinamis. Pada lapisan fotosfer, kita dapat melihat adanya berbagai fitur yang bertanggungjawab atas dinamika materi dan energi di atas permukaan Matahari hingga ruang antarplanet. Salah satu fitur magnetik yang ditemukan pada fotosfer Matahari adalah daerah aktif (*active region*). Daerah aktif merupakan suatu area temporer di atmosfer Matahari yang ditandai dengan medan magnet yang kuat dan kompleks. Medan magnet ini diduga dibangkitkan oleh suatu proses dinamo yang terjadi pada plasma yang berotasi di zona konveksi Matahari.

Dengan konfigurasi medan magnet yang kompleks, daerah aktif dapat menyimpan dan mengakumulasi energi magnetik dalam kuantitas yang masif. Energi ini dapat dilepaskan melalui manifestasi gelombang maupun rekoneksi magnetik yang memicu suar surya dan lontaran massa korona (Woods dkk., 2017).

II.1.2. Suar surya

Daerah aktif sering diasosiasikan dengan bintik Matahari dan merupakan sumber dari suar surya dan lontaran massa korona (Zell, 2015). Suar surya atau *solar flare* merupakan suatu fenomena yang disebabkan oleh asosiasi berbagai proses dinamis yang berada di atmosfer Matahari (Parker, 1963). Selain itu, suar

surya juga biasanya didefinisikan sebagai peristiwa yang berasal dari pelepasan energi dari medan magnet yang berada di korona, baik melalui proses rekoneksi ataupun bentuk disipasi magnetik lainnya (Dennis dan Schwartz, 1989). Oleh karena itu, konfigurasi medan magnet pada atmosfer Matahari sangat penting untuk diteliti untuk memahami dan memprediksi kemunculan suar surya.

Kuat suar surya sendiri sering diklasifikasikan berdasarkan kecerlangannya pada panjang gelobang sinar-X yaitu pada 1 – 8 Å. Sistem klasifikasi ini membagi suar surya ke dalam lima kelas yang dinotasikan oleh huruf-huruf alfabet: A, B, C, M , X. Rentang energi setiap kelas suar surya dapat dilihat pada tabel II.1

Tabel II.1: Sistem klasifikasi suar surya berdasarkan kecerlangannya pada panjang gelombang sinar-X. Sumber: Su, 2007.

Kelas	Intensitas Maksimum pada 1 – 8 Å (W/m^2)
X	$I \geq 10^{-4}$
M	$10^{-5} \leq I < 10^{-4}$
C	$10^{-6} \leq I < 10^{-5}$
B	$10^{-7} \leq I < 10^{-6}$
A	$I < 10^{-7}$

Selain dari notasi alfabet, kekuatan suar surya juga ditandai dengan angka 1 – 9 yang mengikuti huruf kelas suar surya. Angka ini menunjukkan kekuatan relatif suar surya pada suatu kelas terhadap suar surya dengan angka 1 pada kelas yang sama, contohnya suar surya kelas M3 tiga kali lebih kuat daripada suar surya kelas M1.

Proses rekoneksi yang sering ditemukan sebagai pemicu suar surya biasa terjadi pada atmosfer bagian atas Matahari (kromosfer & korona). Sayangnya hingga saat ini, pemetaan langsung terhadap konfigurasi medan magnet untuk atmosfer bagian ini masih terbatas (Ishikawa dkk., 2021). Namun konfigurasi medan magnet untuk lapisan fotosfer telah terpetakan dan terparametrisasi dengan baik, bahkan secara kontinu (Bobra dkk., 2014). Oleh karena itu, model prediksi suar surya pada masa kini didasari dari nilai pengukuran medan magnet di fotosfer untuk memparametrisasi daerah aktif yang berperan dalam pembentukan suar surya (Bobra dan Couvidat, 2015).

II.1.3. Prediksi Cuaca Antariksa

Secara umum, prediksi adalah prakiraan mengenai kejadian yang belum terjadi ketika prakiraan dilakukan. Suatu model prediksi haruslah memiliki tiga aspek sebagai berikut:

- **Aktual.** Suatu model prediksi dapat memprediksi suatu peristiwa

sedini mungkin sebalum peristiwa tersebut terjadi sehingga preparasi dan mitigasi terhadap peristiwa tersebut secara praktis dapat dilakukan.

- **Akurat**. Artinya hasil prediksi sesuai dengan kenyataan dalam batasan yang dapat diterima.
- **Andal**. Artinya model prediksi dapat diandalkan pada segala kondisi

Tak terbatas pada suar surya saja, prediksi terhadap cuaca antariksa merupakan tantangan besar karena proses cuaca antariksa melibatkan proses elektrohidrodinamika yang kompleks dan lingkungan yang tak terbatas. Model cuaca antariksa pada dasarnya dapat dideskripsikan melalui dua sudut pandang yang berbeda. Sudut pandang pertama adalah sudut pandang berbasis informasi. Sudut pandang ini bersifat empiris dan bergantung pada data deret waktu yang tersedia. Sudut pandang ini mengasumsikan adanya hubungan antarparameter cuaca antariksa. Dengan mengamati evolusi tiap parameter, kita dapat memperoleh hubungan antarparameter secara empiris berdasarkan data yang tersedia. Biasanya *modeling* melalui sudut pandang ini dilakukan dengan memulai dari analisis data deret waktu, parameter mana sajakah yang ditinjau dan dinilai memiliki pengaruh terhadap suatu fenomena cuaca antariksa, kemudian diperoleh dinamika antarparameter dalam data, lalu lakuakan interpretasi fisis terhadap hasil yang diperoleh (Vassiliadis, dkk., 1995, 1996).

Sudut pandang lain dalam memodelkan cuaca antariksa adalah sudut pandang berbasis fisis. Pada sudut pandang ini, fenomena cuaca antariksa dianggap sebagai transport kuantitas fisis (energi, momentum, helisitas, dll) dalam lingkungan antariksa. Sudut pandang ini menggunakan teori fisis seperti magnetohidrodinamika dan kinematika plasma dalam mengaproksimasi aktivitas cuaca antariksa. Model prediksi cuaca antariksa berbasis fisis memiliki keunggulan dalam batasan variabel masukan serta keluaran yang digunakan. Model ini dapat memprediksi pada set variabel yang lebih beragam. Selain itu, tidak seperti model prediksi berbasis informasi, model prediksi berbasis fisis dapat melakukan prediksi di luar domain *training*.

Penelitian ini secara khusus menggunakan pendekatan berbasis informasi (empiris) dalam memprediksi suar surya sebagai suatu fenomena cuaca antariksa. Pertimbangan atas pilihan ini adalah model empiris tidak terbatasi oleh hukum kekekalan seperti pada model berbasis fisis. Oleh karena itu, model empiris cenderung lebih cepat dan akurat daripada model fisis dalam memprediksi berdasarkan suatu set variabel yang dilatih (Vassiliadis, dkk., 2007).

Untuk kasus suar surya, secara khusus penelitian ini menggunakan pendekatan secara empiris berdasarkan fakta bahwa suar surya merupakan proses

pelepasan energi secara tiba-tiba. Oleh karena itu, wajar apabila diasumsikan bahwa terdapat suatu proses penumpukan energi sebelum terjadi suar surya. Proses ini disebut sebagai proses pra-*flare*.

Saat ini suar surya sebagai fenomena yang erat kaitannya dengan aktivitas megnetik Matahari telah diterima secara luas sebagai fakta (Hudson dan Ryan, 1995). Dengan demikian, proses pra-*flare* semestinya tercermin dalam evolusi parameter magnetik daerah aktifnya (Duan, dkk., 2019). Di sisi lain, daerah aktif memiliki berbagai parameter magnetik yang masing-masing berpotensi merepresentasikan proses pra-*flare*.

Penelitian ini bertujuan untuk meninjau potensi parameter magnetik tersebut dalam hubungannya dengan proses pra-*flare* dengan pendekatan empiris. Dengan diketahuinya peran masing-masing parameter magnetik terhadap proses pra-*flare*, model prediksi empiris dapat dibangun berdasarkan signifikansi peran parameter tersebut sebagai fitur.

II.2. Aspek Komputasi

Pembelajaran mesin berperan penting dalam penyelesaian masalah di berbagai bidang, khususnya statistika dan kecerdasan buatan serta penerapannya di bidang teknik dan sains. Beberapa contoh kasus dimana pembelajaran mesin dapat diterapkan dalam penyelesaian problema di beberapa bidang antara lain:

- Prediksi apakah seorang pasien menderita penyakit jantung berdasarkan demografi, diet, dan pengukuran medis,
- Prediksi pasar saham dalam 6 bulan ke depan berdasarkan performa perusahaan dan data kondisi ekonomi,
- Identifikasi apakah suatu surel merupakan surel penting atau *spam* berdasarkan isi surel.

Pada masing-masing kasus di atas, kita memerlukan suatu *dataset* yang memuat hasil pengukuran lampau yang dapat bersifat kuantitatif seperti nilai saham, atau dapat pula bersifat kategori seperti pengidap/bukan pengidap penyakit. Dalam dataset ini terdapat nilai pengukuran yang dikenal sebagai fitur serta nilai keluaran yang ingin diprediksi yang biasa disebut sebagai label. Contohnya pada kasus prediksi pasien penderita penyakit jantung, fitur dapat berupa demografi, diet, dan berbagai pengukuran medis yang dilakukan terhadap satu objek (dalam kasus ini manusia) seperti berat badan dan riwayat penyakit. Sedangkan label yang ingin diprediksi dalam kasus ini adalah berupa kategori yaitu pengidap/bukan pengidap penyakit jantung. Dataset ini kemudian dibagi menjadi dua bagian yaitu *training dataset* dan *test dataset*.

Dengan menggunakan *training dataset*, kita dapat melatih suatu model prediksi agar model tersebut dapat mengenali pola tertentu antara fitur dengan label. Dengan demikian, diharapkan kita dapat memperoleh suatu model prediksi yang dapat mengestimasi nilai keluaran dari pengukuran baru. Sedangkan *test dataset* digunakan untuk mengevaluasi performa dari prediksi yang dilakukan oleh model. Evaluasi dilakukan dengan membandingkan hasil prediksi model dengan label sesungguhnya pada *test dataset*. Pemisahan dataset menjadi *training* dan *test dataset* dilakukan untuk menghindari *overfitting* terhadap *dataset*. *Overfitting* terjadi ketika performa model sangat baik hanya pada kasus data tertentu saja sehingga objektivitas model tergolong rendah. Model prediksi yang baik adalah model yang dapat meprediksi nilai keluaran dengan akurat dan konsisten.

Deskripsi di atas merupakan konsep dasar dari *supervised learning*. Kata "*supervised*" menunjukkan adanya variabel nilai keluaran atau label yang telah tersedia dalam *training dataset* dan *test dataset* untuk membantu proses pembelajaran model. Sedangkan pada *unsupervised learning*, kita hanya memiliki dataset yang mengandung fitur tanpa adanya label. Pada *unsupervised learning*, fokus utama model adalah untuk mendeskripsikan bagaimana data terorganisasi dan terkelompokkan.

Di balik penerapan pembelajaran mesin, diperlukan suatu proses pengolahan dataset yang kompleks hingga diperoleh dataset *training* maupun *test* yang sesuai untuk diproses oleh model prediksi. Tak jarang data perlu dibersihkan, ditransformasikan, bahkan dikurangi atau dihapus agar model pembelajaran mesin dapat mempelajari pola dengan lebih efisien. Selain pembersihan data dari titik data tak sah atau *nan*, penelitian ini sendiri menerapkan normalisasi, transformasi menggunakan *short time fourier transform* (STFT), dan analisis fitur dengan *principal component analysis* (PCA) sebelum menjalankan algoritma *supervised learning* terhadap data.

Penelitian ini secara khusus menggunakan model *supervised learning* yaitu *support vector machine* untuk memprediksi kemunculan suar surya. Model prediksi yang ini merupakan model klasifikasi biner (*binary classifier*), yaitu model yang dapat mengklasifikasikan setiap titik data ke salah satu dari dua kelas yaitu positif dan negatif. Titik data dengan label positif mengindikasikan akan terjadi suar surya dalam waktu dekat pada daerah aktif yang sama dengan titik data tersebut. Sebaliknya, titik data dengan label negatif mengindikasikan tidak akan terjadi suar surya dalam waktu dekat pada daerah aktif yang sama dengan titik data tersebut.

II.2.1. *Principal Component Analysis* (PCA)

Principal Component Analysis (PCA) adalah suatu teknik yang digunakan dalam menganalisis data dengan dimensi fitur yang tinggi dengan meninjau kontribusi setiap fitur terhadap variasi total data. Fitur dengan nilai variasi yang tinggi akan memiliki kontribusi terhadap variansi total data yang lebih tinggi daripada fitur dengan nilai variasi yang rendah. Hal ini didasari dari intuisi bahwa fitur dengan distribusi yang lebih menyebar bersifat lebih informatif daripada fitur dengan distribusi yang sempit.

Adapun algoritma untuk melakukan PCA adalah sebagai berikut.

1. Siapkan dataset:
    a. Pastikan dataset dalam bentuk matriks dengan setiap baris mewakili satu sampel dan setiap kolom mewakili satu fitur.
    b. Jika dataset memiliki rentang data atau skala yang berbeda, lakukan normalisasi atau standarisasi data agar semua fitur memiliki skala yang mirip.
2. Menghitung rata-rata setiap fitur:
    a. Hitung nilai rata-rata dari setiap kolom (fitur) dalam dataset.
    b. Kurangi setiap nilai dalam dataset dengan rata-rata yang sesuai untuk memindahkan pusat distribusi fitur ke rata-ratanya.
3. Menghitung matriks kovarian:
    a. Hitung matriks kovarian dari dataset yang telah diubah pusatnya dengan menggunakan rumus: $\boldsymbol{Cov(X) = \frac{1}{n} X^T X}$ di mana X adalah dataset yang telah diubah pusatnya dan n adalah jumlah sampel dan X^T adalah transpos dari X.
4. Menghitung nilai dan vektor eigen:
    a. Hitung nilai vektor eigen dari matriks kovarian.
    b. Vektor eigen adalah vektor yang memberikan sumbangan terbesar dalam variasi dataset.
    c. Vektor eigen dapat dihitung menggunakan metode Dekomposisi Nilai Singular (*Singular Value Decomposition, SVD*) atau metode iteratif seperti Metode Daya Lanjutan (*Power Iteration*).
5. Mengurutkan nilai eigen:
    a. Urutkan nilai eigen dalam urutan menurun sehingga komponen utama (*principal component*) pertama adalah komponen utama dengan nilai eigen terbesar yang memberikan sumbangan terbesar terhadap variasi data.
6. Memilih jumlah komponen utama:
    a. Tentukan jumlah komponen utama yang akan dipertahankan

berdasarkan jumlah variasi yang ingin dipertahankan atau dengan menggunakan aturan seperti metode persentase variasi kumulatif.

7. Menghasilkan matriks transformasi:
    a. Pilih jumlah komponen utama yang dipilih.
    b. Ambil vektor eigen terkait dengan komponen utama yang dipilih.
    c. Bentuk matriks transformasi dengan menggunakan vektor eigen ini sebagai kolom-kolomnya.
8. Transformasi dataset:
    a. Kalikan dataset yang telah diubah pusat dengan matriks transformasi yang dihasilkan.
    b. Hasilnya adalah dataset yang telah ditransformasi ke dalam ruang fitur yang baru, di mana setiap fitur merupakan kombinasi linear dari komponen utama.
9. Opsional: Rekonstruksi dataset awal:
    a. Jika perlu, dataset yang telah ditransformasi dapat direkonstruksi ke dalam ruang fitur asal.
    b. Untuk itu, dapat dikalikan kembali dengan matriks transformasi transposenya dan ditambahkan kembali dengan rata-rata fitur awal.

Pada penelitian ini sendiri, PCA digunakan untuk menganalisis signifikansi setiap fitur terhadap dataset total.

II.2.2. *Short Time Fourier Transform* (STFT)

Dengan proses pra-*flare* yang direpresentasikan sebagai fluktuasi atau perubahan yang signifikan secara tiba-tiba yang tampak pada parameter magnetik daerah aktif, identifikasi fluktuasi ini menjadi pekerjaan yang penting. Dalam rangka mengidentifikasi proses ini, peneltian sebelumnya (Fernanda, 2022) mengukur panjang waktu fluktuasi secara kualitatif dan menggunakannya sebagai jendela waktu prediksi. Penelitian ini secara khusus menggunakan *Short Time Fourier Transform* (STFT) dalam mengidentifikasi proses pra-*flare* pada masing-masing parameter magnetik yang ditinjau.

Seperti namanya, *Short Time Fourier Transform* merupakan suatu bentuk transformasi Fourier yang digunakan untuk menganalisis sinyal pada domain frekuensi. Namun transformasi Fourier biasa mengasumsikan bahwa keseluruhan sinyal bersifat stasioner. Hal ini menjadi batasan untuk sinyal non-stasioner yang mana kebanyakan deret waktu parameter magnetik daerah aktif Matahari merupakan sinyal non-stasioner. Di sisi lain, STFT melakukan segmentasi terhadap keseluruhan sinyal pada domain waktu dan menerapkan transformasi Fourier terhadap setiap segmen.

STFT memiliki persamaan

$$\boldsymbol{STFT}\{\boldsymbol{x}(\boldsymbol{t})\}(\boldsymbol{\omega},\boldsymbol{\tau}) = \int_{-\infty}^{\infty} \boldsymbol{x}(\boldsymbol{t})\,\boldsymbol{w}(\boldsymbol{t}-\boldsymbol{\tau})\boldsymbol{e}^{-\boldsymbol{i}\boldsymbol{\omega}\boldsymbol{t}}\boldsymbol{dt}\,, \qquad \text{(II.I)}$$

dengan $x(t)$ merupakan sinyal deret waktu, w sebagai *window function*, ω merupakan tengah bin sumbu frekuensi yang disampel, dan τ merupakan tengah bin sumbu waktu. Hasil dari STFT merupakan suatu *colormap* atau spektrogram yang merepresentasikan perubahan konten frekuensi terhadap waktu dalam suatu sinyal dengan sumbu warna sebagai koefisien korelasi spektrum dan dua sumbu lainnya merepresentasikan waktu dan frekuensi.

Kelebihan utama dari STFT adalah kemampuannya dalam memberikan koefisien korelasi spektrum pada waktu tertentu dalam sinyal. Kelebihan ini menjadi penting dalam analisis sinyal non-stasioner seperti deret waktu parameter magnetik pada penelitian ini. Kelebihan lain dari STFT adalah STFT memberikan keleluasaan dalam mengatur resolusi sampling domain frekuensi dengan mengatur panjang segmen dalam domain waktu. Selain itu, dengan sifat STFT yang memperbolehkan sebagian dari setiap segmen yang berdekatan untuk tumpang tindih, spektrogram yang dihasilkan oleh STFT memberikan transisi yang lebih halus antar segmen.

Sebelumnya, STFT telah diterapkan secara luas di bidang penelitian lain seperti pemrosesan audio dan gambar, kedokteran dan medis, dan elektronik. STFT digunakan dalam mendeteksi variasi durasi pendek dalam tegangan listrik (Anggriawan, dkk., 2020). STFT juga berhasil diterapkan dalam mendeteksi keabnormalan jantung manusia (Bustami, dkk., 2007).

Pada penelitian ini, fluktuasi signifikan yang tampak pada parameter magnetik sebelum terjadi suar surya diasumsikan sebagai suatu proses pra-*flare*. Dengan kelebihan STFT yang mampu memberikan representasi lokal korelasi spektrum pada domain waktu, STFT diharapkan dapat mendeteksi fluktuasi yang signifikan. Hal ini dikarenakan fluktuasi yang signifikan akan memberikan korelasi spektrum yang lebih besar daripada proses latar belakang.

II.2.3. *Support Vector Machine*

Support vector machine (SVM) adalah suatu model klasifikasi biner yang menggunakan fungsi pemisah dalam pengkategorian titik data. Fungsi pemisah pada *support vector machine* dapat memisahkan data yang terpisahkan secara linier maupun non-linier. SVM menggunakan pemetaan non-linier untuk mentransformasikan data

training awal ke dimensi yang lebih tinggi.

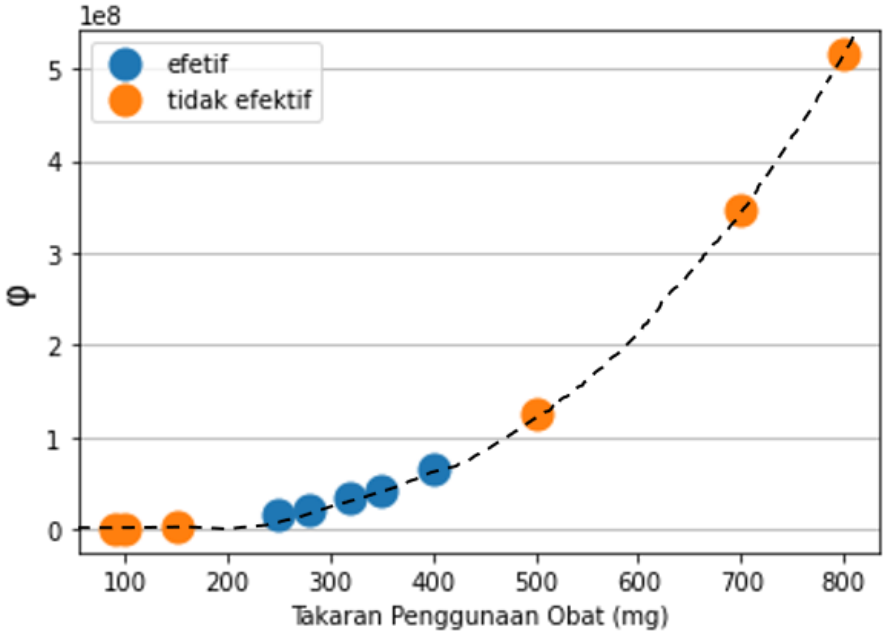
Mari kita tinjau suatu kasus sederhana penerapan SVM dalam klasifikasi biner berdasarkan satu fitur. Misal kita ingin memprediksi dosis efektif penggunaan suatu obat. *Training dataset* yang kita miliki adalah kumpulan data penggunaan obat dengan takaran tertentu serta apakah penggunaan obat tersebut efektif atau tidak. Pada kasus ini, takaran penggunaan obat berperan sebagai fitur, sedangkan efektif tidaknya penggunaan obat berperan sebagai label. Misal kita memiliki suatu dataset dengan distribusi seperti pada gambar II.3.

Gambar II.3. Distribusi data penggunaan obat dengan takaran tertentu serta efektifitasnya. Pada kasus ini, efektifitas yang berperan sebagai label adalah suatu kategori biner yaitu efektif/tidak efektif.

Untuk menghasilkan fungsi pemisah, kita perlu menaikkan dimensi data dengan mengalikan fitur dengan suatu fungsi kernel $\varphi(x_1,\ x_2, \dots, x_n)$, dengan x adalah fitur dan n adalah jumlah fitur. Salah satu contoh fungsi kernel adalah fungsi polinomial

$$\boldsymbol{\varphi = x^3}. \tag{II.2}$$

Dengan menerapkan kernel ini, kita mentrasnformasi data ke dimensi yang lebih tinggi seperti yang ditunjukkan oleh gambar II.4



Gambar II.4. Data diplot terhadap fungsi kernel $\boldsymbol{\varphi}$ (persamaan (II.1)) sebagai sumbu-y.

Kemudian dapat diperoleh fungsi pemisah yang berupa satu garis lurus untuk kasus fitur 1 dimensi ini. Fungsi pemisah ini sering disebut sebagai *support vector classifier* (SVC). Untuk kasus fitur 2 dimensi, fungsi pemisah berupa bidang, sedangkan fitur dengan dimensi lebih dari 2 memiliki fungsi pemisah berupa *hyperplane*. Salah satu keunggulan dari SVM adalah SVC sebagai fungsi pemisah memperbolehkan adanya misklasifikasi dengan menerapakan *soft margin* di sekitar SVC.

Support vector classifier diperoleh melalui persamaan

$$\boldsymbol{min}\left(\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\omega}^{\boldsymbol{T}}\boldsymbol{\omega}+\boldsymbol{C}\sum_{\boldsymbol{k=1}}^{\boldsymbol{m}}\boldsymbol{\varepsilon_k}\right), \tag{II.3}$$

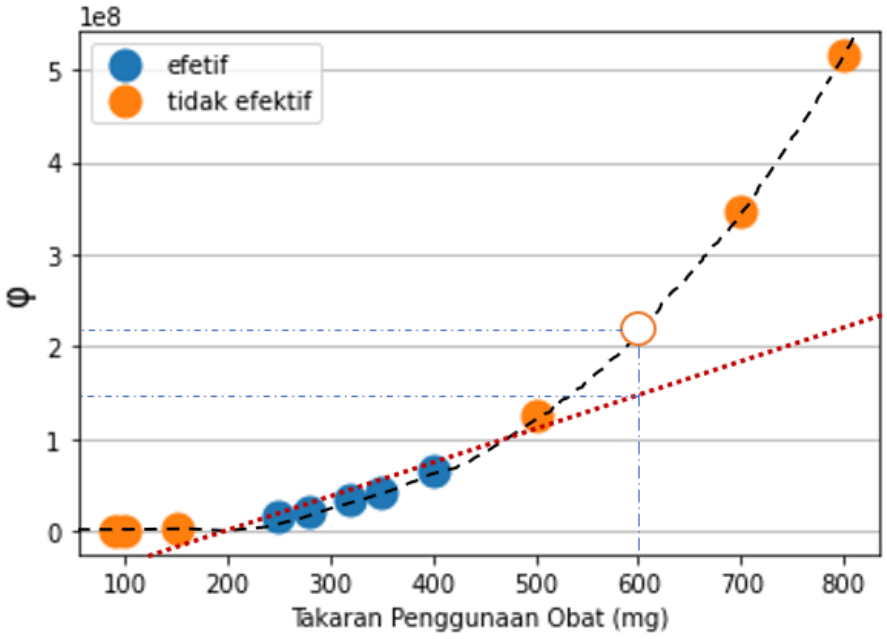
minimum untuk ε dan ω yang memenuhi

$$\boldsymbol{y_k}(\boldsymbol{\omega}^{\boldsymbol{T}}\boldsymbol{\varphi}(\boldsymbol{x_k})+\boldsymbol{b})\geq \mathbf{1}-\boldsymbol{\varepsilon_k}, \tag{II.4}$$

dan

$$\boldsymbol{\varepsilon_k}\geq \mathbf{0}, \tag{II.5}$$

dengan $\boldsymbol{\omega}$: vektor normal ke SVC, $\boldsymbol{\omega}^{\boldsymbol{T}}$: $\boldsymbol{\omega}$ transpos, $\boldsymbol{C}$: koefisien *soft margin*, $\boldsymbol{k}$: indeks titik data ke-k, $\boldsymbol{m}$: jumlah data *training*, $\boldsymbol{x}$: vektor titik data fitur, $\boldsymbol{\varphi}$: fungsi kernel. $\boldsymbol{\varepsilon}$: derajat misklasifikasi, $\boldsymbol{y}$: label titik data (Bobra dan Couvidat, 2015).



Gambar II.5. *Support vector classifier* (garis titik-titik merah) memisahkan data berdasarkan kelasnya. Sehingga apabila diperoleh pengukuran baru (lingkaran putih) kita dapat memprediksi klasifikasi pengukuran tersebut.

Contoh prediksi seperti yang diilustrasikan oleh gambar II.5, apabila kita ingin memprediksi efektivitas takaran obat sebanyak ~600 mg (lingkaran putih). Kita

kalilkan nilai ini dengan fungsi kernel. Oleh karena nilai fungsi kernel ini lebih besar daripada nilai SVC pada absis yang sama, kita peroleh hasil prediksi yang menyatakan bahwa takaran obat sebanyak ~600 mg tidak efektif. Sebaliknya apabila kita peroleh nilai ordinat yang lebih kecil daripada nilai SVC pada absis yang sama, model prediksi akan menyimpulkan bahwa takaran obat tersebut efektif.

Pada penelitian ini, performa model prediksi support vector machine dievaluasi dalam lima metrik skor yaitu akurasi, presisi, recall, f1, dan TSS yang memiliki formula

$$\boldsymbol{akurasi} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{N}}, \tag{II.6}$$

$$\boldsymbol{presisi} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}, \tag{II.7}$$

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}, \tag{II.8}$$

$$\boldsymbol{f1} = \boldsymbol{2}\,\frac{\boldsymbol{presisi} \cdot \boldsymbol{recall}}{\boldsymbol{presisi} + \boldsymbol{recall}}, \tag{II.9}$$

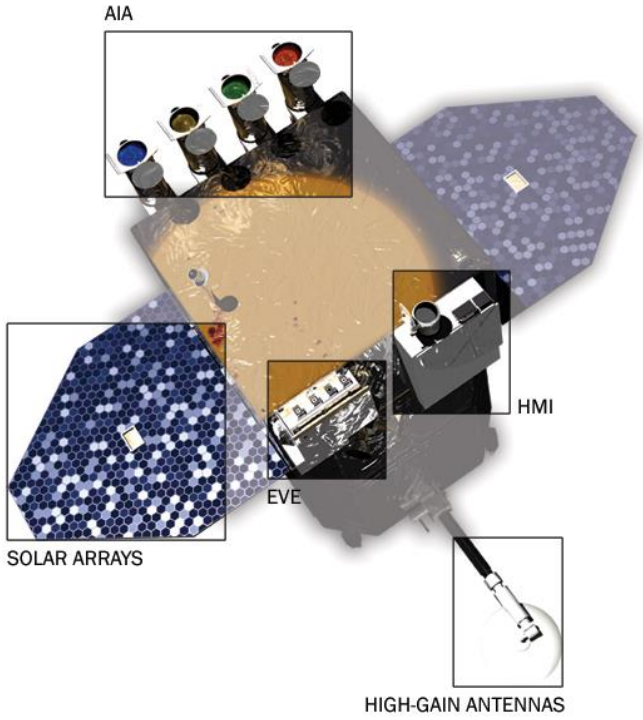
$$\boldsymbol{TSS} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} + \frac{\boldsymbol{FP}}{\boldsymbol{TN} + \boldsymbol{FP}}, \tag{II.10}$$

dengan N adalah jumlah prediksi, TP adalah jumlah positif benar (*true positive*), TN adalah jumlah negatif benar (*true negative*), FP adalah jumlah positif salah (*false positive*), dan FN adalah jumlah negatif salah (*false negative*).

BAB III

DATA DAN PERANGKAT PENELITIAN

III.1. *Solar Dynamic Observatory*



Gambar III.1. Ilustarsi bagan wahana antariksa *Solar Dynamics Observatory*. Sumber: https://sdo.gsfc.nasa.gov/mission/spacecraft.php

Solar Dynamic Observatory (SDO) merupakan misi pertama dalam rangka program *Living With Stars* milik NASA. Program ini adalah salah satu program NASA yang didesain untuk memahami sebab variabilitas Matahari serta dampaknya terhadap Bumi. SDO sendiri dirancang untuk mempelajari atmosfer Matahari pada berbagai panjang gelombang dengan resolusi spasial dan temporal yang tinggi.

SDO diluncurkan pada 11 Februari 2010 pukul 10:23 EST dengan menumpang roket Atlas V dari SLC 41 di Cape Canaveral. dan masih beroperasi hingga saat ini. Atlas V mengantarkan wahana antariksa SDO hingga ke *Geosychronous Transfer Orbit* (GTO). Kemudian dengan sistem penggerak miliknya sendiri, SDO mergerak menuju *Geosynchronous Orbit* (GEO).

SDO diharapkan dapat membantu kita dalam memahami medan magnet Matahari secara lebih mendalam, bagaimana struktur dan proses pembangkitan medan

magnet Matahari, serta bagaimana proses pelepasan energi mangnetik di atmosfer Matahari ke heliosfer baik dalam bentuk angin Matahari, partikel energetik, maupun gelombang elektromagnetik. Sedangkan tujuan saintifik SDO sendiri adalah demi mengembangkan pemahaman tentang tujuh pertanyaan saintifik:

1. Apa mekanisme yang menyebabkan siklus kuasi periodik 11 tahunan aktivitas Matahari?
2. Bagaimana fluks medan magnet di daerah aktif terbentuk, terkonsentrasi, dan terdispersi di permukaan Matahari?
3. Bagaimana rekoneksi magnetik skala kecil dapat mengatur ulang sistem arus dan medan topologi skala besar serta efeknya terhadap pemanasan korona dan percepatan angin Matahari?
4. Dimanakah lokasi terjadinya peningkatan variasi irandiasi pada spektrum EUV Matahari serta hubunganya dengan siklus aktivitas magnetik?
5. Bagaimanakah konfigurasi medan magnet yang dapat menghasilkan lontaran massa korona, erupsi filamen, dan suar surya yang dapat menghasilkan radiasi dan partikel energetik?
6. Dapatkah struktur serta dinamika angin Matahari di lingkungan geoantariksa ditentukan dari konfigurasi medan magnet dan struktur atmosfer di lingkungan dekat permukaan Matahari?
7. Dapatkah kita membuat prediksi yang akurat dan konsisten terhadap kapan suatu fenomena aktivitas Matahari terjadi dan pengaruhnya kepada cuaca antariksa dan iklim antariksa?

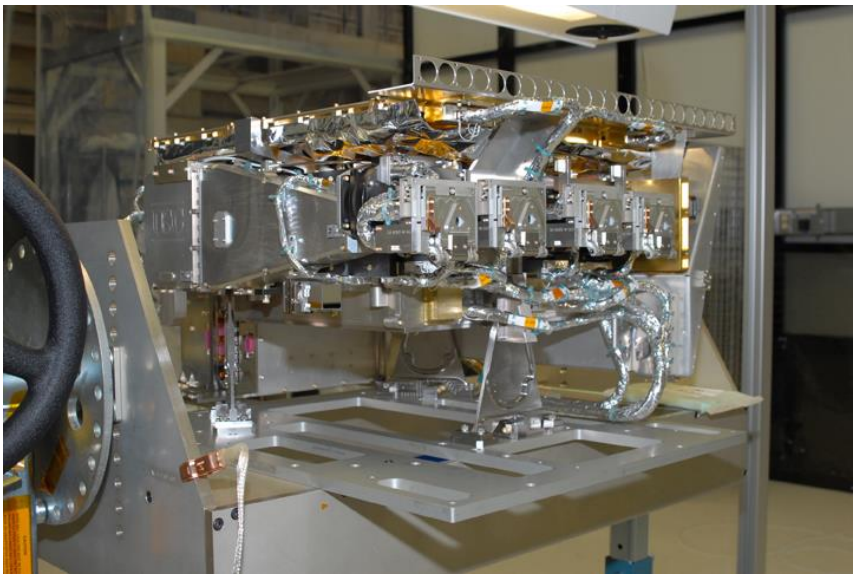
SDO dilengkapi oleh tiga instrumen, yaitu *Atmospheric Imaging Assembly* (AIA), *EUV Variability Experiment* (EVE), dan *Helioseismic and Magnetic Imager* (HMI).

1. ***Atmospheric Imaging Assembly* (AIA)**

Gambar III.2. Kenampakan instrumen *Atmospheric Imaging Assembly* **(**AIA). Sumber: https://sdo.gsfc.nasa.gov/mission/instruments.php

AIA adalah instrumen yang didesain khusus untuk melakukan pengamatan terhadap lapisan korona Matahari. Instrumen ini mengamati keseluruhan piringan Matahari dalam medan pandang 1,3 kali diameter sudut Matahari pada berbagai panjang gelombang secara bersamaan. Data pengamatan AIA memiliki resolusi 1 detik busur dan selang waktu antar-pengamatan selama 10 detik. Tujuan utama instrumen ini adalah untuk membantu memahami proses fisis yang terjadi di atmosfer Matahari. AIA dikelola oleh *Lockheed Martin Solar Astrophysics Laboratory*.

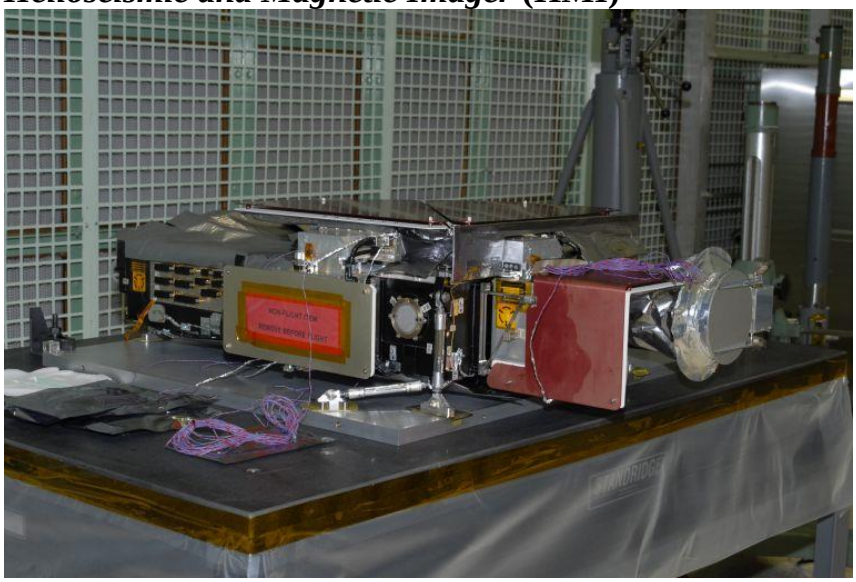
2. ***Extreme Ultraviolet Variability Experiment* (EVE)**



Gambar III.3. Kenampakan instrumen *Extreme Ultraviolet Variability Experiment* **(**EVE).
Sumber: https://sdo.gsfc.nasa.gov/mission/instruments.php

EVE adalah instrumen yang mengukur iradiansi Matahari pada panjang gelombang ultraviolet ekstrem (EUV), yaitu pada 0,1 – 105 nm. Radiasi pada panjang gelombang ini bertanggungjawab atas pemanasan termosfer dan pembentukan ionosfer Bumi. Rentang panjang gelombang yang lebar ini mengharuskan pengamatan EUV oleh EVE dilakukan melalui beberapa *channel*. Data pengamatan EVE memiliki resolusi spektrum sebesar 0,1 nm dan selang waktu antar-pengamatan selama 20 detik serta akurasi iradiansi minimal 25%. Data keluaran dari EVE adalah berupa spektograf untuk keseluruhan EUV. EVE dikelola oleh *Laboratory of Atmospheric and Space Physics University of Colorado*.

3. ***Helioseismic and Magnetic Imager* (HMI)**

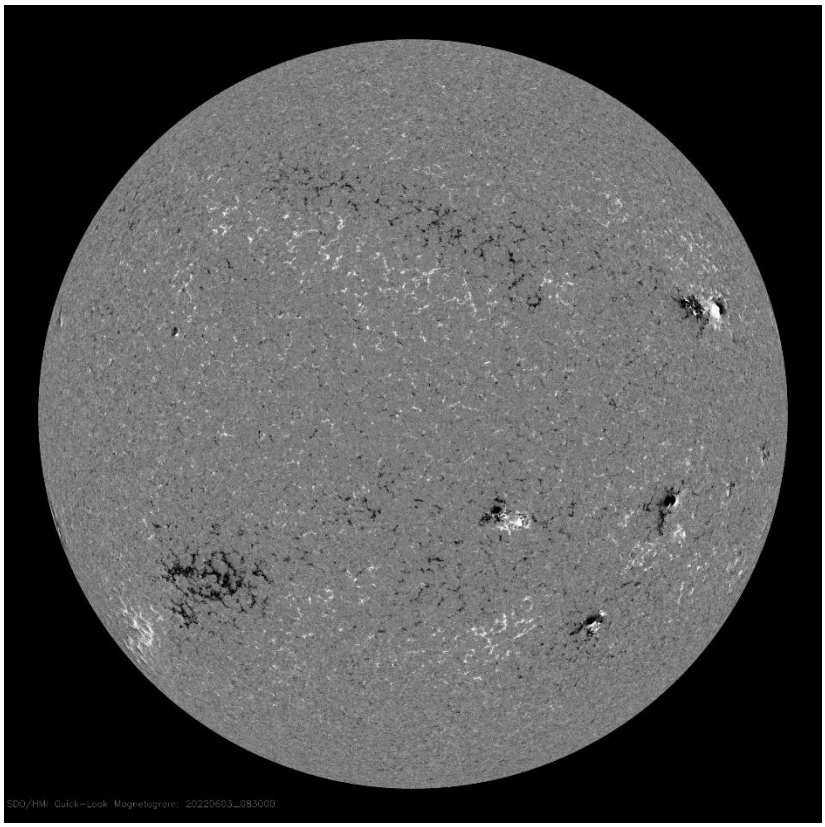


Gambar III.4. Kenampakan instrumen *Helioseismic and Magnetic Imager* (HMI).
Sumber: https://sdo.gsfc.nasa.gov/mission/instruments.php

HMI adalah instrumen yang didesain untuk mempelajari osilasi dan medan magnet di fotosfer. Pengamatan HMI dilakukan pada panjang gelombang 6173 Å dengan resolusi 1 detik busur. Data keluaran dari HMI berupa dopplergram atau pemetaan kecepatan permukaan Matahari, filtergram kontinum atau citra fotosfer pada rentang panjang gelombang lebar, dan magnetogram atau pemetaan vektor medan magnet fotosfer. HMI dikelola oleh *Stanford University*.

Penelitian ini menggunakan hasil olahan data magnetogram dari HMI sebagai salah satu sumber data utama. Magnetogram adalah pemetaan vektor medan magnet piringan Matahari. Nilai piksel pada magnetogram

menunjukkan kuat medan magnet pada arah tertentu. Contohnya pada Gambar III.5, warna hitam menunjukkan arah medan magnet menjauh dari pengamat, sedangkan warna putih menunjukkan arah medan magnet mendekat pengamat. Daerah dengan warna abu-abu menunjukkan daerah dengan kuat medan magnet yang lebih lemah daripada daerah dengan warna hitam maupun putih. Semakin hitam atau putih suatu daerah menunjukkan semakin besar kuat medan magnet pada daerah tersebut.



Gambar III.5. Magnetogram pada arah medan pandang yang diambil pada 3 Juni 2022. Sumber: https://sdo.gsfc.nasa.gov/assets/img/latest/latest_2048_HMIB.jpg

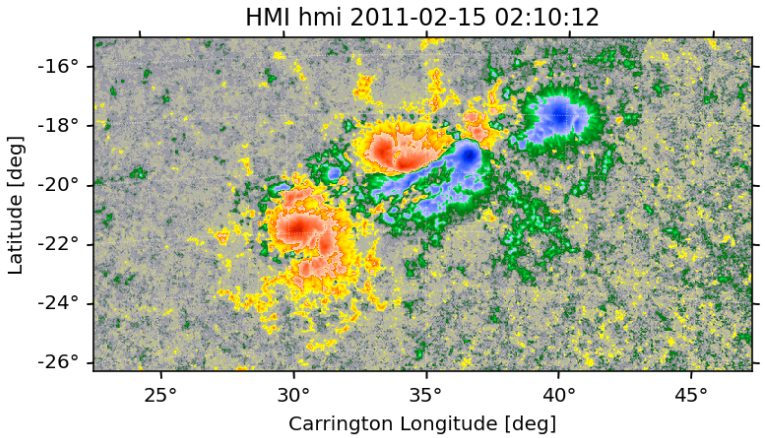
III.2 *Spaceweather HMI Active Region Patch* (SHARP)

Spaceweather HMI Active Region Patch (SHARP) adalah seri DRMS atau dapat dikatakan sebagai database yang memuat data deret waktu dari berbagai parameter cuaca antariksa yang diturunkan dari magnetogram hasil pengamatan HMI. Data ini disimpan dalam bentuk header FITS. Selain itu, SHARP juga memuat 31 segmen data (Tabel III.1) untuk setiap daerah aktif yang terdeteksi oleh algoritma *pipeline* HMI. Terdapat empat macam seri DRMS SHARP:

- **hmi.sharp_720s**: data definitif dalam koordinat CCD dimana vektor medan magnet telah didekomposisi ke komponen azimuth, inklinasi, dan arah medan

pandang,

- **hmi.sharp_cea_720s**: data definitif dimana vektor medan magnet telah dipetakan ke proyeksi Lambert Equal-Area dan didekomposisi menjadi B_radial, B_phi, dan B_theta,
- **hmi.sharp_720s_nrt:** data *near-real time* dalam koordinat CCD dimana vektor medan magnet telah didekomposisi ke komponen azimuth, inklinasi, dan arah medan pandang,
- **hmi.sharp_cea_720s_nrt**: data *near-real time* dimana vektor medan magnet telah dipetakan ke proyeksi *Lambert Equal-Area* dan didekomposisi menjadi B_radial, B_phi, dan B_theta.



Gambar III.6. Contoh daerah aktif yang dideteksi melalui algoritma SHARP. Color map menunjukkan fluks total tanpa tanda (*total unsigned flux*). Sumber: Bobra, 2014.

Tabel III.1: Daftar segmen data dalam SHARP

Nama Segmen	Deskripsi
MAGNETOGRAM	Magnetogram
BITMAP	Label identifikasi piksel terhadap *smooth margin* pembatas: • 1 : medan lemah di luar margin pembatas • 2 : medan kuat di luar margin pembatas • 33: medan lemah di dalam margin pembatas • 34: medan kuat di dalam margin pembatas
DOPPLERGRAM	Dopplergram
CONTINUUM	Intensitas kontinum
INCLINATION	Inklinasi medan magnet terhadap arah pandang
AZIMUTH	Medan magnet pada arah azimuth dengan nol didefinisikan sebagai arah atas dari piksel kolom dalam CCD, nilai meningkat

	berlawanan arah jarum jam
FIELD	Kuat medan magnet
VLOS_MAG	Kecepatan plasma pada arah pandang. Positif menandakan pergeseran merah
DOP_WIDTH	Lebar garis spektrum akibat efek doppler, asumsi Gaussian.
ETA_0	Koefisien *center-to-continuum absorption*
DAMPING	Osilasi dipol elektron sebagai silator harmonik sederhana. Default=0,5
SRC_CONTINUUM	Fungsi sumber pada dasar fotosfer
SRC_GRAD	Gradien fungsi sumber terhadap tebal optis
ALPHA_MAG	Porsi resolusi elemen yang terisi oleh plasma termagnetisasi
CHISQ	Tingkat kecocokan profil terhadap *fitting* iterasi *least square*
CONV_FLAG	Flag dan indeks proses *ME-inversion convergence*
INFO_MAP	Indeks kualitas tiap piksel terhadap proses inversi *output*
CONFID_MAP	Indeks tingkat keyakinan tiap piksel terhadap proses inversi *output*
INCLINATION_ERR	Standar deviasi dari error inklinasi medan magnet terhadap arah pandang
AZIMUTH_ERR	Standar deviasi dari error medan magnet arah azimuth
FIELD_ERR	Standar deviasi dari error kuat medan magnet
VLOS_ERR	Standar deviasi dari error kecepatan plasma pada arah pandang
ALPHA_ERR	Standar deviasi dari alpha
FIELD_INCLINATION_ERR	Koefisien korelasi dari error inklinasi medan magnet total
FIELD_AZ_ERR	Koefisien korelasi dari error medan magnet arah azimuth
INCLIN_AZIMUTH_ERR	Koefisien korelasi dari error inklinasi medan magnet arah azimuth
FIELD_ALPHA_ERR	Koefisien korelasi error alpha arah azimuth
INCLINATION_ALPHA_ERR	Koefisien korelasi error inklinasi alpha
AZIMUTH_ALPHA_ERR	Koefisien korelasi alpha arah azimuth
DISAMBIG	Flag untuk perubahan 180^o pada arah azimuth
CONF_DISAMBIG	Tingkat keyakinan DISAMBIG
Bp	Komponen medan magnet arah barat bujur Matahari (CEA)
Bt	Komponen medan magnet arah selatan lintang Matahari (CEA)
Br	Komponen medan magnet arah radial (CEA)

Berbeda dengan magnetogram yang memuat data seluruh piringan Matahari, SHARP hanya berfokus pada daerah dengan medan magnet kuat atau daerah aktif saja (gambar III.6). Daerah aktif diidentifikasi secara otomatis oleh algoritma *pipeline* HMI. Algoritma ini mendeteksi daerah di piringan Matahari yang mengindikasikan medan

magnet kuat dan mengikuti daerah ini selama pergerakannya di piringan Matahari. Setiap bagian ini dinotasikan sebagai *HMI Active Region Patch* (HARP) dan diberi nomor (*HARP Number*), mirip seperti penomoran daerah aktif yang dilakukan oleh NOAA (*NOAA Active Region Number*). SHARP menggunakan informasi geometrik dari setiap HARP untuk menurunkan parameter cuaca antariksa pada setiap tempo pengukuran.

Dalam mengidentifikasi HARP, algoritma SHARP mendefinisikan suatu kotak dengan *smooth boundary* dalam koordinat CCD sebagai pembatas. Pembatas ini berpusat pada lokasi *flux-weighted centroid* daerah aktif yang telah terdeteksi oleh *pipeline* HMI. Kotak pembatas inilah yang didefinisikan sebagai HARP. Area dalam pembatas ini didefinisikan sebagai daerah aktif.

Parameter cuaca antariksa dalam HARP dihitung setiap 12 menit dan vektor medan magnet ditransformasikan ke proyeksi *Lambert Cylindrical Equal-Area* serta didekomposisi ke komponen B_x, B_y, dan B_z. Beberapa parameter dapat langsung diperoleh dari data. Untuk parameter yang memerlukan penurunan secara numerik, penurunan dilakukan dengan metode *finite difference* dengan 9 titik setensilan. Piksel yang diperhitungkan dalam memperoleh parameter cuaca antariksa adalah piksel dengan nilai yang logis dan di atas batas *noise* serta berada dalam *smooth boundary* kotak pembatas HARP. Piksel-piksel ini kemudian diintegrasikan sehingga dapat diturunkan berbagai parameter cuaca antariksa. Daftar parameter cuaca antariksa yang dapat diturunkan oleh algoritma SHARP tertera pada Tabel III.2.

Tabel III.2: Daftar parameter cuaca antariksa yang tersedia dalam database SHARP. Sumber: Bobra dan Couvidat, 2015.

Kata Kunci	Deskripsi Fisis	Formula
TOTUSJH	Total helisitas arus tanpa tanda dalam G^2/m	$H_{c_{total}} \propto \sum \lvert B_z \cdot J_z \rvert$
TOTPOT	Total densitas energi magnetik fotosfer dalam erg/cm^3	$\rho_{tot} \propto \left(\vec{\boldsymbol{B}}^{Obs} - \vec{\boldsymbol{B}}^{Pot}\right)^2 dA$
TOTUSJZ	Total arus vertikal tanpa tanda dalam mA/m^2	$J_{z_{total}} = \sum \lvert J_z \rvert \, dA$
ABSNJZH	Nilai absolut helisitas arus bersih dalam G^2/m	$H_{c_{obs}} \propto \sum \lvert B_z \cdot J_z \rvert$
SAVNCPP	Jumlah nilai absolut dari arus bersih per polaritas dalam Ampere	$J_{z_{sum}} \propto \left\lvert \sum^{B_z^+} J_z dA \right\rvert + \left\lvert \sum^{B_z^-} J_z dA \right\rvert$
USFLUX	Total fluks tanpa tanda dalam Maxwell	$\phi = \sum \lvert B_z \rvert \, dA$
AREA_ACR	Area dengan medan kuat dalam piksel	$Area = \sum pixel$
MEANPOT	Rata-rata ekses densitas energi megnetik fotosfer dalam erg/cm^3	$\bar{\rho} \propto \frac{1}{N} \sum \left(\vec{\boldsymbol{B}}^{Obs} - \vec{\boldsymbol{B}}^{Pot}\right)^2$
R_VALUE	Total fluks dekat garis inversi polaritas dalam Maxwell	$\phi = \sum \lvert B_{LoS} \rvert \, dA \;\; dalam \; R \; mask$

SHRGT45	Persentase piksel dengan shear angle > 45° dalam persen	$\frac{Area\ dengan\ shear > 45^\circ}{Total\ area} \times 100\%$
MEANSHR	Rata-rata shear angle diukur dari B_{total} dalam derajat	$\vec{\Gamma} = \frac{1}{N}\sum cos^{-1}\left(\frac{\vec{B}^{Obs} \cdot \vec{B}^{Pot}}{\lvert\vec{B}^{Obs}\rvert\lvert\vec{B}^{Pot}\rvert}\right)$
MEANGAM	Rata-rata sudut inklinasi, gamma dalam derajat	$\vec{\gamma} = \frac{1}{N}\sum tan^{-1}\left(\frac{B_h}{B_z}\right)$
MEANGBT	Rata-rata gradien medan magnet total dalam Gauss/Mm	$\overline{\lvert\nabla B_{tot}\rvert} = \frac{1}{N}\sum\sqrt{\left(\frac{\partial B}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2}$
MEANGBZ	Rata-rata gradien medan magnet vertikal dalam Gauss/Mm	$\overline{\lvert\nabla B_z\rvert} = \frac{1}{N}\sum\sqrt{\left(\frac{\partial B_z}{\partial x}\right)^2 + \left(\frac{\partial B_z}{\partial y}\right)^2}$
MEANGBH	Rata-rata gradien medan magnet horizontal dalam Gauss/Mm	$\overline{\lvert\nabla B_h\rvert} = \frac{1}{N}\sum\sqrt{\left(\frac{\partial B_h}{\partial x}\right)^2 + \left(\frac{\partial B_h}{\partial y}\right)^2}$
MEANJZH	Rata-rata helisitas arus dalam G^2/m	$\overrightarrow{H_c} \propto \frac{1}{N}\sum B_z \cdot J_z$
MEANJZD	Rata-rata densitas arus vertikal dalam mA/m^2	$\overrightarrow{H_c} \propto \frac{1}{N}\sum\left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}\right)$
MEANALP	Rata-rata parameter puntiran, alpha dalam 1/Mm	$\alpha_{total} \propto \frac{\sum J_z \cdot B_z}{\sum {B_z}^2}$

III.3. *Scikit-learn dan Scipy*

Pengolahan data pada penelitian ini memanfaatkan dua pustaka dalam bahasa pemrograman Python, yaitu *Scikit-learn* dan *Scipy*. *Scikit-learn* merupakan pustaka pembelajaran mesin yang tersedia secara bebas dalam bahasa pemrograman *Python*. *Scikit-learn* mendukung pembelajaran mesin berbasis *supervised* maupun *unsupervised learning*. Pustaka ini menyediakan berbagai algoritma untuk klasfikasi, regresi, maupun *clustering* dan mendukung berbagai proses yang biasa digunakan dalam pembelajaran mesin seperti *fitting* dan optimisasi model, praproses data, dan reduksi dimensi data (Pedregosa, dkk., 2011). Di sisi lain, *Scipy* merupakan pustaka yang menyediakan algoritma untuk berbagai kasus komputasi saintifk dan teknis (Virtanen, dkk., 2020). Penelitian ini secara khusus menggunakan salah satu fitur *Scipy* yaitu pemrosesan sinyal untuk melakukan algoritma STFT terhadap data. Berikut merupakan fungsi-fungsi yang digunakan dari kedua pustaka ini pada penelitian ini.

III.3.1 Normalisasi *Robust Scaler*

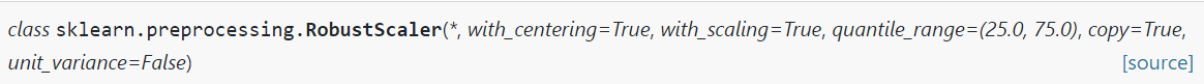
Salah satu fitur dari Scikit-learn yang dipakai dalam penelitian ini adalah normalisasi. Metode normalisasi yang dipilih untuk data yang digunakan dalam penelitian ini adalah normalisasi *robust scaler*. Normalisasi ini memiliki formula

$$X_{norm} = \frac{(X_i - X_{median})}{X_{iqr}} \ . \qquad \text{(III.1)}$$

Normalisasi *robust scaler* mentransformasi distribusi data terhadap mediannya dan menskalakan rentang data terhadap jarak antara kuartil 1 dan 3 (*interquartile range*). Normalisasi ini dipilih karena ketahanannya terhadap data degan outlier. Selain itu,

normalisasi ini cenderung tidak mengubah distribusi awal data.

Normalisasi dilakukan dikarenakan rentang nilai di setiap parameter magnetik yang berbeda (gambar III.15) dikhawatirkan dapat mempengaruhi performa model prediksi. Selain itu, normalisasi perlu dilakukan agar algoritma STFT dapat bekerja dengan baik. Normalisasi juga merupakan hal yang wajib dilakukan guna menganalisis lebih lanjut kontribusi setiap parameter magnetik dalam distribusi total data (analisis PCA).

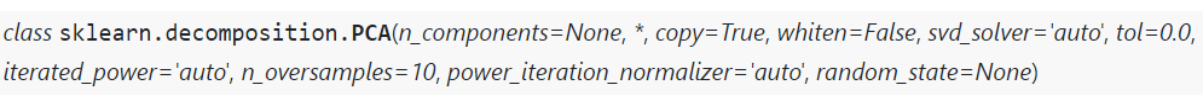


Gambar III.7. Fungsi normalisasi *robust scaler* pada Scikit-learn dengan *hyperparameter* serta nilai bawaannya: *with_centering* (pemusatan distribusi data), *with_scaling* (penskalaan terhadap rentang kuantil), *quantile_range* (rentang kuantil yang dipilih), *copy*, *unit_variance* (mentransformasi hingga variansi dari distribusi menjadi 1). Sumber: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html .

Pada Scikit-learn sendiri, normalisasi *robust scaler* dipanggil melalui kelas *sklearn.preprocessing.RobustScaler*. Fungsi ini memiliki beberapa *hyperparameter* dengan nilai bawaan seperti yang tampak pada gambar III.7.

III.3.2. Analisis Parameter menggunakan PCA

Analisis PCA dilakukan dengan menggunakan dua pustaka yaitu Scikit-learn dan pustaka PCA (Taskesen, 2020) yang tersedia dalam bahasa pemrograman Python. Pustaka PCA digunakan untuk menentukan parameter terbaik yang merepresentasikan setiap komponen utama dan plot presentasi variasi kumulatif. Parameter terbaik pada setiap komponen utama belum dapat merepresentasikan kontribusi total fitur terhadap variasi data. Oleh karena itu, digunakan kelas *sklearn.decomposition.PCA* pada pustaka Scikit-learn untuk menentukan *loading score* atau kontribusi fitur dalam komponen utama pada setiap fitur dan presentasi variasi setiap komponen utama. Nilai *loading score* setiap fitur pada setiap komponen utama kemudian dikalikan dengan presentasi variasi pada setiap komponen utama dan dijumlahkan. Nilai ini diambil sebagai skor total yang merepresentasikan kontribusi total fitur terhadap variasi dataset. Kelas PCA dalam Scikit-learn memiliki beberapa hyperparameter seperti ditunjukkan oleh gambar III.8.



Gambar III.8. Fungsi PCA pada Scikit-learn dengan *hyperparameter* bawaannya: *n_components* (jumlah komponen utama), *copy*, *whiten*, *svd_solver* (*solver* untuk *Singular Value Decomposition*), *tol* (nilai toleransi untuk *svd_solver*), *iterated_power* (jumlah iterasi untuk *power method* yang diterapkan ketika *svd_slover*=”*randomized*”), *n_oversamples*, *power_iteration_normalizer*, *random_state*. Sumber: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html .

III.3.3. STFT (Scipy)

Dalam rangka memperoleh nilai korelasi spektrum yang merepresentasikan fluktuasi atau proses pra-*flare*, penulis menggunakan algoritma STFT yang tersedia dalam pustaka Scipy. STFT pada Scipy dapat diakses melalui kelas *scipy.signal.stft*. Kelas ini memiliki beberapa *hyperparameter* dengan nilai bawaan seperti pada gambar III.9.

```
scipy.signal.stft(x, fs=1.0, window='hann', nperseg=256, noverlap=None, nfft=None,
detrend=False, return_onesided=True, boundary='zeros', padded=True, axis=-1,
scaling='spectrum')
```

Gambar III.9. Fungsi STFT pada Scipy dengan *hyperparameter* bawaannya: *x* (data deret waktu yang akan ditransform), *fs* (frekuensi sampling), *window* (jenis window function yang digunakan), *nperseg* (jumlah data deret waktu per segmen), *noverlap* (jumlah titik data yang tumpang tindih), *nfft* (panjang dari *fast fourier transform* yang dilakukan), *detrend* (cara men-detrend data), *return_onesided* (pilihan apakah akan menghasilkan spektrum dengan satu sisi), *boundary* (tambahan titik data untuk menghaluskan awal dan akhir sinyal), *padded* (penambahan titik data nol pada sinyal agar panjang titik data kelipatan *nperseg*), *axis*, *scaling*. Sumber: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.stft.html .

III.3.4. *Fitting Support Vector Machine* Menggunakan *LinearSVC*

Dalam rangka membangun model prediksi *support vector machine*, penulis menggunakan kelas *sklearn.svm.LinearSVC* yang tersedia dalam pustaka Scikit-learn. Berbeda dengan algoritma *support vector machine* yang lebih umum di Scikit-learn yaitu *sklearn.svm.SVC*, *LinearSVC* menggunakan optimisasi *liblinear* dengan kernel linier. *LinearSVC* dipilih karena optimisasi *liblinear* bekerja jauh lebih cepat daripada optimisasi *libsvm* yang digunakan pada *sklearn.svm.SVC* (Fan, dkk., 2008). Optimisasi *libsvm* memiliki kompleksitas waktu *training* yang sebanding dengan n^2 hingga n^3 dengan n sebagai jumlah sampel *training* . Ini jauh lebih lambat daripada *liblinear* dengan kompleksitas waktu *training* yang sebanding dengan n.

Kompleksitas waktu *training* yang tinggi pada *libsvm* muncul akibat algoritma *libsvm* yang menghitung jarak setiap titik data *training* satu sama lain ($n \times n$) (Chang dan Lin, 2011). Jarak ini dapat disimpan dalam *cache* untuk mempercepat perhitungan. Namun untuk dataset dengan jumlah titik data yang besar, memori *cache* atau RAM yang diperlukan untuk menyimpan perhitungan ini sangat besar. Perkiraan memori RAM yang diperlukan untuk menyimpan matriks ($n \times n$) dengan jumlah titik data $n = 500.000$ yang disimpan dalam array bilangan dengan format *float* 64-bit adalah sekitar 1 TB. Oleh karena itu, meskipun *liblinear* membatasi kita dalam pilihan kernel yang tersedia (hanya dapat diimplementasikan dengan kernel linier), *liblinear* menjadi pilihan yang lebih masuk akal untuk diimplementasikan pada penelitian ini yang menggunakan dataset yang besar.

Kelas LinearSVC memiliki beberapa *hyperparameter* seperti pada gambar III.9.

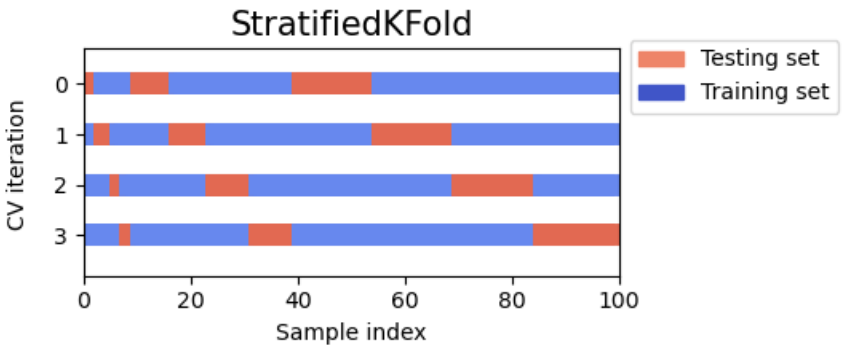
```
class sklearn.svm.LinearSVC(penalty='l2', loss='squared_hinge', *, dual='warn', tol=0.0001, C=1.0, multi_class='ovr',
fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0, random_state=None, max_iter=1000)
```

Gambar III.10. Fungsi LinearSVC pada Scikit-learn yang mengimplementasikan *liblinear* dengan *hyperparameter* bawaannya: *penalty* (aturan penalti atau regularisasi yang digunakan), *loss* (*loss function* yang digunakan), *dual* (pilihan untuk menyelesaikan problema dual atau primal), *tol* (faktor toleransi untuk *stopping criteria*), *C* (parameter regularisasi), *multi_class* (strategi apabila kasus multi kelas), *fit_intercept* (pilihan apakah perlu ditentukan nilai intercept dari model support vector classifier), *intercept_scaling*, *class_weight* (bobot kelas), *verbose*, *random_state*, *max_iter*. Sumber: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html .

III.3.5. Validasi Silang Menggunakan *Stratified K-fold*

Dalam rangka memperoleh suatu model *supervised learning* berdasarkan satu dataset, perlu dilakukan pemisahan dataset menjadi *training* dan *test* dataset. Pemisahan ini dapat dilakukan secara sederhana dengan menggunakan data dengan rentang tertentu dalam dataset sebagai *training dataset* dan sisanya sebagai *test dataset*. Namun metode ini dinilai kurang dapat memberikan hasil atau performa yang konsisten secara statistik. Oleh karena itu, dilakukanlah pemisahan dataset dengan cara validasi silang atau *cross validation*. Validasi silang adalah suatu metode pemisahan *training* dan *test dataset* dengan memvariasikan rentang pemisahan *training* dan *test*. Setiap varian pemisahan kemudian dijalankan terhadap model prediksi atau estimator.

Penelitian tugas akhir ini secara khusus menerapkan kelas *sklearn.model_selection.StratifiedKFold* yang tersedia dalam pustaka *Scikit-learn* sebagai metode pemisah *training* dataset dan *test* dataset. *Stratified k-fold* adalah metode *cross validation* yang membagi dataset ke dalam *training* dan *test* dataset sehingga diperoleh sejumlah k varian pembagian dataset dengan masing-masing varian memiliki distribusi label yang sama dengan distribusi label dataset awal. Secara bawaan, *StratifiedKFold* memisahkan *train* dan *test dataset* dengan proporsi 1:4. *Sampling* terhadap dataset untuk dipilih sebagai *training* maupun *test* dataset juga dilakukan secara bertingkat terhadap keseluruhan rentang dataset. Metode ini dipilih karena *stratified k-fold* dinilai sangat baik digunakan untuk kasus dataset dengan distribusi label yang tidak seimbang (Pedregosa, dkk., 2011). Skema dari *startified k-fold cross validation* ditunjukkan oleh gambar III.11. Sedangkan hyperparameter untuk *StratifiedKFold* dapat dilihat pada gambar III.12.



Gambar III.11. Skema *stratified k-fold cross validation* dalam membagi dataset manjadi *training* dan *test* dataset. Pada skema ini, iterasi *cross validation* menghasilkan tiga varian pembagian dataset (k = 3). Sumber: https://scikit-learn.org/stable/modules/cross_validation.html .

class `sklearn.model_selection.StratifiedKFold`(*n_splits=5, *, shuffle=False, random_state=None*)

Gambar III.12. Fungsi *StratifiedKFold* dengan *hyperparameter* bawaannya: n_spits (jumlah split atau pemisahan *training* dan *test)*, *shuffle*, *random_state*. Sumber: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

III.3.6. *Tuning Hyperparameter* Menggunakan *Grid Search*

Untuk memaksimalkan performa prediksi yang dilakukan oleh model, sebaiknya dilakukan pengaturan atau tuning terhadap *hyperparameter* yang mempengaruhi kinerja model. Pada penelitian ini yang menggunakan model prediksi *LinearSVC* pada Scikit-learn, *hyperparameter* yang mempengaruhi kinerja model dan dapat diubah atau diatur dengan leluasa adalah *hyperparameter C*. *C* merupakan parameter regularisasi yang berkaitan dengan regularisasi l2. Nilai *C* berbanding terbalik terhadap kekuatan regularisasi. Ketika nilai *C* kecil, penalti untuk misklasifikasi juga kecil sehingga terbentuk *soft margin* yang cukup lebar pada fungsi *support vector classifier*. Sebaliknya apabila nilai *C* besar, penalti untuk misklasifikasi juga besar sehingga terbentuk *soft margin* yang sempit pada fungsi *support vector classifier*.

Dalam rangka menemukan nilai *C* yang optimal, dilakukan pengaturan terhadap nilai ini dengan menggunakan metode *grid search* yang tersedia dalam kelas *sklearn.model_selection.GridSearchCV* pada Scikit-learn. Algoritma *grid search* cukup sederhana yaitu nilai *C* yang optimal diperoleh dengan mencoba setiap nilai *C* pada rentang tertentu dengan rentang nilai yang semakin halus. Kelas GridSearchCV memiliki beberapa *hyperparameter* seperti yang ditunjukkan pada gambar III.13.

```
class sklearn.model_selection.GridSearchCV(estimator, param_grid, *, scoring=None, n_jobs=None, refit=True, cv=None,
verbose=0, pre_dispatch='2*n_jobs', error_score=nan, return_train_score=False)                                    [source]
```

Gambar III.13. Fungsi *GridSearchCV* dengan *hyperparameter* bawaannya: *estimator* (model prediksi yang digunakan), *param_grid* (rentang parameter yang ingin dioptimasi), *scoring* (metrik skor yang digunakan), *n_jobs* (jumlah *thread* yang digunakan), *refit* (metrik skor yang ingin diutamakan untuk dioptimasi), *cv* (metode validasi silang yang digunakan), *verbose*, *pre_dispatch* (berkaitan dengan *multithreading*), *error_score* (skor yang diberikan ketika terjadi eror), *return_train_score*.

III.4. Data

Data yang digunakan pada penelitian ini berasal dari dua sumber utama, yaitu SHARP dan NOAA.

III.4.1. Data NOAA

Data dari NOAA diambil sebagai data komplemen untuk mengidentifikasi fenomena suar surya pada daerah aktif. Data ini diambil melalui ftp://ftp.swpc.noaa.gov/pub/warehouse/ .

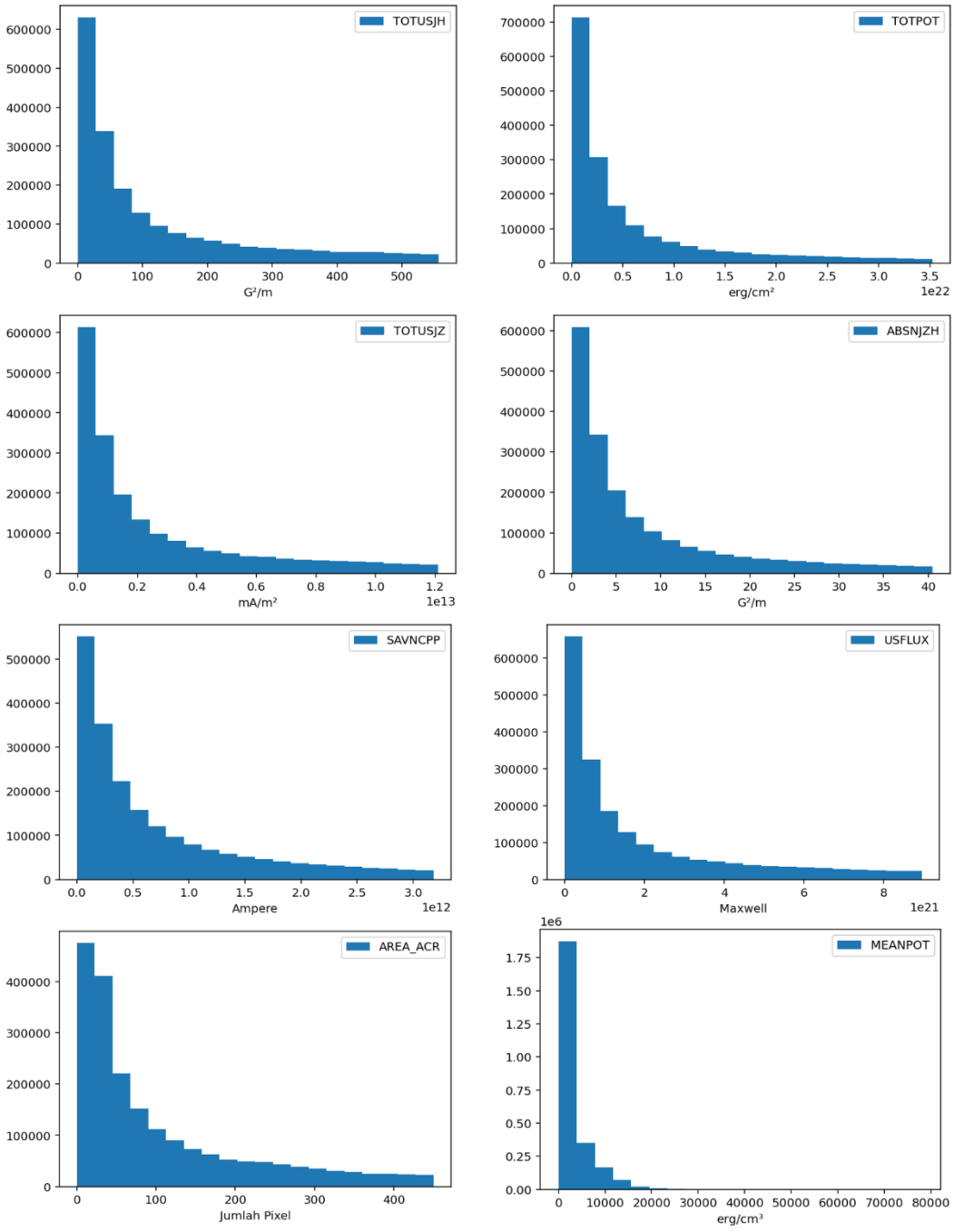
Data mentah yang didapat dari situs ini adalah catatan fenomena aktivitas Matahari yang terjadi setiap hari. Data ini disimpan dalam format txt dan terdapat satu file txt untuk setiap harinya. Data mentah terdiri dari 10 kolom yaitu *Event, Begin, Max, End, Obs, Q, Type, Loc/Frq, Particulars*, dan *Reg#*. Kolom *Event* memuat penomoran setiap fenomena aktivitas Matahari. Kolom *Begin, End,* dan *Max* masing-masing menunjukkan waktu mulai, selesai, dan intensitas maksimal kejadian fenomena aktivitas Matahari. Kolom *Obs* menunjukkan lokasi observatorium yang melaporkan kejadian fenomena aktivitas Matahari. Kolom *Q* menunjukkan kualitas data yang diperoleh. Kolom *Type* memberikan informasi tentang tipe fenomena aktivitas Matahari yang terjadi. Kolom *Loc/Frq* menunjukkan lokasi fenomena aktivitas Matahari terjadi dalam derajat heliografis dari meridian pusat Matahari dan frekuensi kejadian dalam Mhz. Kolom *Particulars* memuat informasi tambahan mengenai fenomena aktivitas Matahari yang terjadi seperti kelas suar surya untuk kasus suar surya. Sedangkan kolom *Reg#* menunjukkan nomor daerah aktif tempat terjadinya fenomena ktivitas Matahari (*NOAA Active Region Number*).

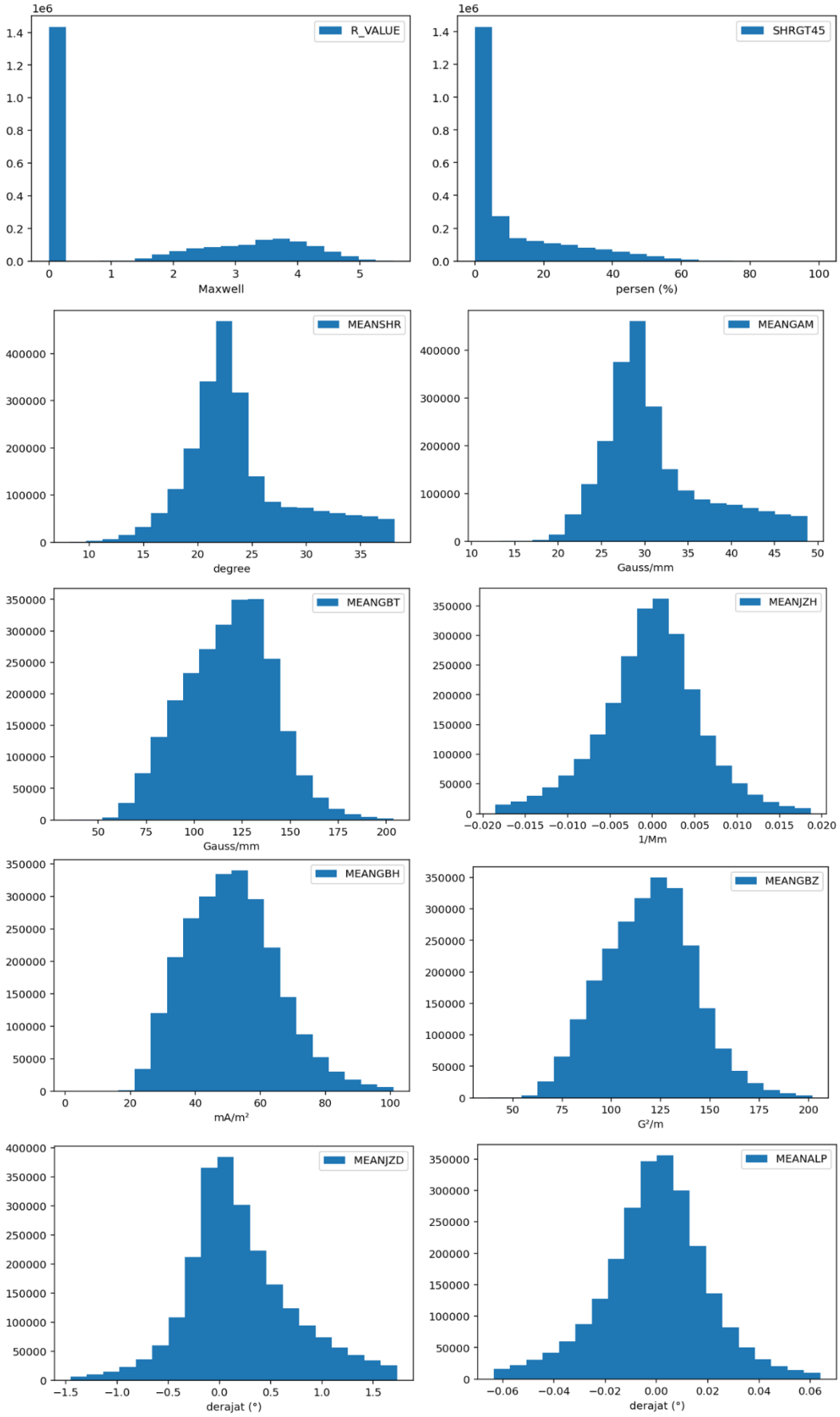
Data ini memuat seluruh fenomena suar surya yang teramati oleh algoritma NOAA dari 1 Mei 2010 hingga 31 Desember 2021. Dataset dari NOAA ini terdiri dari 68.924 baris dan 11 kolom (10 kolom asli data ditambah satu kolom *timestamp*)

III.4.2. Data SHARP

Data dari SHARP adalah berupa deret waktu 18 parameter cuaca antariksa (Tabel III.2) yang diukur dari 2 Mei 2010 pukul 00:00 UTC hingga 31 Desember 2021 pukul 23:48 UTC pada daerah aktif yang berlokasi ±69^{o} dari bujur meridian Matahari. Selain 18 parameter tersebut, diambil juga beberapa metadata sebagai identitas titik data yaitu

T_REC (waktu pengukuran), HARPNUM (*HARP Number*), NOAA_AR dan NOAA_ARS (*NOAA Active Region Number*), LON_FWT dan LAT FWT (lokasi heliografis daerah aktif) . Dengan selang waktu antar titik data SHARP sebesar 12 menit dan tingkat deteksi HMI sebesar 98,4% (Hoeksema dkk., 2014), jumlah titik data yang diperoleh dari SHARP adalah sebanyak 2.900.689 titik data. Proses pengambilan data dilakukan dengan menggunakan modul drms dalam bahasa pemrograman Python. Distribusi data pada masing-masing parameter yang telah dibersihkan terhadap pencilan ditunjukkan oleh gambar III.15.



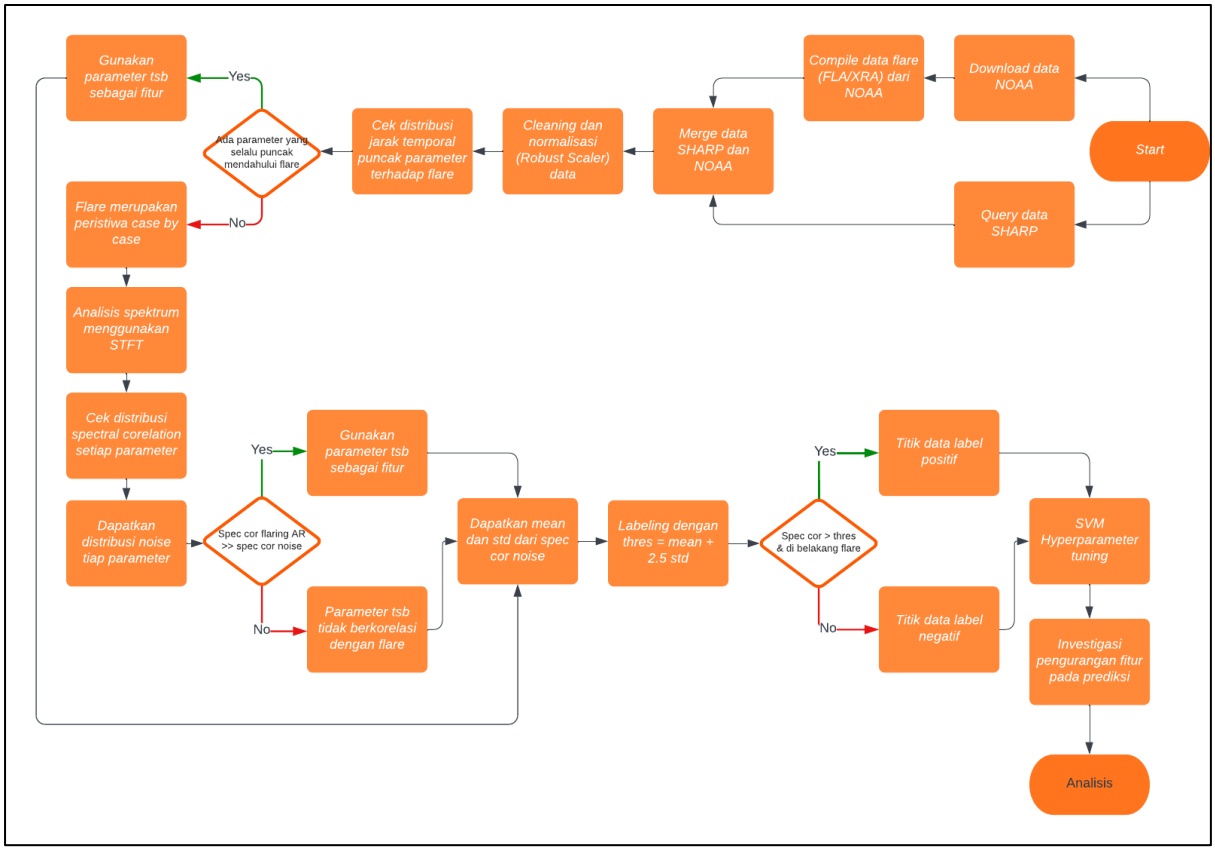


Gambar III.16. Kumpulan histogram yang menunjukkan distribusi data masing-masing parameter cuaca antariksa.

BAB IV

ALUR PENGOLAHAN DATA

Proses pengambilan dan pengolahan data penelitian ini dilakukan secara umum menggunakan bahasa pemrograman *Python* versi 3.7 dengan sedikit bantuan aplikasi *spreadsheet Microsoft Excel*. Pengambilan dan pengolahan data pada penelitian ini secara umum ditunjukkan oleh diagram alir gambar IV.I. Kode asal atau *source code* pengerjaan dapat diakses melalui tautan https://github.com/CAF103/Thesis .



Gambar IV.1. Diagram alir pengambilan dan pengolahan data penelitian ini.

IV.1 Pengumpulan Data SHARP

Data parameter magnetik SHARP dapat diambil melalui dua cara yaitu melalui halaman *JSOC Data Explore Info and Export* (http://jsoc.stanford.edu/ajax/lookdata.html?ds=hmi.sharp_720s) atau melalui perintah *query* yang dijalankan pada bahasa pemrograman *Python*. Pada penelitian ini, penulis menggunakan cara kedua yaitu menjalankan perintah *query* untuk memperoleh data 18 parameter megnetik SHARP. Kode untuk proses *query* ini dapat dilihat pada gambar IV.2.

```
# Import required libraries
import pandas as pd
import drms
import numpy as np
```

```
%%time
# Connecting to the server
c = drms.Client()

# The date range for query stored in date.csv
timestamp = pd.read_csv('date.csv')
date = timestamp['date']

# The date 2010.05.02 - 2010.05.03 as first query
keys = c.query('hmi.sharp_cea_720s[][2010.05.02 - 2010.05.03][? (LON_FWT < 90) AND (LON_FWT > -90) ?]', key='T_REC, HARPNUM, NOAA

# The remaining query being done in a loop
# This is done this way in order to lessen the connection burden
for i in range(len(timestamp) - 1): # If connection trouble occurs, the loop range can be saparated into different loops with sma
    keys1 = c.query('hmi.sharp_cea_720s[][' + date[i] + ' - ' + date[i+1] + '][? (LON_FWT < 90) AND (LON_FWT > -90) ?]', key='T_R
    keys = pd.concat([keys, keys1], ignore_index=True)
    print(i)

keys.drop_duplicates()

# Save the query into a csv file
keys.to_csv('dataset.csv')
```

Gambar IV.2. Kode untuk melakukan query data SHARP dari server JSOC.

Hal pertama yang diperlukan untuk dapat menjalankan perintah ini adalah mengimpor pustaka yang diperlukan yaitu drms sebagai penghubung koneksi ke server JSOC, numpy, dan pandas untuk menyimpan hasil query menjadi satu dataset. Kemudian sambungkan koneksi ke server JSOC. Pada kasus ini, penulis membagi keseluruhan query menjadi query-query kecil yang dilakukan per satu hari agar meringankan beban koneksi ke server. Hal ini juga dilakukan agar query dapat disambung atau dilanjutkan apabila koneksi terputus. Setelah query selesai, data yang telah diperoleh disimpan dengan nama berkas “dataset.csv” menggunakan bantuan pustaka pandas. Dengan demikian, diperolehlah dataset yang memuat deret waktu dari 18 parameter magnetik Matahari beserta metadatanya yang terdiri dari 2.900.689 baris dan 24 kolom.

IV.2. Kompilasi Data NOAA

Di sisi lain**,** pengambilan data NOAA dilakukan langsung dengan menjalankan protokol ftp ke server SWPC NOAA (ftp://ftp.swpc.noaa.gov/pub/warehouse/). Protokol *ftp* merupakan protokol lama yang sudah tidak didukung oleh kebanyakan browser internet modern. Namun protokol ini dapat dijalankan melalui File Explorer Windows 10 dengan memasukkan pranala ke kolom alamat di File Explorer. Kemudian kita dapat mengunduh seluruh data yang kita inginkan dari server SWPC dengan cara menyalinnya ke penyimpanan lokal. Contoh data mentah yang diambil dari NOAA dapat dilihat pada gambar IV.3.

```
EDITED EVENTS for 23 Jun 95

Event     Begin    Max       End  Obs  Q  Type  Loc/Frq   Particulars       Reg#
8880 +     0024   ////      0025  CUL  C   RSP  018-260   III/2

8890       0028   ////      0028  CUL  C   RSP  018-030   III/1

8900 +     0550   ////      0551  CUL  C   RSP  020-080   III/1

8910       0700   ////      0701  SVI  U   RSP  037-062   III/2

8920 +     1251   1253      1253  SVI  G   RBR  245       100
8920 +     1251   ////      1254  SAG  C   RSP  030-080   III/2

8930       1309   ////      1310  SVI  U   RSP  035-065   III/2

8940 +     1358   ////      1403  SAG  C   RSP  030-060   III/1

8950 +     1417   1421      1424  GO8  5   XRA  1-8A      B2.7    7.01E-05  7882
8950 +     1420   1421      1428  SVI  3   FLA  N07E12    SF      ERU       7882

8960 +     1657   1703      1716  GO8  5   XRA  1-8A      B1.9    1.66E-04  7882
8960 +     1701   1703      1707  HOL  3   FLA  N04E14    SF                7882

8970 +     1728   1732      1735  GO8  5   XRA  1-8A      B2.8    9.22E-05  7882
8970       1731   1732      1741  HOL  3   FLA  N04E13    SF                7882

8980 +     1810   1818      1822  GO8  5   XRA  1-8A      B1.7    9.95E-05  7882
8980       1818   1820      1824  RAM  3   FLA  N04E11    SF      ERU       7882

8990       1832   1844      1848  GO8  5   XRA  1-8A      B1.2    1.15E-04
```

Gambar IV.3. Contoh data mentah yang diperoleh dari NOAA pada 23 Juni 1995 (diedit). Informasi mengenai setiap kolom dapat dilihat pada subbab III.4.1.

Data diambil dengan rentang waktu dari 1 Mei 2010 hingga 31 Desember 2021. Seluruh data mentah yang masih dalam bentuk file *txt* per satu hari kemudian dikompilasi menjadi satu dataset dalam format *csv*. Dalam proses penggabungan ini, karena data mentah tidak memberikan *timestamp* kejadian dalam satu format maka perlu dibuat satu kolom yang menyimpan *timestamp* kejadian. *Timestamp* yang direpresentasikan secara bawaan dipilih sebagai waktu puncak suar surya terjadi, dengan kata lain kolom *Max*. Namun dapat kita lihat bahwa tak jarang kolom *Max* memberikan nilai yang tidak sah (////). Jika hal ini terjadi, dipilih waktu awal (*Begin*) atau akhir (*End*) suar surya sebagai waktu yang merepresentasikan *timestamp* kejadian. Kode untuk memproses mengompilasi dataset ini dapat dilihat pada gambar IV.4.

```
# Import required libraries
import pandas as pd
import glob
import numpy as np
from datetime import datetime as dt_obj

# Function to convert into date_time object for max time
def to_date_max(tstr1, tstr2):
    year   = int(tstr1[:4])
    month  = int(tstr1[5:7])
    day    = int(tstr1[8:10])
    hour   = int(tstr2[:2])
    minute = int(tstr2[2:4])
    return dt_obj(year, month, day, hour, minute)
```

```
%%time
# Create events dataset
DF = pd.DataFrame(columns=['Event', 'Begin', 'Max', 'End', 'Obs', 'Q', 'Type', 'Loc/Frq', 'Particulars'
txt_files = glob.glob('Events/*.txt')

# Opening daily event files
for files in txt_files:
    with open(files, 'r') as f:
        f.readline()
        f.readline()
        date = f.readline()[7:17]
        f.close()
    #print(files)
    data = np.genfromtxt(files, dtype=str, skip_header=3, delimiter=(8, 8, 8, 8, 6, 3, 6, 9, 19, 5), au
    df = pd.DataFrame(data, columns=['Event', 'Begin', 'Max', 'End', 'Obs', 'Q', 'Type', 'Loc/Frq', 'Pa
    df['Date'] = [date for i in range(len(df.Reg))]
    DF = pd.concat([DF, df], ignore_index=True)

# Filtering duplicates
DF = DF.drop_duplicates()
# Filtering for invalid values
temp_list = []
date = [i for i in DF.Date]
time = [i for i in DF.Begin]
Max = [i for i in DF.Max]
End = [i for i in DF.End]
# We filter invalid value in timestamp to create Date column which will be used as timestamp of the eve
# First we look at Max time. If the value is '////' (invalid) or its str lenght != 4, replace it with E
# If End time also invalid, replace it with '0000'
for i in range(len(time)):
    if len(time[i]) != 4 or time[i] == '////' or int(time[i]) > 60:
        if len(Max[i]) == 4 and Max[i] != '////':
            time[i] = Max[i]
        elif len(End[i]) == 4 and End[i] != '////':
            time[i] = End[i]
        else:
            time[i] = '0000'

for i in range(len(date)):
    date_time = to_date_max(date[i], time[i])
    temp_list.append(date_time)
DF['Date'] = temp_list

# Saving the event list as csv
DF.to_csv('All NOAA Notes.csv')
DF
```

Gambar IV.4. Proses kompilasi data dari NOAA. Beberapa pustaka yang digunakan dalam proses ini antara lain pandas untuk membuat dan menyimpan dataframe, glob untuk mengakses *path files*, datetime untuk membuat *timestamp*, dan numpy.

Dengan demikian diperolehlah dataset yang memuat seluruh kejadian suar surya yang terdeteksi oleh pengamatan oleh GOES (NOAA) dalam rentang 1 Mei 2010 hingga 31 Desember 2021 yang memuat 68.294 baris dan 11 kolom (10 kolom bawaan dan 1

kolom *timestamp* kejadian). Dataset ini kemudian disimpan dalam suatu file dengan nama "All NOAA Notes.csv". Tampilan sekilas struktur dataset ini dapat dilihat pada tabel IV.1.

Tabel IV.1: Tampilan sekilas dataset akhir yang diperoleh dari NOAA.

	Event	Begin	Max	End	Obs	Q	Type	Loc/Frq	Particulars	Reg	Date
1	4260	1202	1209	1218	G14	5	XRA	1-8A	B1.9 1.4E-04		2010-01-01 12:09:00
3	4270	1233	1243	1300	G14	5	XRA	1-8A	B2.3 3.1E-04		2010-01-01 12:43:00
5	4280	2329	2333	2342	G14	5	XRA	1-8A	B1.1 8.9E-05		2010-01-01 23:33:00
7	4290 +	0310	0313	0319	////	5	XRA	1-8A	B1.1 4.4E-05	1039	2010-01-02 03:13:00
9	4300	0709	0724	0740	G14	5	XRA	1-8A	C1.0 1.2E-03	1039	2010-01-02 07:24:00
...	...	...	...	...	...	...	...	...	...	...	...
118402	9830 +	2018	2024	2034	G16	5	XRA	1-8A	B4.5 3.9E-04		2021-12-30 20:24:00
118404	9840 +	2234	2237	2241	G16	5	XRA	1-8A	B4.2 1.5E-04		2021-12-30 22:37:00
118406	9850 +	2244	2251	2259	G16	5	XRA	1-8A	B4.1 3.4E-04		2021-12-30 22:51:00
118407	9850	2255	2255	2255	LEA	G	RBR	610	110		2021-12-30 22:55:00
118409	9860 +	2325	////	2359	PAL	C	RSP	025-180	VI/1		2021-12-30 23:59:00

IV.3. Penggabungan Data SHARP dengan Data NOAA

Data dari SHARP kemudian digabungkan dengan data dari NOAA. Penggabungan kedua dataset dilakukan dengan memperhatikan nomor NOAA *Active Region Number* (NOAA_AR) yang terdapat di kedua dataset dan waktu atau *timestamp* kejadian. Oleh karena resolusi data SHARP adalah 12 menit, apabila kedua titik data dari dua dataset yang berbeda memiliki NOAA_AR yang sama dan kejadian berada dalam rentang ±6 menit antara satu sama lain, dianggap kedua data tersebut merepresentasikan kejadian yang sama. Kode untuk proses ini dapat dilihat pada *notebook* "(3) Merging SHARP Dataset with Events Dataset (NOAA).ipynb" pada tautan https://github.com/CAF103/Thesis .

Dengan menggabungkan data dari SHARP dengan data dari NOAA, diperoleh dataset lengkap yang memuat deret waktu 18 parameter cuaca antariksa sebagai fitur, penomoran daerah aktif baik dalam HARPNUM maupun NOAA_AR dan NOAA_ARS, dan kelas suar surya. Data yang telah digabungkan terdiri dari 2.900.689 baris dan 24 kolom yang terdiri dari enam metadata yaitu T_REC, HARPNUM, NOAA_AR, NOAA_ARS, LON_FWT, LAT_FWT dan 18 parameter cuaca antariksa dari SHARP (Tabel II.2) sebagai parameter yang berpotensi menjadi fitur.

Dataset ini dapat dipahami sebagai dataset yang memuat deret waktu setiap daerah aktif yang terdeteksi oleh algoritma *pipeline* SHARP dari rentang waktu yang telah disebutkan. Selain itu, dataset ini juga memuat informasi mengenai kejadian suar surya yang diamati oleh NOAA. Data ini memuat 2.900.689 titik data dengan 6608 suar surya kelas

C, 5312 suar surya kelas B, 683 suar surya kelas M, 50 suar surya kelas X, dan 6 suar surya kelas A. Dataset ini disimpan dengan nama file “all_the_data_we_need.csv”

IV.4. Pembersihan *NaN*

Sebelum dapat dilakukan analisis, data yang telah digabungkan perlu dibersihkan dari segala nilai data yang tidak sah. Nilai-nilai ini dapat berupa tak hingga (∞), negatif tak hingga (-∞), dan null atau kosong (*NaN*). Setiap titik data yang memiliki nilai ini kemudian dihapus shingga diperoleh dataset yang bersih. Pembersihan terhadap nilai data yang tidak sah dilakukan dengan menggunakan bantuan pustaka pandas seperti yang dapat dilihat pada gambar IV.5.

```
%%time
# Dropping inf and nan
df.replace([np.inf, -np.inf], 'delete', inplace=True)
df = df[~df.isin(['delete']).any(axis=1)]

for i in range(6, 24):
    df = df[df[df.columns[i]].notna()]

df
```

Gambar IV.5. Kode yang menunjukkan proses pembersihan terhadap nilai data yang tidak sah, meliputi ∞ (np.inf), -∞ (-np.inf), dan NaN (np.nan). Setiap titik data ∞ atau -∞ diubah menjadi string “delete” terlebih dahulu. Kemudian setiap titik data yang mengandung string “delete” dan NaN dihapus.

IV.5. Penerapan Normalisasi terhadap Dataset

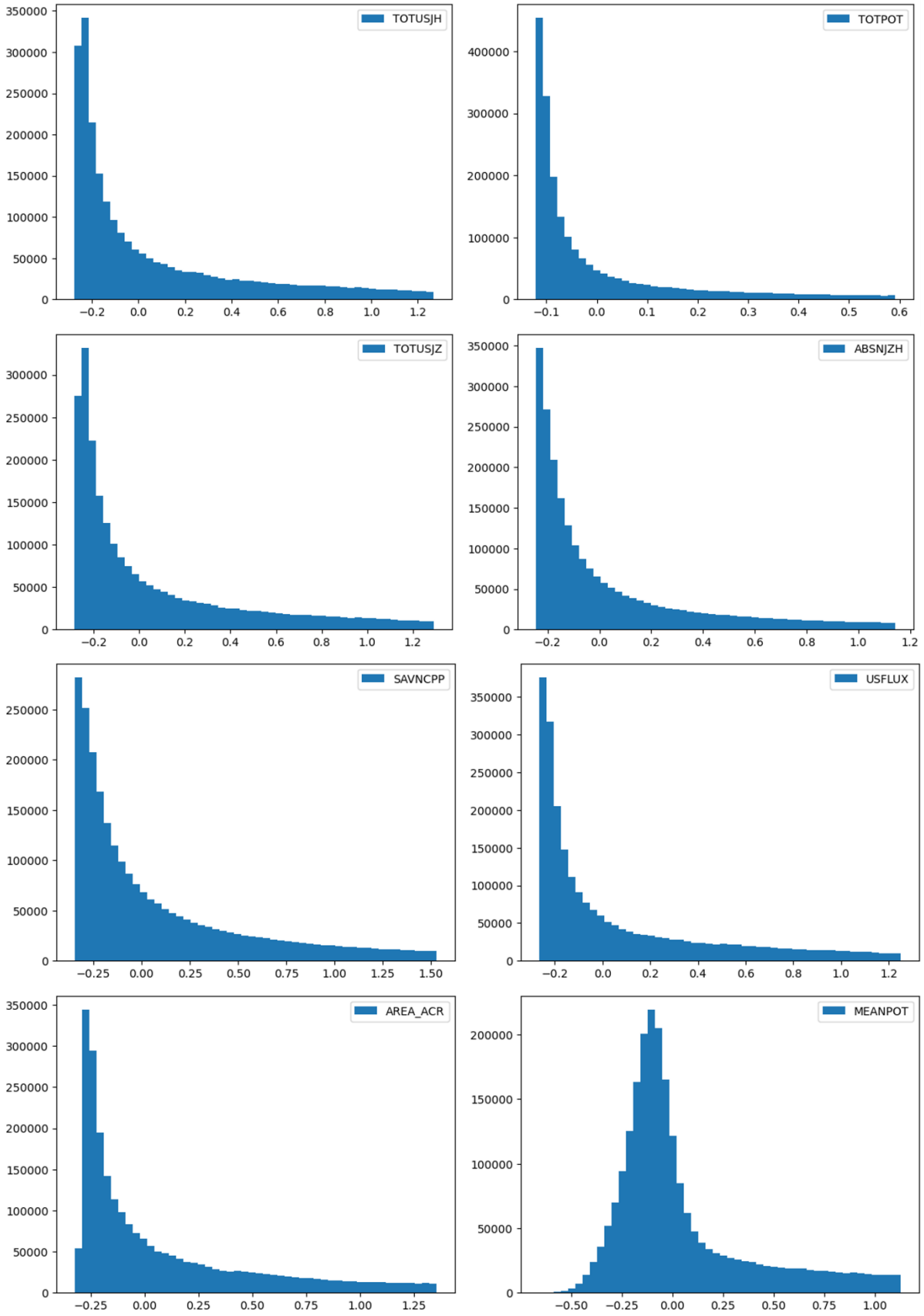
Untuk menyamakan rentang distribusi setiap parameter magnetik dalam dataset, dilakukan normalisasi terhadap data. Normalisasi yang dipilih adalah normalisasi *robust scaler* yang disediakan oleh pustaka Scikit-learn dalam kelas *sklearn.preprocessing.RobustScaler* . *RobustScaler* diterapkan pada setiap parameter magnetik dengan menggunakan seluruh *hyperparameter* bawaannya. Kode untuk menjalankan normalisasi *robust scaler* dapat dilihat pada gambar IV.6. Sedangkan distribusi setiap parameter magnetik yang sudah dinormalisasi dapat dilihat pada gambar IV.7.

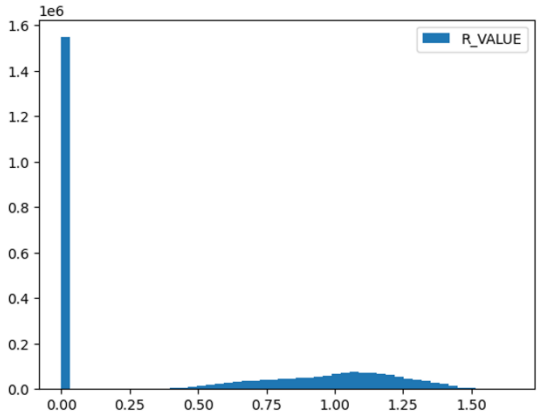
```
%%time
# RobustScaler normalization
RS = df.copy()                                  # Make a copy of df dataset
for i in range(6, 24):
    transformer = RobustScaler()
    x = np.array(RS.iloc[:,i]).reshape(-1, 1)
    transformer.fit(x)
    transformed_data = transformer.transform(x)
    RS.loc[:,RS.columns[i]] = transformed_data
```

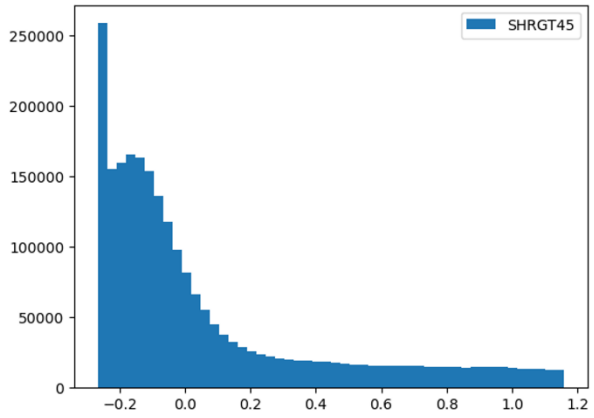
Gambar IV.6. Kode yang menunjukkan proses normalisasi dengan menggunakan *RobustScaler*.

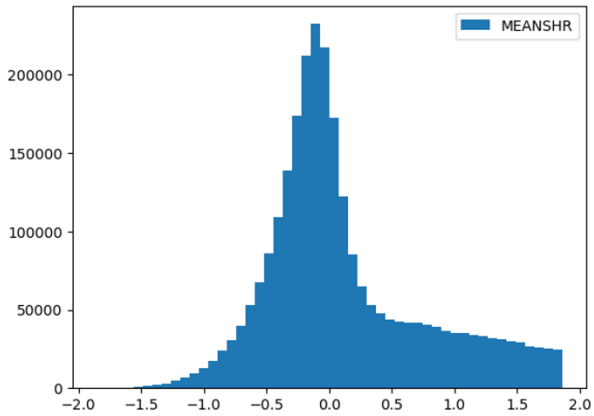
Seperti yang dapat dilihat pada gambar IV.7, rentang data pada distribusi data setiap parameter magnetik berubah sehingga memiliki rentang dengan orde yang mirip. Selain itu, dapat dilihat pula bahwa bentuk distribusinya tidak mengalami perubahan yang tampak. Hal ini dikarenakan normalisasi *robust scaler* hanya mentranslasi median dan mentransformasi rentangnya ke *interquantile range* yang sama. Hal ini penting dilakukan karena distribusi data yang tidak berubah dapat lebih menjamin bahwa fisis setiap parameter magnetik juga tidak

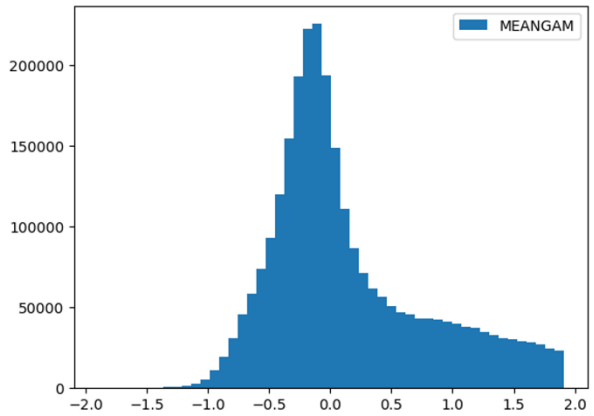
berubah. Hal ini dapat dilihat dari gambar IV.8 yang menunjukkan evolusi salah satu parameter magnetik yaitu TOTUSJH pada daerah aktif dengan HARPNUM 8.

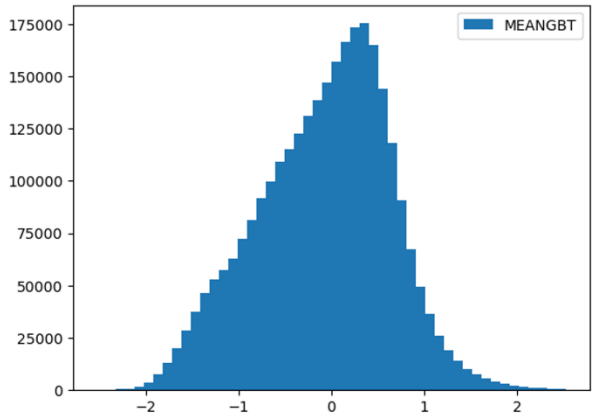


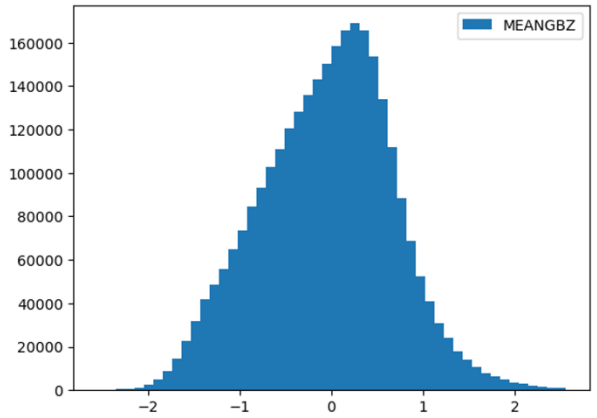

1e6
R_VALUE

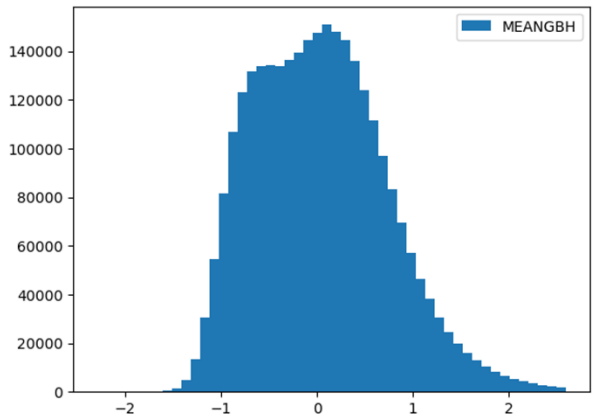

SHRGT45

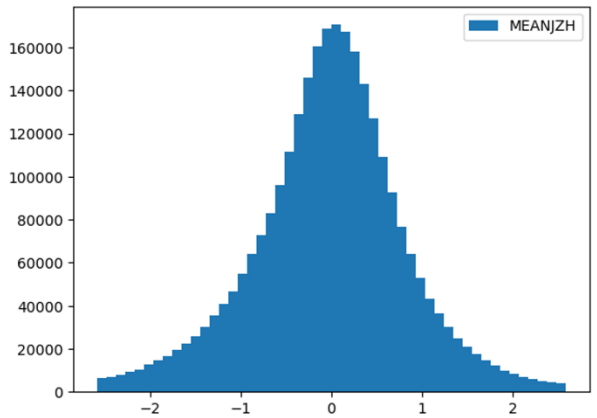

MEANSHR

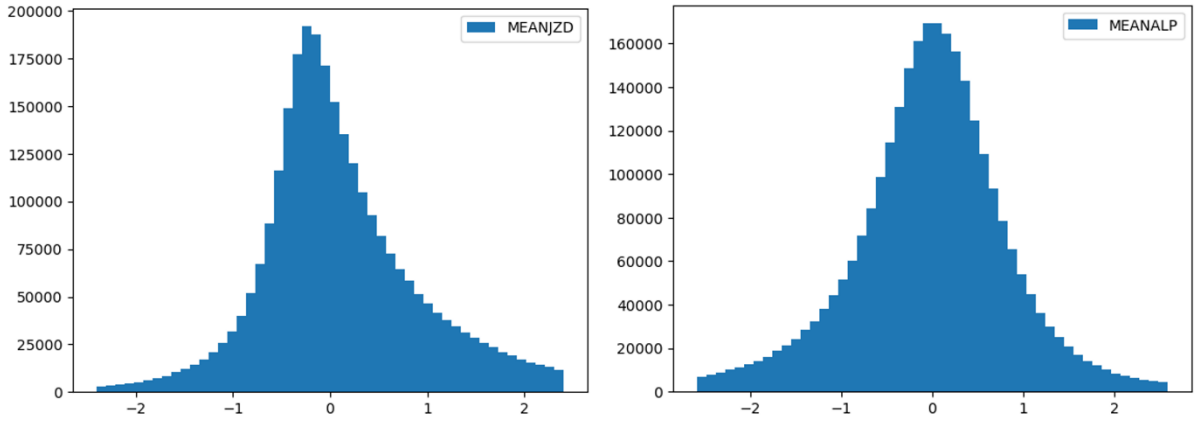

MEANGAM


MEANGBT

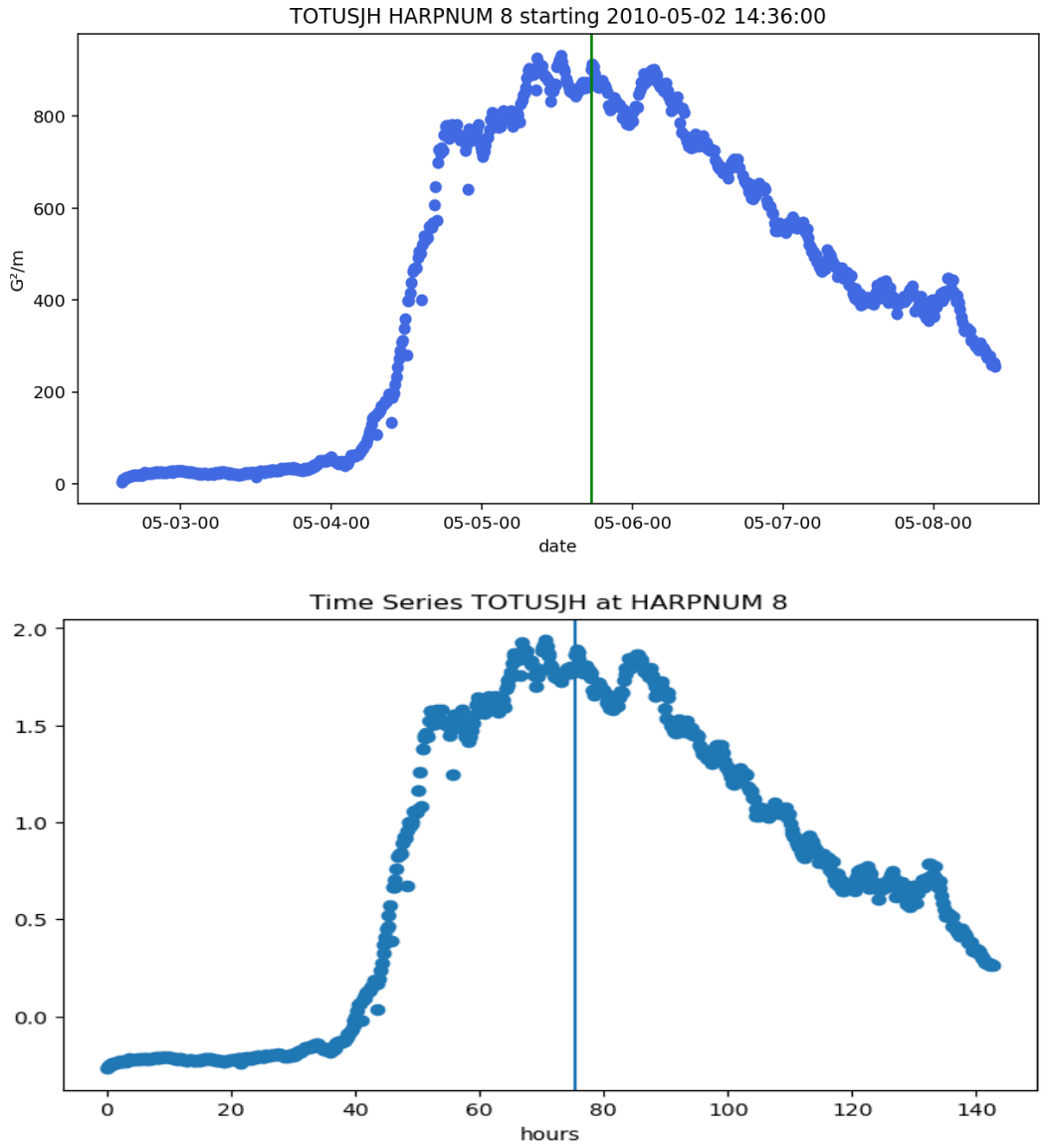

MEANGBZ


MEANGBH


MEANJZH



Gambar IV.7. Distribusi data setiap parameter magnetik yang telah dinormalisasi.



Gambar IV.8. Evolusi parameter TOTUSJH pada HARPNUM 8 tanpa normalisasi (atas) dan dengan normalisasi (bawah) dengan garis vertikal merupakan waktu kejadian suar surya. Tampak bahwa normalisasi tidak mengubah proses fisis parameter.

IV.6. Penerapan Batasan Bujur Heliografis

Pemfilteran data terhadap lokasi bujur heliografis dilakukan setelah dilakukan normalsisasi terhadap data parameter magnetik. Penerapan batasan ini dilakukan dengan melihat kolom LON_FWT yang memuat data posisi bujur daerah aktif yang diamati. Dengan menerapkan batasan ini, dataset yang awalnya terdiri dari 2.900.689 titik data berkurang menjadi 2.651.020 atau mengalami pengurangan sebesar 8% dari jumlah titik data semula. Selain itu, jumlah suar surya yang ditinjau juga mengalami pengurangan yaitu dari yang bermula 50 suar surya kelas X dan 683 suar surya kelas M berkurang menjadi 36 suar surya kelas X dan 535 suar surya kelas M atau mengalami pengurangan sebesar 20% dari jumlah suar surya semula. Hal ini menunjukkan cukup banyak suar surya yang terjadi di luar rentang bujur $\pm 69^{o}$. Namun penerapan batasan ini merupakan hal yang penting dilakukan untuk menghindari error pengukuran yang besar pada tepian piringan Matahari akibat efek proyeksi.

Setelah menerapkan normalisasi dan batasan bujur heliografis pada dataset, dataset disimpan dengan nama file “Robust_Sclaed.csv”.

IV.7. Analisis Fitur terhadap Rumusan Masalah Pertama

Rumusan masalah pertama bertanya, “**Apakah terdapat suatu proses di parameter magnetik tertentu yang selalu mendahului terjadinya suar surya?**”. Untuk mengetahui jawaban dari pertanyaan ini, penulis menganalisis *timestamp* waktu kejadian suar surya dan kapan parameter magnetik tersebut mengalami puncak di daerah aktif yang mengalami suar surya. Alasan di balik analisis ini adalah asumsi dasar bahwa terdapat suatu proses pra-*flare* yang terjadi di daerah aktif yang tercermin dalam evolusi parameter magnetik sebelum suar surya terjadi.

Pada setiap parameter magnetik, setiap suar surya yang terjadi akan dihitung jarak temporalnya terhadap puncak parameter magnetik tersebut pada daerah aktif yang berkaitan. Dengan demikian, diperoleh suatu distribusi jarak teporal suar surya terhadap puncak parameter magnetik. Dengan meninjau distribusi ini untuk masing-masing parameter magnetik, kita dapat melihat apakah terdapat suatu proses di parameter magnetik tersebut yang konsisten mendahului terjadinya suar surya. Apabila terdapat parameter magnetik yang konsisten mengalami puncak sebelum suar surya, proses pra-*flare* dapat direpresentasikan melalui evolusi parameter tersebut. Sebaliknya, jika ternyata tidak ada parameter magnetik yang konsisten mengalami puncak sebelum suar surya, artinya proses pra-*flare* untuk setiap suar surya terjadi pada kombinasi parameter magnetik yang lain satu sama lain. Dapat dikatakan pula bahwa proses pra-*flare* adalah proses yang unik untuk setiap suar surya. Kode yang menunjukkan proses plotting jarak temporal dan distribusinya untuk setiap parameter magnetik dapat dilihat pada gambar IV.9.

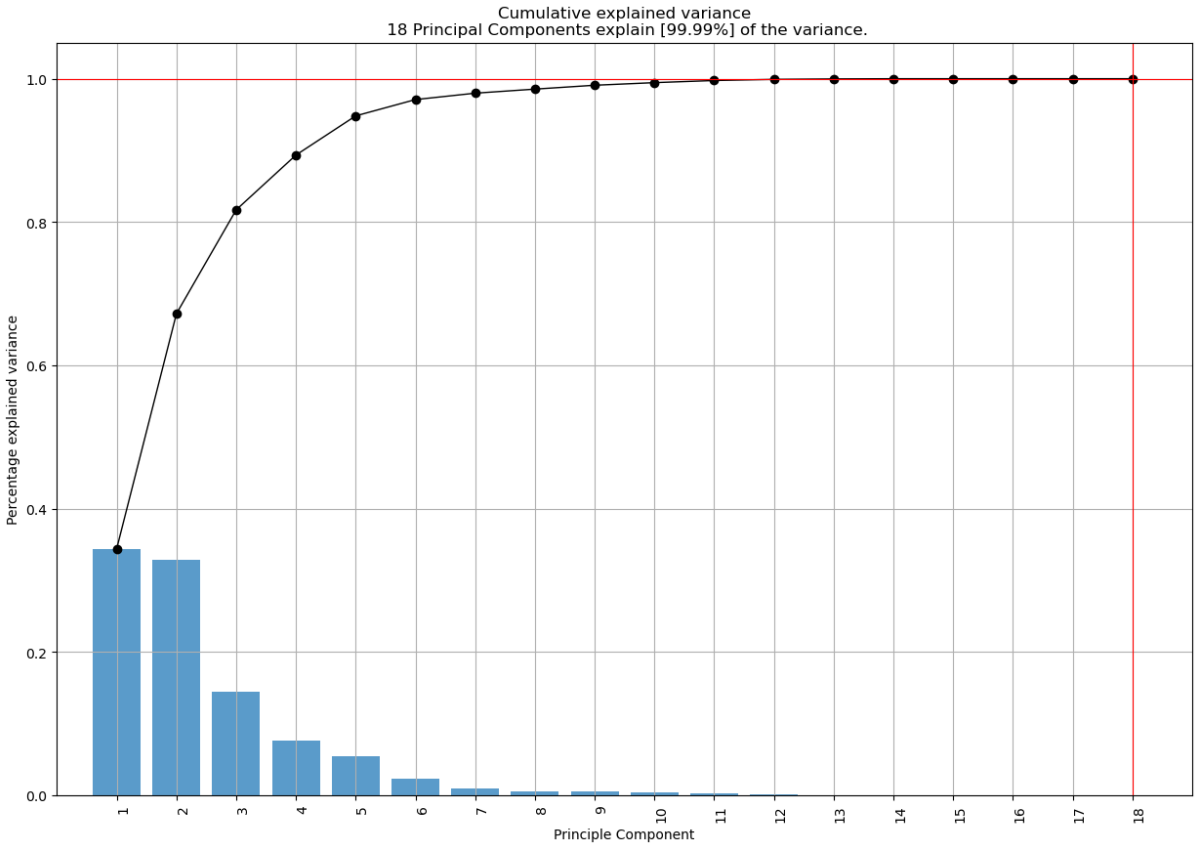
```
%%time
# First get the X and M class flare events
X_and_M = df.loc[df['X_or_M'] == 'Y']
# Get the HARPNUM and flare_index of each flares
HARPNUM = to_array(X_and_M.HARPNUM)
flare_index = to_array(X_and_M['flare_index'])
# Get the names of used columns
cols = ['TOTUSJH', 'TOTPOT', 'TOTUSJZ', 'ABSNJZH', 'SAVNCPP', 'USFLUX',
       'AREA_ACR', 'MEANPOT', 'R_VALUE', 'SHRGT45', 'MEANSHR', 'MEANGAM',
       'MEANGBT', 'MEANGBZ', 'MEANGBH', 'MEANJZH', 'MEANJZD', 'MEANALP']


# Do iteration for each column
for j in cols:
    # Make a blank list to store timestamp of maximum of each parameter
    maxtime = []
    # Do iteration for each flare
    for i in range(len(X_and_M)):
        # Get the time series of the AR which contain the flare
        data = df.loc[df['HARPNUM'] == (df.loc[df['flare_index'] == flare_index[i]].HARPNUM.iloc[0])]
        # Get the timestamp of the flare
        zero = data.loc[data['flare_index'] == flare_index[i]].date_time.iloc[0]
        # Set the timestamp as X axis
        X = to_array(data['date_time'])
        # Transform the time into temporal distance from the flare occurance
        X = to_array([(l - zero).total_seconds()/3600 for l in X])
        # Set the parameter values as Y axis
        Y = to_array(data[j])
        # Get the maximum position with respect to the timestamp of the flare
        Y_max = np.max(Y)
        X_max = X[np.where(Y == Y_max)[0][0]]
        # Plot the maximum position with respect to flare timestamp
        plt.plot(X_max, Y_max, 'o')
        # Store the maximum position into a list
        maxtime.append(X_max)
    plt.title(j)
    plt.xlabel('Hours')
    plt.show()

    # Do a fitting for temporal distance between flare occurance and maximum value of the parameters
    mu, sigma = norm.fit(maxtime)
    n, bins, patches = plt.hist(maxtime, 60, facecolor='green', density=1, alpha=0.5)
    y = norm.pdf(bins, mu, sigma)
    plt.plot(bins, y, 'r--', linewidth=2)
    plt.xlabel('Hours')
    plt.title(j)
    plt.grid(True)
    plt.show()
    print('mean: ' + str(mu) + ' hours')
    print('std:' + str(sigma) + ' hours')

    # Counting the precentage of flare which happens before and after maxtime
    left = sum(1 if (x < 0) else 0 for x in maxtime)
    right = sum(1 if (x >= 0) else 0 for x in maxtime)
    print('Flare happened before peaks: ' + str(100*right/len(maxtime)) + '%')
    print('Flare happened after peaks: ' + str(100*left/len(maxtime)) + '%')
```

Gambar IV.9. Kode yang menunjukkan proses plotting jarak temporal dan distribusinya untuk setiap parameter magnetik.

Selain dari distribusi jarak temporalnya, penulis juga menganalisis kontribusi masing-masing parameter magnetik terhadap keseluruhan variasi data dengan metode *principal component analysis* (PCA). Digunakan pustaka PCA dengan jumlah komponen utama sama dengan jumlah parameter magnetik (18) untuk memperoleh fitur yang paling cocok terhadap masing-masing komponen utama (tabel IV.2) dan plot variasi komponen utama terhadap variasi total data (gambar IV.10). Sedangkan untuk memperoleh skor kontribusi setiap fitur secara objektif (tabel IV.3), digunakan *sklearn.decomposition.PCA* yang tersedia pada pustaka Scikit-learn. Proses pengerjaan tahap ini dapat dilihat pada file “(6) Parameter Ranking with PCA.ipynb” pada repositori Github penelitian ini.

Tabel IV.2. Parameter magnetik terbaik yang terrepresentasikan oleh masing-masing komponen utama.

	PC	feature	loading	type
0	PC1	MEANPOT	-0.364623	best
1	PC2	TOTUSJZ	0.411280	best
2	PC3	MEANGBH	-0.598927	best
3	PC4	AREA_ACR	-0.485993	best
4	PC5	MEANALP	0.994229	best
5	PC6	R_VALUE	-0.840960	best
6	PC7	MEANPOT	0.519831	best
7	PC8	MEANGBH	-0.626952	best
8	PC9	MEANJZH	-0.625122	best
9	PC10	MEANPOT	0.608867	best
10	PC11	USFLUX	-0.757115	best
11	PC12	MEANSHR	0.642448	best
12	PC13	SHRGT45	0.701571	best
13	PC14	MEANGBT	-0.702179	best
14	PC15	MEANJZD	0.760710	best
15	PC16	SAVNCPP	0.564786	best
16	PC17	TOTUSJZ	-0.710970	best
17	PC18	ABSNJZH	0.707118	best
18	PC11	TOTPOT	0.449810	weak
19	PC18	TOTUSJH	-0.707093	weak
20	PC13	MEANGAM	-0.648179	weak
21	PC14	MEANGBZ	0.637066	weak



Gambar IV.10. Grafik variasi komponen utama terhadap variasi total data beserta variasi kumulatifnya relatif terhadap variasi total data (garis tebal hitam).

Tabel IV.3. Skor kontribusi setiap parameter magnetik terhadap variasi total data. Skor ini diperoleh dengan menjumlahkan perkalian *loading score* setiap parameter dan kontribusi variasi komponen utama yang bersangkutan.

Feature	Score
MEANGAM	0.182140
MEANSHR	0.175842
SHRGT45	0.174708
MEANGBT	0.170748
MEANPOT	0.169986
MEANGBZ	0.169937
USFLUX	0.169555
TOTPOT	0.164761
AREA_ACR	0.157467
R_VALUE	0.141715
ABSNJZH	0.137829
TOTUSJH	0.137829
MEANJZD	0.137791
MEANJZH	0.137492
SAVNCPP	0.137305
TOTUSJZ	0.137261
MEANGBH	0.110936
MEANALP	0.073455

IV.8. Tranformasi STFT

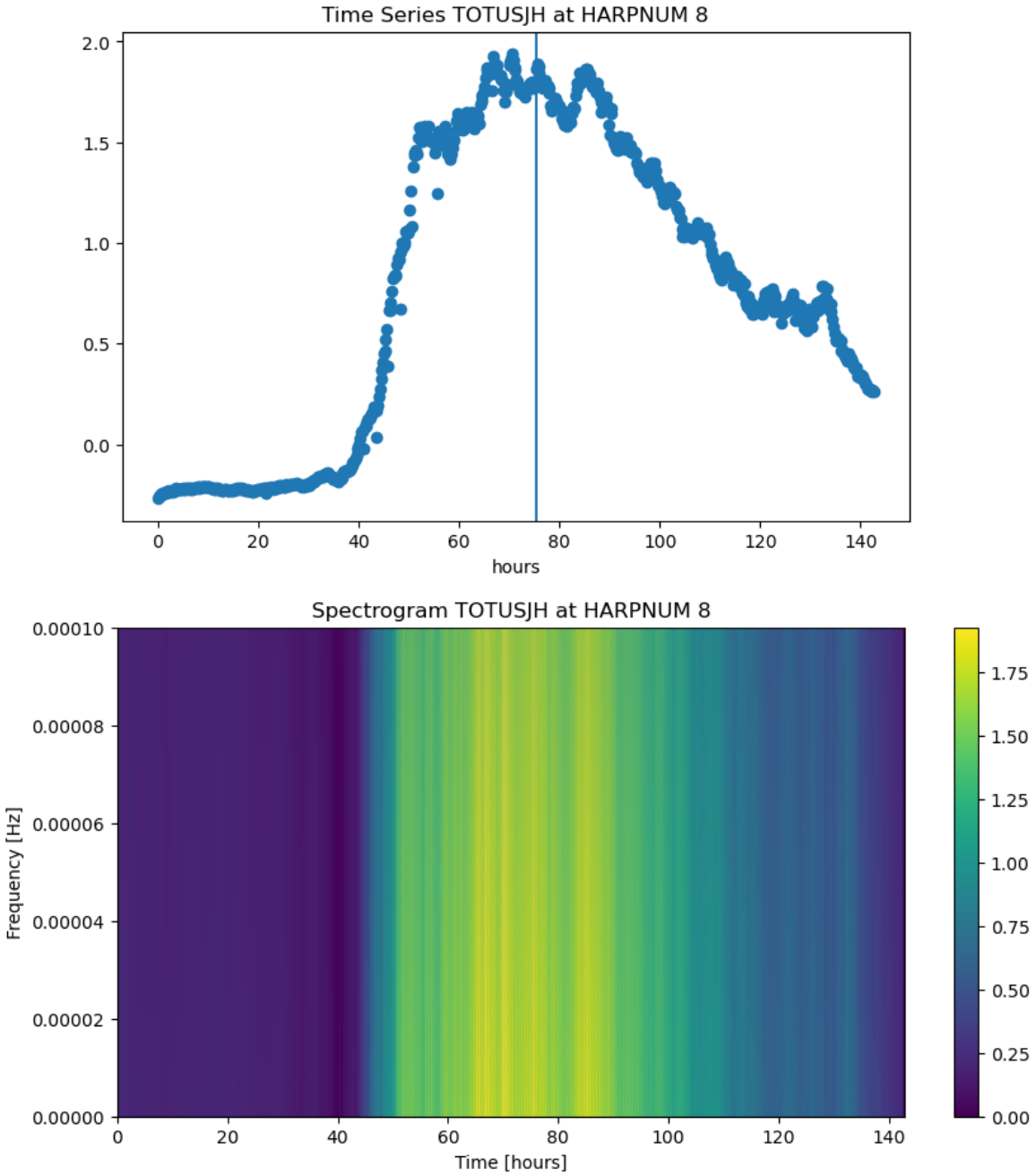
Transformasi data deret waktu domain waktu ke data multidimensi melalui short time fourier transform (STFT) dilakukan untuk mendeteksi fluktuasi yang signifikan atau perubahan yang tiba-tiba pada parameter magnetik. Seperti yang tampak pada gambar IV.8, kita dapat melihat bahwa terdapat fluktuasi yang signifikan atau perubahan secara tiba-tiba pada parameter magnetik TOTUSJH mulai dari 4 Mei 2010 atau jam ke 40 sejak daerah aktif pertama kali diamati hingga kemunculan suar surya. Fluktuasi inilah yang ingin kita deteksi.

Keluaran dari STFT adalah spektrogram dengan sumbu x merupakan waktu, sumbu y merupakan frekuensi yang di sampel, dan sumbu warna menunjukkan korelasi spektrum. Fluktuasi yang tinggi akan memberikan nilai korelasi spektrum yang tinggi pula. Dengan memafaatkan STFT, penulis akan menunjukkan bahwa deteksi fluktuasi ini datap dilakukan.

Kelas *stft* yang tersedia pada pustaka Scipy digunakan dengan memasukkan data deret waktu setiap parameter setiap daerah aktif sebagai data yang akan ditransformasikan.

Kemudian dipilih nilai *fs* atau frekuensi sampling dari deret waktu sama dengan 1/720 hz. Nilai ini berkaitan dengan jarak antar titik data di SHARP yaitu 12 menit atau 720 detik. Kemudian dipilih nilai hyperparameter *nperseg* = 3 untuk mengatur jumlah titik data persegmen sama dengan tiga. Pilihan ini diambil untuk menjaga resolusi temporal yang maksimal setelah transformasi dilakukan. Kode untuk proses ini dapat dilihat pada notebook "(7) Spectrogram Generation (Robust Scaler).ipynb" pada tautan https://github.com/CAF103/Thesis .

Contoh hasil dari transformasi ini dapat dilihat pada gambar IV.10.



Gambar IV.11. Salah satu contoh transformasi STFT yang diterapkan pada parameter magnetik TOTUSJH pada HARPNUM 8.

Dapat dilihat dari gambar IV.11 bahwa spektrogram hasil STFT dapat menunjukkan bahwa

fluktuasi dapat dibedakan terhadap proses latar belakang dari kontras nilai korelasi spektrumnya.

Meskipun transformasi STFT ini dijaga untuk tetap mempertahankan resolusi temporal, dipilihnya nperseg=3 tetap menurunkan resolusi temporal dataset sebesar 50%. Hal ini dikarenakan tidak semua titik data memiliki korelasi spektrum, melainkan hanya titik data yang berada di tengah segmen yang memiliki korelasi spektrum sehingga titik data yang tidak memiliki nilai korelasi spektrum tidak terpakai. Hal ini menurunkan jumlah titik data dalam dataset menjadi 1.328.816 titik data.

IV.9. Analisis Statistik terhadap Korelasi Spektrum

Sebelum dilakukan pelabelan terhadap titik data, penulis juga melakukan analisis statistik terhadap nilai korelasi spektrum pada daerah aktif tanpa suar surya terhadap daerah aktif dengan suar surya. Analisis ini dilakukan dengan mengambil properti statistik dari distribusi korelasi spektrum setiap parameter magnetik pada daerah aktif dengan suar surya dan tanpa suar surya. Kedua kasus daerah aktif ini dianggap sebagai dua kasus yang berbeda sehingga diperoleh dua distribusi yang berbeda pula untuk setiap parameter magnetik.

Untuk membedakan titik data yang merupakan anggota distribusi proses pra-*flare* dengan distribusi proses latar belakang, digunakanlah nilai *modified z-score* (z_m) yang memiliki formula

$$z_m = 0{,}6745\,\frac{|x_i - \bar{x}|}{|x_\iota - \bar{x}|}, \qquad \text{(IV.1)}$$

dengan tanda garis horizontal di atas menunjukkan nilai median. Nilai z_m merupakan suatu skor yang menunjukkan kebolehjadiannya sebagai pencilan atau *outlier* dalam suatu distribusi. Nilai batasan $z_m = 3$ biasa digunakan untuk memfilter pencilan dari suatu distribusi sehingga dalam distribusi yang tersisa adalah titik data dengan nilai $z_m < 3$ (Iglewicz dan Hoaglin, 1993). Pada penelitian ini, digunakan nilai $z_m = 3{,}5$ untuk memfilter pencilan dari distribusi.

Di sini penulis mengasumsikan bahwa proses latar belakang merupakan proses yang dominan lebih banyak terjadi daripada proses pra-*flare*. Asumsi ini dapat dijustifikasi dari jumlah suar surya yang ditinjau jauh lebih kecil daripada jumlah keseluruhan titik data (0,2% dari keseluruhan data). Dengan demikian, proses pra-*flare* dapat dikatakan sebagai pencilan atau *outlier* terhadap keseluruhan distribusi. Distribusi proses latar belakang pada setiap parameter kemudian diwakili dengan distribusi korelasi spektrum yang telah difilter dengan metode *modified z-score*. Sedangkan distribusi proses pra-flare diwakili oleh distribusi nilai korelasi spektrum daerah aktif dengan suar surya.

Dengan membandingkan distribusi proses pra-*flare* terhadap distribusi proses latar

belakang (tabel IV.4), penulis dapat menjawab rumusan masalah yang kedua yaitu, "**Bagaimana kenampakan proses pra-flare pada setiap parameter magnetik?"**. Dengan menjawab rumusan masalah yang kedua ini, penulis dapat memberikan peringkat kontribusi parameter magnetik terhadap kemunculan suar surya berdasarkan kenampakan proses pra-flare-nya (tabel IV.5).

Tabel IV.4. Statistik dari masing-masing distribusi proses pra-flare dan proses latar belakang di setiap parameter magnetik.

Parameter	***Background***	***Flare***	**Rasio**
TOTUSJH			
Rata-rata	0.116751141	2.353719602	20.160142
Median	0.10772625	1.865466028	17.316727
Standar deviasi	0.08386941	2.040444772	24.328832
TOTPOT			
Rata-rata	0.05040029	3.927391202	77.92398
Median	0.049053749	2.60283027	53.060781
Standar deviasi	0.033821744	4.616265051	136.48808
TOTUSJZ			
Rata-rata	0.120867581	2.24794094	18.598378
Median	0.109844953	1.828223285	16.643671
Standar deviasi	0.088938058	1.835742882	20.64069
ABSNJZH			
Rata-rata	0.104101255	4.265759217	40.97702
Median	0.096777323	2.290011234	23.662684
Standar deviasi	0.065305272	5.659425542	86.661082
SAVNCPP			
Rata-rata	0.14571847	3.044961841	20.896197
Median	0.133124834	1.772816573	13.316949
Standar deviasi	0.091235913	3.642374784	39.922599
USFLUX			
Rata-rata	0.114556909	1.943512191	16.965473
Median	0.105899279	1.58047303	14.924304
Standar deviasi	0.081586979	1.602452343	19.64103
AREA_ACR			
Rata-rata	0.127699737	1.961482744	15.360116
Median	0.116209837	1.563400874	13.453258
Standar deviasi	0.094622114	1.651427019	17.452865
MEANPOT			
Rata-rata	0.102284121	1.014137153	9.9149031
Median	0.074478273	0.872818405	11.719101
Standar deviasi	0.092552393	0.768697951	8.3055438
R_VALUE			
Rata-rata	0.215159881	0.581669609	2.7034297
Median	0	0.618949891	-
Standar deviasi	0.253404065	0.156783392	0.6187091

SHRGT45			
Rata-rata	0.10273408	0.710390882	6.9148513
Median	0.087061429	0.673665207	7.7378147
Standar deviasi	0.078708984	0.451558596	5.7370655
MEANSHR			
Rata-rata	0.21201915	0.671422207	3.1667998
Median	0.144667108	0.631539529	4.3654673
Standar deviasi	0.191410658	0.429447656	2.2435932
MEANGAM			
Rata-rata	0.214177266	0.589486071	2.7523279
Median	0.153712813	0.541427298	3.5223303
Standar deviasi	0.184145164	0.385372743	2.092766
MEANGBT			
Rata-rata	0.256746991	0.273040672	1.063462
Median	0.227047193	0.251487106	1.1076424
Standar deviasi	0.17349488	0.177136226	1.0209882
MEANGBZ			
Rata-rata	0.266209046	0.252967697	0.9502596
Median	0.231731568	0.223125078	0.9628601
Standar deviasi	0.183901496	0.174332904	0.9479689
MEANGBH			
Rata-rata	0.266363917	0.254419553	0.9551577
Median	0.23442561	0.231537534	0.9876802
Standar deviasi	0.183707683	0.171201857	0.9319254
MEANJZH			
Rata-rata	0.333357841	0.476618398	1.4297501
Median	0.266834863	0.323369937	1.2118729
Standar deviasi	0.254005381	0.50165777	1.9749887
MEANJZD			
Rata-rata	0.346733687	0.155102733	0.4473253
Median	0.262773112	0.126580862	0.4817116
Standar deviasi	0.274154416	0.131205805	0.4785836
MEANALP			
Rata-rata	0.342227609	0.343053897	1.0024144
Median	0.273901473	0.24612718	0.8985975
Standar deviasi	0.262919143	0.333354397	1.267897

Tabel IV.5. Peringkat signifikansi parameter magnetik berdasarkan rasio rata-rata distribusi proses pra-*flare* terhadap proses latar belakang.

Rank	Parameter	Rasio rata-rata	Keterangan
1	TOTPOT	77.92398	Total densitas energi magnetik fotosfer dalam erg/cm^3
2	ABSNJZH	40.97702	Nilai absolut helisitas arus bersih dalam G^2/m
3	SAVNCPP	20.896197	Jumlah nilai absolut dari arus bersih per polaritas dalam Ampere
4	TOTUSJH	20.160142	Total helisitas arus tanpa tanda dalam G^2/m
5	TOTUSJZ	18.598378	Total arus vertikal tanpa tanda dalam mA/m^2

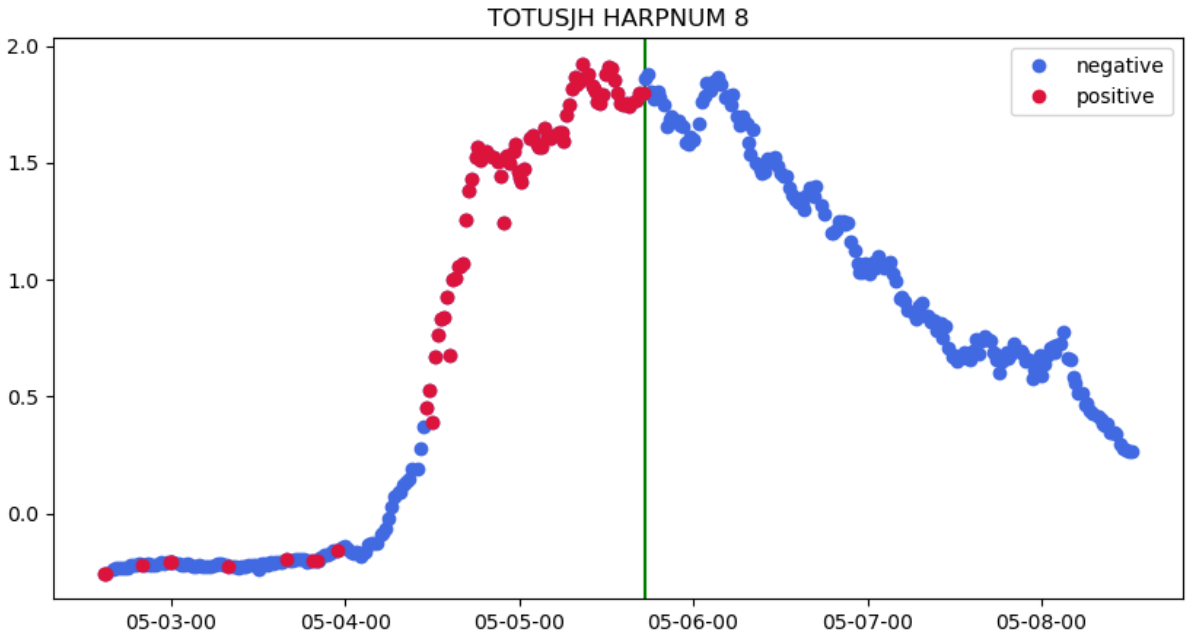
6	USFLUX	16.965473	Total fluks tanpa tanda dalam Maxwell
7	AREA_ACR	15.360116	Area dengan medan kuat dalam piksel
8	MEANPOT	9.9149031	Rata-rata ekses densitas energi megnetik fotosfer dalam erg/cm^3
9	SHRGT45	6.9148513	Persentase piksel dengan shear angle > 45^o dalam persen
10	MEANSHR	3.1667998	Rata-rata shear angle diukur dari B_{total} dalam derajat
11	MEANGAM	2.7523279	Rata-rata sudut inklinasi, gamma dalam derajat
12	R_VALUE	2.7034297	Total fluks dekat garis inversi polaritas dalam Maxwell
13	MEANJZH	1.4297501	Rata-rata helisitas arus dalam G^2/m
14	MEANGBT	1.063462	Rata-rata gradien medan magnet total dalam Gauss/Mm
15	MEANALP	1.0024144	Rata-rata parameter puntiran, alpha dalam 1/Mm
16	MEANGBH	0.9551577	Rata-rata gradien medan magnet horizontal dalam Gauss/Mm
17	MEANGBZ	0.9502596	Rata-rata gradien medan magnet vertikal dalam Gauss/Mm
18	MEANJZD	0.4473253	Rata-rata densitas arus vertikal dalam mA/m^2

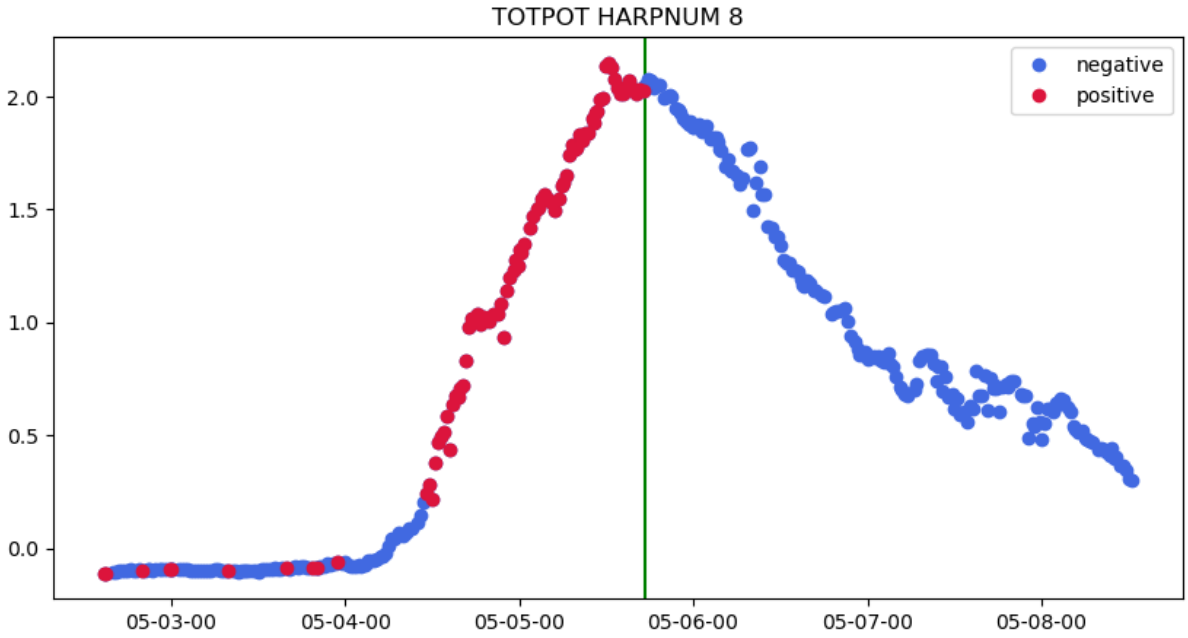
IV.10. Pelabelan Titik Data

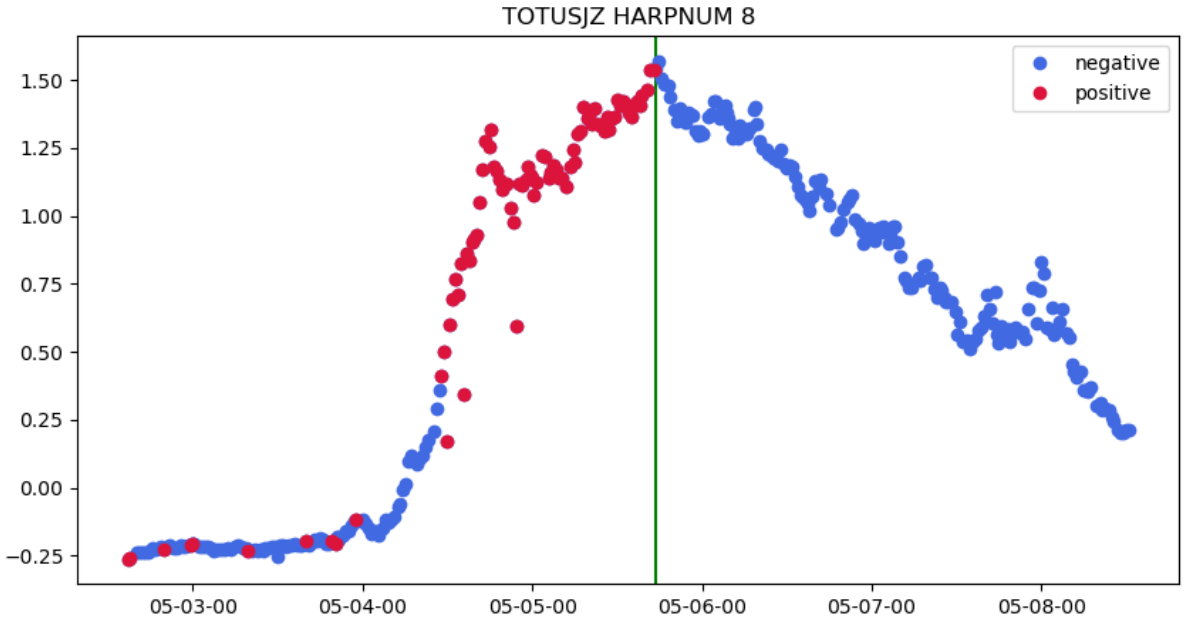
Dalam penelitian ini, label sebagai aspek yang diprediksi oleh estimator terdiri dari dua kelas, yaitu positif (1) dan negatif (-1). Pada kasus model prediksi keunculan suar surya, kelas positif pada suatu titik data menunjukkan bahwa dalam waktu dekat, akan muncul suar surya pada HARP yang sama dengan titik data tersebut. Sedangkan label negatif pada suatu titik data menunjukkan bahwa tidak akan muncul suar surya pada HARP yang sama dengan titik data tersebut dalam waktu dekat. Tujuan dari pelabelan di sini adalah untuk mengidentifikasi proses pra-*flare* sehingga kemunculan suar surya dapat diprediksi. Dalam penelitian ini, pelabelan dilakukan dengan melihat fluktuasi sebelum suar surya yang tercermin dari nilai korelasi spektrum yang tinggi.

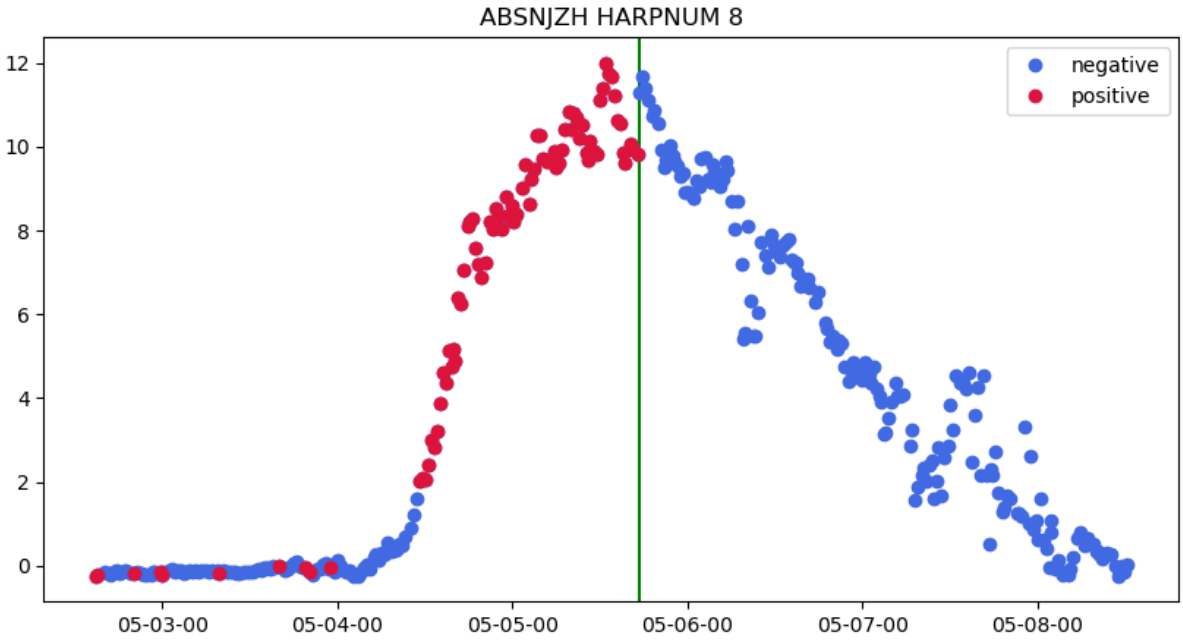
Untuk dapat membedakan proses pra-*flare* dengan proses latar belakang, diperlukan distribusi korelasi spektrum daerah aktif dengan suar surya dan distribusi korelasi spektrum daerah aktif tanpa suar surya pada setiap parameter magnetik. Hasil dari analisis kedua distribusi ini ini dapat memberikan gambaran kontribusi setiap parameter magnetik terhadap proses pra-*flare*. Berdasarkan hasil analisis ini, diambilah 12 parameter magnetik dengan kontribusi teratas (tabel IV.5) sebagai parameter magnetik yang ditinjau sebagai fitur dalam model prediksi.

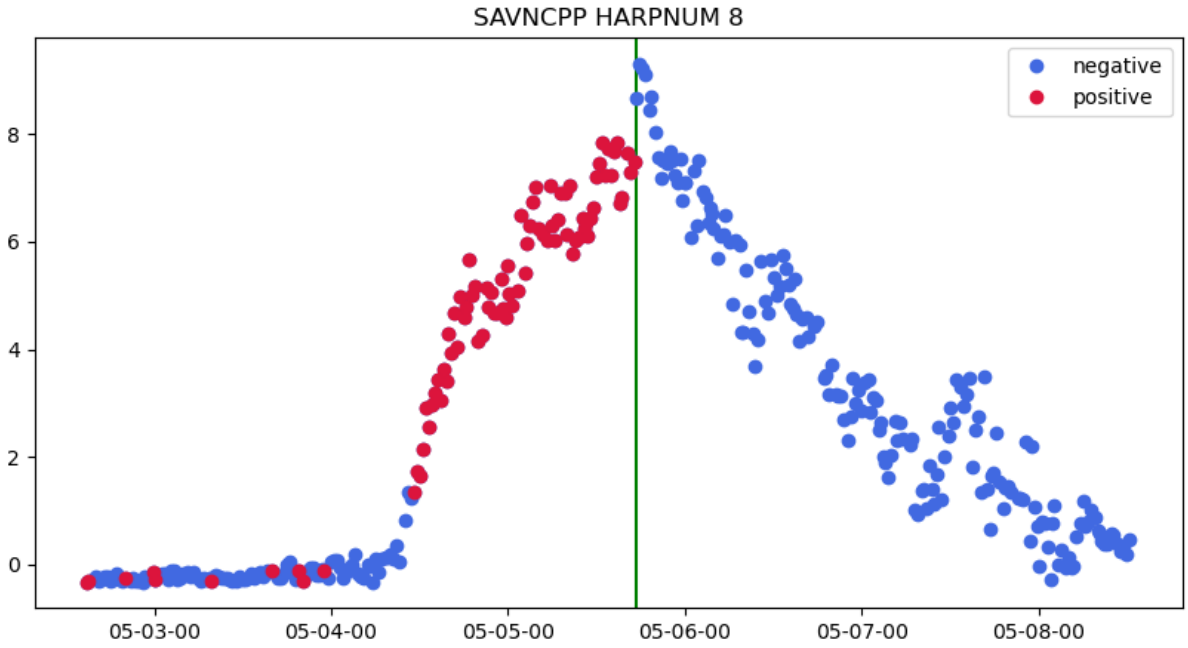
Pelabelan kemudian dilakukan dengan menerapkan suatu nilai batas pada penjumlahan nilai korelasi spektrum keduabelas parameter magnetik yang ditinjau. Nilai batasnya sendiri diperoleh dari penerapan aturan empiris terhadap distribusi penjumlahan nilai korelasi spektrum keduabelas parameter magnetik. Penelitian ini menggunakan batas 2,5 standar deviasi sebagai batas label positif dan negatif. Batasan ini berkorelasi dengan 98,758 % dari keseluruhan distribusi jumlah korelasi spektrum. Untuk titik data dengan nilai jumlah korelasi spektrum yang lebih tinggi dari median + 2,5 standar deviasi, diberikan nilai label positif untuk titik data tersebut. Begitu juga sebaliknya. Contoh hasil pelabelan menggunakan metode ini ditunjukkan oleh gambar IV.12.

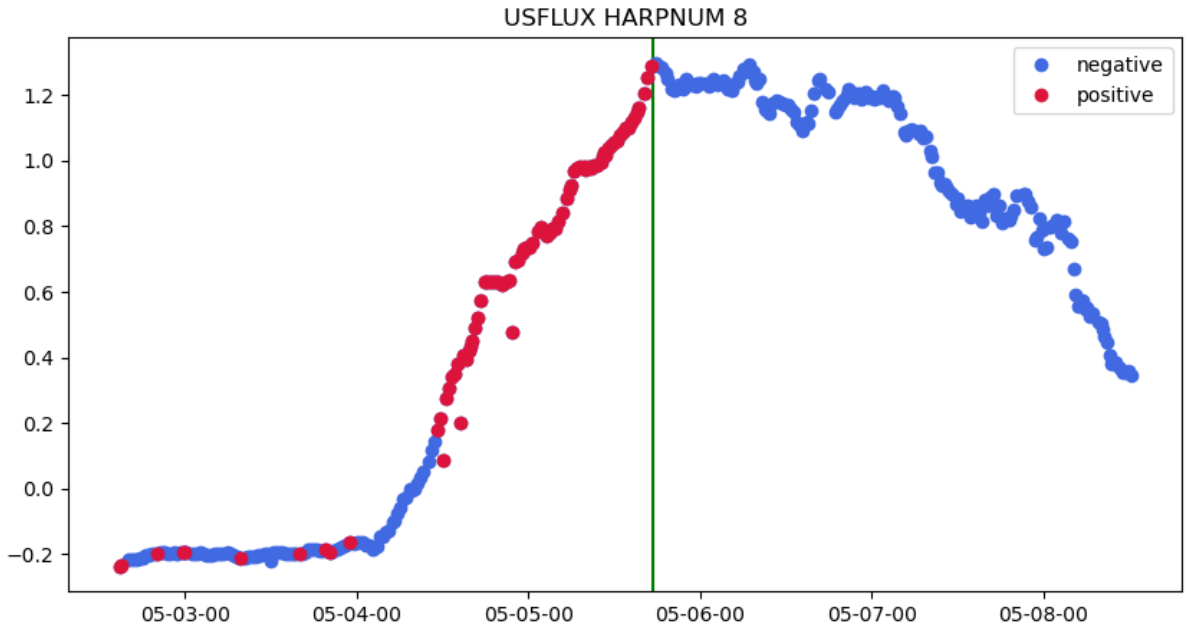

TOTUSJH HARPNUM 8
negative
positive
2.0
1.5
1.0
0.5
0.0
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00

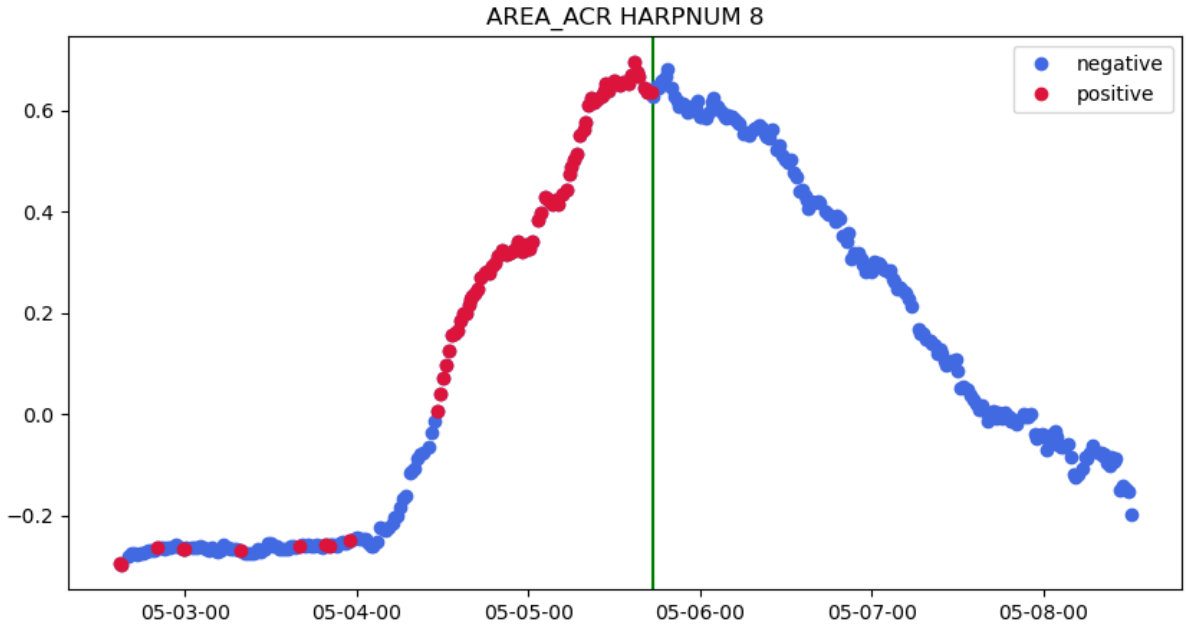

TOTPOT HARPNUM 8
negative
positive
2.0
1.5
1.0
0.5
0.0
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00

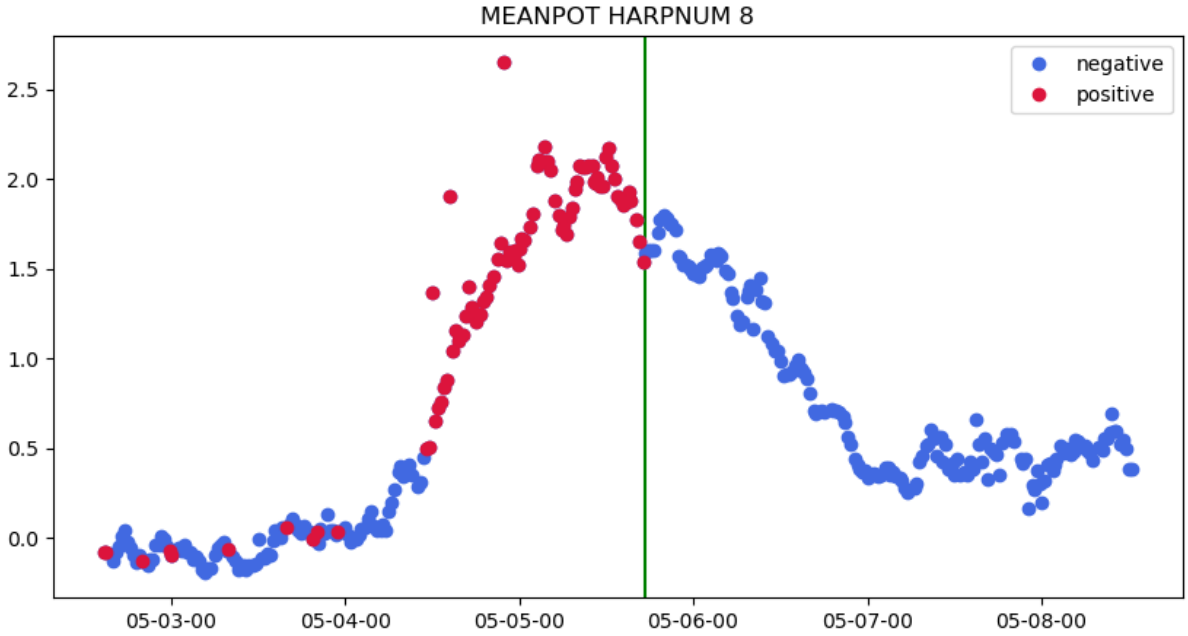

TOTUSJZ HARPNUM 8
negative
positive
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00

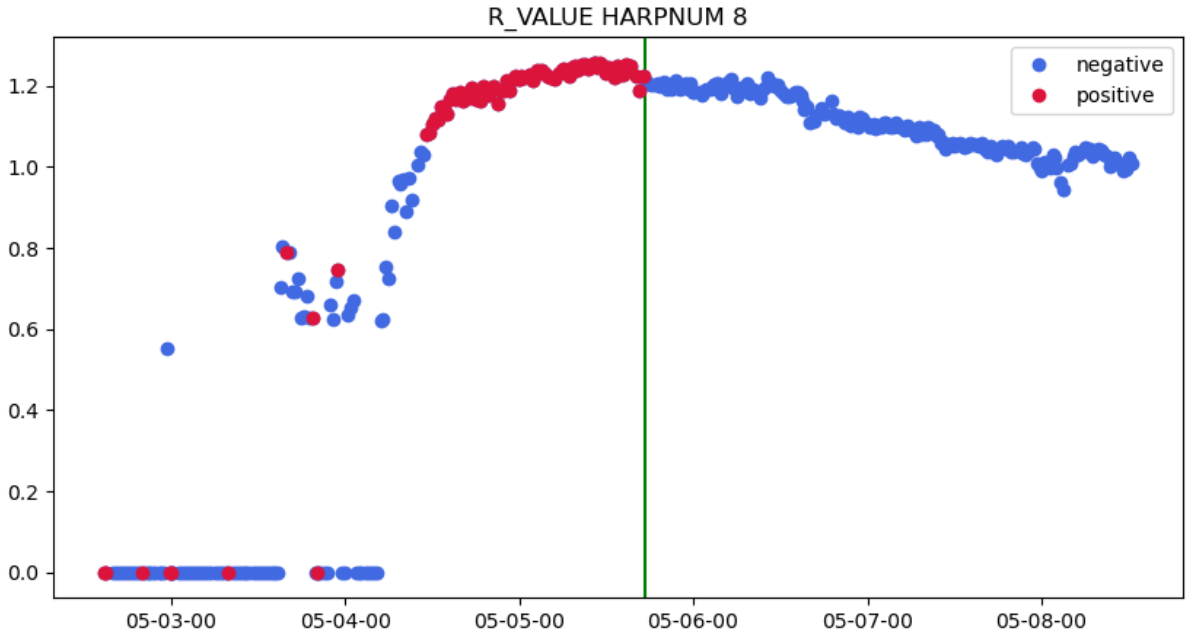

ABSNJZH HARPNUM 8
negative
positive
12
10
8
6
4
2
0
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00

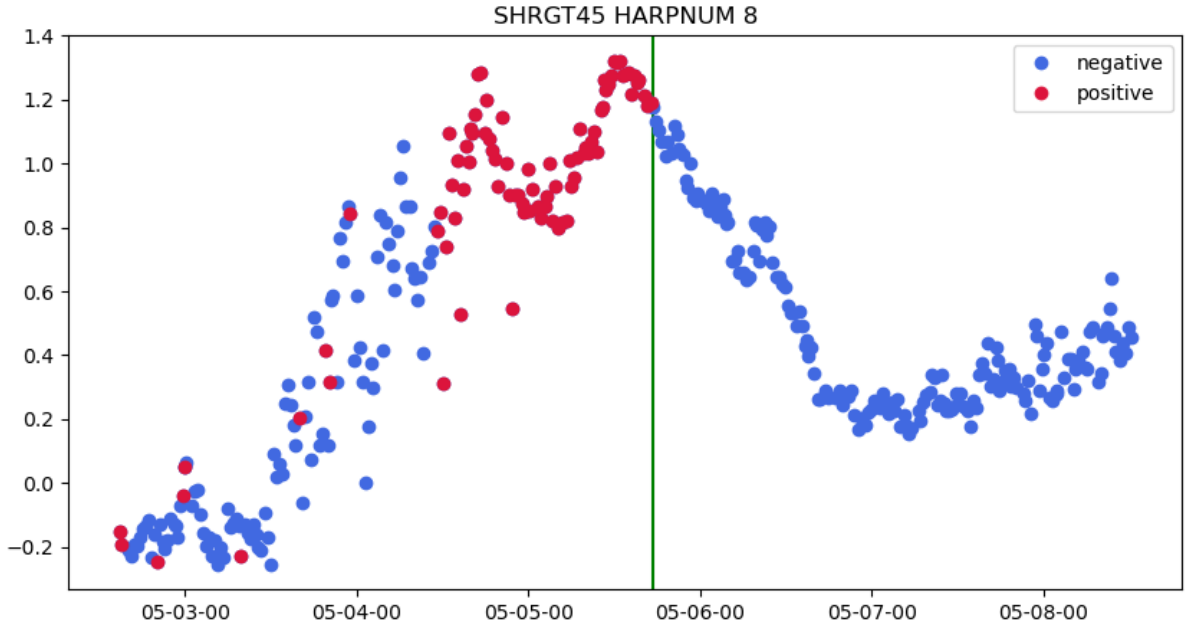

SAVNCPP HARPNUM 8
negative
positive
8
6
4
2
0
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00

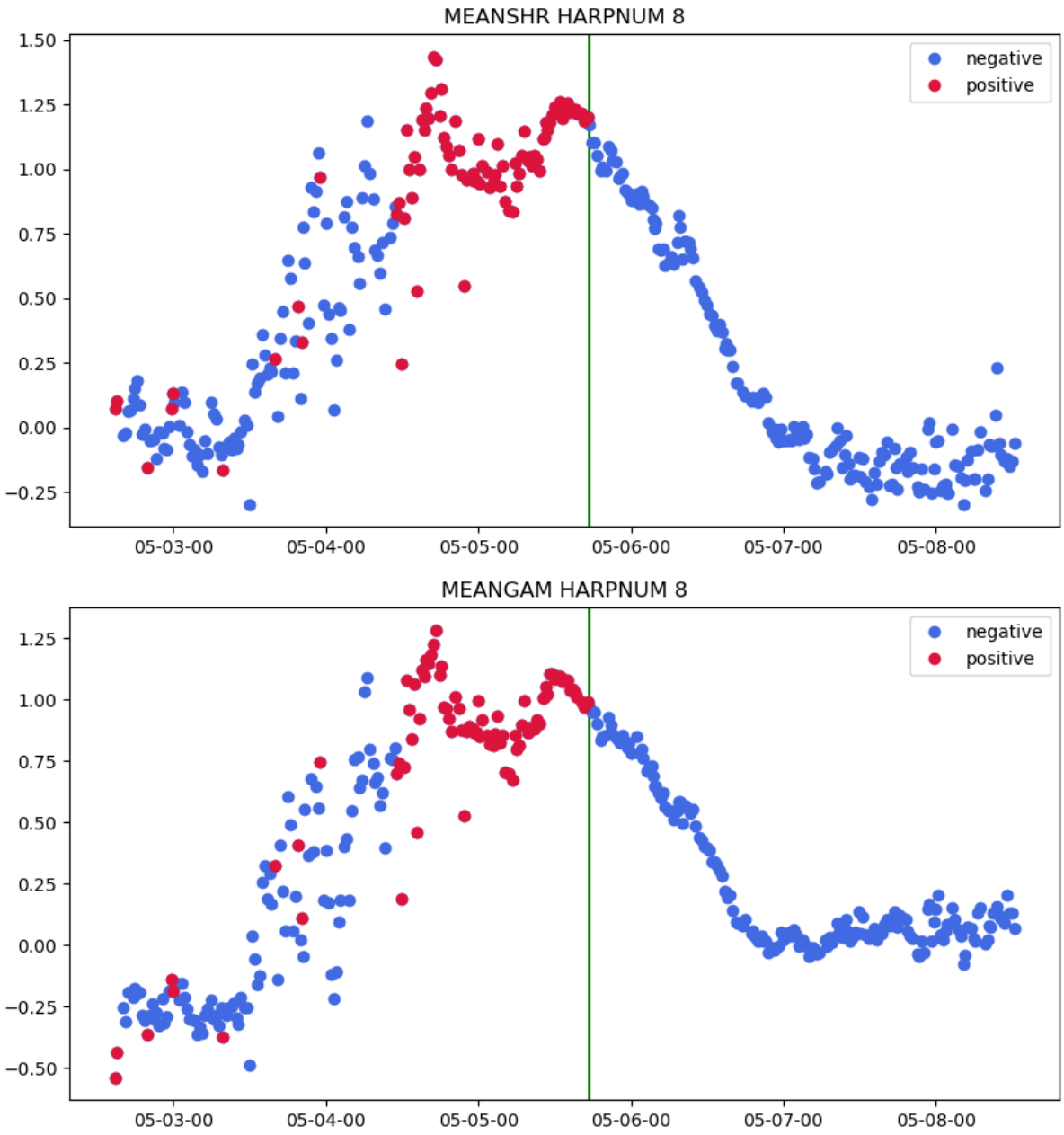

USFLUX HARPNUM 8
negative
positive
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00


AREA_ACR HARPNUM 8
negative
positive
0.6
0.4
0.2
0.0
−0.2
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00


MEANPOT HARPNUM 8
negative
positive
2.5
2.0
1.5
1.0
0.5
0.0
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00


R_VALUE HARPNUM 8
negative
positive
1.2
1.0
0.8
0.6
0.4
0.2
0.0
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00


SHRGT45 HARPNUM 8
negative
positive
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
05-03-00
05-04-00
05-05-00
05-06-00
05-07-00
05-08-00



Gambar IV.12. Contoh pelabelan yang diterapkan pada HARPNUM 8 yang diperlihatkan pada evolusi setiap parameter magnetik.

Dengan dijalankannya sistem pelabelan ini, diperolehlah dataset yang sudah dilabeli yang terdiri dari 53.725 titik data dengan label positif (4,04 %) dan 1.275.091 (95,95 %) titik data dengan label negatif. Dataset ini kemudian disimpan dan diberi nama “Labeled_Dataset.csv”. Proses pengerjaan pelabelan dapat dilihat pada file “(8) Labeling.ipynb” pada repositori Github penelitian ini.

IV.11. Prediksi dengan *Support Vector Machine*

Proses penerapan support vector machine dalam rangka memprediksi label titik data dimulai dari *tuning* atau penyesuaian *hyperparameter* yang ada di kelas LinearSVC yang digunakan. Proses ini dilakukan dengan menggunakan bantuan kelas GridSearchCV yang juga tersedia di pustaka Scikit-learn. *Hyperparameter* pada LinearSVC yang diatur di awal tanpa penyesuaian adalah class_weight=’balanced’. Dengan mengatur class_weight=’balanced’, bobot kelas label lebih diperhitungkan dalam menjalankan support vector machine. Hal ini penting dilakukan karena ketimpangan jumlah label positif terhadap negatif (4,04 % vs 95,95 %). Dengan mengatur class_weight=’balanced’, bobot yang diberikan kepada setiap titik data mengikuti persamaan

$$\boldsymbol{w_i} = \frac{\boldsymbol{n}}{\boldsymbol{n_{kelas}} \times \boldsymbol{n_i}}, \qquad \text{(IV.2)}$$

Dengan $\boldsymbol{n}$ adalah banyaknya titik data, $\boldsymbol{n_{kelas}}$ adalah banyaknya kelas (dalam kasus klasifikasi biner, $\boldsymbol{n_{kelas}}$ = 2), dan $\boldsymbol{n_i}$ adalah banyaknya anggota kelas ke-i.

Untuk menjalankan GridSearchCV, kita perlu memberikan nilai-nilai *tuning* untuk *hyperparameter*, untuk kasus ini adalah nilai *C*. Proses *hyperparameter tuning* ini dilakukan dua kali dengan nilai yang disampel berturut turut (0,001; 0,01; 0,1; 0; 1; 10; 100; 1000) dan (0,01; 0,01467799; 0,02154435; 0,03162278, 0,04641589; 0,06812921; 0,1). Tuning dilakukan dua kali untuk memperoleh nilai *C* yang lebih optimal dan lebih halus. Setelah itu, dideklarasikan pula sistem validasi silang yang digunakan yaitu StratifiedKFold dengan nilai *n_split*=5. GridSearchCV dijalankan dengan mengoptimalkan metrik performa TSS. Dengan demikian, diperolehlah nilai *C* yang paling optimal untuk dataset yang digunakan adalah 0,06812.

Dengan dimilikinya nilai *C* yang optimal, dijalankanlah LinearSVC terhadap dataset yang telah dilabeli. Penjalanan LinearSVC ini juga tetap menerapkan validasi silang StratifiedKFold dengan nilai n_split=5. LinearSVC dijalankan sebanyak 12 kali dengan iterasi pertama menggunakan 12 parameter magnetik dengan peringkat kontribusi tertinggi sebagai fitur (tabel IV.5). Itrasi selanjutnya kemudian dijalankan dengan mengurangi jumlah fitur yang digunakan satu per satu berdasrkan peringkatnya dari yang paling rendah hingga tinggi. Performa model prediksi kemudian dievaluasi pada lima metrik skor yaitu akurasi, presisi, recall, f1, dan TSS.

BAB V

HASIL DAN DISKUSI

Salah satu tujuan penelitian ini adalah menentukan peran masing-masing parameter magnetik terhadap kemunculan suar surya. Parameter magnetik fotosfer memiliki hubungan erat dengan kemunculan suar surya sehingga digunakan sebagai fitur dalam berbagai model prediksi seperti pada Bobra dan Couvidat (2015), dan Zhang dkk. (2022). Dengan mengetahui peranan setiap parameter magnetik terhadap kemunculan suar surya, kita dapat membatasi parameter magnetik apa saja yang sebaiknya ditinjau dalam model prediksi demi memperoleh performa prediksi yang optimal. Selain itu, peranan atau signifikansi parameter terhadap kemunculan suar surya dapat memberikan wawasan terhadap fisis di balik suar surya secara lebih dekat.

Pembatasan terhadap jumlah fitur yang digunakan dalam model prediksi secara umum dilakukan untuk memperoleh model prediksi yang efisien. Dalam pembelajaran mesin sendiri, jumlah fitur yang lebih sedikit umumnya lebih disukai karena beberapa alasan:

- *Curse of dimensionality*: Ketika jumlah fitur (dimensi) meningkat, jumlah data yang dibutuhkan untuk melatih model dengan efektif juga meningkat secara eksponensial. Ini dikenal sebagai "*curse of dimensionality*". Dengan jumlah fitur yang lebih sedikit, kita dapat menghindari masalah ini dan mengurangi jumlah data *training* yang diperlukan.
- Kompleksitas yang lebih rendah: Fitur yang lebih sedikit umumnya menghasilkan model yang lebih sederhana, yang lebih mudah diinterpretasikan, dipahami, dan dipelihara. Model yang kompleks dengan jumlah fitur yang banyak rentan terhadap *overfitting*, yang berarti model mungkin menghafal data latihan daripada mempelajari pola umum, yang mengakibatkan performa yang buruk pada data baru yang belum dilihat sebelumnya.
- Seleksi dan ekstraksi fitur: Dengan mengurangi jumlah fitur, kita dapat fokus pada fitur-fitur yang paling relevan dan informatif untuk tugas yang dihadapi. Teknik seleksi fitur dapat diterapkan untuk mengidentifikasi fitur-fitur yang paling berpengaruh terhadap prediksi, sedangkan metode ekstraksi fitur dapat mentransformasi fitur-fitur asli menjadi representasi berdimensi lebih rendah yang tetap mempertahankan informasi penting. Salah satu teknik ekstraksi fitur adalah PCA.
- Efisiensi komputasi: Bekerja dengan jumlah fitur yang lebih sedikit dapat secara signifikan meningkatkan efisiensi komputasi dalam proses pelatihan dan inferensi. Ruang fitur yang besar memerlukan lebih banyak memori dan sumber daya komputasi,

sehingga sulit untuk mengolah algoritma dalam skala besar atau menerapkannya di lingkungan dengan sumber daya terbatas.

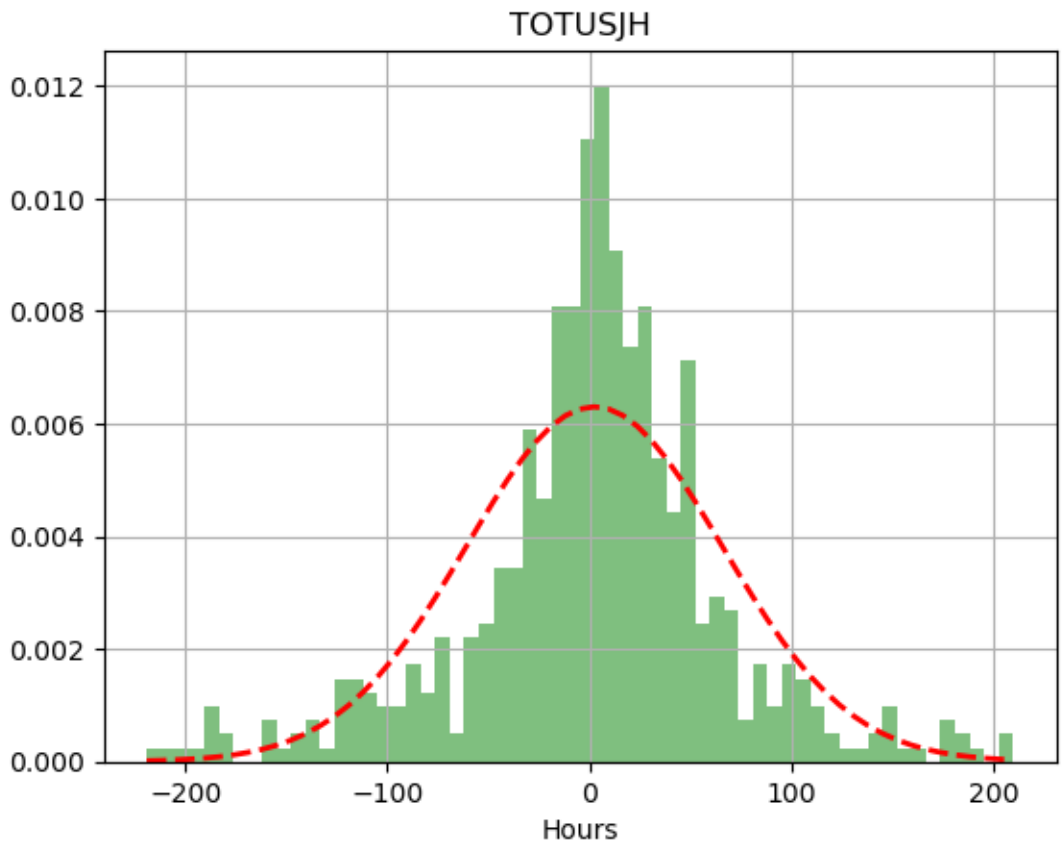
- Menghindari noise dan informasi yang tidak relevan: Jumlah fitur yang lebih besar dapat memperkenalkan *noise* dan informasi yang tidak relevan, yang dapat mengganggu proses pembelajaran. Fitur yang tidak relevan atau redundan dapat menambah kompleksitas yang tidak perlu pada model dan memperkenalkan noise yang mengganggu identifikasi pola yang bermakna dalam data.

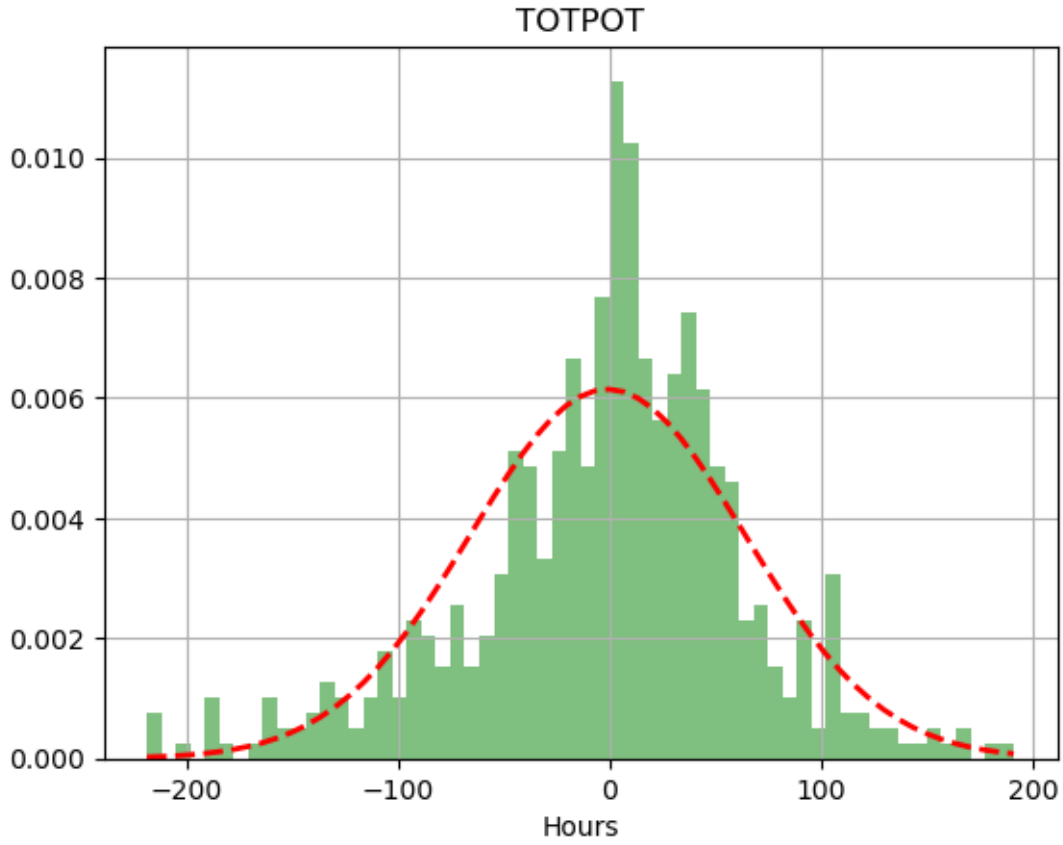
Serangkaian prosedur yang dilaksanakan dalam penelitian ini dijalankan untuk mengoptimalkan penggunaan fitur yang digunakan dalam prediksi. Analisis jarak temporal suar surya, PCA, analisis korelasi spektrum, dan investigasi pengurangan fitur dalam menjalankan LinearSVC, semuanya dilakukan demi memperoleh model prediksi terbaik dengan jumlah fitur paling optimal. Berikut dituliskan pada bab ini hasil dari analisis-analisis tersebut beserta diskusi dan komentar penulis.

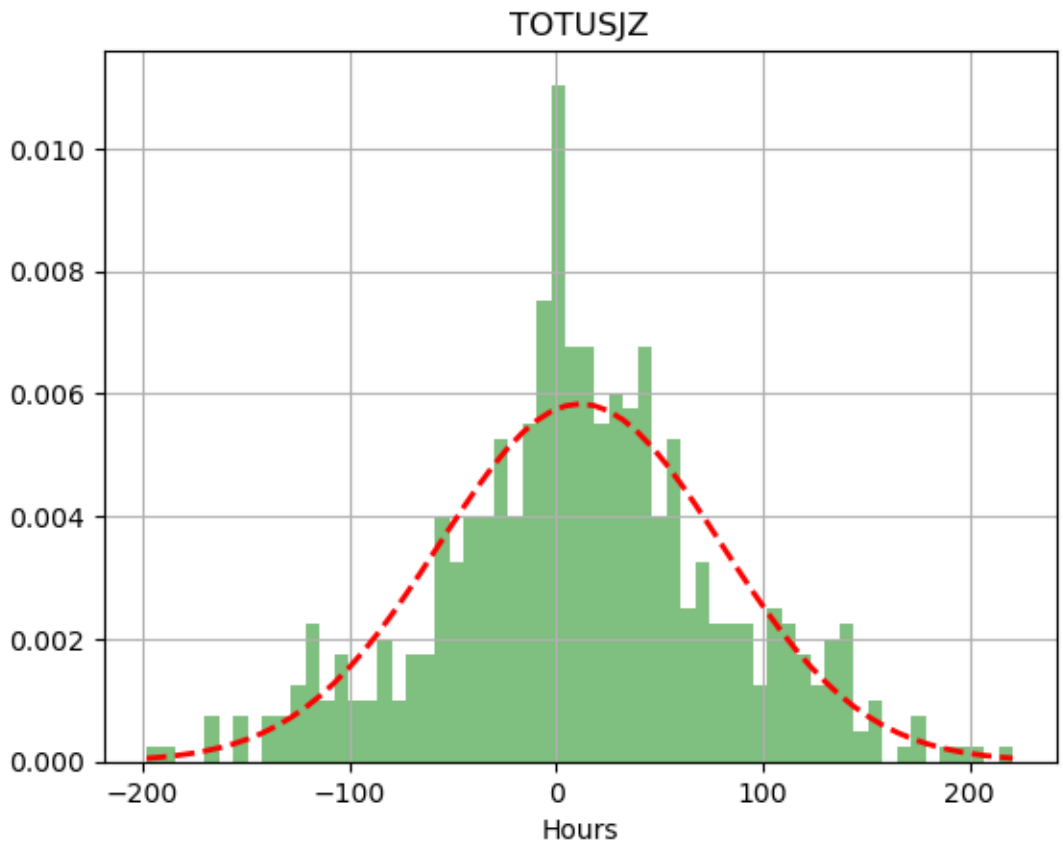
V.1 Distribusi Jarak Temporal Suar Surya terhadap Puncak Parameter Magnetik

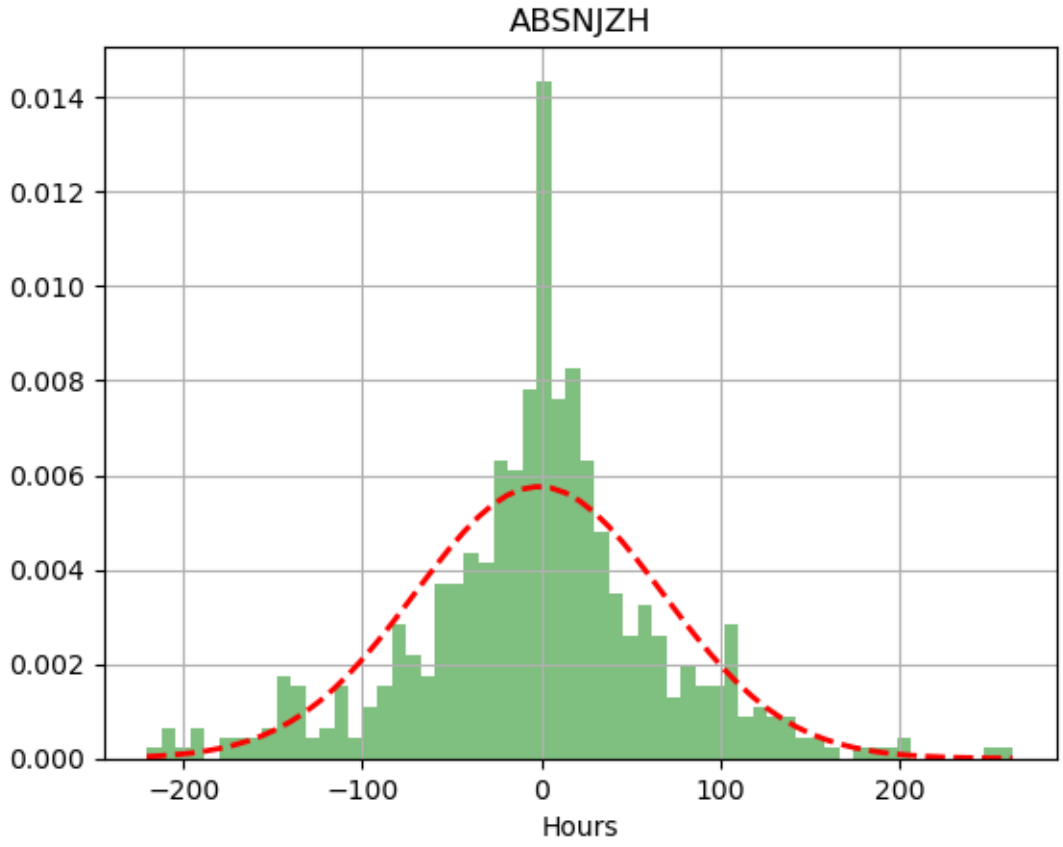
Latar belakang di balik munculnya analisis jarak temporal adalah untuk menginvestigasi rumusan masalah pertama, "**Apakah terdapat suatu proses di parameter magnetik tertentu yang selalu mendahului terjadinya suar surya?"**. Di sini, proses pra-*flare* diasumsikan sebagai kenaikan yang tampak pada parameter magnetik. Apabila ditemukan proses kenaikan ini yang selalu konsisten terjadi sebelum suar surya pada parameter tertentu, kita dapat menyimpulkan bahwa parameter tersebut berkorelasi kuat dengan proses pra-*flare*. Dengan melihat puncak parameter pada daerah aktif dengan suar surya, kita dapat menentukan sampai kapan proses kenaikan tersebut terjadi, dengan asumsi proses kenaikan dominan terjadi hingga puncak dan setelah puncak evolusi parameter magnetik didominasi oleh proses penurunan.

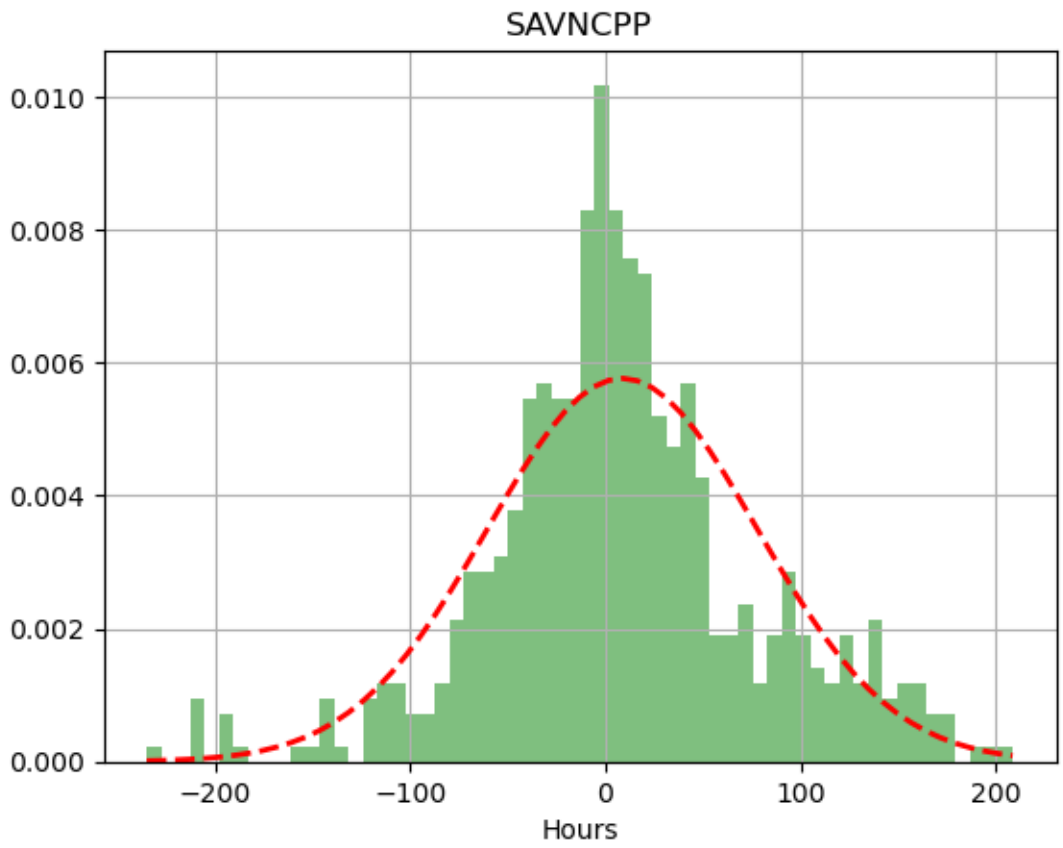
Adapun histogram distribusi jarak temporal suar surya terhadap puncak setiap parameter magnetik ditunjukkan oleh gambar V.1 dengan properti statistiknya ditunjukkan pada tabel V.1.

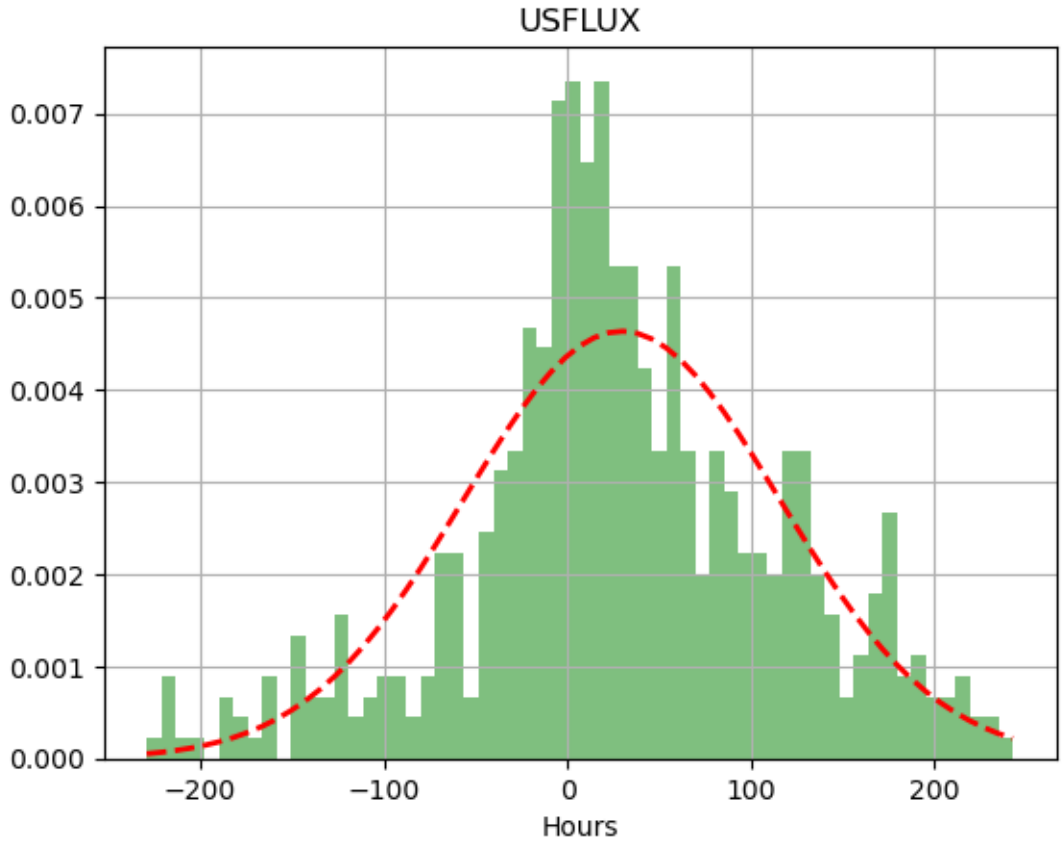

TOTUSJH
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−200
−100
0
100
200
Hours

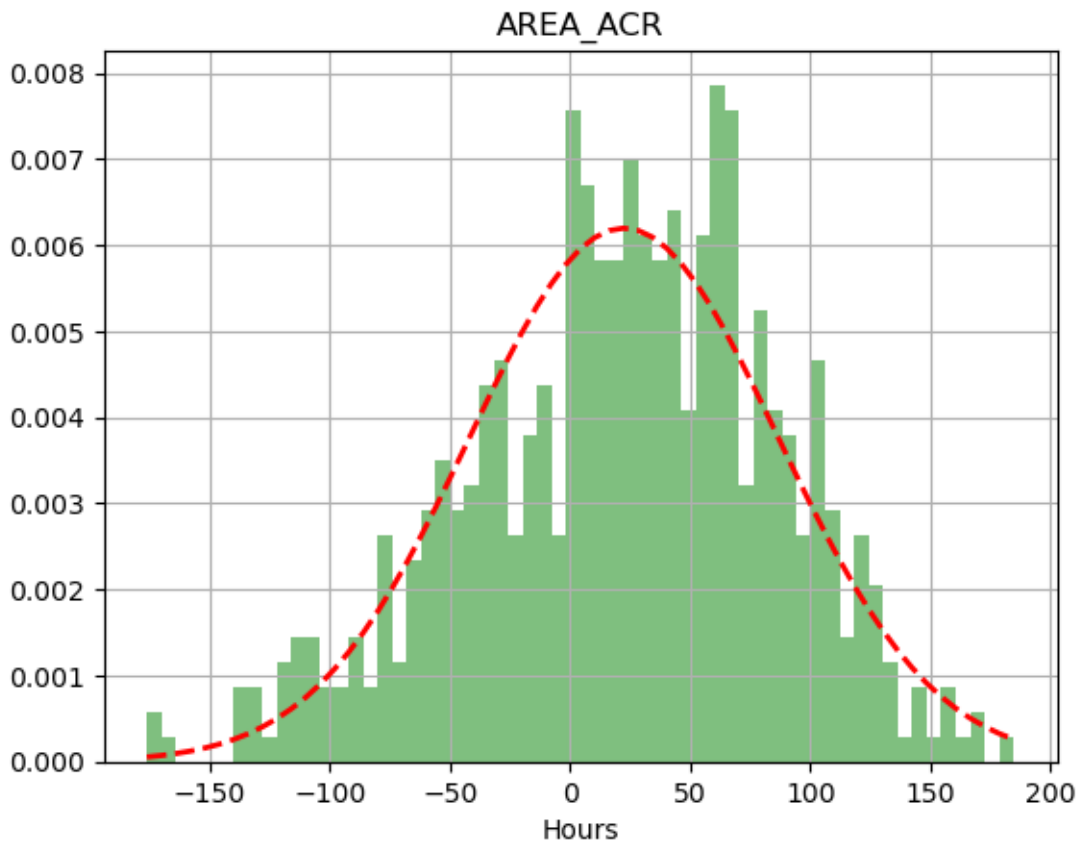

TOTPOT
0.010
0.008
0.006
0.004
0.002
0.000
−200
−100
0
100
200
Hours

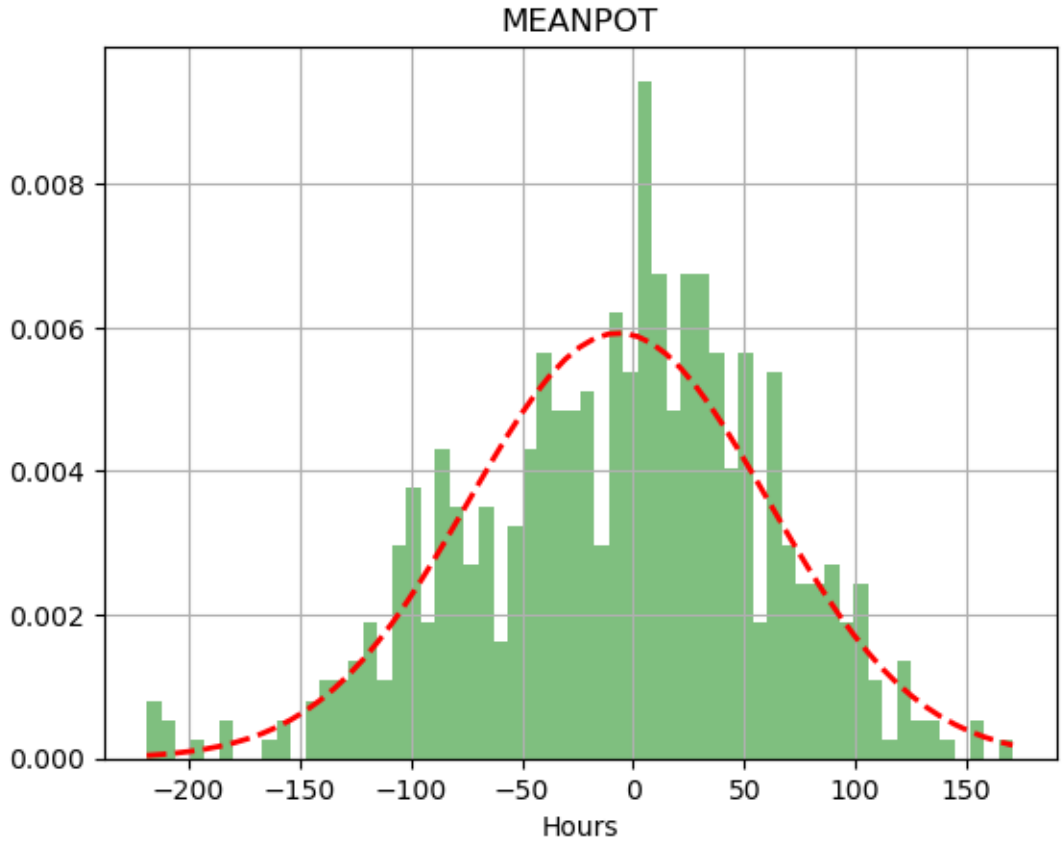

TOTUSJZ
0.010
0.008
0.006
0.004
0.002
0.000
−200
−100
0
100
200
Hours

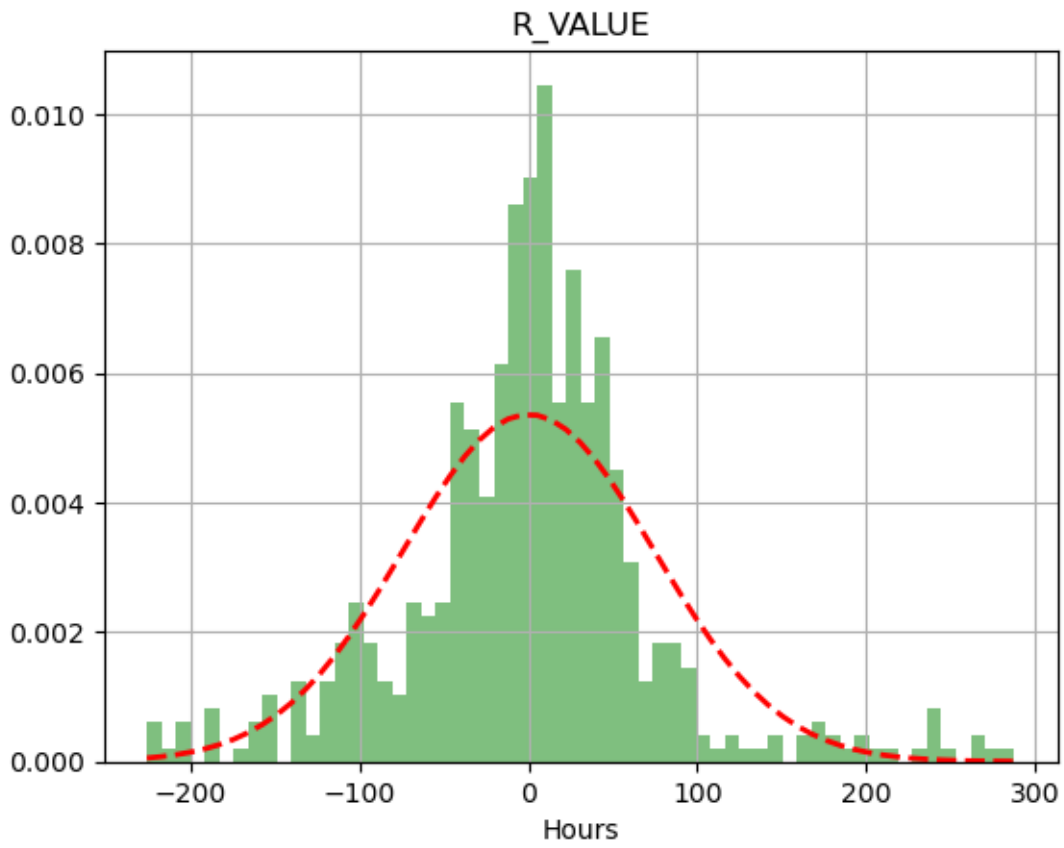

ABSNJZH
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−200
−100
0
100
200
Hours

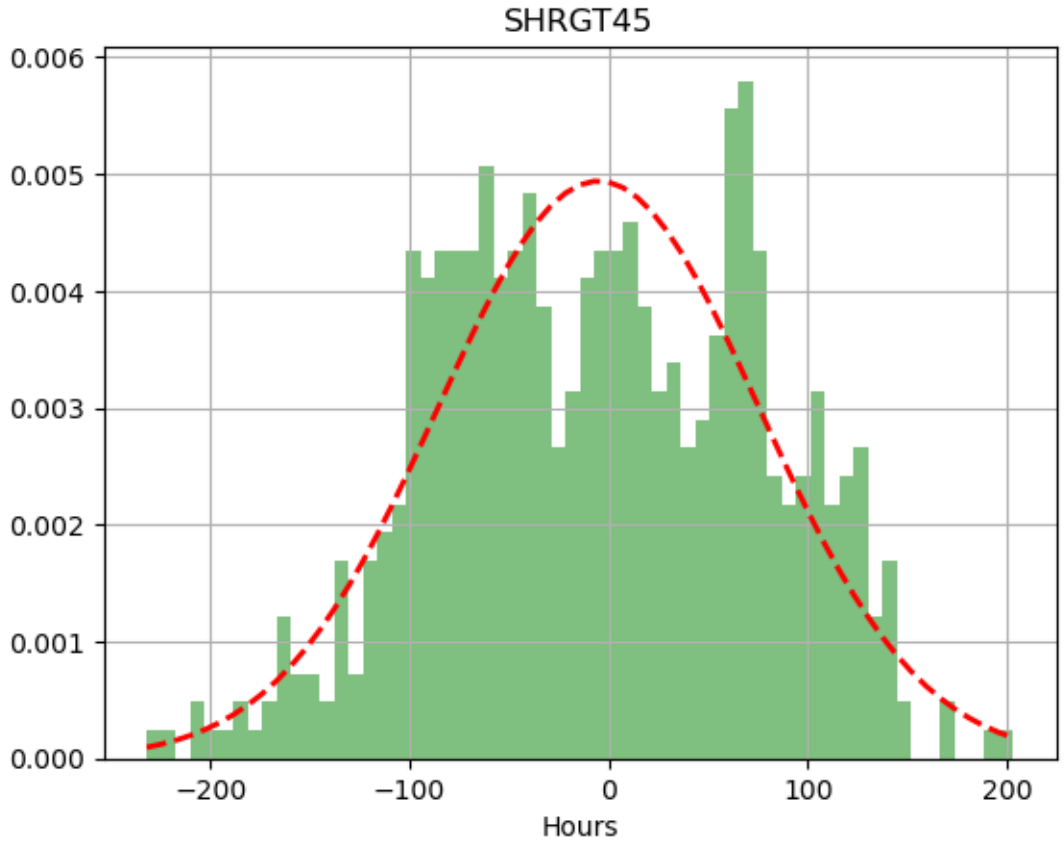

SAVNCPP
0.010
0.008
0.006
0.004
0.002
0.000
−200
−100
0
100
200
Hours

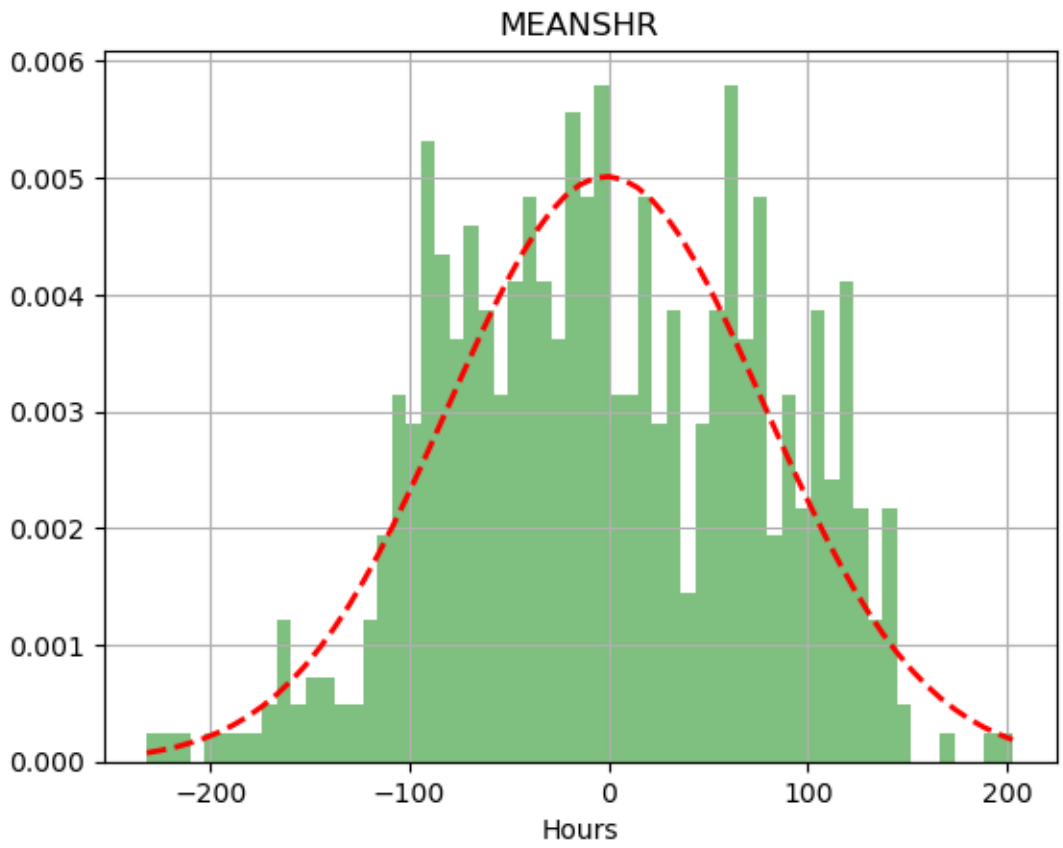

USFLUX
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−200
−100
0
100
200
Hours

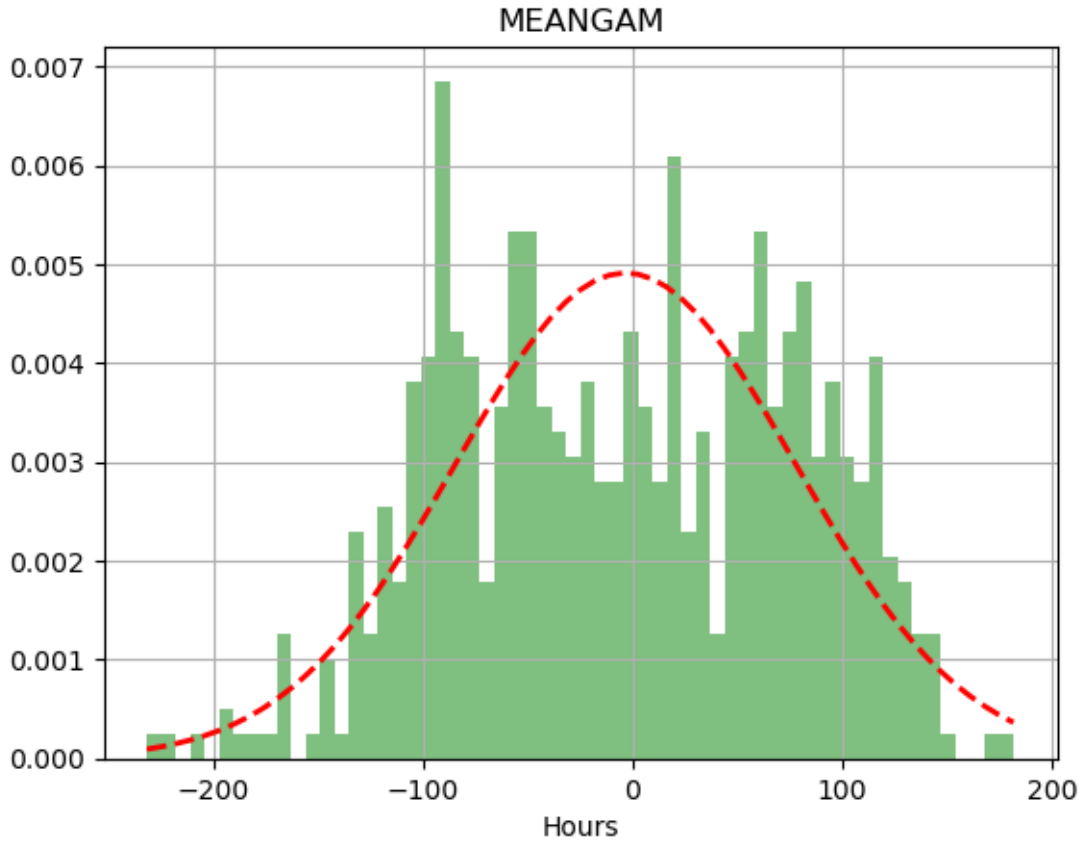

AREA_ACR
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−150
−100
−50
0
50
100
150
200
Hours

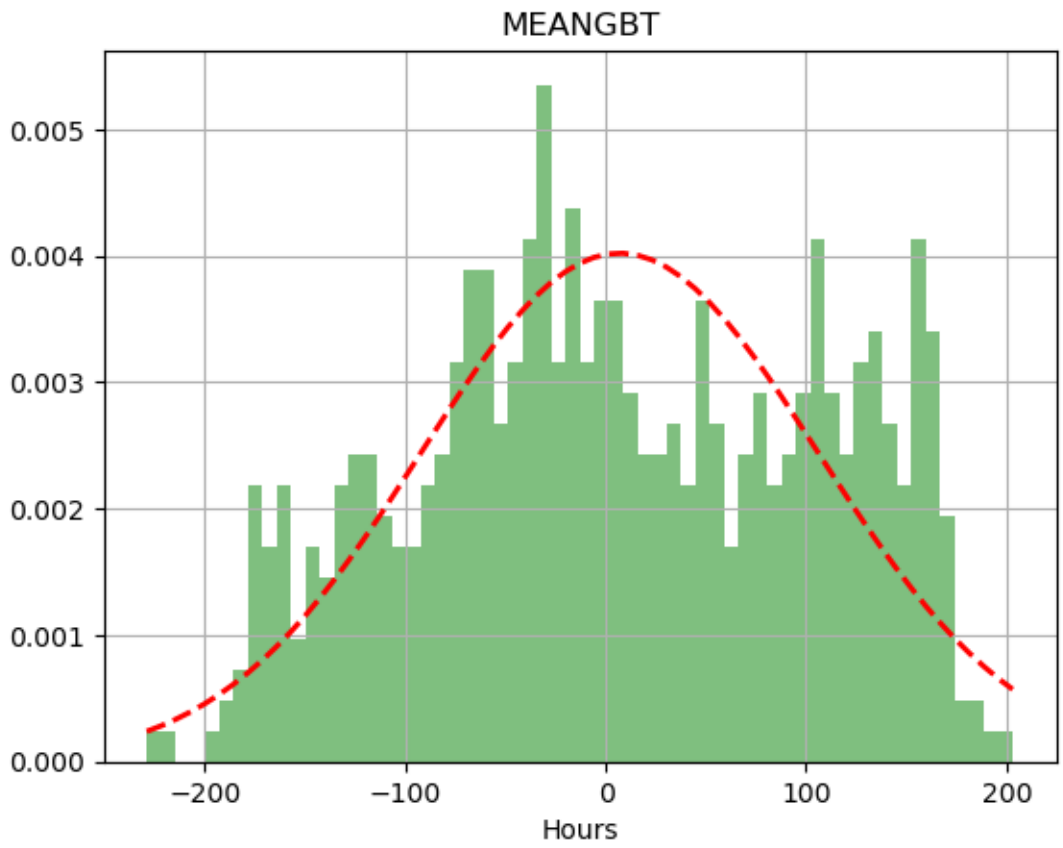

MEANPOT
0.008
0.006
0.004
0.002
0.000
−200
−150
−100
−50
0
50
100
150
Hours

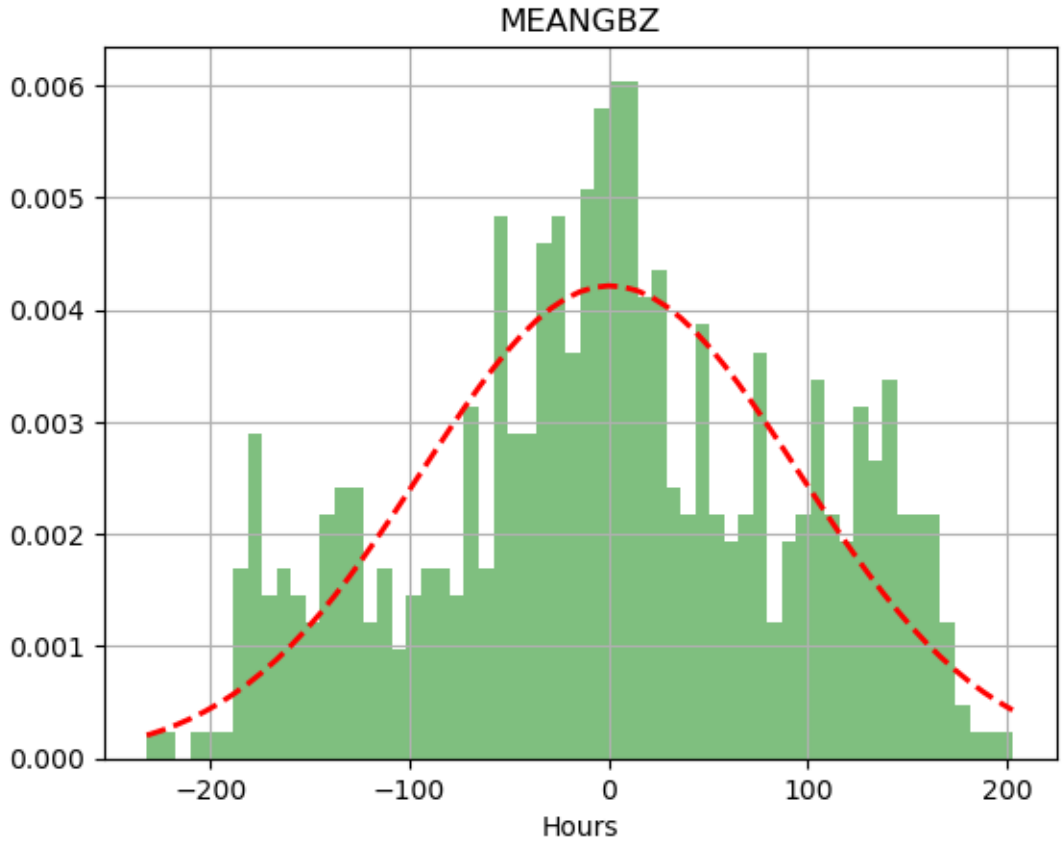

R_VALUE
0.010
0.008
0.006
0.004
0.002
0.000
−200
−100
0
100
200
300
Hours

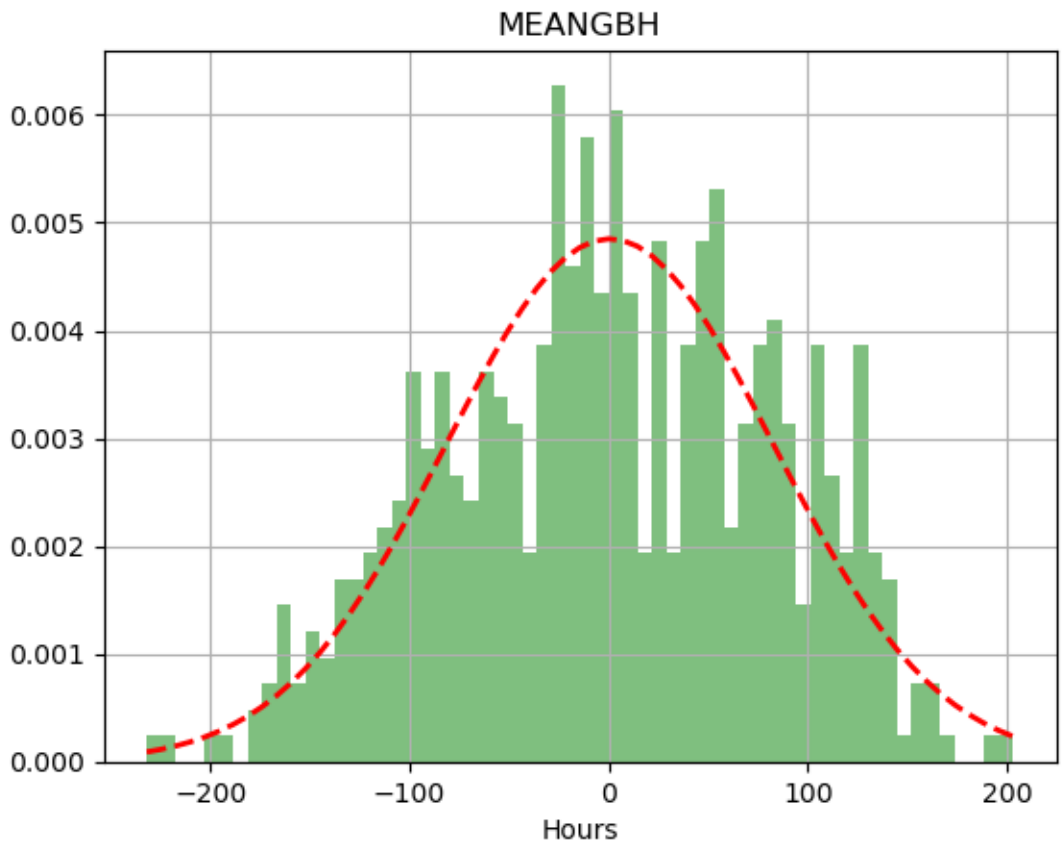

SHRGT45
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−200
−100
0
100
200
Hours

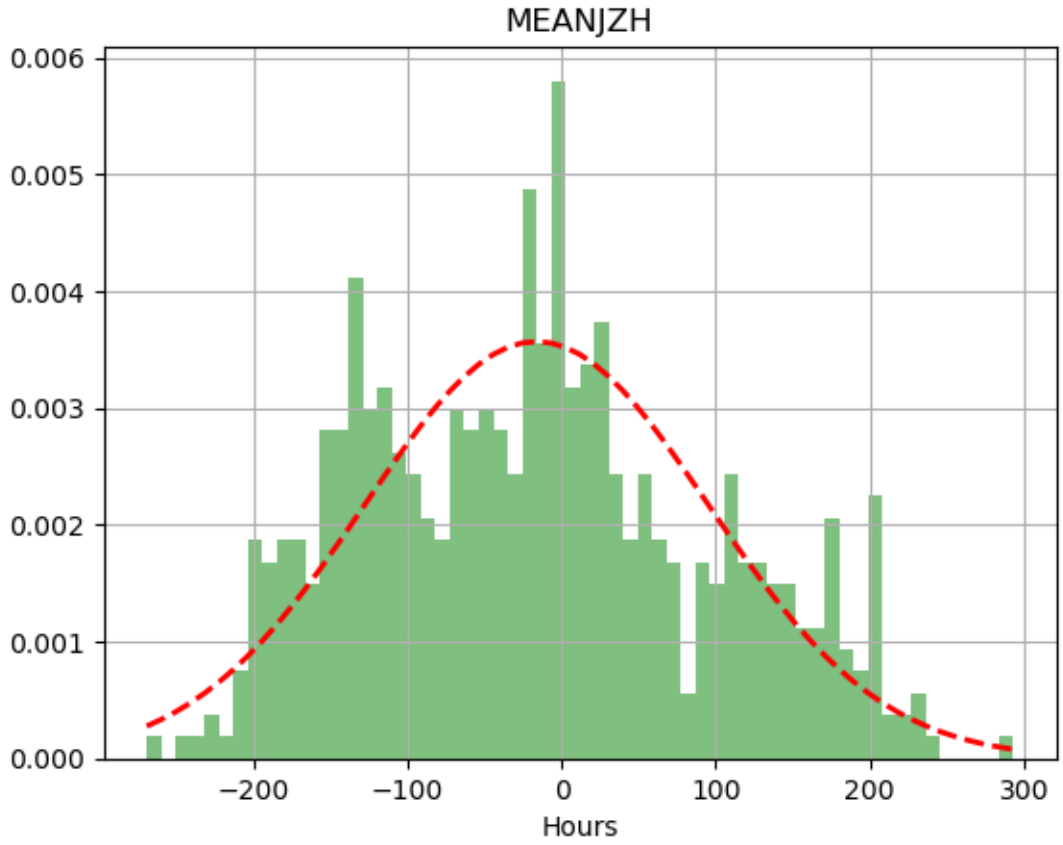

MEANSHR
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−200
−100
0
100
200
Hours

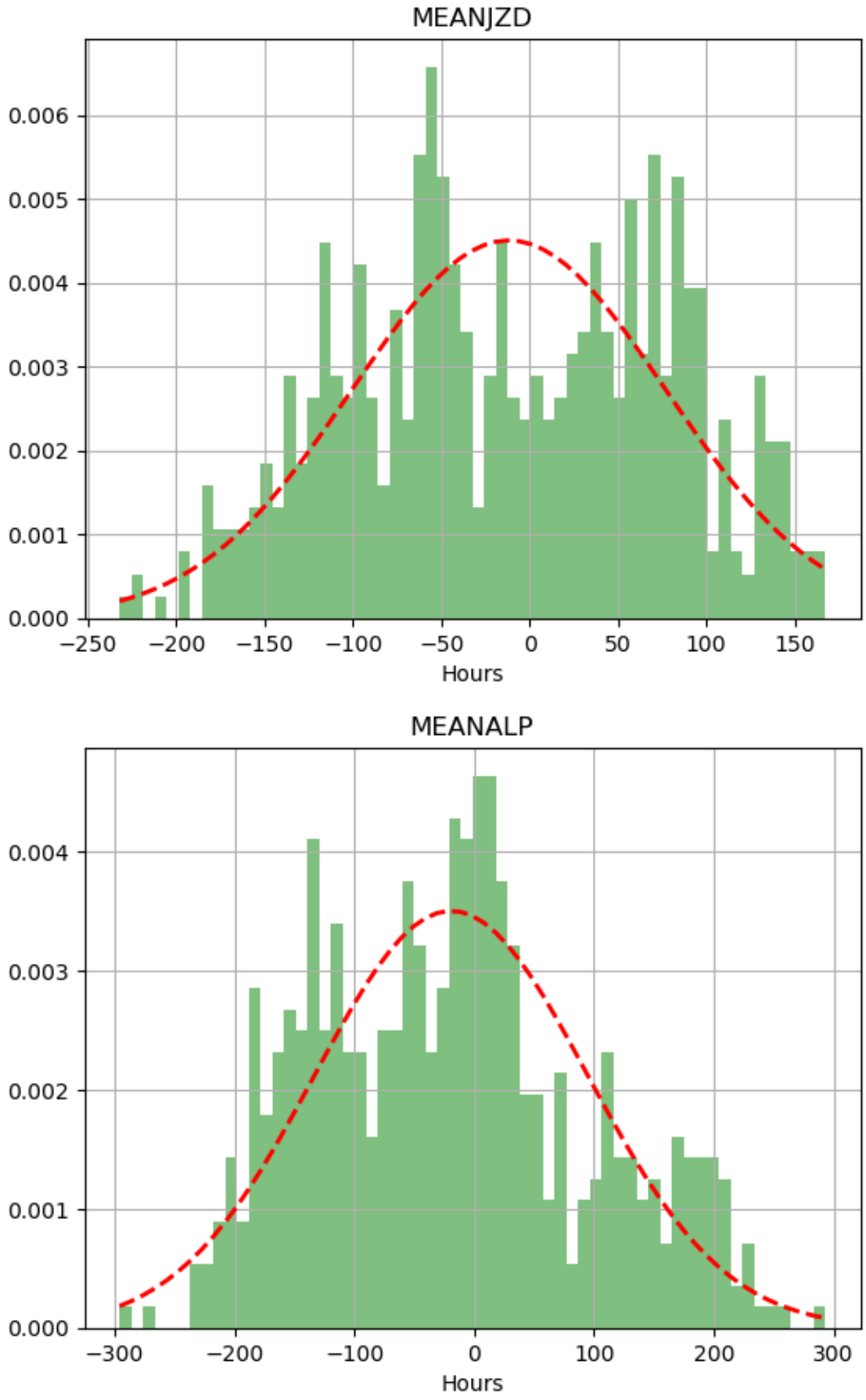

MEANGAM
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−200
−100
0
100
200
Hours


MEANGBT
0.005
0.004
0.003
0.002
0.001
0.000
−200
−100
0
100
200
Hours


MEANGBZ
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−200
−100
0
100
200
Hours


MEANGBH
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−200
−100
0
100
200
Hours


MEANJZH
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−200
−100
0
100
200
300
Hours



Gambar V.1. Histogram distribusi jarak temporal suar surya terhadap puncak parameter. Nilai sumbu x positif menunjukkan suar surya terjadi sebelum puncak parameter. Garis merah

putus-putus merupakan *fit gaussian* terhadap distribusi.

Tabel V.1. Properti statistik distribusi jarak temporal setiap parameter magnetik yang diperoleh dari *fit gaussian*.

Parameter	Rata-rata (jam)	Standar deviasi (jam)	Persentase flare sebelum puncak
TOTPOT	-1,07	64,92	55,52
ABSNJZH	-1,55	69,31	51,66
SAVNCPP	8,63	69,18	55,16
TOTUSJH	2,42	63,31	55,87
TOTUSJZ	11,71	68,38	58,32
USFLUX	29,54	86,02	65,84
AREA_ACR	22,41	64,37	67,42
MEANPOT	-6,6	67,34	51,13
SHRGT45	-5,04	80,74	47,81
MEANSHR	-1,03	79,6	46,93
MEANGAM	-3,44	81,2	49,03
R_VALUE	0,19	74,43	53,23
MEANJZH	-16,19	111,78	41,85
MEANGBT	6,99	99,2	50,08
MEANALP	-19,25	113,78	42,03
MEANGBH	0,59	82,24	50,26
MEANGBZ	0,71	94,7	50,61
MEANJZD	-11,79	88,56	45,88

Distribusi jarak temporal ini menunjukkan hasil yang menarik terhadap pertanyaan rumusan masalah pertama. Dari sini, kita dapat melihat bahwa kebanyakan distribusi jarak temporal pada setiap parameter magnetik menyerupai distribusi gaussian dengan nilai rata-rata yang mendekati nol. Hal ini menandakan bahwa kebanyakan suar surya terjadi di sekitar puncak parameter suar surya. Temuan ini mendukung asumsi bahwa pelepasan energi pada suar surya cenderung terjadi pada puncak energi magnetiknya yang tercermin pada puncak energi parameter magnetik.

Kita juga dapat melihat bahwa standar deviasi distribusi jarak temporal yang lebar (> 63 jam). Nilai standar deviasi yang tinggi dengan nilai rata-rata mendekati nol di setiap parameter magnetik memberikan wawasan bahwa tidak ada parameter magnetik yang konsisten selalu puncak sebelum suar surya. Temuan ini menjawab rumusan masalah pertama, bahwa tidak ada suatu proses di parameter magnetik tertentu yang selalu mendahului terjadinya suar surya. Artinya proses pra-flare adalah proses yang unik untuk setiap suar surya. Proses pra-*flare* muncul pada kombinasi parameter magnetik yang berbeda untuk setiap suar surya.

Di sisi lain, jika memang terdapat suatu proses tertentu yang selalu atau konsisten terjadi sebelum suar surya, seharusnya didapati distribusi jarak temporal yang timpang ke arah positif dengan nilai rata-rata yang sebanding dengan standar deviasinya. Distribusi yang seperti

ini menandakan bahwa kebanyakan suar surya terjadi ketika proses pra-*flare* atau memuncaknya parameter tersebut. Dengan ditemukannya parameter yang konsisten sebagai penanda proses pra-*flare*, kita dapat melakukan seleksi fitur terhadap hasil analisis ini. Tidak tampaknya distribusi yang demikian pada parameter magentik mana pun menguatkan simpulan bahwa proses pra-*flare* merupakan proses yang unik untuk setiap suar surya.

Meskipun demikian, data menyatakan bahwa sebagian besar distribusi parameter magnetik menunjukkan nilai rata-rata yang positif dan >50% suar surya terjadi sebelum parameter magentik memuncak. Hal ini dapat mengisyaratkan bahwa lebih banyak proses pra-*flare* yang terjadi sebelum puncak tertinggi parameter magentik tersebut. Namun data juga menunjukkan bahwa sekitar 40% suar surya juga terjadi setelah puncak tertinggi parameter. Hal ini dapat mengisyaratkan bahwa proses pra-*flare* tidak selalu terjadi setelah puncak tertinggi parameter magnetik tersebut. Asumsi bahwa tidak ada proses pra-*flare* setelah puncak tertinggi parameter tentunya membatasi kasus dimana memang terjadi proses kenaikan parameter atau pra-*flare* setelah puncak tertingginya. Atau suar surya dapat terjadi meskipun parameter magnetik tidak memuncak akibat tercapainya suatu kondisi batas tertentu (energi, kombinasi parameter, dsb.). Untuk itu, diperlukan penelitian lebih lanjut untuk menginvestigasi mengenai kemungkinan ini.

V.2. Peringkat Parameter Magnetik Berdasarkan PCA

Relevansi setiap parameter magnetik terhadap keseluruhan struktur data diinvestigasi melalui metode PCA. Dengan menggunakan PCA, kita dapat melihat peran setiap fitur dalam variasi total data. Paradigma yang mendasari analisis PCA adalah fitur dengan variasi yang tinggi lebih informatif daripada fitur dengan variasi yang rendah sehingga fitur dengan variasi tinggi memberikan kontribusi variasi yang lebih tinggi terhadap data keseluruhan.

Hasil analisis PCA untuk kedelapanbelas parameter magnetik yang digunakan pada penelitian ini dapat dilihat pada tabel IV.2, tabel IV.3, dan gambar IV.10. Pada tabel IV.2, kita dapat melihat parameter dengan *loading score* paling dominan pada setiap komponen utama. Meskipun PCA memberikan komponen utama pertama sebagai komponen utama dengan kontribusi variasi terbesar terhadap data, diikuti oleh komponen utama kedua, ketiga, dst., pemeringkatan berdasarkan *loading score* tertinggi pada setiap komponen utama kurang merepresentasikan kontribusi setiap parameter terhadap variasi keseluruhan data. Hal ini dikarenakan setiap komponen utama tersusun oleh banyak atau bahkan seluruh parameter magnetik dengan kontribusi yang berbeda. Oleh karena itu, penulis memberikan skor terhadap setiap parameter magnetik berdasarkan *loading score* dan presentase variasi komponen utama terhadap variasi total data (gambar IV.10). Dengan menjumlahkan perkalian *loading score* dan presentase variasi komponen utama setiap parameter magnetik, diperolehlah skor yang ditampilkan pada tabel IV.3.

Tabel IV.3 menunjukkan bahwa setiap parameter magnetik memiliki skor yang kompetitif. Hal ini menunjukkan bahwa setiap parameter magnetik memiliki kontribusi terhadap variasi keseluruhan data yang hampir sama. Temuan ini mengisyaratkan bahwa sulit dilakukan pemilihan filter hanya dari nilai skor atau kontribusi variasi parameter magnetik terhadap variasi keseluruhan data. Oleh karena itu, penulis tidak menggunakan pemeringkatan berdasarkan PCA sebagai metode seleksi fitur.

Alasan lain tidak digunakannya PCA sebagai metode seleksi fitur adalah asumsi dasar PCA itu sendiri yaitu fitur yang informatif adalah fitur dengan nilai variasi yang tinggi. Sedangkan untuk kasus suar surya sendiri, fitur dengan variasi yang tinggi belum tentu berkorelasi kuat dengan kemunculan suar surya. Fitur yang memiliki korelasi kuat dengan suar surya adalah fitur yang menunjukkan perilaku tertentu ketika proses pra-*flare*. Dengan menyeleksi fitur yang demikian, model prediksi dapat disusun dengan lebih optimal dan efisien. Penulis menilai bahwa penerapan PCA dapat memberikan wawasan yang lebih luas apabila data dikelompokkan terlebih dahulu sebelum pelaksanaan PCA. Dengan mengelompokkan data dari daerah aktif dengan suar surya saja dan daerah aktif tanpa suar surya, mungkin hal ini dapat memberikan wawasan baru yang inspiratif.

Namun demikian, analisis PCA tetap memberikan wawasan mengenai jumlah fitur yang cukup untuk merepresentasikan seluruh data. Berdasarkan gambar IV.10, nilai variasi keseluruhan data dapat direpresentasikan hingga >90% dengan 5 komponen utama dan >95% dengan 6 komponen utama. Sedangkan peningkatan jumlah fitur >6 tidak memberikan kontribusi yang signifikan terhadap keseluruhan struktur data. Hal ini dapat mengartikan bahwa berdasarkan strukturnya, dataset yang digunakan dapat direpresentasikan oleh 6 fitur dengan tingkat representasi ~95%. Namun penulis merasa simpulan ini perlu diteliti lebih lanjut, terutama dari balik proses PCA yang dilakukan.

V.3. Perbandingan Distribusi Korelasi Spektrum Daerah Aktif

Dengan pengurangan fitur melalui metode PCA dinilai kurang dapat merepresentasikan fitur dengan korelasi yang kuat terhadap kemunculan suar surya, penulis meninjau distribusi korelasi spektrum pada daerah aktif. Pada penelitian ini, korelasi spektrum merepresentasikan tinggi dan fluktuasi yang terjadi pada evolusi parameter magnetik. Dengan membandingkan distribusi korelasi spektrum pada daerah aktif dengan suar surya terhadap distribusi korelasi spektrum proses latar belakang, kita dapat mendapatkan wawasan mengenai kenampakan proses pra-*flare* yang tampak sebagai fluktuasi pada evolusi daerah aktif relatif terhadap proses latar belakang. Dengan demikian, kita dapat memperoleh jawaban untuk rumusan masalah kedua, "**Bagaimana kenampakan proses pra-flare pada setiap parameter magnetik?**".

Proses latar belakang di sini merupakan proses yang terjadi pada evolusi parameter

magnetik daerah aktif tanpa suar surya. Meskipun demikian, pada penelitian ini, suar surya yang ditinjau hanya suar surya kelas tinggi yaitu kelas M dan X. Oleh karena itu, proses latar belakang belum tentu dapat dikatakan sebagai bukan proses pra-*flare*. Boleh jadi proses latar belakang merupakan proses pra-*flare* untuk suar surya dengan kelas yang lebih rendah. Namun demikian, penulis berpendapat bahwa proses pra-*flare* suar surya kelas yang lebih rendah tidak sedominan proses pra-*flare* untuk suar surya kelas tinggi. Oleh karena itu, pembagian proses dalam evolusi daerah aktif menjadi proses latar belakang dan proses pra-*flare* yang dilakukan dapat dijustifikasi.

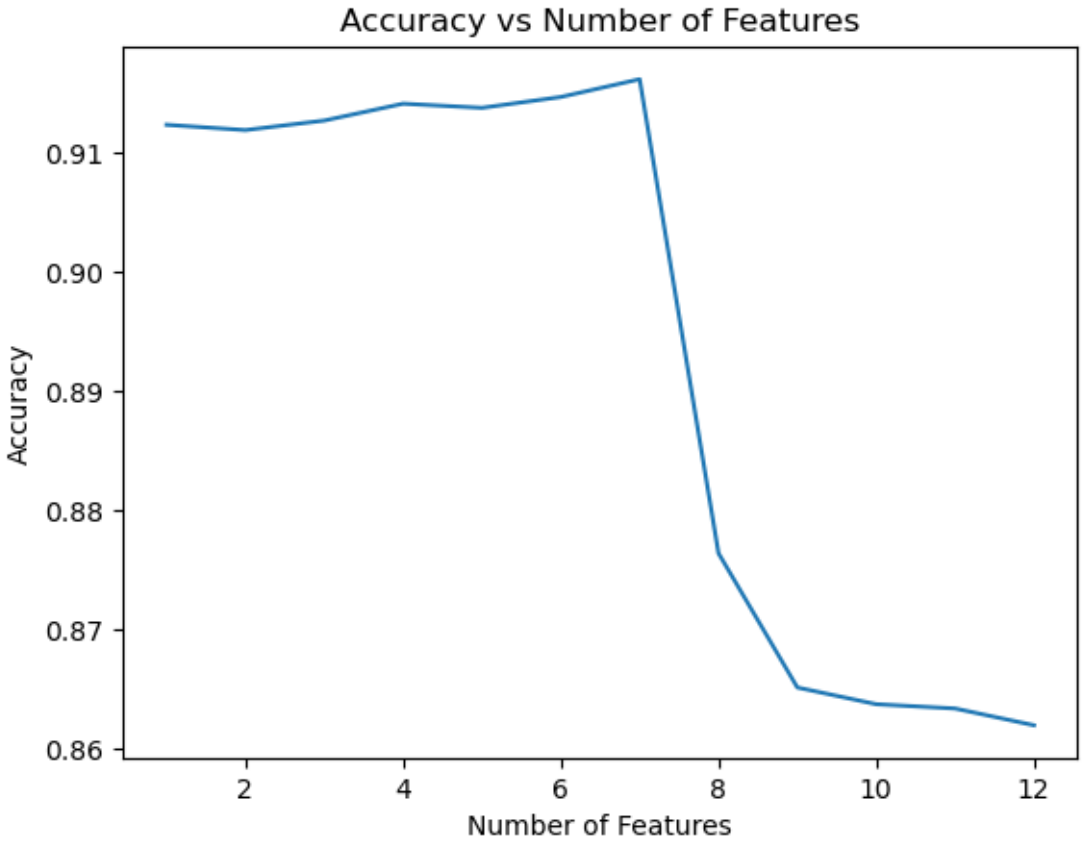
Hasil dari analisis ini ditunjukkan oleh tabel IV.4 dan tabel IV.5. Tabel IV.4 menunjukkan statistik dari distribusi proses pra-*flare* dan proses latar belakang untuk setiap parameter. Dari tabel ini, kita dapat melihat bahwa terdapat parameter yang menunjukkan kontras antara proses pra-flare dengan proses latar belakang. Hal ini tampak terutama pada 12 parameter pertama yang tersusun dalam tabel IV.5. Keduabelas parameter tersebut menunjukkan kontras dari kedua proses dalam rasio rata-rata kedua distribusi (rasio rata-rata proses pra-*flare* terhadap proses latar belakang). Keduabelas parameter tersebut memberikan nilai rasio rata-rata > 1. Hal ini menunjukkan bahwa proses pra-*flare* tampak lebih tinggi pada evolusi parameter magnetik daripada proses latar belakang pada parameter-parameter tersebut. Sedangkan untuk parameter sisanya memiliki rasio rata-rata ~1 yang menandakan bahwa proses pra-*flare* tidak dapat dibedakan tingginya terhadap proses latar belakang. Artinya pada parameter-parameter ini, proses pra-*flare* tidak tampak.

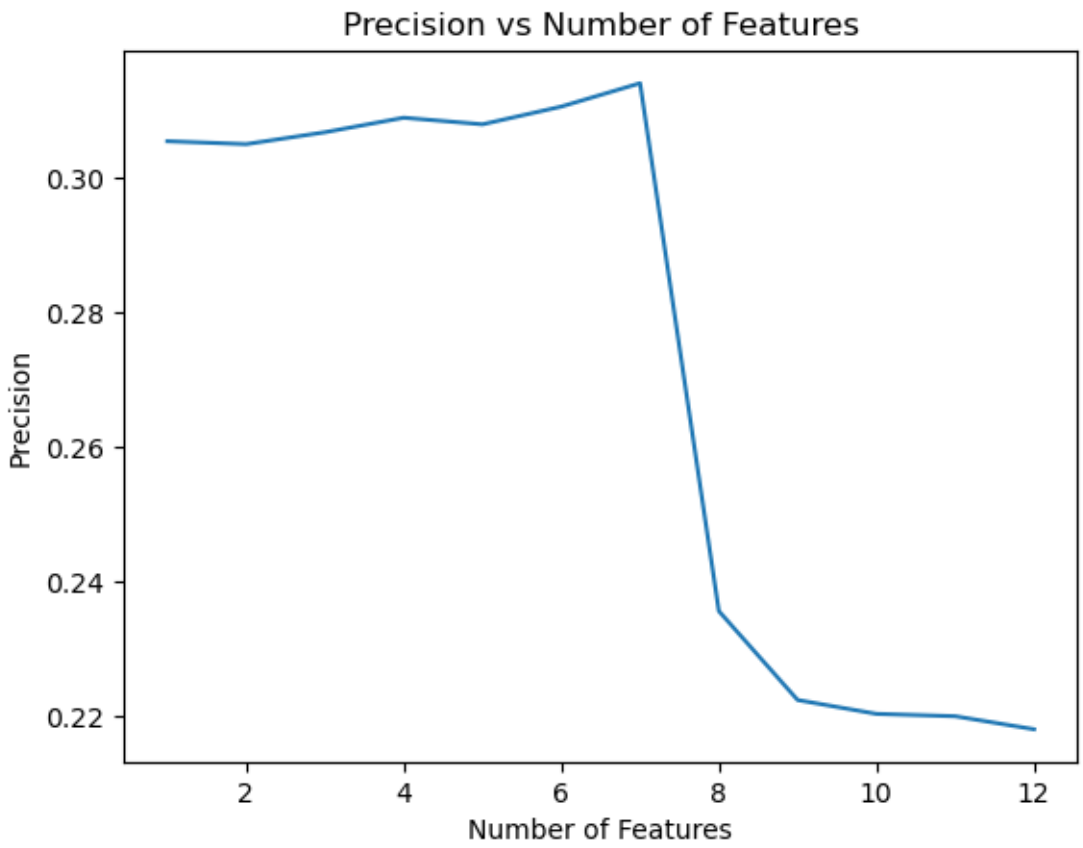
Selain dilihat dari rata-rata distribusinya, kita juga dapat melihat dari standar deviasi distribusi korelasi spektrum di setiap parameter. Tampak bahwa dua belas parameter pertama pada tabel IV.5 menunjukkan nilai standar deviasi yang tinggi relatif terhadap proses latar belakangnya. Hal ini juga menandakan fluktuasi yang dominan terlihat pada proses pra-*flare* dibandingkan proses latar belakang. Oleh karena itu, dalam meninjau parameter sebagai fitur dan penentuan nilai batas untuk pelabelan, dipilihlah dua belas parameter ini karena terlihatnya proses pra-*flare*.

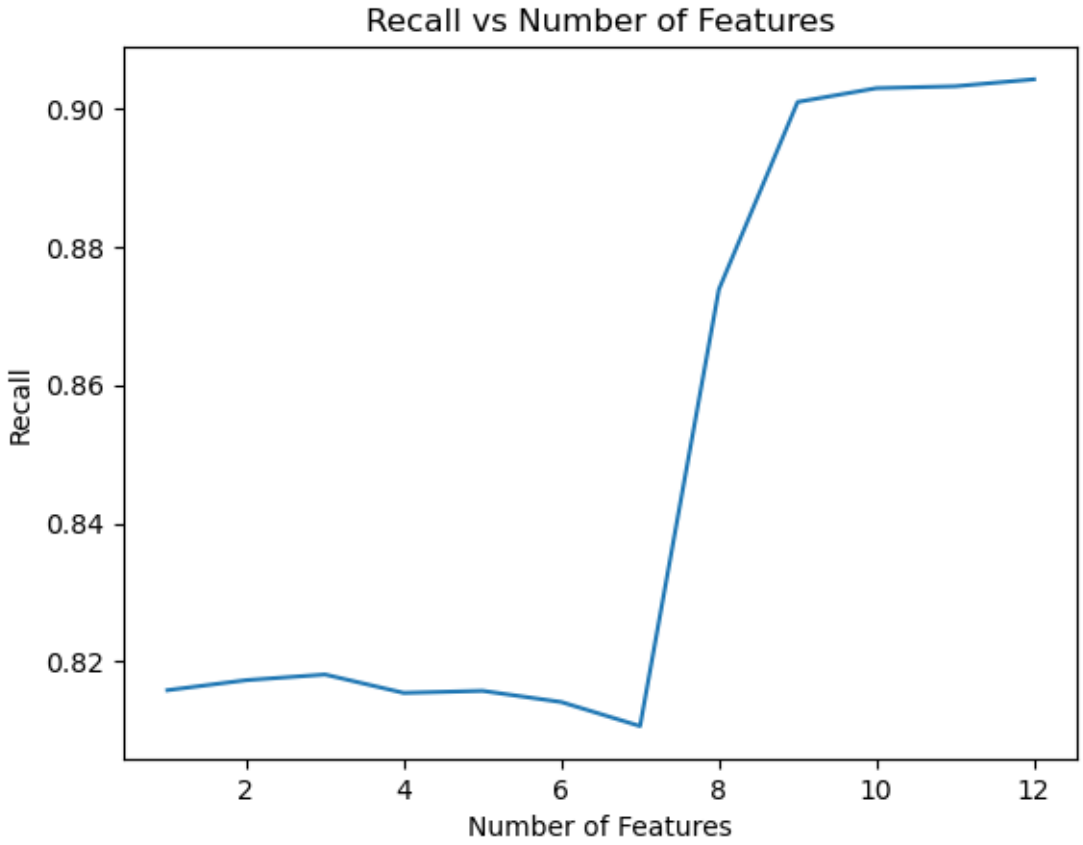
Keduabelas parameter magnetik yang menunjukkan proses pra-*flare* pada evolusinya didominasi oleh parameter total atau nilai integrasi parameter pada seluruh daerah aktif seperti TOTPOT, TOTUSJH, dan TOTUSJZ. Di sisi lain, parameter-parameter yang menunjukkan nilai rata-rata parameter magnetik pada daerah aktif seperti MEANJZD, MEANGBZ, dan MEANGBH bukan merupakan parameter yang menunjukkan korelasi dengan proses pra-*flare*. Artinya, nilai rata-rata parameter magnetik pada daerah aktif kurang dapat merepresentasikan evolusi daerah aktif tersebut dalam proses pra-*flare*. Penulis berpendapat bahwa hal ini berkaitan dengan suar surya yang merupakan peristiwa lokal bahkan dalam daerah aktif sekalipun. Temuan ini dapat memberikan wawasan baru pada paradigma pemilihan parameter yang berkorelasi dengan proses pra-*flare*.

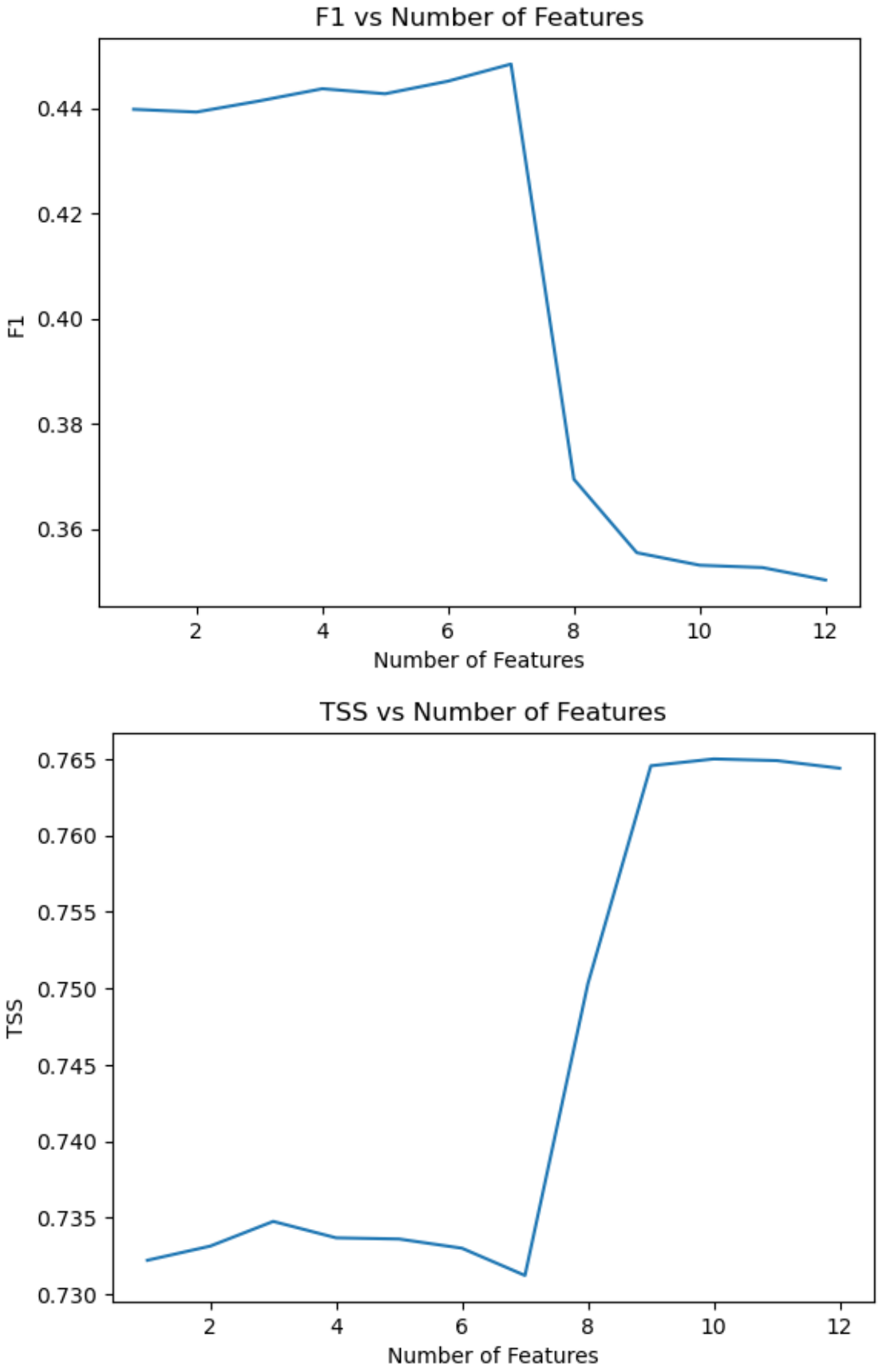
V.4. Pengaruh Pengurangan Fitur terhadap Performa *Support Vector Machine*

Dengan dataset yang telah dilabeli, penulis mencoba menjalankan model prediksi *support vector machine* menggunakan *hyperparameter* C yang telah dioptimasi. Penjalanan model prediksi ini dilakukan sebanyak dua belas kali dengan setiap iterasi menggunakan fitur yang lebih sedikit. Pengurangan fitur setiap iterasi dilakukan berdasarkan peringkat pada tabel IV.5 dengan setiap iterasi mengurangi parameter dengan peringkat yang paling bawah. Performa pembelajaran mesin pada setiap iterasi ditampilkan dalam lima metrik skor akurasi, presisi, recall, f1, dan TSS, ditunjukka oleh gambar V.2.




Precision vs Number of Features
Precision
0.30
0.28
0.26
0.24
0.22
2
4
6
8
10
12
Number of Features


Recall vs Number of Features
Recall
0.90
0.88
0.86
0.84
0.82
2
4
6
8
10
12
Number of Features



Gambar V.2. Plot metrik skor yang menunjukkan performa model prediksi terhadap pengurangan jumlah fitur.

Dari gambar V.2, kita dapat melihat bahwa terdapat perubahan metrik skor terhadap pengurangan jumlah fitur. Ada metrik yang menunjukkan peningkatan, ada pula yang menunjukkan penurunan terhadap pengurangan jumlah fitur. Jika dilihat lebih detail, metrik skor terhadap jumlah fitur tampak terkopel satu sama lain terutama pada jumlah fitur sama dengan 7, 8, dan 9. Di satu sisi, akurasi, presisi, dan f1 menunjukkan perilaku yang mirip satu sama lain yaitu kenaikan metrik terhadap pengurangan jumlah fitur. Di sisi lain, TSS dan recall menunjukkan penurunan metrik terhadap pengurangan fitur. Penurunan atau kenaikan pada pemilihan jumlah fitur sama dengan 7, 8, dan 9 mencerminkan penurunan atau kenaikan kontras rasio rata-rata proses pra-flare terhadap proses latar belakang pada parameter dengan peringkat ke-7, 8, dan 9 (tabel IV.5).

Secara absolut, rentang kenaikan atau penurunan di setiap metrik tampak tidak signifikan. Tidak ada perubahan berupa kenaikan atau penurunan metrik sebesar > 0,1. Namun secara relatif, perubahan ini cukup signifikan di beberapa metrik. Terhadap penurunan jumlah fitur, akurasi mengalami kenaikan sekitar 5,8%, presisi mengalami kenaikan sekitar 40%, recall mengalami penurunan sekitar 8,8%, f1 mengalami kenaikan sekitar 25,7%, dan TSS mengalami penurunan sekitar 4,5%. Pada kasus ini, presisi dan f1 tampak meningkat secara signifikan. Hal ini wajar dikarenakan f1 dan presisi adalah metrik yang terikat satu sama lain (persamaan II.9). Meskipun f1 juga terikat dengan recall, karena kenaikan presisi jauh lebih signifikan maka nilai f1 juga ikut naik.

Pertanyaan selanjutnya adalah bagaimana cara kita mengurangi fitur berdasarkan metrik-metrik skor ini. Hal yang pasti adalah jumlah fitur yang optimal berada pada titik belok kenaikan atau penurunan metrik terhadap jumlah fitur (jumlah fitur = 7 atau 9) karena pada ekstremanya, performa model prediksi tidak berbeda jauh dengan nilai ini. Untuk menjawab pertanyaan ini, mari kita tinjau dua metrik yang saling berkebalikan responnya terhadap pengurangan fitur yaitu presisi dan *recall*.

Presisi merupakan keandalan model dalam model dalam memprediksi positif. Nilainya akan semakin rendah jika model banyak salah mengklasifikasi titik data negatif ke positif (FP) (persamaan II.7). Recall juga menunjukkan seberapa andal model dalam meprediksi positif. Namun perbedaannya dengan presisi adalah nilai recall akan semakin rendah apabila model banyak salah mengklasifikasikan titik data positif ke negatif (FN) (persamaan II.8). Lantas, metrik manakah yang lebih diutamakan? Jawabannya bergantung pada kasus yang ditinjau oleh model prediksi. Contoh model yang mementingkan minimalnya positif salah adalah model identifikasi nasabah dengan skor kredit baik. Pada kasus ini, pihak bank tentunya tidak ingin nasabah pengaju utang memiliki reputasi buruk atau penunggak sehingga prediksi dibobotkan untuk tidak memberikan positif salah. Sedangkan contoh model yang mementingkan minimalnya negatif salah adalah model identifikasi penyakit jantung. Pada kasus ini, dokter pendiagnosa tidak ingin ada pasien penyakit jantung yang tidak terdeteksi mengidap penyakit

jantung sehingga bobot untuk meminimalkan negatif salah lebih tinggi.

Untuk kasus dimana kedua metrik ini dianggap penting, kita menggunakan metrik keempat yaitu f1. f1 merupakan rata-rata harmonik antara presisi dan recall. Oleh karena itu, nilai f1 yang baik juga menandakan nilai presisi dan recall yang optimal. Sehingga berdasarkan gambar V.2, dapat dikatakan bahwa jumlah fitur sama dengan 7 adalah jumlah fitur yang optimal pada kasus ini. Tapi jika kita lihat grafik yang menunjukkan hubungan antara TSS dengan jumlah fitur, jumlah fitur sama dengan 7 memberikan nilai TSS paling rendah. Pemilihan jumlah fitur sama dengan 7 tetap dapat dijustifikasi terhadap fakta ini karena jika kita lihat penurunan skor TSS cukup insignifikan (~4,5% secara relatif). Selain itu, nilai TSS juga memiliki rentang yang lebar yaitu -1 hingga 1 sehingga penurunan sebesar 0,03 dinilai kurang signifikan.

Temuan bahwa jumlah fitur optimal minimal sama dengan 7 cukup selaras dengan analisis PCA yang menyebutkan bahwa jumlah fitur lebih dari 6 memberikan representasi data >95%. Namun korelasi ini perlu diinvestigasi lebih lanjut. Hal menarik lainnya yang tampak pada plot metrik skor terhadap pengurangan jumlah fitur adalah performa model prediksi yang tidak jatuh bebas pada jumlah fitur yang sedikit. Pada jumlah fitur sama dengan satu, kita masih memperoleh performa yang sebanding dengan jumlah fitur lebih banyak. Penulis berpendapat bahwa hal ini mungkin terjadi karena fitur dengan peringkat pertama yaitu TOTPOT memiliki rasio proses pra-flare terhadap proses latar belakang yang sangat tinggi (~77) sehingga fitur ini saja cukup untuk melakukan prediksi yang layak. Selain itu, TOTPOT menempati peringkat pertama dalam relevansinya terhadap proses pra-*flare* juga dianggap wajar secara fisis. Hal ini dikarenakan TOTPOT memuat informasi mengenai densitas energi magnetik yang erat kaitannya dengan kemunculan suar surya.

V.5. Perbandingan dengan Referensi

Sub-bab ini memberikan perbandingan performa prediksi ledakan Matahari menggunakan metode *support vector machine* yang dilakukan dalam penelitian ini terhadap referensi utama (Bobra dan Couvidat, 2015) dan penelitian sebelumnya (Fernanda, 2022). Bobra dan Couvidat menggunakan dataset deret waktu dengan rentang Mei 2010 – Mei 2014. Dengan menggunakan sistem pelabelan statis dengan jendela prediksi 24 jam, dataset ini memuat 285 ledakan Matahari kelas M dan 18 ledakan Matahari kelas X. Bobra dan Couvidat juga hanya meninjau ledakan kelas X dan M saja sebagai ledakan Matahari yang signifikan. Bobra dan Couvidat meninjau 13 parameter magnetik daerah aktif sebagai fitur yang digunakan dalam model prediksinya. Daftar parameter ini dapat dilihat pada Tabel V.2.

Tabel V.2: Tabel daftar parameter yang dipakai sebagai fitur pada penelitian Bobra (2015)

Kata Kunci	Deskripsi Fisis	Formula
TOTUSJH	Total helisitas arus tanpa tanda dalam G^2/m	$H_{c_{total}} \propto \sum \lvert B_z \cdot J_z \rvert$
TOTBSQ	Total magnitudo gaya Lorentz	$F \propto \sum B^2$
TOTPOT	Total densitas energi magnetik fotosfer dalam erg/cm^3	$\rho_{tot} \propto \left(\vec{\boldsymbol{B}}^{Obs} - \vec{\boldsymbol{B}}^{Pot}\right)^2 dA$
TOTUSJZ	Total arus vertikal tanpa tanda dalam mA/m^2	$J_{z_{total}} = \sum \lvert J_z \rvert \, dA$
ABSNJZH	Nilai absolut helisitas arus bersih dalam G^2/m	$H_{c_{obs}} \propto \sum \lvert B_z \cdot J_z \rvert$
SAVNCPP	Jumlah nilai absolut dari arus bersih per polaritas dalam Ampere	$J_{z_{sum}} \propto \left\lvert \sum^{B_z^+} J_z dA \right\rvert + \left\lvert \sum^{B_z^-} J_z dA \right\rvert$
USFLUX	Total fluks tanpa tanda dalam Maxwell	$\phi = \sum \lvert B_z \rvert \, dA$
TOTFZ	Jumlah gaya Lorentz pada arah-z	$\partial F_z \propto \sum \left({B_x}^2 + {B_y}^2 - {B_z}^2\right) dA$
AREA_ACR	Area dengan medan kuat dalam piksel	$Area = \sum pixel$
MEANPOT	Rata-rata ekses densitas energi megnetik fotosfer dalam erg/cm^3	$\bar{\rho} \propto \frac{1}{N} \sum \left(\vec{\boldsymbol{B}}^{Obs} - \vec{\boldsymbol{B}}^{Pot}\right)^2$
R_VALUE	Total fluks dekat garis inversi polaritas dalam Maxwell	$\phi = \sum \lvert B_{Los} \rvert \, dA \;\; dalam \; R \; mask$
EPSZ	Jumlah gaya Lorentz ternormalisasi pada arah-z	$\partial F_z \propto \frac{\sum \left({B_x}^2 + {B_y}^2 - {B_z}^2\right)}{\sum B^2}$
SHRGT45	Persentase piksel dengan shear angle > 45^o dalam persen	$\frac{Area \; dengan \; shear > 45°}{Total \; area} \times 100\%$

Sedangkan penelitian sebelumnya melakukan perbandingan performa prediksi terhadap beberapa model *supervised learning*, salah satunya adalah *support vector machine*. Dataset yang digunakan pada penelitian sebelumnya disusun dari data 10 parameter magnetik yang diambil dari 10 Mei 2010 hingga 31 Desember 2020 yang terdiri dari 2.642.196 titik data. Kesepuluh parameter magnetik tersebut merupakan 10 parameter yang sama yang digunakan oleh Bobra dan Couvidat pada tabel V.3 namun tanpa parameter turunan (TOTBSQ, TOTFZ, dan EPSZ).

Dari penelitian sebelumnya, penulis membandingkan terhadap performa support vector machine dengan sistem pelabelan dinamis yang mana merupakan model dengan performa tertinggi. Sedangkan untuk penelitian ini, penulis mengambil performa dari model dengan jumlah fitur sama dengan 7. Perbandingan performa setiap model *support vector machine* beserta rasio labelnya ditunjukkan oleh tabel V.3.

Tabel V.3: Tabel yang menunjukkan perbandingan performa prediksi yang dilakukan oleh model SVM terhadap referensi utama.

	SVM 7 fitur	Pelabelan Dinamis (Fernanda, 2022)	Bobra dan Couvidat, 2015
Rasio negatif/positif	23,75	50,3813	16,5
Akurasi	0,91	0,85	0,96
Presisi	0,31	0,12	0,69
Recall	0,82	0,94	0,63
TSS	0,73	0,79	0,61

Dari tabel V.3, dapat dilihat bahwa performa model prediksi pada penelitian ini menunjukkan peningkatan relatif yang cukup drastis terhadap penelitian sebelumnya. Khususnya pada metrik presisi terjadi peningkatan relatif sebesar 150% tanpa adanya penurunan metrik lain yang signifikan. Di sisi lain, performa penelitian pada beberapa metrik ini masih di bawah Bobra dan Couvidat terutama bagian akurasi dan presisi. Namun seperti yang telah dijelaskan oleh Bobra dan Couvidat (2015) dan penelitian sebelumnya (Fernanda, 2022), kedua metrik ini rentan terhadap ketimpangan kelas sehingga kurang cocok untuk digunakan sebagai representasi performa model. Meskipun demikian, tampak adanya peningkatan di bagian presisi sudah memberikan isyarat bahwa kita sudah berada pada arah yang benar.

Nilai akurasi yang tinggi menunjukkan bahwa model andal dalam mengidentifikasi positif benar dan negatif benar terhadap keseluruhan prediksi. Nilai *recall* yang tinggi menunjukkan bahwa model cukup andal dalam mengidentifikasi positif atau sedikit titik data positif yang teridentifikasi sebagai negatif. Namun nilai presisi yang rendah menunjukkan bahwa model kurang andal dalam megidentifikasi positif benar terhadap keseluruhan tebakan positif. Artinya model terlalu berani dalam memprediksi label positif. Nilai TSS dapat dianggap sebagai nilai atau skor umum performa model prediksi dengan rentang -1 hingga 1. Nilai TSS -1 menandakan model yang selalu salah dalam memprediksi, 0 menandakan prediksi acak, dan 1 menandakan model yang selalu benar dalam memprediksi.

Dari segi performanya saja, penulis berpendapat bahwa model prediksi yang dihasilkan sudah cukup baik. Nilai akurasi >80%, recall > 80%, dan TSS >70% menunjukkan bahwa model prediksi tidak menebak secara acak dan tidak menebak monoton negatif (recall !~0%). Namun kekurangan dari model yang dihasilkan oleh penelitian ini adalah model terlalu percaya

diri dalam memprediksi positif. Penulis berpendapat bahwa sulitnya diperoleh model prediksi suar surya yang andal karena terkendala ketersediaan data yang tidak menyeluruh. Saat ini, data permukaan Matahari yang dapat diekstrak informasinya dengan baik hanya sekitar 1/3 dari keseluruhan permukaan Matahari. Hal ini akan sangat mempengaruhi dalam sistem pelabelan yang berdampak pada ketepatan pengidentifikasian proses pra-*flare*. Oleh karena itu, penulis menyarankan di masa depan untuk membangun proyek pengamatan keseluruhan bola Matahari secara kontinu.

BAB VI

SIMPULAN DAN SARAN

VI.1 Simpulan

- Proses pra-*flare* merupakan proses yang unik pada setiap suar surya. Hal ini dapat dilihat dari distribusi jarak temporal suar surya terhadap puncak tertinggi parameter magnetiknya. Proses yang memicu suar surya atau proses pra-*flare* terjadi sebagai kombinasi unik berbagai parameter magnetik pada daerah aktif.
- Tidak semua parameter magnetik yang ditinjau dalam penelitian ini berkorelasi dengan kemunculan suar surya. Hal ini ditunjukkan pada nilai rasio distribusi korelasi spektrum daerah aktif dengan suar surya terhadap korelasi spektrum latar belakang. Enam parameter magnetik yang tidak menunjukkan adanya perubahan ketika proses pra-*flare* terjadi adalah MEANJZH, MEANGBT, MEANALP, MEANGBH, MEANGBZ, dan MEANJZD. Hasil ini sejalan dengan hasil pemeringkatan parameter yang dilakukan oleh Bobra dan Couvidat (2015) yang menunjukkan bahwa parameter yang menunjukkan rata-rata memiliki signifikansi yang rendah terhadap model prediksi. Peran atau kontribusi setiap parameter magnetik terhadap proses pra-flare dapat dilihat pada tabel IV.5.
- Penjalanan *support vector machine* dengan pengurangan fitur menunjukkan bahwa jumlah fitur sama dengan 7 yaitu TOTPOT, ABSNJZH, SAVNCPP, TOTUSJH, TOTUSJZ, USFLUX, dan AREA_ACR merupakan kombinasi fitur yang optimal dengan meninjau nilai rata-rata harmonik recall dan presisinya. Model ini menghasilkan performa metrik akurasi, presisi, recall, f1, dan TSS masing-masing 0,91; 0,31; 0,82; 0,44; dan 0,73.

VI.2. Saran

- Penulis menyarankan untuk menerapkan model normalisasi lain yaitu *power transform* dalam menormalisasi data. Sifat dari normalisasi power transform selain ketahanannya terhadap pencilan adalah mentransformasikan data ke dalam distribusi gaussian. Saran murni ini bersifat eksploratif yang mendorong penelitian selanjutnya untuk menginvestigasi pengaruh dari penerapan normalisasi ini.
- Saran lain adalah mengerucutkan kasus suar surya yang ditinjau. Contohnya hanya

meninjau suar surya yang terjadi pada proses pra-*flare* yang naik saja di sebagian besar parameter. Dengan menerapkan pengerucutan ini, diharapkan performa model prediksi dapat naik secara lebih signifikan dan terhindar dari keterbatasan data suar surya yang hanya berada pada 1/3 bola Matahari.

- Penerapan pengelompokkan terhadap data sebelum pelaksanaan PCA. Dengan mengelompokkan data menjadi dua kategori: daerah aktif dengan suar surya, daerah aktif tanpa suar surya, diharapkan PCA dapat memberikan wawasan tentang kontras dua proses ini (pra-*flare* dan latar belakang)
- Penulis sangat menyarankan untuk memperluas rentang dan hyperparameter yang disesuaikan untuk model prediksi. Saat ini, hal ini tidak dilakukan karenak keterbatasan perangkat keras. Namun di masa depan ketika hal ini bukan lagi batasan, penulis menyarankan untuk mencoba menerapkan support vector machine dengan *solver* LibSVC yang mendukung kernel yang lebih banyak.

DAFTAR PUSTAKA

- Hudson, H. dan J. Ryan. 1995. *High-Energy Particles in Solar Flares*. Annual Review of Astronomy and Astrophysics. 33 (1). hlm. 239-282. https://doi.org/10.1146/annurev.aa.33.090195.001323
- Priest, E. dan T. Foorbes. 2000. *Magnetic Reconnection MHD Theory and Applications*. Cambridge: Cambridge University Press. ISBN 0 521 48179 1
- Jyothi, S. A. 2021. *Solar Superstorms: Planning for an Internet Apocalypse*. Proceedings of the 2021 ACM SIGCOMM 2021 Conference. New York: 23 - 27 Agustus 2021. hlm. 692 - 704. https://doi.org/10.1145/3452296.3472916
- McIntosh, P. S. 1990. *The classification of sunspot groups*. Solar Physics. 125 (2). hlm. 251-267. https://doi.org/10.1007/BF00158405
- Kálmán, B. 1997. *Flow Patterns Around Old Sunspots and Flare Activity*. Astronomy and Astrophysics. 327. hlm. 779-785. https://ui.adsabs.harvard.edu/abs/1997A&A...327..779K
- Herdiwijaya, Dhani dan Sherly Imelda. 2006. *The Probability of Flare Occurrences Based on Sunspot Group and Magnetic Configurations*. Applied Mathematics & Information Sciences. 11. hlm. 37-43.
- Carmichael, H. 1964. *A process for flares*. NASA Spec. Publ. 50. hlm. 451. https://ui.adsabs.harvard.edu/abs/1964NASSP..50..451C/
- Sturrock, P.A. 1966. *Model of the high-energy phase of solar flares*. Nature. 211. hlm. 695-697. https://doi.org/10.1038/211695a0
- Hirayama, T. 1974. *Theoretical model of flares and prominences. I: Evaporating flare model*. Solar Physics. 34 (2). hlm. 323-338. https://doi.org/10.1007/BF00153671
- Kopp, R.A. dan G.W Pneuman. 1976. *Magnetic reconnection in the corona and the loop prominence phenomenon*. Solar Physics. 50. hlm. 85-98. https://doi.org/10.1007/BF00206193
- Janvier, M., G. Aulanier, V. Bommier, B. Schmieder, P. Démoulin, E. Pariat. 2014. *Electric Currents in Flare Ribbons: Observations and Three-dimensional Standard Model*. The Astrophysical Journal. 788. hlm. 11-23. https://doi.org/10.1088/0004-637X/788/1/60
- Bobra, M. G., X. Sun, J. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, K. Leka. 2014. *The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs - Space-Weather HMI Active Region Patches*. Solar Physics. 289. hlm. 3549-3578. https://doi.org/10.1007/s11207-014-0529-3
- Qahwaji, R. dan T. Colak. 2007. *Automatic Short-Term Solar Flare Prediction Using*

Machine Learning and Sunspot Associations. Solar Physics. 241. hlm. 195–211. https://doi.org/10.1007/s11207-006-0272-5

- Bobra, M. G. dan S. Couvidat. 2015. *Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-Learning Algorithm*. The Astrophysical Journal. 798. hlm. 135. https://doi.org/10.1088/0004-637X/798/2/135
- Nishizika, N., Y. Kubo, K. Sugiura, M. Den, M. Ishii. 2021. *Operational solar flare prediction model using Deep Flare Net*. Earth, Planets and Space. 73. Artikel no. 64. https://doi.org/10.1186/s40623-021-01381-9
- Zhang, H., Q. Li, Y. Yang, J. Jing, J. T. L Wang, H. Wang, Z. Shang. 2022. *Solar Flare Index Prediction Using SDO/HMI Vector Magnetic Data Products with Statistical and Machine-learning Methods*. The Astrophysical Journal Supplement Series. 263. hlm. 28. https://doi.org/10.3847/1538-4365/ac9b17
- Fernanda, C. A. 2022. *Prediksi Ledakan Matahari Kelas X dan M Menggunakan Supervised Learning*. (Skripsi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung: Bandung). Diakses dari https://digilib.itb.ac.id/gdl/view/65977/?rows=166&per_page=45
- William, D. R. 2018. *Sun Fact Sheet (Online)*. NASA's Goddard Space Flight Center. Tersedia di: https://hesperia.gsfc.nasa.gov/sftheory/index.htm
- Woods, M. M., L. K. Harra, S. A. Matthews, D. H. Mackay, S. Dacie, D. M. Long. 2017. *Observations and Modelling of the Pre-flare Period of the 29 March 2014 X1 Flare*. Solar Physics. 292. hlm. 24-51. https://doi.org/10.1007/s11207-017-1064-9
- Zell, Holly. 2015. *Active Regions on the Sun (Online)*. NASA. Tersedia di: https://www.nasa.gov/image-feature/active-regions-on-the-sun
- Parker, E. N. 1963. *The Solar-Flare Phenomenon and the Theory of Reconnection and Annihilation of Magnetic Fields*. The Astrophysical Journal Supplement Series. 8. hlm. 177-178. https://doi.org/10.1086/190087
- Dennis, B. dan R. Schwartz. 1989. *Solar Flares: The Impulsive Phase*. International Astronomical Union Colloquium. 104. hlm. 75-94. https://doi.org/10.1007/BF00161688
- Su, Yingna, L. Golub, A. Van Ballegooijen, J. McCaughey, E. Deluca, K. Reeves, M. Gros. 2007. *Magnetic Shear in Two-ribbon Solar Flares*. Acta Astronomica Sinica. 210. hlm. 37. https://ui.adsabs.harvard.edu/abs/2007AAS...210.3702S/
- Ishikawa, Ryohko, dkk. 2021. *Mapping solar magnetic fields from the photosphere to the base of the corona*. Science Advances. 7 (8). hlm. 8406. https://doi.org/10.1126/sciadv.abe8406
- Vassiliadis, D., A. J. Klimas, D. N. Baker, D. A. Roberts. 1995. *A Description of The Solar Wind-Magnetosphere Coupling Based on Nonlinear Filters*. Journal of Geophysical Research. 100. hlm. 3495-3512. https://doi.org/10.1029/94JA02725

- Vassiliadis, D., A. J. Klimas, D. N. Baker, D. A. Roberts. 1996. *The Nonlinearity of Models of The vB South-AL Coupling*. Journal of Geophysical Research. 101. hlm. 19779-19787. https://doi.org/10.1029/96JA01408
- Vassiliadis, D, V. Bothmer, I. A. Daglis. 2007. *Forecasting Space Weather*. Space Weather Physics and Effects. Chichester: Praxis Publishing Ltd. ISBN 10: 3-540-23907-3
- Duan, A., C. Jiang, W. He, X. Feng, P. Zhou, J. Cui. 2019. *A Study of Pre-flare Solar Coronal Magnetic Fields: Magnetic Flux Ropes*. The Astrophysical Journal. 884. hlm 73. https://doi.org/10.3847/1538-4357/ab3e33
- Anggriawan, D. O., E. Wahjono, I. Sudiharto, A. A. Firdaus, D. N. N. Putri, A. Budikarso. 2020. *Identification of Short Duration Voltage Variations Based on Short Time Fourier Transform and Artificial Neural Network*. 2020 International Electronics Symposium (IES). Surabaya, Indonesia. hlm. 43-47, https://doi.org/10.1109/IES50839.2020.9231815
- Bustami, F. R. A., M. H. M. Saad, M. J. M. Nor, B. B. Aziz, O. Inayatullah. 2007. *The Application Of Short Time Fourier Transform And Image Processing Techniques To Detect Human Heart Abnormalities*. Proceedings of the International Conference onElectrical Engineering and InformaticsInstitut Teknologi Bandung, Indonesia June 17-19, 2007. hlm/ 450-453. ISBN 978-979-16338-0-2
- Pedregosa, F., dkk. 2011. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research. 12. hlm. 2825-2830. https://doi.org/10.48550/arXiv.1201.0490
- Virtanen, P. dkk. 2020. *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*. Nature Methods. 17. hlm. 261-272. https://doi.org/10.1038/s41592-019-0686-2
- Taskesen, E. (2020). *pca: A Python Package for Principal Component Analysis*. (Version 1.8.4) [Computer software]. https://erdogant.github.io/pca
- Fan, R. E., K. W. Chang, C. J. Hsieh, X. R. Wang, C. J. Lin. 2008. *LIBLINEAR: A Library for Large Linear Classification*. Journal of Machine Learning Research. 9. hlm. 1871-1874.
- Chang, C. C. dan C. J. Lin. 2011. *LIBSVM: A library for support vector machines*. ACM Transactions on Intelligent Systems and Technology. 2 (3). hlm. 1-27. https://doi.org/10.1145/1961189.1961199
- Hoeksema, J. Todd, dkk. 2014. *The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: Overview and Performance*. Solar Physics. 289. hlm. 3483-3530. https://doi.org/10.1007/s11207-014-0516-8
- Iglewicz, B. dan D. C. Hoaglin. 1993. *Volume 16: How to Detect and Handle Outliers*. Wisconsin: American Society for Quality Control. ISBN 0-87389-247-X.