Implementasi Metode *Random Forest* Dalam Memprediksi Peristiwa *Flare* Di Siklus Ke-23 Dan 24 Menggunakan WEKA *Data Mining*

Mohamad Dena Nugraha[1*], Judhistira Aria Utama[1], Santi Sulistiani[2]

[1]*Program Studi Fisika, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudhi 229 Bandung 40154, Indonesia*
[2]*Pusat Sains Antariksa, Lembaga Penerbangan dan Antariksa Nasional, Jl. Dr. Djundjunan 133 Bandung 40173, Indonesia*
** Corresponding author. E-mail: arzein12@student.upi.edu*
hp: +62-89-656501626

ABSTRAK

Saat ini model operasional untuk peramalan aktivitas Matahari masih didasarkan pada hubungan statistik antara aktivitas Matahari dan evolusi medan magnet Matahari. Langkah-langkah konvensional didasarkan pada klasifikasi kelompok bintik Matahari (*sunspot group*) yang memberikan informasi terbatas dari bintik Matahari (*individual sunspot*). Oleh karena itu, penelitian ini menggunakan data bintik Matahari (*individual sunspot*) untuk memprediksi peristiwa *flare* menggunakan metode *Random Forest* dalam aplikasi WEKA *Data Mining*. Model prediksi *flare* diperoleh untuk memprediksi cuaca antariksa yang berdampak langsung pada lapisan ionosfer dan dapat mengakibatkan kegagalan komunikasi radio HF. Penulis menggunakan *Supplied Test Set* dengan pembagian data latihnya yaitu 10%, 30%, 50%, 70%, 90% dan *Cross-validation Folds 10*. Hasil dari model ini menunjukan model prediksi terbaik dengan metode *Random Forest* yaitu menggunakan opsi *Supplied Test Set* dengan nilai *Correlation coefficient* = 0.8, *Mean absolute error* = 0.1 dan *Root mean squared error* = 1.06, model ini tepat untuk memprediksi peristiwa *flare* menggunakan metode *Random Forest*.

Kata Kunci: Bintik Matahari; *Flare*; *Random Forest*; WEKA.

ABSTRACT

Nowadays the operational model for predicting the sun activity is based on the statistical relation between the sun activity and the sun magnetic fields evolution. The conventional steps are based on the classification of sunspot group which give limited information from the individual sunspot. So, this study uses the data of individual sunspot to predict flare phenomenon using random forest method from WEKA Data Mining Application. The flare prediction model acquired to predict outer space weather which directly impact to ionosphere and might cause the failed of HF radio communication. The author used supplied test set with the training data distribution of 10%, 30%, 50%, 70%, 90%, dan cross-validation folds 10. The result of this model have led the best predictive model by the random forest method of using supplied test set option with correlation coefficient value of 0.8, Mean absolute error of 0.1, and root mean square error of 1.06. This model is suitable to predict the flare phenomenon using random forest method.

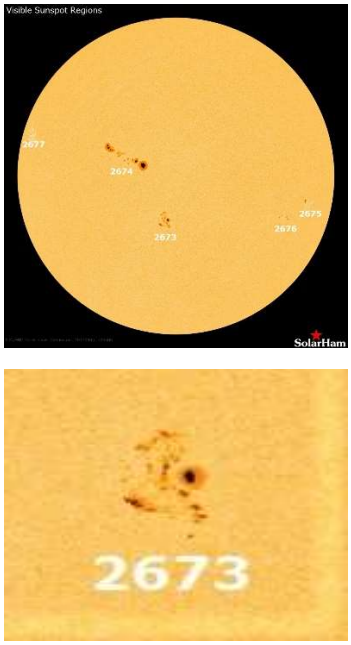
Keywords : Sunspot, Flare, Random Forest, WEKA.

ISBN: 978-602-74598-2-3

1. Pendahuluan

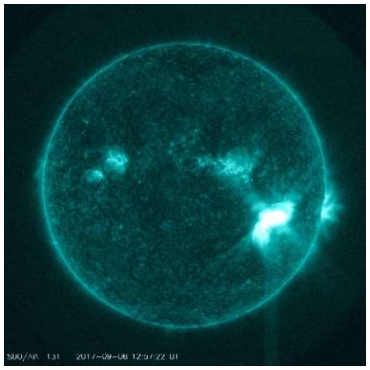
Matahari merupakan sumber energi terbesar yang dimiliki Bumi. Matahari secara terus-menerus membombardir Bumi dan seluruh Tata Surya dengan energi cahaya, partikel bermuatan listrik, dan medan magnet. Angin surya membawa jutaan ton materi ke angkasa setiap detik dan dapat melampaui planet-planet di Tata Surya. Energi dari Matahari tersebut dapat tiba-tiba mengalami peningkatan jika aktivitas Matahari tinggi. Aktivitas ini muncul karena semakin kompleksnya medan magnet di Matahari dalam satu siklusnya.

Salah satu aktivitas matahari yang penulis teliti yaitu bintik Matahari yang merupakan indikator tingkat aktivitas Matahari. Bintik Matahari adalah bagian dari permukaan Matahari yang dipengaruhi aktivitas magnetis hebat yang terpuntir karena rotasi diferensial dan mengakibatkan terhambatnya konveksi membentuk daerah yang bersuhu lebih rendah (4000-4500 K) daripada daerah di sekelilingnya (6000 K). Walau begitu, bintik Matahari itu sangat panas hingga berkilau 10 kali lebih cerah dari pada Bulan [1]. Bintik Matahari yang merupakan aktivitas magnetis hebat, juga merupakan tempat terjadinya lengkung-lengkung korona (*coronal loops*) [2].



Gambar 3. Bintik Matahari
3 Sep 2017 (sumber: solarham.net)

Bintik Matahari biasanya menjadi sumber asal ledakan Matahari (*flare*) dapat dilihat perkembangan bintik Matahari AR 2673 dari 3 september 2017 (Gambar 1) menjadi *flare* 6 september 2017 (gambar 2). *Flare* kadang-kadang dapat disertai oleh semburan partikel energetik atau lontaran massa korona (CME) yang dapat membahayakan teknologi tinggi dan kehidupan, baik di Bumi maupun di lingkungan sekitar Bumi. Karena *flare* sebagai salah satu jenis badai elektromagnetik Matahari, berdasarkan Wheatland (2005) gelombang elektromagnetik tiba di Bumi lebih awal dari energi proton (SPE) dan lebih awal dari CME maka akan sangat sulit untuk memprediksi bahayanya dan dapat menghasilkan kerugian yang sangat besar [3].



Gambar 4. Flare 6 sep 17,
kelas X9.3 (sumber: SDO)

Oleh karena itu, dalam penelitian ini parameter bintik Matahari yang dibuat secara berurutan digunakan sebagai prediktor yang nantinya bertujuan untuk merancang model prediksi agar dapat membuat prediksi *flare* yang akurat dan dapat lebih dini. Dalam sistem ini, data deret waktu ditambahkan dengan parameter bintik Matahari dan metode *random forest* dalam aplikasi WEKA *Data Mining* digunakan sebagai metode prediksi. Untuk memperkirakan hal tersebut, model ini dibangun di atas kumpulan data berskala besar yang mencakup data dari tanggal 20 Desember 1997 hingga 27 Juni 2018.

Hasil penelitian ini diharapkan dapat menjadi referensi yang baik untuk memprediksi *flare* menggunakan aplikasi WEKA *Data Mining*, dapat mengetahui model yang pas untuk memprediksi *flare* dan dapat bermanfaat ke depannya untuk memprediksi cuaca

ISBN: 978-602-74598-2-3

antariksa khususnya *flare* yang berdampak langsung pada lapisan ionosfer dan dapat mengakibatkan kegagalan komunikasi radio HF.

2. Bahan dan Metode Penelitian

2.1. Data

Data yang digunakan dalam kegiatan penelitian ini adalah data sekunder berupa data yang dihasilkan dari Space Weather Prediction Center of the National Oceanic and Atmospheric Administration (NOAA). Data ini merupakan data bintik Matahari yang kolomnya terdiri dari tanggal, nomor daerah aktif, posisi lintang dan bujur heliografis (koordinat di Matahari), luas area, klasifikasi McIntosh, jumlah bintik, kelas magnetik dan *flare* kelas C, M, X. Data ini dibuat deret waktu atau berurut langsung setiap empat hari kebelakang dimulai dari tanggal 20 Desember 1997 hingga 27 Juni 2018, yang terdiri dari 43.740 data.

2.2. *Random Forest*

Random Forest (RF) merupakan salah satu metode yang digunakan untuk klasifikasi dengan membangun banyak pohon klasifikasi dimana data acak terdistribusi sama dan dari setiap pohon dipilih nilai yang paling banyak muncul di kelasnya. Kesalahan generalisasi dari RF untuk penggolongan pohon tergantung pada keakuratan masing-masing pohon di RF dan korelasi di antara mereka [4]. Menggunakan pemilihan fitur acak untuk membagi setiap node (simpul di atasnya) menghasilkan tingkat kesalahan yang lebih baik dibandingkan dengan klasifikasi Adaboost [5].

RF dapat meningkatkan akurasi karena adanya pemilihan secara acak dalam membangkitkan anak simpul untuk setiap node (simpul diatasnya) dan diakumulasikan hasil klasifikasi dari setiap pohon, kemudian dipilih hasil klasifikasi yang paling banyak muncul. Banyaknya pohon yang akan dibentuk sangat berpengaruh terhadap tingkat akurasi hasil klasifikasi. Semakin banyak pohon, semakin akurat hasil klasifikasinya. Selain itu juga RF dapat menangani input variabel yang besar, menyeimbangkan *error* dalam *unbalanced dataset* [6].

2.3. WEKA *Data Mining*

Machine learning mempelajari bagaimana sebuah mesin atau komputer dapat belajar dari pengalaman atau bagaimana cara memprogram mesin untuk dapat belajar. *Machine learning* membutuhkan data untuk belajar sehingga biasa juga diistilahkan dengan *learn from data* [7]. Terdapat beberapa aplikasi *machine learning* yang telah dikembangkan oleh universitas-universitas ternama di dunia. Salah satu yang popular adalah *machine learning* WEKA (*Waikato Environment for Knowledge Analysis*). WEKA merupakan perangkat lunak *Data Mining* yang memiliki sekumpulan algoritma standar *Machine Learning* untuk kebutuhan praproses, klasifikasi, pengelompokan, regresi, *Association Rules Mining* (ARM) dan visualisasi [8].

WEKA menyediakan *Library* pada kelas *weka.classifiers* yang dapat langsung digunakan dalam pemrograman Java [9]. Beberapa algoritma *machine learning* yang terdapat pada kelas *weka.classifiers* tersebut antara lain: *Support Vector Machine* (functions.SMO), *K-Nearest Neighbor* (lazy.IBk), *Naive Bayes* (bayes.NaiveBayes), *C4.5 Decision Tree* (trees.J48), *Logistic Regression* (functions.Logistic) dan *Random Forest* (trees.RandomForest) [10].

Dalam aplikasi ini pula selain terdapat algoritma *machine learning*, terdapat opsi *training-test* atau pelatihan-pengujian. Opsi-opsi tersebut diantaranya ada *Use Training Set*, *Supplied Test Set*, *Cross-validation Folds 'k'* dan *Split Percentage 'k'%*. Di sini yang penulis gunakan adalah *Supplied Test Set* dan *Cross-validation Folds 10*.

2.3.1. *Supplied Test Set*

Pengklasifikasi dievaluasi pada seberapa akurat ia memprediksi kelas dari set test atau data uji yang dimuat dari file terpisah. Dengan mengeklik tombol 'Set', maka akan menampilkan jendela yang mengharuskan memilih file untuk diuji. Jadi, file pelatihan terpisah dan berbeda dengan file uji. Tetapi sebelumnya penulis melatih file pelatihan dengan menggunakan opsi *Use Training Set* terlebih dahulu yaitu pengklasifikasi yang dievaluasi tentang seberapa akurat mereka memprediksi kelas dari contoh yang dilatih atau data ujinya sama dengan data yang dilatihnya. Setelah itu, model yang dihasilkan baru dievaluasi ulang menggunakan *Supplied Test Set* dengan memasukan data uji terpisah. Di sini penulis membuat data latih yang bervariasi yaitu mulai dari 10%, 30%, 50%, 70% dan 90%

ISBN: 978-602-74598-2-3

dari data keseluruhan. Sehingga penulis memiliki 5 model yang diperoleh dari *Supplied Test Set* dimasing-masing siklus.

2.3.2. *Cross-validation Folds 'k'*

Pengklasifikasi dievaluasi dengan validasi silang, menggunakan jumlah lipatan (k) yang dimasukkan bebas di kotak teks '*Folds*'. Jumlah lipatan (k) berguna untuk membagi data latih dan uji. Penulis di sini menggunakan jumlah lipatan 10 (k = 10), yang artinya data akan dibagi 10 bagian dimana 1 bagian merupakan data uji dan sisanya data latih. Kemudian diambil 1 bagian lain menjadi data uji dan 9 lain menjadi data latih lagi, terus seperti itu hingga 10 bagian tersebut pernah dijadikan data uji.

Tabel 3. Hasil model masing-masing siklus

Opsi test		**CC**	**MAE**	**RMSE**
Training siklus 23	Testing siklus 24	0.2463	0.2938	1.9736
	Seluruh data	0.3093	0.2792	1.7807
Training siklus 24	Testing siklus 23	0.2537	0.3187	2.4448
	Seluruh data	0.4688	0.2352	1.9683
CV Folds 10	Seluruh Data	0.2999	0.2831	2.1459

Tabel 4. Hasil model seluruh data

Siklus 23			
Opsi test	CC	MAE	RMSE
CV Folds 10	0.2228	0.3167	2.5978
Training 10%	0.2844	0.2669	2.4179
Training 30%	0.2998	0.2351	2.4111
Training 50%	0.4689	0.2433	2.2236
Training 70%	0.5993	0.185	2.0758
Training 90%	0.8004	0.1428	1.5895
Siklus 24			
Opsi test	CC	MAE	RMSE
CV Folds 10	0.3239	0.2361	1.5578
Training 10%	0.2902	0.2335	1.5757
Training 30%	0.5001	0.231	1.4243
Training 50%	0.574	0.1793	1.3378
Training 70%	0.7447	0.1309	1.1311
Training 90%	0.7858	0.1036	1.0624

3. Hasil dan Pembahasan

3.1. Hasil

Diperoleh berbagai model yang dibuat menggunakan metode *Random Forest* dengan variasi opsi data latih-uji untuk masing-masing siklus yaitu *Cross-validation Folds 10* dan *Supplied Test Set* untuk data latih 10%, 30%, 50%, 70%, 90%, data uji masing-masing siklus. Untuk keseluruhan data, model yang digunakan yaitu *Cross-validation Folds 10* dan *Supplied Test Set* dengan data latih masing-masing siklus data ujinya lawan masing-masing siklus ditambah seluruh data. Korelasi serta galat setiap model dapat dilihat di Tabel 1 dan untuk keseluruhan data dapat dilihat di Tabel 2. Dengan CC (*Correlation coefficient*) yaitu bila nilai mendekati 1 maka korelasi model baik, MAE (*Mean absolute error*) yaitu bila nilai mendekati 0 maka model semakin akurat dan RMSE (*Root mean squared error*) yaitu bila nilai mendekati 0 maka model semakin akurat.

ISBN: 978-602-74598-2-3

3.2. Pembahasan

3.2.1. Masing-masing siklus

Dari hasil yang diperoleh untuk masing-masing siklus, terdapat 6 model yang dapat dibandingkan. Dengan korelasi yang paling baik maupun galat yang lebih rendah yaitu menggunakan *Supplied Test Set* dibandingkan *Cross-validation Folds 10*. Korelasi yang diperoleh untuk *Supplied Test Set* sendiri yaitu untuk siklus 23 CC = 0.8 dan untuk siklus 24 CC = 0.79. Untuk *Cross-validation Folds 10* siklus 23 CC = 0.22 dan siklus 24 CC = 0.32. Galat MAE maupun RMSE yang paling rendah pun diperoleh dari *Supplied Test Set* yaitu siklus 23 MAE = 0.14, RMSE = 1.59 dan siklus 24 MAE = 0.1, RMSE = 1.06. Sedangkan untuk *Cross-validation Folds 10* siklus 23 MAE = 0.32, RMSE = 2.59 dan siklus 24 MAE = 0.24, RMSE = 1.56. Dari model-model yang dibuat untuk kedua siklus tersebut, siklus 23 dan siklus 24 diperoleh bahwa tidak adanya keidentikan aktivitas Matahari antara kedua siklus tersebut. Dapat dilihat pula dari korelasi maupun galatnya yang berbeda jauh, hal tersebut dapat menandakan kedua siklus ini tidak dapat dinyatakan identik atau aktivitas Matahari kedua siklus ini berbeda.

Maka disimpulkan bahwa prediksi *flare* yang baik menggunakan model *Supplied Test Set* dengan memisahkan data latih dan ujinya. Sedangkan untuk kedua siklus ini tidak adanya keidentikan, maka untuk model prediksi *flare* yang baik tidak disatukan antar siklus atau bila disatukan diperlukan siklus-siklus lain agar didapat tingkat keakuratan yang baik untuk memprediksi *flare* kedepannya. Korelasi yang kecil maupun galat yang tinggi dari setiap model dapat disebabkan beberapa faktor seperti input data yang kurang tepat dengan aplikasi, kekurangan aplikasi untuk mengolah data yang besar, ataupun kesalahan dari pengguna. Untuk ketidak identikan kedua sikluspun dikarenakan setiap siklus memiliki karakteristik tersendiri, sehingga yang nanti akan dijadikan data latih untuk memprediksi *flare* baiknya dari berbagai siklus agar lebih akurat.

3.2.2. Keseluruhan data

Tidak hanya untuk masing-masing siklus, penulis pun melakukan pengolahan untuk seluruh data. Dari hasil di atas diperoleh 5 model yang digunakan, dan sama seperti sebelumnya model yang baik yaitu menggunakan model *Supplied Test Set* dengan memisahkan data latih dan uji. Untuk akurasinya sendiri CC = 0.47 dan galatnya MAE = 0.24, RMSE = 1.78. Sedangkan untuk *Cross-validation Folds 10* akurasinya CC = 0.29 dan galatnya MAE = 0.28, RMSE = 2.15.

Dari hasil ini pun diperoleh bahwa model yang baik dengan menggunakan *Supplied Test Set*, dan seperti disebutkan di atas besarnya nilai galat dan kecilnya akurasi karena ketidak identikan kedua siklus atau terlalu bervariasinya data, sehingga baiknya referensi atau data siklusnya ditambah agar aplikasi dapat melatih berbagai data. Kesalahan lainpun seperti pemasukan input data yang kurang sesuai dengan aplikasi ataupun ketidak mampuan aplikasi untuk mengolah data besarpun memungkinkan membuat hasil akurasi kecil dan galatnya menjadi besar.

4. Simpulan

Dari penelitian yang dilakukan diperoleh berbagai model dengan metode yang sama. Untuk masing-masing data siklus dan keseluruhan data, model yang digunakan yaitu *Supplied Test Set* dengan berbagai variasi dan *Cross-validation Folds 10*. Variasi *Supplied Test Set* untuk masing-masing siklus adalah memisahkan data latih 10%, 30%, 50%, 70%, 90% dan data ujinya masing-masing siklus. Lalu untuk keseluruhan data, variasi *Supplied Test Set* yaitu data latihnya masing-masing siklus dan data ujinya yang pertama siklus yang lainnya, yang kedua seluruh data.

Dari hasil tersebut, setiap model dibandingkan dan diperoleh model yang akurat adalah model dengan opsi latih-uji *Supplied Test Set* yaitu dengan memisahkan data latih dan data uji. Dari kedua siklus tersebutpun dapat dilihat hasilnya bahwa kedua siklus tidak identik atau memiliki karakter aktivitas Matahari yang berbeda.

Lalu, untuk penelitian lebih lanjut mengenai prediksi *flare* menggunakan *machine learning* disarankan lebih banyak mencari referensi agar dapat lebih mengetahui mengenai Matahari, *Random Forest* maupun aplikasi *machine learning* lain. Lebih baik lagi dapat membuat aplikasi sendiri agar lebih mudah menganalisis hasilnya dan usahakan mampu menampung data yang besar. Bila menggunakan WEKA *Data Mining*, perbanyak uji coba menggunakan variasi yang lain karena

ISBN: 978-602-74598-2-3

metode algoritma dalam aplikasi ini cukup banyak. Lebih baik lagi dapat mengetahui cara kerja algoritma metode-metode yang ada agar dapat menentukan metode yang baik untuk pengolahan berbagai data.

5. Ucapan Terima Kasih

Terimakasih kepada Ibu Santi Sulistiani, M.Si., yang telah membimbing selama penelitian di Pussainsa LAPAN Bandung juga kepada Bapak Dr. Judhistira Aria Utama, M.Si., selaku pembimbing dari UPI yang selalu memberi saran dan koreksi serta hibah dana yang diberikan. Terimakasih pula kepada Harbi Setyo Nugroho dan pihak-pihak lain yang membantu.

6. Referensi

[1] Cyndia, A. (2015). Proses timbulnya bintik matahari. [online]. Diakses di: http://nationalgeographic.co.id/berita/2015/07/proses-timbulnya-bintik-matahari. Diakses pada: 30 januari 2018.

[2] Wandani, F.M. (2015). Pengamatan Sunspot. Universitas Islam Negeri Sunan Kalijaga: Yogyakarta.

[3] Wheatland, M. 2005. A statistical solar flare forecast method. Space Weather, 3(7):S07003-1-S07003-11.

[4] Breiman, L. (2001). Random Forests. Machine Learning. 45: 5-32.

[5] Freund, Y., Schapire, R. (1996). Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, 148–156.

[6] Kurniawan, F.A. (2011). Analisis dan Implementasi Random Forest fan CART untuk Klasifikasi pada Misuse Intrusion Detection System. Telkom University.

[7] Alpaydin, E. (2010). Introduction to Machine Learning, Second Edition, London: MIT Press.

[8] Desai, A., Rai, S. (2012). Analysis of Machine Learning Algorithms using WEKA. In Proceedings ICWET 2012, 27-32.

[9] Witten, I.H., Fank, E. (2011). Data Mining Practical Machine Learning Tools and Techniques, Third Edition, Burlington: Morgan Kaufmann Publishers.

[10] Lukman, A. (2014). Machine Learning Multi Klasifikasi Citra Digital. Konferensi Nasional Ilmu Komputer (KONIK) 2014.

ISBN: 978-602-74598-2-3