REPOSITOR, Vol. 2, No. 11, November 2020, Pp. 1529-1534
ISSN : 2714-7975
E-ISSN : 2716-1382

Seleksi Fitur Relieff Pada Klasifikasi Malware Android Menggunakan Support Vector Machine(SVM)

Irma Fitriani*[1], Setio Basuki[2], Agus Eko Minarno[3]
[1,2,3]Universitas Muhammadiyah Malang
e-mail: irmafiitriani21@gmail.com*[1], setio_basuki@umm.ac.id[2] , minarno@umm.ac.id[3]

Abstrak

Seiring berkembangnya jaman perkembangan malware android terus mengalami peningkatan setiap tahunnya.Machine Learning adalah salah satu teknik yang bisa kita gunakan dalam melakukan analisa malware android dengan 2 model pendekatan statis dan dinamis.Penulis menggunakan Support Vector Machine(SVM) untuk proses klasifikasiannya dan menggunakan kernel RBF. Fitur yang digunakan dalam penelitian ini adalah Permission dan Broadcast Receiver. Untuk menambah hasil akurasinya digunakan metode Seleksi Fitur Relieff. Dan Seleksi Fitur pembandingnya adalah Chi-Square(CHI),Correlation-based Feature Selection(CFS), dan Gain Ratio(GR). Hasil dari Seleksi Fitur Relieff akan di evaluasi dengan Seleksi Fitur pembandingnya serta juga dengan hasil klasifikasi tanpa menggunakan Seleksi Fitur. Akurasi klasifikasi Seleksi Fitur Relieff menghasilkan 33.33333%, hasil akurasi Seleksi Fitur pembanding lainnya juga memberikan hasil sama dengan Seleksi Fitur Relieff. Sedangkan hasil klasifikasi tanpa Seleksi Fitur memberikan hasil yang cukup tinggi yaitu 95%. Hasil pengujian menunjukkan bahwa Seleksi Fitur tidak cocok digunakan dengan data yang sedikit karna memberikan hasil yang jauh lebih rendah dari tanpa menggunakan Seleksi Fitur.

Kata kunci: *Seleksi Fitur Relieff, Klasifikasi Malware Android,dan SVM*

Abstract

As the development evolve android malware continues to increase every year. Machine Learning is one technique that we can use in analyzing android malware with 2 static and dynamic approach models. The writer uses Support Vector Machine (SVM) for the classification process and uses the RBF kernel . The features used in this study are Permission and Broadcast Receiver. To add to its accuracy the Relieff Feature Selection method is used. And the Comparison Feature Selection is Chi-Square (CHI), Correlation-based Feature Selection (CFS), and Gain Ratio (GR). The results of Relieff Feature Selection will be evaluated by comparison Feature Selection as well as by classification results without using Feature Selection. The accuracy classification of Relieff Feature Selection produces 33,33333%, the results of the accuracy of other Comparative Feature Selection also give the same results as Relieff Feature Selection. While the classification results without Feature Selection give quite high results, namely 95%. The test results show that Feature Selection is not suitable for use with little data because it gives results that are much lower than without using Feature Selection.

Keywords*: Relieff Feature selection, Android Malware Classification and SVM*

1. Pendahuluan

Malware merupakan perangkat lunak yang dirancang untuk melakukan aktifitas b erbahaya atau perusak perangkat lainnya, seperti Trojan, virus, exploit ,dan spyware.[1]Dari data yang dirilis oleh G Data Security Labs pada tahun 2015, terdapat 3,045,722 macam malware baru[2]Seperti yang dikutip dari CRN terdapat 7 malware android paling berbahaya yang sering ditemukan pada smartphone di Indonesia yaitu basebridge, JIFake, Kungfu, Fakedolphin, SNDApps, FakeInst, dan VDLoader. Menurut laporan F-secure baru-baru ini, 97% dari malware mobile dirancang untuk platform Android yang memiliki banyak konsumen[3]Untuk perangkat seluler yang menggunakan Android sebagai platformnya, cara resmi untuk menginstal aplikasi adalah dengan google playstore[4]

Google mempublikasikan Bouncer, yaitu layanan scanning secara otomatis pada aplikasi yang diunggah di playstore untuk mengetahui kemungkinan terdapat adanya malware. Layanan Bouncer untuk menekan adanya malware belum bisa mengurangi masalah tersebut. Layanan Bouncer bekerja berdasarkan pada run-time dynamic behavioural analysis, dan langkah untuk menghindari proses analisisnya sudah dibuktikan oleh Oberheide dan Miller[5]

Hal tersebut memperlihatkan bahwa sangat diperlukan adanya suatu metode deteksi yang lebih efektif untuk mengurangi dampak berkembangnya malware Android. Metode yang diusulkan ialah berlandaskan pada pendekatan berbasis machine learning. Metode klasifikasi dipilih karna dalam proses klasifikasi dapat menciptakan suatu model yang dapat membedakan data kedalam kelas-kelas yang berbeda sesuai aturan atau fungsi tertentu[6]

Oleh karna itu banyak dilakukan penelitian klasifikasi malware pada aplikasi android. Pada penelitian Hendra Saputra dkk melakukan penerapan seleksi fitur pada klasifikasi malware android menggunakan algoritma SVM, penelitian ini menggunakan 3 seleksi fitur yaitu (CFS), (GR),dan (CHI). Berdasarkan hasil pengujian yaitu menghasilkan akurasi yang cukup tinggi yaitu pada hasil dataset seleksi fitur (GR), dan (CHI) kelemahan pada penelitian ini adalah nilai akurasi menggunakan seleksi fitur sedikit lebih rendah dari hasil akurasi data normal (tidak menggunakan seleksi fitur) dimana hasil seleksi fitur tidak sesuai harapan penulis[7]

Pada penelitian lainnya yang dilakukan oleh anang fahmi ridho & Aisyatul Karima melakukan analisis implementasi metode klasifikasi bayes untuk deteksi malware android. Penelitian ini menggunakan metode naïve bayes classifier untuk klasifikasinya, dataset berjumlah 200 APK yang terdiri dari 100 aplikasi aman dan 100 aplikasi malware. Model yang diguankan dalam proses klasifikasi ada 3 yaitu property berbasis permission, berbasis code, dan gabungan dari keduanya. Pada penelitian ini memiliki kekurangan yaitu pada hasil klasifikasi menggunakan model berbasis permission mendapatkan hasil akurasi yang rendah yaitu hanya 54% sedangkan hasil model code dan gabungan menghasilkan akurasi yang cukup tinggi yaitu 89,5% dan 88%. [8]

Maka pada penelitian ini penulis ingin melakukan penelitian analisis implementasi seleksi fitur dengan metode ReliefF pada klasifikasi malware android menggunakan algoritma SVM. Metode seleksi fitur ReliefF dipilih karna pada penelitian sebelumnya yang dilakukan oleh Ugur Pehlivan dkk metode seleksi fitur tersebut memiliki nilai akurasi tertinggi dengan algoritma SVM [9]Dan algoritma klasifikasi yang dipilih yaitu SVM, SVM dipilih karna memiliki akurasi klasifikasi yang lebih stabil dibanding dengan algoritma lainnya [10]. Penelitian ini menggunakan 3 jenis malware yaitu Wapsx, FakeInst, dan Dowgin dimana setiap malwarenya berjumlah 200 file apk. Parameter yang akan digunakan adalah permission dan broadcast receiver. Sehingga diharapkan akan menghasilkan hasil akurasi yang lebih akurat.

2. Metode Penelitian

2. 1Pengumpulan data

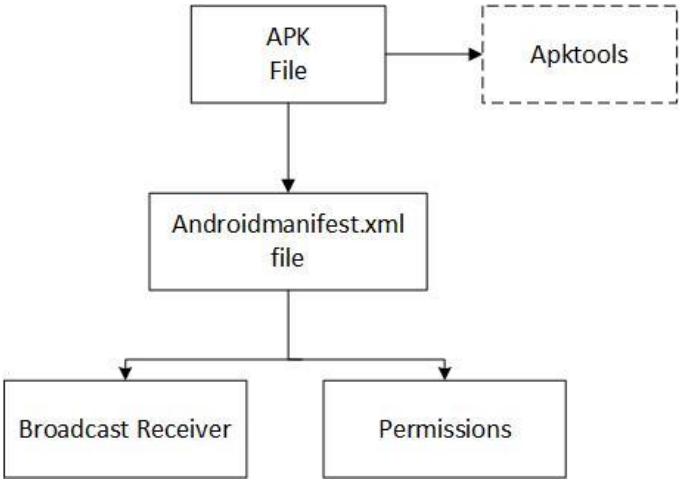
Pengumpulan data adalah tahap awal dari penelitian ini, data yang digunakan dalam penelitian ini adalah APK file, file APK didapat dari situs VirusShare , namun file apk yang telah dikumpulkan dari situs Virusshare belum diketahui label malwarenya, disini peneliti menggunakan situs VirusTotal untuk mengetahui label malwarenya. Hasil uploadfile apk menggunakan situs VirusTotal. Pembagian data malware dapat dilihat pada Tabel 1 Total malware yang digunakan dalam penelitian ini berjumlah 600 data, dalam penelitian ini menggunakan 3 jenis malware yaitu fakeinst, dowgin dan wapsx masing-masing berjumlah 200 malware.

Tabel 1 Pembagian data malware untuk proses klasifikasi

No	Nama Jenis Malware	Jumlah malware	Data Latih 70%	Data Uji 30%
1	Fakeinst	200	140	60
2	Dowgin	200	140	60
3	Wapsx	200	140	60
	Total =	600	420	180

2.2 Ekstraksi file

Pada tahap ini file APK yang sudah diketahui labelnya perlu di ekstrak untuk mendapatkan file *AndroidManifest.xml* dimana didalam file tersebut terdapat data property permission dan broadcast receiver,disini peneliti menggunakan apktool untuk mengekstrak file apk.



Gambar 1 Alur Proses ekstraksi apk file

Dari hasil ekstraksi terdapat beberapa file akan tetapi peneliti hanya menggunakan file *AndroidManifest.xml* dimana didalamnya terdapat permission dan broadcast receiver yang nantinya akan digunakan untuk penentuan jenis atribut yang dib utuhkan dalam proses klasifikasi.

2.3 Parsing data kedalam file txt

Setelah semua apk selesaai diekstrak, selanjutnya file dari *androidManifest.xml* akan dipindahkan kedalam file yang berformat txt untuk mengambil property yang dibutuhkan dalam prosesklasifikasi. Poperti yang digunakan dalam penelitian ini adalah permission dan broadcast receiver.

2.4 Penentuan Atribut

Pada tahap ini akan dilakukan penentuan atribut yang akan digunakan dalam proses klasifikasi nantinya. Pada tahap ini semua property setiap apk pada file berformat txt tadi disusun kedalam file excel untuk dilakukan remove duplicate atau penghapusan data yang duplicat sehingga data property tidak ada yang sama.

2.5 Seleksi Fitur

Seleksi fitur digunakan untuk mengurangi atribut yang tidak penting atau tidak releva n dalam data hasil dari ekstrasi yang bisa mempengaruhi hasil kinerja pada *machine learning.* Metode seleksi fitur yang akan digunakan yaitu *ReliefF.* Algortima ini merupakan algoritma pemilihan atribut yang berbasis pada instan atau record.Pemilihan atribut dilakukan deng an menghitung perbedaan bobot untuk tiap instan yang terpilih secara acak (random sampling)dengan instan yang terpili sebagai near hit (tetangga terdekat instan terpilih pa da kelas yang sama) dan near miss (tetangga terdekat instan terpilih pada kelas yang berbeda).

Algoritma ReliefF :

Input : setiap latih vector dari jumlah atribut dan jumlah class

Output : vector w dari estimasi kualitas atribut

1. Set all weights W[A] :=0,0;
2. For I := 1 to m do begin
3. Randomly select an instance R_i;
4. Find k neare9st hits H_j;
5. For each class C ≠ class (R_i) do

6. From class C find k nearest misses $M_j(C)$;
7. For0A := 1 to a do
8. W[A] := W[A]- $\sum_{j=1}^{k} diff(A, R_i, H_j)$/(m.k)+
9. $\sum_{C \neq class(R_i)} \left[\left(\frac{P(C)}{1-P(cl(R_i))} \right) \sum_{j=1}^{k} diff(A, R_i, M_j)(C)) \right]$/(m . k);
10. end;

Gambar 2 algoritma ReliefF

Tabel. 2 Seleksi Fitur Yang Digunakan

No	Nama Seleksi Fitur
1	Relieff
	Seleksi Fitur Pembanding [7]
1	Correlaton-based Feature Selection (CFS)
2	Gain Ratio (GR)
3	Chi Square (CHI)

2.6 Klasifikasi

Pada tahap ini, hasil seleksi fitur akan diolah menggunakan algoritma klasifikasi SVM. Pertama data dibagi menjadi 20kategori yaitu data hasil seleksi fitur *ReliefF* dan data original (data tanpa seleksi fitur). Selanjutnya dari jumlah 600 data yang digunakan data dibagi menjadi 70% data latih dan 30% data uji. Sebelum melakukan klasifikasi, peneliti harus menentukan parameter kernel terlebih dahulu.peneliti menggunakan kernel radial basis fu nction (RBF), dari penelitian sebelumnya kernel ini memberikan hasil yang stabil[8].Setelah menentukan parameter kernelnya selanjutnya menentukan parameter cost, pada penelitian ini akan menggunakan nilai default dari cost nilai default cost adalah 1.0 . Selanjutnya masuk ke proses pemodelan klasifikasi SVM.Setelah data terpisah secara linier maka ak an terbentuk hyperlane sebagai pemisah pola secara linier.Hyperlane pemisah terbaik da pat ditemukan dengan mengukur margin disekitar garis pemisah hypelane.Setelah hyperla ne terbentuk maka bisa dilakukan proses klasifikasi dengan data yang baru untuk memp rediksi kelas dari data tersebut.dari hasil prediksi klasifikai kita bisa membandingkan dengan kelas yang sebenarnya apakah data tersebut diprediksi kelas yang benar atau salah. Sehingga bisa dianalisa apakah hasil akurasi tinggi atau tidak.

2.7 Pengujian Metode

Tahap ini adalah tahap pengujian, sistem akan dievaluasi dengan menghitung hasil prediksi akurasi dari klasifikasi. Klasifikasi dibagi menjadi dua cara yaitu:

1. Menghitung akurasi tanpa melakukan seleksi fitur
2. Menghitung akurasi menggunakan teknik seleksi fitur

Untuk mengukur akurasi evaluasi model klasifikasi yaitu menggunakan persamaan (1) untuk menganalisa lebih jauh pengaruh seleksi fitur. Dalam penelitian ini peneliti menggunakan metode ranker untuk proses pada seleksi fiur Relieff. Skema pengujian menghitung akurasi menggunakan persamaan (1) dapat dilihat pada Tabel 3.

Tabel 3 Skema pengujian akurasi evaluasi klasifikasi

No	Keterangan	Jumlah
1	Total Atribut yang digunakan	
2	Data atribut yang diklasifikasi secara benar	
3	Perhitungan Akurasi=	$\frac{jumlah\ yang\ diklasifikasi\ secara\ benar}{Total\ testing\ sample\ yang\ di\ uji}$x100% Persamaan (1)

3. Hasil Dan Pembahasan

Sistem akan dievaluasi dengan menghitung hasil prediksi akurasi dari klasifikasi. Klasifikasi dibagi menjadi dua cara yaitu:

1. Menghitung akurasi tanpa melakukan seleksi fitur
2. Menghitung akurasi menggunakan teknik seleksi fitur

Untuk mengukur akurasi evaluasi model klasifikasi yaitu menggunakan persamaan (1) untuk menganalisa lebih jauh pengaruh seleksi fitur. Dalam penelitian ini peneliti menggunakan metode ranker untuk proses pada seleksi fiur Relieff. Dan untuk seleksi fitur pembandingnya CFS menggunakan menggunakan Search Methode Best First, seleksi fitur GR menggunakan Ranked, dan seleksi fitur CHI menggunkan metode ranked.

a. Hasil Klasifikasi Data Original

Tabel 4 Hasil Klasifikasi Data Original

No	Keterangan	Jumlah
1	Total Atribut yang digunakan	356
2	Data Atribut yang diklasifikasikan secara benar	171
3	Perhitungan Akurasi=	$\frac{171}{180}$ x 100% = **95%**

b. Hasil Klasifikasi Data Seleksi Fitur *Relieff*

Tabel 5 Hasil Klasifikasi Data Seleksi Fitur *Relieff*

No	Keterangan	Jumlah
1	Total Atribut yang digunakan	269
2	Data Atribut yang diklasifikasikan secara benar	60
3	Perhitungan Akurasi=	$\frac{60}{180}$ X 100% = **33,33333%**

c. Hasil Klasifikasi Data Seleksi Fitur CFS

Tabel 6 Hasil Klasifikasi Data Seleksi Fitur CFS

No	Keterangan	Jumlah
1	Total Atribut yang digunakan	12
2	Data Atribut yang diklasifikasikan secara benar	60
3	Perhitungan Akurasi=	$\frac{60}{180}$ X 100% = **33,33333%**

d. Hasil Klasifikasi Data Seleksi Fitur GR

Tabel 7 Hasil Klasifikasi Data Seleksi Fitur GR

No	Keterangan	Jumlah
1	Total Atribut yang digunakan	97
2	Data Atribut yang diklasifikasikan secara benar	60
3	Perhitungan Akurasi=	$\frac{60}{180}$ X 100% = **33,33333%**

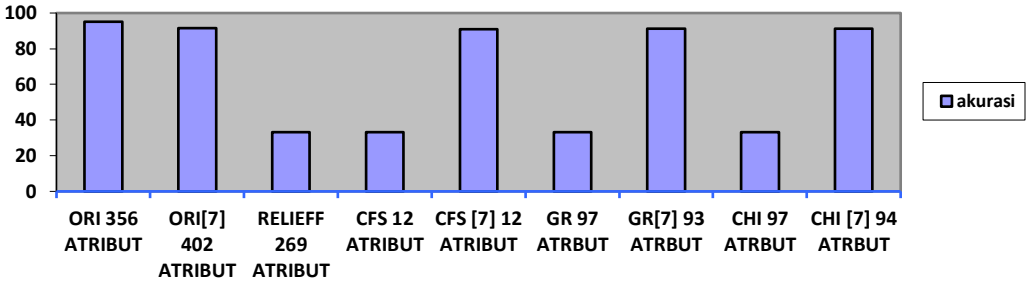
e. Hasil Klasifikasi Data Seleksi Fitur Chi-Square

Tabel 8 Hasil Klasifikasi Data Seleksi Fitur Chi-Square

No	Keterangan	Jumlah
1	Total Atribut yang digunakan	97
2	Data Atribut yang diklasifikasikan secara benar	60
3	Perhitungan Akurasi=	$\frac{60}{180}$ X 100% = **33,33333%**

Tabel 9 Perbandingan Hasil Akurasi

No	Seleksi Fitur	Jumlah Atribut yang digunakan	Jumlah Atribut Yang diklasifikasi Secara Benar	Hasil Akurasi Klasifikasi
1	Data Ori	356	171	**95%.**
2	Data Ori [7]	402	220	**91.67%.**
3	Relieff	269	60	**33.33333%.**
4	CFS	12	60	**33.33333%.**
5	CFS [7]	12	218	**90.83%.**
6	GR	97	60	**33.33333%.**
7	GR [7]	93	219	**91.25%.**
8	CHI	97	60	**33.33333%.**
9	CHI [7]	94	219	**91.25%.**



Gambar 3 perbadigan hasil akurasi klasifiasi

4. Kesimpulan

Dari evaluasi hasil akurasi yang didapat diatas menunjukkan bahwa pada penelitian ini klasifikasi jenis *Malware Android* menggunakan *Permission dan broadcast receiver* dapat berkerja dengan optimal, tetapi metode Seleksi Fitur *Relieff* yang digunakan dalam penelitian ini tidak memberikan hasil akurasi yang diharapkan. Karena hasil akurasi Seleksi Fitur *Relieff* berada sangat jauh dibawah hasil akurasi dari data original (dataset tanpa seleksi fitur).Tetapi hasil klasifikasi tanpa seleksi fitur memberian hasil akurasi yang lebih tinggi dari penelitian sebelumnya yang dilakukan oleh Hendra dkk [7]. Walaupun Seleksi Fitur dapat mengurangi atribut yang tidak relevan atau mubazir dengan cukup signifikan. Selain itu peneliti juga menyimpulkan turunnya akurasi yang menggunakan Seleksi Fitur dikarenakan adanya beberapa faktor lain yaitu pada saat penentuan label jenis *Malware* menggunakan situs *VirusTotal* dimana pada *VirusTotal* terdapat beberapa jenis antivirus yang membaca virus Android dengan jenis *Malware* yang berbeda-beda. Sehingga pada penelitian ini peneliti menentukan jenis *Malware* dengan mengambil jumlah terbanyak yang dibaca oleh antivirus pada *VirusTotal.*

Referensi

[1] V. Wahanggara and Y. Prayudi, “Sistem Deteksi Malicious Software Berbasis System Call untuk Klasifikasi Barang Bukti Digital Menggunakan Metode Support Vector Machine,” *SENTRA (Seminar Nas. Teknol. dan Rekayasa)*, no. July, pp. 1–8, 2015.

[2] A. H. Muhammad, B. Sugiantoro, A. Luthfi, M. Teknik, I. Universitas, and I. Indonesia, “Abstrak,” no. 1, 2004.

[3] L. Sayfullina *et al.*, “Improved Naive Bayes Classifier for Android Malware Classification.”

[4] T. S. Barhoom and M. I. Nasman, “Malware Detection Based on Permissions on Android Platform Using Data Mining,” *J. Eng. Res. Technol.*, vol. 3, no. 3, pp. 51–57, 2016.

[5] I. Martín, J. A. Hernández, and S. de los Santos, “Machine-Learning based analysis and classification of Android malware signatures,” *Futur. Gener. Comput. Syst.*, vol. 97, pp. 295–305, 2019, doi: 10.1016/j.future.2019.03.006.

[6] Bustami, “Penerapan Algoritma Naive Bayes,” *J. Inform.*, vol. 8, no. 1, pp. 884–898, 2014.

[7] H. Saputra, S. Basuki, and M. Faiqurahman, “Implementasi teknik seleksi fitur pada klasifikasi malware Android menggunakan support vector machine (SVM),” *Repositor*, vol. 1, no. 1, p. 1, 2019, doi: 10.22219/repositor.v1i1.1.

[8] & K. Ridlo. A. F., “Analisis Implementasi Metode Klasifikasi Bayes Untuk Deteksi Malware Android,” no. x, pp. 1–11, 2012.

[9] U. Pehlivan, N. Baltaci, C. Acarturk, and N. Baykal, “The analysis of feature selection methods and classification algorithms in permission based Android malware detection,” *IEEE SSCI 2014 2014 IEEE Symp. Ser. Comput. Intell. - CICS 2014 2014 IEEE Symp. Comput. Intell. Cyber Secur. Proc.*, no. December, 2014, doi: 10.1109/CICYBS.2014.7013371.

[10] Q. C. and Y. J. D. Zhang, H. Huang, *A Comparison Study of Credit Scoring Models*. Haikou: Third International Conference on Natural Computation (ICNC 2007), 2007.