Seminar Nasional Teknologi dan Rekayasa (SENTRA) 2020
ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

SELEKSI FITUR *MALWARE FAMILY* MENGGUNAKAN METODE C5.0

Denar Regata Akbi, S.Kom, M.Kom.[1], Vinna Rahmayanti Setyaning Nastiti, S.Si., M.Si.[2], Achmad Rizal Yogaswara[3]
[1,2,3]Universitas Muhammadiyah Malang, Malang

Kontak Person:
Denar Regata Akbi
Jalan Raya Tlogomas 246, Universitas Muhammadiyah Malang
E-mail: dnarregata@umm.ac.id

Abstrak

Malicious software (Malware) merupakan kode program berbahaya yang dapat mengganggu kinerja dari sistem komputer sehingga dapat mengakibatkan kerusakan sistem, ataupun terjadinya pencurian data yang dilakukan oleh malware. Semakin hari perkembangan malware semakin berbagai macam dan mengalami evolusi semakin canggih. Untuk meminimalisir terkena serangan dari malware peneliti dan pengembang antivirus melakukan beberapa penelitian salah satunya adalah melakukan analisis terhadap fitur – fitur yang terdapat dalam malware, salah satu teknik yang banyak digunakan untuk melakukan analisis fitur adalah metode C5.0. Analisis fitur dilakukan dengan melihat fitur – fitur apa saja yang paling berpengaruh terhadap data malware yang dianalisis, CICInvesAndMal2019 merupakan data malware yang digunakan pada penelitian ini dan didapatkan hasil bahwa dari 305 data malware yang memiliki 918 fitur, tereduksi dan terseleksi menjadi 129 fitur yang paling berpengaruh serta hasil evaluasi yang didapatkan dari penelitian ini adalah nilai accuracy 0,085 dan nilai kappa 0,058.

Kata kunci: *malware, C5.0, analisis fitur, accuracy, kappa*

1. Pendahuluan

Malicious software atau dikenal dengan *malware* merupakan perangkat lunak berbahaya yang dirancang untuk merusak atau menyusup ke sistem komputer tanpa persetujuan pemilik komputer. Virus, worm, trojan, keylogger, dan spyware merupakan beberapa contoh dari malware[1]**.** Teknik untuk mendeteksi *malware* terbagi menjadi dua teknik besar, yaitu *anomaly-based* dan *signature-based*[2]. Analisa *malware* yang dilakukan pada kedua teknik tersebut dapat dilakukan menggunakan beberapa pendekatan diantaranya *static analysis*, *dynamic analysis*, dan *hybrid analysis*[3][4].

Perkembangan *malware* semakin hari semakin pesat, sehingga diperlukan analisis lebih mendalam untuk bisa menghasilkan sistem pendeteksi yang baik, analisis yang bisa dilakukan salah satunya pada bagian fitur dari data malware, dimana data tersebut mencerminkan salah satu karakterisktik dari malware. Metode yang dapat digunakan untuk melakukan analisis terhadap fitur dari data malware salah satunya menggunakan machine learning[5]. Seleksi fitur merupakan salah satu teknik machine learning yang dapat digunakan untuk melakukan penyeleksian fitur – fitur yang penting terhadap data fitur yang ada[6].

Algoritma C5.0 merupakan salah satu algoritma dalam klasifikasi di *data mining* dan juga termasuk teknik dari *decision tree*. Atribut dengan nilai informasi terbesar akan dijadikan inti untuk *node* selanjutnya[7].

Penelitian yang dilakukan oleh Taheri, Laya, dkk[8]., menggunakan beberapa dataset dari tahun 2012 sampai tahun 2019, dimana collect data yang dilakukan menggunakan beberapa macam lingkungan, terdapat dataset yang lingkungan collect datanya menggunakan simulator, ada pula yang menggunakan real-phone. Dari tahun 2012 sampai 2019, peneliti selalu melakukan penambahan attribute untuk melakukan analisis data malware yang digunakan, hal tersebut bertujuan untuk melihat attribute apa saja yang terdapat dan melekat pada suatu malware. Beberapa attribute yang digunakan salah satunya, intent, permissions, API Calls, dll.

Penelitian yang dilakukan oleh Feizollah, Ali, dkk[9, melakukan review 100 penelitian yang terkait dengan perspektif seleksi fitur terhadap data malware pada Android, kesimpulan yang didapatakan peneliti melakukan pengkategorian data seleksi fitur berdasarkan 4 kelompok, kelompok pertama, selekesi fitur yang dianalisis dari hasil collect data menggunakan teknik static, kelompok kedua, seleksi fitur yang dianalisis dari hasil collect data menggunakan teknik dynamic, kelompok ketiga, seleksi fitur yang dianalsis dari hasil collect data menggunakan teknik hybrid, dan kelompok

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

yang terakhir seleksi fitur yang dianalisis dari hasil collect data pada metadata yang terdapat pada aplikasi yang berhubungan dengan Google Play.

Penelitian Wang, Wei, dkk tahun 2019[10], mencoba untuk melakukan ektraksi fitur kemudian melakukan rekonstruksi dari fitur yang ada, untuk digunakan sebagai referensi dari framework yang dibuat peneliti yang digunakan untuk mendeteksi suatu malware yang ada. Peneliti telah melakukan survey paper sebanyak 1947 paper, dan terseleksi sebanyak 236 paper utama yang akhirnya digunakan sebagai acuan oleh peneliti. Pada penelitian ini, peneliti mencoba untuk melakukan survey secara komprehensif terkait beberapa malapps dan beberapa fitur yang berpengaruh, ketepatan pemilihan fitur – fitur tersebut, dapat mempengaruhi output dari hasil klasifikasi terhadap suatu malware yang nantinya digunakan sebagai referensi untuk pembuatan framework baik untuk pendeteksian, analisis, ataupun evaluasi.

Maka dalam penelitian kali ini, penulis ingin melakukan seleksi fitur dari 305 data malware yang memiliki 918 fitur, dimana data malware di dapat dari penelitian yang dilakukan oleh Laya Taheri dkk pada tahun 2019, dengan judul "Extensible Android Malware Detection and Family Classification Using Network-Flows and API-Calls"[8], pada penelitian yang dilakukan oleh Laya Taheri dkk, tidak melakukan tahapan fitur seleksi, sehingga fitur – fitur yang mempunyai nilai mayoritas 0 ikut diproses.

Tujuan dari penelitian ini adalah untuk mengetahui fitur – fitur apa saja yang paling berpengaruh dalam dataset malware yang digunakan, sehingga nantinya diharapkan pada penelitian lanjutan, data fitur malware yang telah terseleksi dapat digunakan sebagai acuan untuk proses pengklasifikasian malware berdasarkan family dengan lebih baik lagi.

2. Metode Penelitian

Data pada penelitian ini didapatkan dari website Canadian Institute for Cybersecurity sebuah web yang dimiliki oleh The University of New Brunswick. Canadian Institute for Cybersecurity menyediakan dataset malware yang diberi nama CICInvesAndMal2019. Pada dataset ini terdapat 305 sampel malware yang terdiri dari 39 jenis malware family. Dataset memiliki 918 fitur, dalam 918 fitur tersebut tidak diketahui mana fitur yang benar-benar penting dan tidak. Jika terdapat fitur yang tidak penting dan digunakan dalam proses klasifikasi maka nantinya dapat mempengaruhi hasil akhir klasifikasi, untuk itu maka dilakukan proses preprocessing data guna menyeleksi fitur dalam dataset menjadi lebih sedikit. Metode yang digunakan untuk proses preprocessing adalah metode C5.0. Dataset yang dimiliki terdiri dari 305 data malware. Berikut proses yang dilakukan pada algoritma C5.0: Atribut dengan nilai informasi terbesar akan dijadikan inti untuk *node* selanjutnya. Informasi sangat diperlukan untuk mengklasifikasi tiap sampel yang digunakan dengan cara persamaan (1). Setelah itu, nilai himpunan bagian dari atribut A akan dihitung menggunakan persamaan (2). Kemudian untuk perolehan informasi dihitung menggunakan persamaan (3) [referensi 13 di buku yoga]. Rumus persamaan (1), (2) dab (3) dapat dilihat pada gambar 1.

$$I\,(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i) \qquad (1)$$

$$E(A) = \sum_{j=1}^{y} \frac{s_{1j}+\cdots+s_{mj}}{s}\, I\,(s_{1j}, \ldots, s_{mj}) \qquad (2)$$

$$Gain(A) = I\,(s_1, s_2, \ldots, s_m) - E(A) \qquad (3)$$

Gambar 1 Persamaan C5.0

Dataset malware yang ada dilakukan seleksi fitur menggunakan persamaan C5.0, setelah diproses menggunakan persamaan tersebut, hasil yang didapat nantinya akan terlihat persentase fitur – fitur mana saja yang paling berpengaruh dan fitur – fitur mana saja yang kurang berpengaruh dalam menentukan kelas dari suatu malware. Pengujian dilakukan dengan membandingkan data sebelum dan setelah diproses menggunakan C5.0, selain itu dilakukan juga evaluasi menggunakan nilai accuracy dan nilai kappa.

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

Implematasi fitur seleksi menggunakan C5.0 dilakukan dengan bantuan *software* Rstudio dan menggunakan bahasa pemrograman R. Proses fitur seleksi terdapat beberapa langkah sebagai berikut :

a. Melakukan *Import* dataset *malware* ke dalam R.
b. Membaca dataset yang telah di *import.*
c. Merubah tipe data atribut *Family* dari char menjadi *factor*.
d. Melakukan pembagian dataset menjadi data *training* dan data *test*
e. *Training* metode menggunakan data *training.*
f. *Training* metode dengan data *training* sebanyak sepuluh kali percobaan.
g. Proses klasifikasi.

3. Hasil dan Pembahasan

Data malware yang terdapat pada dataset berjumlah 305 data dengan masing masing data memiliki 918 fitur. Setelah melalui proses feature selection menggunakan metode C5.0, fitur yang tadinya berjumlah 918 berkurang menjadi 129 fitur. Metode C5.0 pada penelitian ini hanya digunakan untuk proses seleksi fitur, training data dilakukan sebanyak 10 kali. Dalam **Gambar 2** merupakan hasil dari *training* data pertama menghasilkan 40 fitur yang berpengaruh dalam proses klasifikasi dengan nilai error 9.7%. Daftar nilai error dari setiap percobaan pertama hingga kesepuluh terdapat pada **Gambar 3**.

```
Evaluation on training data (185 cases):

          Decision Tree
        ----------------
        Size      Errors

          48   18( 9.7%)   <<
```

Gambar 2 Nilai *error training* data percobaan pertama

```
Evaluation on training data (185 cases):

Trial       Decision Tree
-----     ----------------
          Size      Errors

   0        48   18( 9.7%)
   1        41   27(14.6%)
   2        46   42(22.7%)
   3        45   30(16.2%)
   4        43   31(16.8%)
   5        45   48(25.9%)
   6        40   45(24.3%)
   7        47   31(16.8%)
   8        44   29(15.7%)
   9        40   40(21.6%)
```

Gambar 3 Daftar nilai *error training* data 10 kali

Presentase masing-masing fitur pada Attribute Usage dalam **Gambar 3** menunjukkan seberapa besar pengaruh fitur tersebut terhadap jenis family malware. Fitur yang tidak terdaftar dalam Attribut Usage ini dianggap tidak memiliki pengaruh yang signifikan dan dapat dihapus nantinya.

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

```
Attribute usage:

100.00% Down.Up_Ratio
 80.54% X2Grammed_APICalls_.416
 74.59% X2Grammed_APICalls_.41
 71.89% X2Grammed_APICalls_.30
 68.11% X2Grammed_APICalls_.411
 65.95% X2Grammed_APICalls_.142
 64.86% X2Grammed_APICalls_.59
 60.54% X2Grammed_APICalls_.50
 50.27% Bwd_Packets.s
 28.65% X2Grammed_APICalls_.125
 27.03% X2Grammed_APICalls_.86
 24.32% X2Grammed_APICalls_.301
 22.70% Idle_Std
 21.62% X2Grammed_APICalls_.35
 20.00% X2Grammed_APICalls_.264
 19.46% X2Grammed_APICalls_.213
 18.92% X2Grammed_APICalls_.283
 18.38% X2Grammed_APICalls_.77
 17.30% X2Grammed_APICalls_.109
 17.30% X2Grammed_APICalls_.170
 17.30% Fwd_PSH_Flags
 16.76% X2Grammed_APICalls_.408
 15.68% Total_Length_of_Fwd_Packets
 14.59% X2Grammed_APICalls_.40
 14.05% X2Grammed_APICalls_.24
 12.97% X2Grammed_APICalls_.18
 10.27% X2Grammed_APICalls_.4
 10.27% X2Grammed_APICalls_.99
 10.27% X2Grammed_APICalls_.217
 10.27% X2Grammed_APICalls_.357
  8.65% X2Grammed_APICalls_.221
  7.57% X2Grammed_APICalls_.12
  7.57% Fwd_Packet_Length_Min
  7.03% X2Grammed_APICalls_.1
  7.03% X2Grammed_APICalls_.230
  5.41% X2Grammed_APICalls_.197
  4.86% Flow_Bytes.s
  4.86% Active_Std
  3.78% ACK_Flag_Count
  3.24% PSH_Flag_Count
```

Gambar 3 *Attribute Usage*

Gambar 3 menunjukan bahwa atribut down.Up_Ratio merupakan atribut yang paling berpengaruh dengan nilai 100% pada proses *training*, tetapi atribut di atas dapat berubah pada percobaan - percobaan training berikutnya, oleh karena itu training akan diulang sebanyak sepuluh kali guna mendapatkan perubahan hasil fitur mana yang lebih berpengaruh.

Hasil dari proses *training* sebanyak sepuluh kali memperlihatkan bahwa atribut – atribut yang berpengaruh ada sebanyak 129 atribut dengan nilai *error* yang berbeda pada setiap percobaan, tertuang pada **Tabel 1**.

Tabel 1 Hasil Proses Training Sepuluh Kali Percobaan

No	Persentase	Atribut
1.	100.00%	X2Grammed_APICalls_.7
2.	100.00%	X2Grammed_APICalls_.53
3.	100.00%	X2Grammed_APICalls_.416
4.	100.00%	Min_Packet_Length
5.	100.00%	PSH_Flag_Count
6.	100.00%	Down.Up_Ratio
7.	97.30%	Bwd_Packets.s
8.	94.05%	X2Grammed_APICalls_.142
9.	94.05%	X2Grammed_APICalls_.411

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

No	Persentase	Atribut
10.	94.05%	Fwd_Packets.s
11.	82.70%	X2Grammed_APICalls_.357
12.	81.08%	X2Grammed_APICalls_.41
13.	81.08%	X2Grammed_APICalls_.50
14.	81.08%	X2Grammed_APICalls_.329
15.	80.00%	X2Grammed_APICalls_.30
16.	75.68%	X2Grammed_APICalls_.301
17.	71.89%	X2Grammed_APICalls_.59
18.	66.49%	X2Grammed_APICalls_.263
19.	65.41%	X2Grammed_APICalls_.125
20.	64.32%	X2Grammed_APICalls_.342
21.	63.78%	X2Grammed_APICalls_.364
22.	61.62%	Fwd_PSH_Flags
23.	57.84%	Total_Length_of_Fwd_Packets
24.	57.84%	Idle_Min
25.	55.68%	X2Grammed_APICalls_.109
26.	55.14%	X2Grammed_APICalls_.217
27.	53.51%	X2Grammed_APICalls_.796
28.	52.97%	Fwd_IAT_Min
29.	52.43%	X2Grammed_APICalls_.90
30.	49.19%	X2Grammed_APICalls_.1
31.	47.57%	X2Grammed_APICalls_.255
32.	44.32%	X2Grammed_APICalls_.532
33.	40.00%	act_data_pkt_fwd
34.	38.92%	X2Grammed_APICalls_.48
35.	35.68%	X2Grammed_APICalls_.160
36.	35.68%	Idle_Std
37.	35.14%	X2Grammed_APICalls_.408
38.	34.05%	X2Grammed_APICalls_.4
39.	34.05%	X2Grammed_APICalls_.89
40.	33.51%	X2Grammed_APICalls_.331
41.	31.89%	X2Grammed_APICalls_.14
42.	31.89%	X2Grammed_APICalls_.38
43.	31.35%	X2Grammed_APICalls_.116
44.	31.35%	Active_Std
45.	29.19%	X2Grammed_APICalls_.138
46.	28.65%	X2Grammed_APICalls_.22
47.	28.65%	X2Grammed_APICalls_.34
48.	28.11%	X2Grammed_APICalls_.111
49.	27.57%	X2Grammed_APICalls_.5
50.	27.57%	X2Grammed_APICalls_.35
51.	27.03%	X2Grammed_APICalls_.86
52.	25.95%	X2Grammed_APICalls_.193
53.	25.95%	X2Grammed_APICalls_.222
54.	25.95%	X2Grammed_APICalls_.353
55.	25.41%	X2Grammed_APICalls_.121
56.	25.41%	X2Grammed_APICalls_.296
57.	25.41%	X2Grammed_APICalls_.344
58.	24.86%	X2Grammed_APICalls_.240
59.	24.86%	X2Grammed_APICalls_.283
60.	24.86%	X2Grammed_APICalls_.303
61.	23.78%	X2Grammed_APICalls_.13

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

No	Persentase	Atribut
62.	23.78%	X2Grammed_APICalls_.363
63.	23.24%	X2Grammed_APICalls_.99
64.	23.24%	X2Grammed_APICalls_.320
65.	22.70%	X2Grammed_APICalls_.2
66.	22.16%	X2Grammed_APICalls_.12
67.	22.16%	Flow_Packets.s
68.	21.62%	X2Grammed_APICalls_.24
69.	21.62%	X2Grammed_APICalls_.66
70.	21.62%	X2Grammed_APICalls_.309
71.	21.62%	Flow_Bytes.s
72.	21.62%	Init_Win_bytes_backward
73.	21.08%	X2Grammed_APICalls_.213
74.	21.08%	Fwd_Packet_Length_Min
75.	21.08%	Idle_Max
76.	20.54%	X2Grammed_APICalls_.56
77.	20.00%	X2Grammed_APICalls_.264
78.	19.46%	X2Grammed_APICalls_.101
79.	19.46%	X2Grammed_APICalls_.108
80.	19.46%	X2Grammed_APICalls_.343
81.	18.92%	X2Grammed_APICalls_.265
82.	18.38%	X2Grammed_APICalls_.77
83.	17.30%	X2Grammed_APICalls_.0
84.	17.30%	X2Grammed_APICalls_.74
85.	17.30%	X2Grammed_APICalls_.170
86.	16.22%	X2Grammed_APICalls_.225
87.	15.68%	X2Grammed_APICalls_.3
88.	15.68%	X2Grammed_APICalls_.211
89.	15.14%	X2Grammed_APICalls_.28
90.	14.59%	X2Grammed_APICalls_.17
91.	14.59%	X2Grammed_APICalls_.40
92.	14.59%	X2Grammed_APICalls_.192
93.	13.51%	X2Grammed_APICalls_.8
94.	12.97%	X2Grammed_APICalls_.18
95.	12.97%	X2Grammed_APICalls_.165
96.	12.97%	X2Grammed_APICalls_.197
97.	12.97%	X2Grammed_APICalls_.293
98.	12.97%	URG_Flag_Count
99.	12.43%	Bwd_IAT_Mean
100.	11.89%	X2Grammed_APICalls_.311
101.	10.27%	X2Grammed_APICalls_.221
102.	10.27%	Flow_Duration
103.	9.73%	ACK_Flag_Count
104.	9.19%	X2Grammed_APICalls_.36
105.	9.19%	X2Grammed_APICalls_.393
106.	9.19%	X2Grammed_APICalls_.524
107.	8.65%	X2Grammed_APICalls_.19
108.	8.65%	X2Grammed_APICalls_.64
109.	8.65%	X2Grammed_APICalls_.84
110.	8.11%	X2Grammed_APICalls_.281
111.	7.57%	X2Grammed_APICalls_.137
112.	7.57%	X2Grammed_APICalls_.209
113.	7.57%	X2Grammed_APICalls_.220

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

No	Persentase	Atribut
114.	7.03%	X2Grammed_APICalls_.43
115.	7.03%	X2Grammed_APICalls_.91
116.	7.03%	X2Grammed_APICalls_.230
117.	7.03%	X2Grammed_APICalls_.328
118.	7.03%	Bwd_IAT_Total
119.	6.49%	X2Grammed_APICalls_.171
120.	5.95%	X2Grammed_APICalls_.51
121.	5.95%	X2Grammed_APICalls_.87
122.	5.95%	X2Grammed_APICalls_.205
123.	5.95%	X2Grammed_APICalls_.368
124.	5.41%	X2Grammed_APICalls_.212
125.	4.32%	X2Grammed_APICalls_.39
126.	4.32%	X2Grammed_APICalls_.229
127.	4.32%	Fwd_IAT_Total
128.	3.78%	X2Grammed_APICalls_.275
129.	3.78%	X2Grammed_APICalls_.663

Hasil evaluasi dilakukan setelah melalui tahapan sepuluh kali percobaan training, hasil yang didapatkan nilai accuracy 0,0857 sedangkan nilai kappa didapatkan 0,058. Untuk fitur yang paling berpengaruh dengan persentase 100% terdapat 6 fitur, sedangkan fitur yang paling tidak berpengaruh dengan persentase dibawah 5% terdapat 5 fitur.

4. Kesimpulan

Dari hasil penelitian seleksi fitur yang telah dilakukan pada data *CICInvesAndMal2019*, didapatkan hasil nilai accuracy 0,085 dan nilai kappa 0,058, sedangkan untuk hasil seleksi fitur, yang semula fitur dari data malware berjumlah 918, setelah dilakukan proses seleksi didapatkan hasil, fitur yang paling berpengaruh untuk penentuan kelas dari data malware tersebut berjumlah 129 fitur.

Referensi

[1] Gadhiya, Savan, and Kaushal Bhavsar. "Techniques for malware analysis." *International Journal of Advanced Research in Computer Science and Software Engineering* 3.4 (2013)
[2] Chowdhury, Mozammel, Azizur Rahman, and Rafiqul Islam. "Malware analysis and detection using data mining and machine learning classification." *International Conference on Applications and Techniques in Cyber Security and Intelligence*. Edizioni della Normale, Cham, 2017.
[3] Sihwail, Rami, Khairuddin Omar, and KA Zainol Ariffin. "A survey on malware analysis techniques: Static, dynamic, hybrid and memory analysis." *International Journal on Advanced Science, Engineering and Information Technology* 8.4-2 (2018): 1662.
[4] Damodaran, Anusha, et al. "A comparison of static, dynamic, and hybrid analysis for malware detection." *Journal of Computer Virology and Hacking Techniques* 13.1 (2017): 1-12
[5] Ranveer, Smita, and Swapnaja Hiray. "Comparative analysis of feature extraction methods of malware detection." *International Journal of Computer Applications* 120.5 (2015).
[6] Babaagba, Kehinde Oluwatoyin, and Samuel Olumide Adesanya. "A study on the effect of feature selection on malware analysis using machine learning." *Proceedings of the 2019 8th InternationalConference on Educational and Information Technology*. 2019.
[7] Kurniawan, Erwin, et al. "C5. 0 algorithm and synthetic minority oversampling technique (SMOTE) for rainfall forecasting in Bandung regency." *2019 7th International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2019.

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

[8] Taheri, Laya, Andi Fitriah Abdul Kadir, and Arash Habibi Lashkari. "Extensible android malware detection and family classification using network-flows and api-calls." *2019 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2019

[9] Feizollah, Ali, et al. "A review on feature selection in mobile malware detection." *Digital investigation* 13 (2015): 22-37.

[10] Wang, Wei, et al. "Constructing features for detecting android malicious applications: Issues, taxonomy and directions." *IEEE Access* 7 (2019): 67602-67631