JIKB
Jurnal Ilmu Komputer dan Bisnis

Jurnal Ilmu Komputer dan Bisnis (JIKB)
Mei-2021, Vol.XII, No.1, hal.60-69
ISSN(P): 2087-3921; ISSN(E): 2598-9715

Pemodelan *Proxy* Anonim Menggunakan Algoritme *Expectation Maximization* Dengan *Data Balancing*

[1]Nur Witdi Yanto*, [2]Heru Sukoco, [3]Shelvie Nidya Neyman
[1]FTI Universitas Jayabaya, [2,3]FMIPA Institut Pertanian Bogor

Alamat Surat
Email: [1]nur.witdi@gmail.com*, [2]hrskom@ipb.ac.id, [3]shelvie@ipb.ac.id

Article History:
Diajukan: 30-03-2021; Direvisi: 14-04-2021; Diterima: 28-04-2021

ABSTRAK

Penggunaan internet untuk mengakses situs-situs tertentu yang tidak berhubungan dengan pekerjaan dibatasi akses nya oleh perusahaan atau organisasi. Perusahaan atau organisasi melakukan pemblokiran untuk tujuan mengamankan jaringan mereka terhadap ancaman virus, *spyware*, *hacker* dan ancaman lainnya yang dapat merugikan perusahaan dengan cara menerapkan *firewall*, filter URL serta sistem deteksi intrusi. Namun, pengamanan tersebut masih dapat ditembus dengan menggunakan layanan *proxy* anonim. Penggunaan *proxy* anonim memungkinkan user untuk melakukan *bypass* sebagian besar sistem penyaringan. Dalam penelitian ini, data *proxy* anonim diperoleh dengan cara menangkap (*capture*) paket data menggunakan aplikasi wireshark. Data tersebut dimodelkan dengan algoritme *expectation maximization* sehingga diperoleh akurasi model sebesar 71.22% pada pembagian data yang seimbang. Hasil ini menunjukkan bahwa model mampu mengenali penggunaan *proxy* anonim pada traffic internet.

Kata kunci : *Expectation maximization*; *firewall*; pemodelan; *proxy anonim*; *wireshark*

ABSTRACT

Internet usage to access sites that are not related to work are limited by companies or organizations. They are blocked that access for securing their network against threats of viruses, spyware, hackers, and other threats that can harm the company by implementing firewalls, URL filters, and intrusion detection systems. However, securing the network can still be penetrated by using anonymous proxy services. The anonymous proxy allows users to bypass most filtering systems. In this research, the data of anonymous proxies obtained by capturing the packet data using the Wireshark application. The data is modeled with an expectation maximization algorithm so that we can get the accuracy of the model is 71.22% on a balanced distribution of data. These results indicate that the model is able to recognize the use of anonymous proxies on the internet traffic.

Keywords: *Anonymous proxy; expectation maximization; firewall; modeling; wireshark*

1. PENDAHULUAN

Banyak organisasi atau perusahaan menyaring situs-situs internet yang sering diakses oleh pekerja mereka ketika jam sibuk bekerja. Organisasi atau perusahaan melakukan hal tersebut untuk menjaga jaringan mereka tetap aman dalam mengembangkan kebijakan penggunaan jaringan yang sesuai dan menerapkan *firewall*, filter URL dan solusi deteksi

https://doi.org/10.47927/jikb.v12i1.91

intrusi. Namun, layanan *proxy* anonim memungkinkan pengguna untuk melakukan *bypass* sebagian besar sistem penyaringan. *Proxy* anonim memberikan kesempatan bagi pengguna untuk mengakses situs yang dilarang atau illegal serta konten dan aplikasi yang dapat mengekspos suatu perusahaan. *Proxy* anonim juga membahayakan keamanan data dan privasi pengguna dengan cara membuat celah keamanan pada jaringan.

Sistem komunikasi anonim adalah contoh populer dari sistem berbasis *proxy*, yang memungkinkan pengguna untuk menyembunyikan alamat IP mereka dari layanan yang mereka gunakan dan sering di disain dengan menggunakan enkripsi (Chakravarty et al. 2015). Dengan *proxy* anonim, pengguna mengakses halaman website *proxy* anonim kemudian pada halaman yang ditampilkan oleh website tersebut pengguna dapat melakukan akses ke situs yang dilarang atau diblokir, *proxy* akan menampilkan situs yang diblokir tersebut pada halaman *proxy* anonim. Biasanya, dilakukan teknik enkripsi untuk menyembunyikan situs tujuan ketika mengakses server *proxy* melalui parameter URL, hal tersebut menyebabkan sangat sulit sekali untuk dideteksi (Brozycki 2008).

Server *proxy* anonim digunakan untuk menyembunyikan identitas pengguna. Salah satu pendekatan yang digunakan biasa nya dengan *open proxy*. *Open proxy* merupakan server *proxy* yang dapat diakses oleh setiap pengguna internet. *Open proxy* anonim memungkinkan setiap pengguna internet untuk menyembunyikan alamat IP mereka serta identitas dan lokasi dari layanan yang diakses (Dar et al. 2016). Jenis server ini secara teratur digunakan sebagai sarana untuk menyembunyikan identitas penjahat sehingga mereka dapat melakukan berbagai kejahatan di internet tanpa tertangkap. Seorang pengguna yang menggunakan *proxy* anonim melalui jaringan perusahaan untuk menembus filter jaringan, tanpa disadari membocorkan informasi rahasia tentang perusahaan mereka. Situs *proxy* anonim bertindak sebagai perantara, meneruskan permintaan dan menampilkan hasil pada situs *proxy* anonim tersebut, sementara itu juga menyembunyikan identitas pengguna dengan menyembunyikan alamat IP mereka dari server web di internet (Miller et al. 2016). Ada empat tipe utama dari server *proxy* yang digunakan untuk mendapatkan tingkat anonimitas, diantaranya adalah *Transparent proxy*, *Anonymous proxy*, *Distorting proxy* dan *High Anonimity Proxy* (Fashoto et al. 2016).

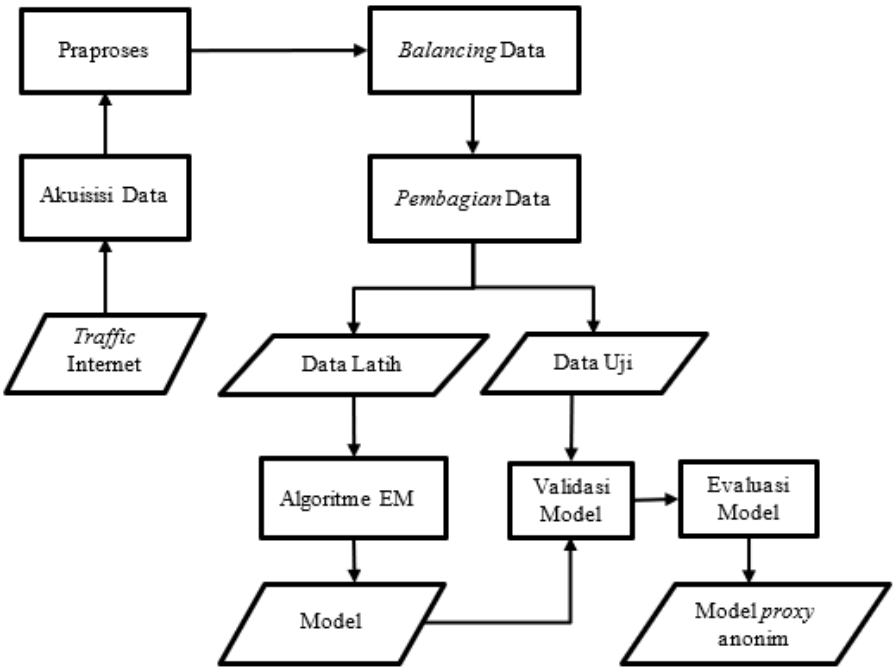
Klasifikasi untuk traffic terenkripsi secara realtime diteliti oleh Bar-Yanai et al. (2010). Dalam penelitian nya, penulis menggunakan metode *hybrid* yaitu dengan mengkombinasikan algoritme *k-means* dan *k-nearest neighbor*. *Classifier* yang diusulkan pada penelitian tersebut menunjukkan hasil yang sangat baik untuk *traffic* terenkripsi. Alshammari dan Zincir-Heywood (2011) meneliti tentang identifikasi *traffic* yang terenkripsi dengan pendekatan berbasis *machine learning* menggunakan fitur *packet header* sederhana dan fitur aliran statistik tanpa menggunakan nomor *port*, alamat IP, *port* sumber/tujuan serta informasi *payload*. Dalam penelitiannya hanya mendeteksi dua aplikasi yang terenkripsi yaitu *secure shell* (SSH) dan Skype. Hasil dari penelitian tersebut membuktikan bahwa untuk mendeteksi aplikasi yang terenkripsi dengan akurasi yang tinggi dapat dilakukan tanpa memeriksa informasi *payload*, alamat IP dan nomor *port*. Mc Keague (2013) meneliti tentang deteksi penggunaan *proxy* anonim. Penelitian ini juga menjelaskan metode-metode yang digunakan oleh *proxy* anonim, karakteristik *proxy* anonim dan mekanisme-mekanisme yang berpotensi untuk mendeteksi ketika *proxy* sedang digunakan. Hasil dari penelitian tersebut sukses mendeteksi penggunaan *Glype*, *PHPProxy*, *Unsecure CGI proxy* dan *Tor Browser*, namun hasil dari sistem yang dibangun belum mencapai akurasi 100%. Penelitian Miller et al. (2016) membahas tentang klasifikasi traffic jaringan untuk mendeteksi *routing web proxy* anonim. Sistem yang dibangun yaitu dengan mendeteksi gangguan serta mengklasifikasikan *traffic* secara spesifik, sistem akan melakukan pengecekan karakteristik yang nampak pada paket yang dibangkitkan oleh *proxy* anonim dan kemudian membuat aturan untuk memutuskan koneksi pengguna *proxy* anonim. Cha dan Kim (2017) membandingkan metode klasifikasi

untuk mendeteksi *traffic* yang terenkripsi. Penelitiannya menggunakan empat metode klasifikasi (*Naive Bayesian*, *Support Vector Machine*, *Classification and Regression Tree* dan *Adaptive Boosting*) dengan tiga tes secara random (*Entropy*, *Chi-Square* dan *Arithmetic Mean*). Rekomendasi metode dari penelitian tersebut adalah *Classification and Regression Tree* (CART) dengan akurasi mencapai 99.9% dan lebih efisien dibanding dengan metode lain nya.

Penelitian ini bertujuan untuk memodelkan traffic penggunaan *proxy* anonim dengan algoritme *Expectation Maximization*. Algoritme *Expectation Maximization* atau yang biasa disebut dengan algoritme EM adalah suatu algoritme yang digunakan untuk menemukan *mixture of Gaussian* yang dapat memodelkan dataset (Dempster et al. 1977). EM adalah salah satu algoritme *clustering* yang berdasarkan model, dimana pendekatannya menggunakan model yang ada untuk mengelompokkan dan mengoptimalkan kecocokan antara data dengan model.

2. METODE

Metode yang digunakan pada penelitian ini dimulai dari pengumpulan data traffic internet, kemudian data tersebut dilakukan praproses untuk diekstraksi atribut-atribut nya kemudian atribut-atribut tersebut diseleksi berdasarkan fitur yang berhubungan dengan penelitian. Untuk mendapatkan pola, maka fitur yang telah terpilih dilakukan klasifikasi dengan algoritme EM. Hasil dari *classifier* atau model tersebut diukur akurasi nya sehingga bisa ditarik kesimpulan dari hasil tersebut. Adapun tahapan-tahapan penelitian yang akan dilakukan seperti ditunjukkan Gambar 1.



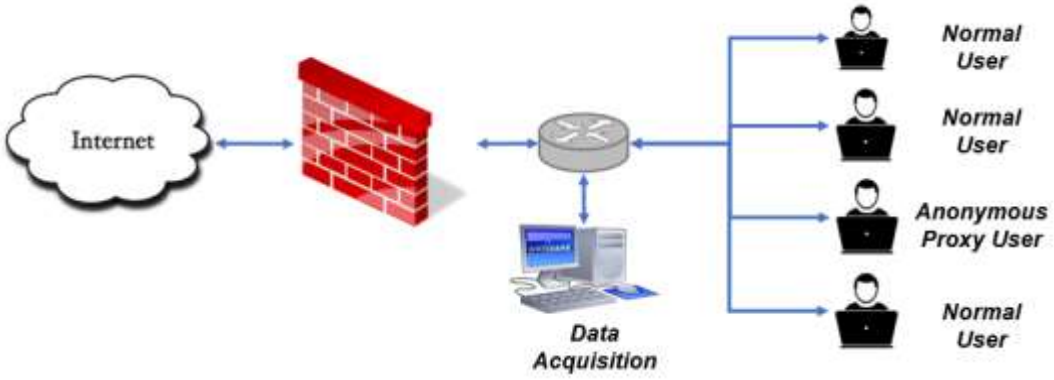
Gambar 1 Tahapan penelitian

2.1 *Traffic Internet*

Traffic internet merupakan aliran paket data yang lewat melalui jaringan internet. Data *traffic* yang digunakan pada penelitian ini merupakan paket data yang mengandung penggunaan *proxy* anonim, dimana ketika melakukan akses ke situs-situs tertentu dilakukan melalui perantara situs *proxy* anonim.

2.2 Akuisisi Data

Akuisisi data adalah proses pengambilan *raw data* pada jaringan internet yang berisikan semua informasi mengenai aktivitas pada *traffic internet*. Pada tahapan akuisisi data, penggunaan jaringan internet oleh pengguna dalam satu jaringan direkam menggunakan satu komputer melalui aplikasi Wireshark. Ilustrasi proses akuisisi data dalam satu jaringan diperlihatkan pada Gambar 2.



Gambar 2 Proses akuisisi data

2.3 Seleksi Fitur

Pada tahapan seleksi fitur menggunakan dua tahap, tahap pertama seleksi fitur manual dimana fitur-fitur yang tidak memiliki nilai observasi dihapus, tahap kedua yaitu seleksi fitur menggunakan algoritme Boruta. Boruta merupakan algoritme seleksi fitur yang bekerja sebagai algoritme pembungkus disekitar random forest (Kursa dan Rudnicki 2010). Boruta secara iteratif membandingkan atribut asli dengan atribut bayangan (yaitu data acak dari salinan semua atribut). Atribut yang memiliki kepentingan lebih rendah daripada atribut bayangan ditandai sebagai *rejected* dan dihapus dari sistem. Di sisi lain, atribut yang memiliki kepentingan lebih tinggi daripada atribut bayangan ditandai sebagai *confirmed*. Atribut bayangan diciptakan kembali pada setiap iterasi. Algoritme akan berhenti saat hanya tersisa atribut *confirmed*, atau ketika mencapai batas dari iterasi yang ditentukan.

2.4 *Data Labeling*

Agar memudahkan algoritme klasifikasi dalam menentukan kelas dan pembagian data, selanjutnya data yang telah melewati hasil praproses diberikan label kelas normal dan kelas *proxy*. Pemberian label pada paket data dengan cara menyaring pada fitur *Destination* ke alamat ip dari situs yang diakses. Ketika situs yang diakses merupakan *web proxy* anonim, maka label paket tersebut masuk ke kelas *proxy*.

2.5 *Balancing Data*

Dataset penelitian yang telah dibersihkan pada tahap praproses terkadang tidak seimbang jumlah antar kelasnya. Dalam klasifikasi, satu set data dikatakan tidak seimbang ketika jumlah data yang mewakili satu kelas lebih kecil dibanding dengan kelas lain (Galar et al. 2011). Kelompok kelas data yang lebih sedikit dikenal dengan kelompok minoritas, kelompok kelas data yang lainnya disebut dengan kelompok mayoritas (Siringoringo R 2018). Teknik *resampling* dapat dikategorikan menjadi tiga kelompok. Metode *undersampling*, yang membuat subset dari kumpulan data asli dengan menghilangkan beberapa data (biasanya kelas mayoritas); metode *oversampling*, yang menciptakan *superset* kumpulan data asli dengan mereplikasi beberapa data atau membuat contoh baru

dari yang sudah ada; dan metode *hybrid* yang menggabungkan kedua metode pengambilan sampel (Galar et al. 2011).

2.6 Pembagian Data

Set data yang telah melewati tahapan balancing data akan dibagi menjadi data latih dan data uji. Data latih digunakan untuk membentuk sebuah model oleh algoritme EM. Model ini merupakan representasi pengetahuan yang akan digunakan untuk prediksi kelas data baru yang belum pernah ada. Data uji digunakan untuk mengukur sejauh mana model berhasil melakukan klasifikasi dengan benar.

2.7 Algoritme EM

Klasifikasi merupakan proses pemodelan yang menggambarkan dan membedakan kelas data dengan tujuan agar dapat digunakan untuk memprediksi kelas dari suatu objek yang tidak diketahui label kelasnya (Han et al. 2012). Algoritme EM adalah salah satu algoritme yang digunakan untuk klasifikasi atau pengelompokan data, pertama kali diperkenalkan oleh Dempster, Laird, dan Rubin pada tahun 1977. EM merupakan salah satu metode untuk menemukan estimasi maximum likelihood dari sebuah dataset dengan distribusi tertentu. EM termasuk algoritme partitional yang berbasiskan model yang menggunakan perhitungan probabilitas, bukan jarak seperti umumnya algoritme clustering yang lainnya (Safuan et al. 2015).

Sesuai namanya, ada 2 proses utama dalam algoritme ini, yaitu proses expectation (E-step), yaitu fungsi untuk memperkirakan evaluasi likelihood berdasarkan beberapa parameter yang ada, dan proses maximization (M-step), yaitu untuk memaksimalkan nilai likelihood yang ditemukan pada proses E. Kedua proses ini digunakan untuk menentukan distribusi data untuk proses E pada perulangan selanjutnya.

2.8 Validasi Model

Validasi model merupakan proses implementasi atau pengujian antara data uji dengan model. Model yang dibangun berdasarkan informasi yang akurat dan diverifikasi seperti yang diharapkan maka model tersebut dapat dikatakan valid (Law dan Kelton 1991).

2.9 Evaluasi Model

Hasil dari validasi model dengan data uji selanjutnya akan dilakukan evaluasi perhitungan seperti presisi, *recall* dan akurasi melalui perhitungan menggunakan *confusion matrix*. *Confusion matrix* memberikan rincian terperinci dari kesalahan klasifikasi. Kelas prediksi ditampilkan di bagian atas matriks, dan kelas yang diamati di sisi kiri. Setiap sel berisi angka yang menunjukkan berapa banyak kasus yang sebenarnya dari pengamatan kelas yang ditetapkan oleh model ke kelas prediksi yang diberikan (Gorunescu 2011). Tingkat akurasi menunjukkan tingkat kebenaran pengklasifikasian data terhadap kelas yang sebenarnya. Semakin rendah nilai akurasi maka semakin tinggi kesalahan klasifikasi (Safuan et al. 2015).

3. HASIL DAN PEMBAHASAN

3.1 Hasil

3.1.1. Data Traffic Internet

Data hasil dari akuisisi dengan menggunakan aplikasi Wireshark dalam bentuk *packet capture* (pcap) dikonversi ke bentuk *Comma Separated Values* (CSV) dan dijadikan satu dataset sehingga menghasilkan sebanyak 763924 jumlah observasi dan 55 jumlah fitur. Praproses data dilakukan untuk menghapus fitur yang tidak memiliki

nilai observasi, memberikan nilai nol (0) pada observasi yang tidak memiliki nilai pada fitur-fitur tertentu, melakukan konversi nilai observasi dari nilai heksadesimal menjadi desimal, melakukan transformasi data dari string menjadi numerik dsb. Setelah dilakukan praproses, fitur source IP difilter pada alamat IP komputer/jaringan yang digunakan oleh user agar didapatkan data keluaran (out). Data out merupakan data dimana user melakukan *request* atau melakukan akses ke situs tertentu. Pada penelitian ini, data yang digunakan hanya data out saja, karena untuk melakukan pemblokiran *web proxy* anonim yang diakses oleh user cukup melalui paket data yang akan diakses. Total observasi yang digunakan pada data out adalah sebanyak 204207 data yang terdiri dari 29 fitur.

3.1.2. Data Labeling

Pemberian label pada data bertujuan agar memudahkan algoritme klasifikasi dalam menentukan kelas dan pembagian data (data latih dan data uji). Label yang benar sangat penting untuk kinerja metode klasifikasi traffic jaringan (Rezaei dan Liu 2019). Kelas yang diberikan hanya ada dua yaitu kelas *proxy* dan normal.

Web proxy anonim merupakan alamat situs penyedia layanan *proxy* anonim yang ada di internet. Sementara alamat IP jaringan adalah alamat IP yang digunakan pada domain *web proxy* anonim ketika diakses. Observasi data yang mengandung alamat IP jaringan proxy anonim pada fitur *destination* diberi label kelas *proxy*, selain itu diberi label normal.

Setelah dilakukan praproses dan pemberian label, maka data yang digunakan pada penelitian ini berjumlah 204207 jumlah observasi, yang terdiri dari 29 fitur dan 2 kelas. Distribusi data penelitian ditunjukkan pada Tabel 1.

Tabel 1 Distribusi dataset

Kelas	Jumlah data
Normal	185938
Proxy	18269
Jumlah	204 207

3.1.3. Seleksi fitur

Seleksi fitur merupakan tahapan untuk mengurangi besarnya dimensi data dengan menghilangkan fitur-fitur yang tidak relevan, sehingga proses klasifikasi dapat dilakukan dengan lebih efektif dan efisien. Setelah melewati proses ekstraksi fitur, berikutnya dilakukan analisis pada setiap fitur, yaitu dengan tujuan untuk mencari fitur mana saja yang memiliki informasi terpenting terhadap karakteristik dari traffic jaringan yang kemudian dapat dipilih sebagai penciri. Pada penelitian ini, metode seleksi fitur yang digunakan terdiri dari 2 tahap, yaitu tahap manual dan tahap seleksi fitur dengan algoritme Boruta. Pada tahap manual, dataset yang masih raw data dilakukan penghapusan fitur yang tidak memiliki nilai (kosong) pada kolom fitur nya dan menghapus fitur dengan tipe data *float* atau desimal pada observasi nya. Sehingga, hasil dari seleksi fitur manual berjumlah 29 fitur. Pada tahap seleksi fitur dengan algoritme Boruta, hasil dari seleksi fitur manual dilanjutkan dengan seleksi fitur menggunakan algoritme Boruta.

Hasil dari seleksi fitur menggunakan algoritme Boruta, dari 29 fitur diperoleh 26 fitur yang ditandai dengan *confirmed* yang kemudian akan digunakan dalam

klasifikasi. Hasil seleksi fitur tersebut terdapat beberapa fitur yang tidak dipilih (*rejected*) yaitu yang berada diantara *shadowMin* dan *shadowMax*.

3.1.4. Balancing Data

Teknik *resampling* secara luas digunakan untuk memecahkan masalah data yang tidak seimbang. Teknik ini dilakukan dengan mencoba menyeimbangkan data asli berdasarkan serangkaian algoritme *sampling* dengan menyesuaikan jumlah sampel dalam kelas yang berbeda, kemudian melatih data "seimbang" baru dengan mengadopsi algoritme klasifikasi (Syukron dan Subekti 2018). Teknik yang digunakan untuk menyeimbangkan data yaitu *undersampling*. Jumlah data pada kelas normal di penelitian ini jauh lebih banyak dibandingkan dengan kelas proxy. Sehingga pada kelas normal harus dihapus sebagian data agar memiliki jumlah yang seimbang dengan kelas proxy.

3.1.5. Pembagian Data

Sebelum dilakukan proses klasifikasi, data yang telah diseleksi fiturnya dilakukan proses pembagian data menjadi data latih dan data uji. Pembagian data antara data latih dan data uji dilakukan dengan mengacak dan membagi data menjadi 50%, 60%, 70%, 80% dan 90% pada data latih nya.

3.1.6. Klasifikasi dengan EM

Setelah data terbagi antara data latih dan data uji, selanjutnya data latih akan dilakukan pelatihan dengan algoritme EM sehingga akan menghasilkan model.

3.1.7. Validasi Model

Model yang terbentuk dari pelatihan dengan algoritme EM selanjutnya divalidasi akurasinya dalam menentukan kelas data dengan menggunakan data uji agar dapat diketahui kinerja algoritme dalam menentukan kelas data.

3.1.8. Evaluasi Model

Model yang telah divalidasi dengan data uji kemudian dievaluasi dengan menggunakan *confusion matrix* sehingga dapat dilakukan perhitungan akurasi, presisi dan *recall* dari model tersebut. Evaluasi dilakukan untuk mengukur keakuratan model terhadap data uji. Tahap evaluasi model dilakukan dengan membandingkan antara data hasil prediksi dengan data yang sebenarnya pada data uji. Hasil perhitungan evaluasi model terhadap data uji ditunjukkan pada Tabel 2.

Tabel 2 Hasil evaluasi model (dalam %)

Data Latih	Data Uji	Akurasi	Presisi	Recall
50	50	71.22	58.90	83.80
60	40	70.90	59.18	84.34
70	30	68.65	63.67	78.61
80	20	68.19	52.39	84.52
90	10	70.66	50.24	89.84

Evaluasi model dilakukan dengan pembagian antara data latih dengan data uji yang masing-masing model hasil pembagian data diuji sebanyak 10 kali perulangan. Hasil perulangan selanjutnya di rata-rata sehingga didapatkan nilai akurasi, presisi dan *recall* model untuk masing-masing pembagian data.

3.2 Pembahasan

Hasil perhitungan evaluasi model pada Tabel 2 menunjukkan bahwa pemodelan dengan menggunakan algoritme EM memiliki akurasi tertinggi pada pembagian data sebesar 50% data latih dan 50% data uji yaitu 71.22%. Selain perhitungan akurasi, untuk mengukur kinerja suatu model klasifikasi juga dihitung presisi dan *recall*. Pada model dengan pembagian data 50% didapati tingkat ketepatan antara kelas yang terklasifikasi benar (presisi) adalah sebesar 58.90% dan kemampuan model untuk memanggil kembali kelas yang terklasifikasi benar (*recall*) adalah sebesar 83.80%.

4. KESIMPULAN

Penggunaan data jaringan internet ketika melakukan akses ke web proxy anonim dapat dimodelkan dengan algoritme EM. Model hasil pelatihan dengan algoritme EM mampu mengidentifikasi kelas dengan akurasi tertinggi sebesar 71.22% pada data latih dengan proporsi 50% data, namun tidak cukup baik ketika model diminta untuk memberikan informasi yang diminta oleh pengguna (presisi) yaitu sebesar 58.90%, hal tersebut disebabkan masih terdapat nya data yang cukup besar melakukan salah klasifikasi pada *False Positif* (FP) dan *False Negatif* (FN). Kemudian untuk memanggil kembali kelas yang terklasifikasi benar (*recall*), model mampu melakukan *recall* sebesar 83.80%. Hasil tersebut memberikan indikasi bahwa metode ini dapat melakukan klasifikasi terhadap *web proxy* anonim.

5. DAFTAR PUSTAKA

Alshammari R, Zincir-Heywood AN. 2011. Can encrypted traffic be identified without port numbers, IP addresses and payload inspection? Comput Networks. 55(6):1326–1350. doi:10.1016/j.comnet.2010.12.002.

Bar-Yanai R, Langberg M, Peleg D, Roditty L. 2010. Realtime classification for encrypted traffic. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 6049 LNCS(46109):373–385. doi:10.1007/978-3-642-13193-6_32.

Bing L. 2009. Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. Second Edi. (NY): Springer Heidelberg Dordrecht London New York.

Brozycki J. 2008. Detecting and preventing anonymous proxy usage. SANS Inst.

Cha S, Kim H. 2017. Detecting encrypted traffic: A machine learning approach. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 10144 LNCS:54–65. doi:10.1007/978-3-319-56549-1_5.

Chakravarty S, Portokalidis G, Polychronakis M, Keromytis AD. 2015. Detection and analysis of eavesdropping in anonymous communication networks. Int J Inf Secur. 14(3):205–220. doi:10.1007/s10207-014-0256-7.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc. 39(1):1–38. doi:10.1177/019262339101900314.

Erman J, Mahanti A, Arlitt M. 2006. Internet traffic identification using machine learning. GLOBECOM - IEEE Glob Telecommun Conf. doi:10.1109/GLOCOM.2006.443.

Fashoto SG, Adekoya A, Owolabi O, Tomori R, Ogunleye O, Adediran S. 2016. Development of an identity management system for a web proxy server in a tertiary institution using anonymity technology. Int J Phys Sci. 11(13):157–167. doi:10.5897/ijps2016.4482.

Fielding R, Gettys J, Mogul JC, Frystyk H, Masinter L, Leach P, Berners-Lee T. 1999. Hypertext Transfer Protocol -- HTTP/1.1.

Galar M, Fernandez A, Barrenechea E, Bustince H. 2011. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. Inst Electr Electron Eng.:1–22. doi:10.1109/TSMCC.2011.2161285.

Gorunescu F. 2011. Data Mining Concepts, Models and Techniques. Springer-Verlag Berlin Heidelberg.

Gulyás G, Schulcz R, Imre S. 2008. Comprehensive analysis of Web privacy and anonymous Web browsers: Are next generation services based on collaborative filtering? CEUR Workshop Proc. 362.

Han J, Kamber M, Pei J. 2012. Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems). Third. Morgan Kaufmann.

Hu Z. 2015. Initializing the EM Algorithm for Data Clustering and Sub-population Detection.
Kohavi R, John GH. 1997. Wrappers for feature subset selection. Artif Intell. 97:273–324.
Kursa MB, Rudnicki WR. 2010. Feature Selection with the Boruta Package. J Stat Softw. 36(11).

Law AM, Kelton DW. 1991. Simulation Modelling & Analysis. (NY): McGraw-Hill, Inc.
Mc Keague J. 2013. Detecting Anonymous Proxy Usage. University of Ulster at Magee.
McGregor A, Hall M, Lorier P, Brunskill J. 2004. Flow Clustering Using Machine Learning Techniques. :205–214. doi:10.1007/978-3-540-24668-8_21.

Mckeague J, Curran K. 2018. Detecting the Use of Anonymous Proxies. Int J Digit Crime Forensics. 10(2). doi:10.4018/IJDCF.2018040105.

Miller S, Curran K, Lunney T. 2015. Securing the internet through the detection of anonymous proxy usage . World Congr Internet Secur.

Miller S, Curran K, Lunney T. 2016. Traffic Classification for the Detection of Anonymous Web Proxy Routing. Int J Inf Secur Res. 5(1):538–545. doi:10.20533/ijisr.2042.4639.2015.0061.

Neto MÂS. 2013. Traffic Classification Based on Statistical Tests for Matching Empirical Distributions of Lengths of IP packets.

Rezaei S, Liu X. 2019. Deep Learning for Encrypted Traffic Classification: An Overview. IEEE Commun Mag. 57(5):76–81. doi:10.1109/MCOM.2019.1800819.

Siringoringo R. 2018. Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor. J ISD. 3(1):44–49.

Safuan, Wahono RS, Supriyanto C. 2015. Penanganan Fitur Kontinyu dengan Feature Discretization Berbasis Expectation Maximization Clustering untuk Klasifikasi Spam Email Menggunakan Algoritma ID3. J Intell Syst. 1(2):148–155.

Syukron A, Subekti A. 2018. Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit. J Inform. 5(2):175–185. doi:10.31311/ji.v5i2.4158.

Yengi YK, Karayel M, Omurca Sİ. 2015. An Alternative Method for Sentiment Classification with Expectation Maximization and Priority Aging. Int J Sci Res Inf Syst Eng. 1(2):91–96.