ANALISIS MALICIOUS URL PADA FILE MENGGUNAKAN METODE K-MEANS CLUSTERING BERBASIS HOST-BASED FEATURE EXTRACTION

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

OLEH :

MUHAMMAD IMAM RAFI
09011281823065

JURUSAN SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2023

LEMBAR PENGESAHAN

ANALISIS MALICIOUS URL PADA FILE MENGGUNAKAN METODE K-MEANS CLUSTERING BERBASIS HOST-BASED FEATURE EXTRACTION

TUGAS AKHIR

Program Studi Sistem Komputer

Jenjang S1

Oleh

MUHAMMAD IMAM RAFI

09011281823065

Indralaya, 8 Mei 2023

Mengetahui,

Ketua Jurusan Sistem Komputer	Pembimbing Tugas Akhir
Dr. Ir. H. Sukemi, M.T.	**Ahmad Heryanto, S. Kom, M.T**
NIP. 196612032006041001	**NIP. 198701222015041002**

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada :

Hari : Jum'at

Tanggal : 14 April 2023

Tim Penguji :

1. **Ketua** : **Sarmayanta Sembiring, M.T.**
2. **Sekretaris** : **Nurul Afifah, M.Kom.**
3. **Penguji** : **Ahmad Fali Oklilas, M.T.**
4. **Pembimbing I** : **Ahmad Heryanto, S.Kom., M.T.**

Mengetahui, 9/5/23

Ketua Jurusan Sistem Komputer

UNIVERSITAS SRIWIJAYA FAKULTAS ILMU KOMPUTER JURUSAN SISTEM KOMPUTER

Dr.Ir.H. Sukemi, M.T.

NIP. 196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Muhammad Imam Rafi

NIM : 09011281823065

Judul : Analisis Malicious Url Pada File Menggunakan Metode K-Means Clustering Berbasis Host-Based Feature Extraction

Hasil Pengecekan Software *iThenticate/Turnitin* : 14%

Menyatakan bahwa laporan tugas akhir saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsru penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari universitas sriwijaya

Demikian, pernyatan ini saya buat dalam keadaan sadar dan tanpa paksaan dari siapapun.

Indralaya, Mei 2023

Muhammad Imam Rafi.

NIM.09011281823065

HALAMAN PERSEMBAHAN

Ku persembahkan Skripsi ini kepada kedua Orang Tua dan Adikku Tercinta

Kepada ayahanda Nazaruddin Effendi dan Ibunda Isro Aini yang telah terus memberikan semangat dan motivasi kepada ku dalam menyelesaikan skripsi ini. Dan Adikku M. Nata Yasa dan Nayla Sakinah yang selalu memberikan dukungan dan motivasi.

"Setetes keringat orangtuaku seribu langkahku untuk maju"

Ku Persembahkan Skripsi Ini Untuk Yang Selalu Bertanya:

"kapan skripsimu selesai ?"

"kapan kamu lulus ?"

Terlambat lulus atau lulus tidak tepat waktu bukanlah sebuah kejahatan, bukan pula sebuah aib. Alangkah kerdilnya jika mengukur kecerdasan seseorang hanya dari siapa yang paling cepat lulus. Bukankah sebaik – baiknya skripsi adalah skripsi yang selesai ?

Karena mungkin ada suatu hal dibalik terlambatnya mereka lulus, dan percaya alasan saya disini merupakan alasan yang sepenuhnya baik.

"Hidup itu bukan berlomba dengan orang lain.

Berlombalah dengan diri sendiri.

Menjadi versi terbaik dari versi diri kamu sendiri.

Mulailah dari selalu mencoba menjadi lebih baik dari hari kemarin."

-Christina Lie-

"*Semua Orang Memiliki Cerita Dan Jalan Hidupnya masing – masing Maka Janganlah Kamu Bandingkan Cerita dan Jalan Hidupmu dengan Orang Lain, Karena Setiap Makhluk Punya Ujian Dan Nikmatnya Masing – Masing"*

KATA PENGANTAR

Assalamu'alaikum Wr.Wb.

Puji syukur atas kehadirat Allah SWT yang telah memberikan karunia dan rahmat-Nya, sehingga penulis dapat menyelesaikan penulisan Tugas Akhir dengan judul "**Analisis Malicious *Url* pada *File* Menggunakan Metode *K-Means Clustering* Berbasis *Host-Based Feature Extraction*".**

Dalam pelaksanaan dan penulisan Tugas Akhir ini tidak akan berhasil tanpa adanya bantuan dari berbagai pihak yang telah membantu penulis berupa bantuan spiritual berupa do'a dari kedua orangtua dan keluarga besar dari penulis, serta bantuan berupa dukungan moral seperti bimbingan, mental, dan juga nasihat kepada penulis agar dalam pelaksanaan dan penulisan Tugas Akhir ini dapat terselesaikan. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih yang sebesar – besarnya kepada :

1. Allah Subhanahu Wa Ta'ala yang telah memberikan berkah, karunia dan hidayah-Nya yang tidak terhitung.
2. Ayah dan Ibu serta adik – adikku yang telah memberikan do'a, restu dan dukungan yang sangat besar selama penulis berjuang dalam menyelesaikan perkuliahan ini dan Tugas akhir ini.
3. Bapak Jaidan Jauhari, M.T., selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Bapak Dr. Ir. H. Sukemi, M.T., selaku Ketua Jurusan Sistem Komputer Fakultas Ilmu Komputer.
5. Bapak Ahmad Heryanto, S.Kom, M.T, selaku Dosen Pembimbing Tugas Akhir dan Dosen Pembimbing Akademik yang telah berkenan meluangkan waktunya untuk membimbing, memberikan saran dan motivasi terbaik kepada penulis dalam menyelesaikan Tugas Akhir ini.
6. Bapak Tri Wanda Septian, S.Kom, M.Sc., yang telah memberikan arahan, masukan, serta saran kepada penulis dan kepada Tim Riset Malicious URL.
7. Mbak Renny Virgasari selaku Admin Jurusan Sistem Komputer yang telah membantu penulis dalam hal – hal administrasi.

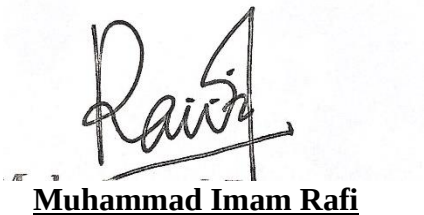
8. Kepada Keluarga Besar Arief Husin dan Ali Imron, yang telah memberikan bantuan berupa do'a, dukungan, dan dorongan semangat kepada penulis dalam menyelesaikan Tugas Akhir ini.
9. Kepada Teman saya Dimas Aditya Kristianto. S.Kom yang telah membantu penulis dalam riset untuk Tugas Akhir. Dan teman lainnya, Muhammad Farhan Alharits, Muhammad Realdi, M.Taufik, Alifah Fidela, Indah Cahya Resti, Rizki Valen Mafaza, Arif Tumpal Leonardo Siaturi. S.Kom, Ades Harafi Duri, Rachmawati Dwinanti Putri, Novi Yuningsih, Muhammad Furqon Rabbani. S.Kom, dan Teman seperjuangan dari SK 2018, SK Reguler B Indralaya, teman grup kacamata dan Fresh yang telah banyak memberikan bantuan dan menghibur penulis selama menjalani perkuliahan ini.
10. Dan kepada orang – orang baik yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan semangat dan do'a selama pelaksanaan dan penulisan Tugas Akhir ini.
11. Dan kepada semua civitas akademika Universitas Sriwijaya dan nama almamater Universitas Sriwijaya, penulis ucapkan terima kasih.

Dalam penulisan Tugas Akhir ini, penulis menyadari bahwa masih terdapat banyak kekurangan. Maka dari itu, penulis memohon maaf dan menerima kritik dan saran sebagai bahan evaluasi penulis untuk di masa mendatang. Harapan penulis, Tugas Akhir ini dapat bermanfaat dan berguna bagi setiap yang membaca Tugas Akhir ini.

Wassalamu'alaikum Wr.Wb.

Indralaya, Mei 2023

Penulis



Muhammad Imam Rafi

NIM. 09011281823065

ANALYSIS MALICIOUS URL IN FILE USING K-MEANS CLUSTERING METHOD AND HOST – BASED FEATURE EXTRACTION

Muhammad Imam Rafi (09011281823065)
Department of Computer System, Faculty of Computer Science, University of Sriwijaya
Palembang, Indonesia
Email : imamrafi45@gmail.com

ABSTRACT

The attacks or threats faced by internet users today have various types of attacks. These attacks are attacks in the form of phishing, malware, spyware, and ransomware. One of the most effective means of cyber attack carried out by attackers is by using URLs. URL (Uniform Resource Locator) is an address used to find the location of a file on the Internet. This makes URLs used as a method for carrying out cyber attacks referred to as Malicious URLs. Malicious URLs or dangerous sites on the internet contain a lot of content in the form of spam, phishing, which is used to initiate attacks. In this study, generate a URL dataset with URL features in the form of DNS records from URLs that will be used as data in clustering with K - Means. And produce a visualization of data clustering results with K - Means using a value of k = 2, namely in the form of benign clusters and malicious URLs. And analyze the visualization results of clustering with K - Means using the clustering validation test using the Silhouette Score with a result of 72.62% for k=2. In this study, generate model validation by training the URL dataset on machine learning and applying Hypeparameter tuning so that the performance results for each cluster are benign (0) 85% precision, 97% Recall, 91% F1-Score, and malicious (1) precision clusters. 96%, Recall 83%, F1-score 89%, and the accuracy of the model used is 89.94%.

Keyword : URL, Uniform Resource Locator, Malicious URL, Host – Based Feature Extraction, K – Means Clustering

ANALISIS MALICIOUS URL PADA FILE MENGGUNAKAN METODE K-MEANS CLUSTERING BERBASIS HOST-BASED FEATURE EXTRACTION

Muhammad Imam Rafi (09011281823065)
Department of Computer System, Faculty of Computer Science, University of Sriwijaya
Palembang, Indonesia
Email : imamrafi45@gmail.com

ABSTRAK

Serangan atau ancaman yang dihadapi oleh pengguna internet saat ini memiliki tipe serangan yang berbagai macam. Serangan – serangan tersebut adalah serangan berupa phising, malware, spyware, dan ransomware. Salah satunya sarana serangan siber yang sangat efektif dilakukan oleh penyerang yaitu dengan menggunakan URL. URL (Uniform Resource Locator) adalah sebuah alamat yang digunakan untuk menemukan lokasi dari sebuah file yang berada di Internet. Hal ini membuat URL digunakan sebagai salah satu metode untuk melakukan serangan siber disebut sebagai Malicious URL. Malicious URL atau situs berbahaya di internet memuat banyak konten berupa spam, phising, yang digunakan untuk memulai serangan. Pada penelitian ini, menghasilkan sebuah dataset URL dengan fitur URL berupa DNS Record dari URL yang akan digunakan sebagai data dalam melakukan clustering dengan K – Means. Dan menghasilkan sebuah visualisasi dari data hasil clustering dengan K – Means dengan menggunakan nilai k=2 yaitu berupa kluster benign dan malicious URL. Dan melakukan analisis terhadap hasil visualisasi dari clustering dengan K – Means dengan menggunakan uji validasi clustering dengan menggunakan Silhouette Score dengan hasil 72,62 % untuk k=2. Pada penelitian ini, menghasilkan validasi model dengan melakukan trainning dataset URL pada machine learning dan menerapkan tuning Hypeparameter sehingga hasil performa setiap kluster yaitu benign (0) presisi 85%, Recall 97%, F1-Score 91%, dan kluster malicious (1) presisi 96%, Recall 83%, F1-score 89%, dan hasil akurasi dari model yang digunakan yaitu dengan nilai 89.94%.

Kata Kunci : URL, Uniform Resource Locator, Malicious URL Host – Based Feature Extraction, K – Means Clustering.

DAFTAR ISI

LEMBAR PENGESAHAN ..... **i**
HALAMAN PERSETUJUAN ..... **ii**
HALAMAN PERNYATAAN ..... **iii**
HALAMAN PERSEMBAHAN ..... **iv**
KATA PENGANTAR ..... **v**
ABSTRACT ..... **vii**
ABSTRAK ..... **viii**
DAFTAR ISI ..... **ix**
DAFTAR GAMBAR ..... **xi**
DAFTAR TABEL ..... **xiii**
BAB I **1**
PENDAHULUAN **1**
1.1. Latar Belakang 1
1.2. Rumusan Masalah 4
1.3. Batasan Masalah 4
1.4. Tujuan 5
1.5. Manfaat 5
1.6. Metodologi Penelitian 5
1.7. Sistematika Penulisan 6
BAB II **8**
TINJAUAN PUSTAKA **8**
2.1. Pendahuluan 8
2.2. URL (Uniform Resource Locator) 15
2.3. Malicious URL 16
2.4. Host – Based Feature Extraction 16
2.5. *Artificial Intelligence* 19
2.5.1. *Machine Learning* 19
2.5.2. *K – Means Clustering* 20
2.6. *Confusion Matrix* 25
2.7. Metode *Elbow* 27
2.8. *Silhouette Coefficient* 29
2.9. Exploratory Data Analysis 30
BAB III **32**

METODOLOGI PENELITIAN **32**
3.1. Pendahuluan 32
3.2. Kerangka Kerja Penelitian 32
3.3. Kebutuhan Perangkat 34
3.4. Skenario Eksperimen 35
3.5. Skenario Riset 36
3.6. Pengolahan Data 36
3.6.1. Pengumpulan RAW data (*.pdf*) 37
3.6.2. Ekstraksi file pdf 37
3.6.3. Ektraksi fitur URL 37
3.7. Seleksi Atribut/Fitur Dataset dari Hasil Ektraksi URL 41
3.8. Menghapus URL yang *Duplicates* 42
3.9. Pengujian Clustering dengan K – Means 43
3.10. Analisis dan Validasi Hasil 46
3.10.1. *Silhouette Coefficient* 46
3.10.2. Exploratory Data Analysis 47
3.10.3. Tuning Hyperparameter 47
BAB IV **49**
HASIL DAN ANALISA **49**
4.1. Pendahuluan 49
4.2. Hasil Ekstraksi File (.pdf) 49
4.3. Hasil Ekstraksi Fitur URL 50
4.4. Hasil Seleksi Fitur URL berupa *Timestamp* 61
4.5. Hasil Menghapus URL yang Sama (*Duplicates*) 62
4.6. Clustering dengan K – Means 63
4.7. *Silhouette Score* 65
4.8. Exploratory Data Analysis 66
4.9. Tuning Hyperparameter 67
4.10. Hasil Validasi & Confusion Matrix 68
BAB V **70**
KESIMPULAN DAN SARAN **70**
5.1. Kesimpulan 70
5.2. Saran 71
DAFTAR PUSTAKA **72**

DAFTAR GAMBAR

Gambar 2.1. Komponen dari URL 15
Gambar 2.2. Flowchart K-Means menggunakan Euclidean Distance..... 22
Gambar 2.3. Flowchart K-Means menggunakan Manhattan Distance..... 23
Gambar 2.4. Flowchart K – means Clustering 25
Gambar 2.5. Confusion Matrix[35] 25
Gambar 2.6. Flowchart Optimalisasi K – Means Clustering menggunakan Elbow 28
Gambar 2.7. Silhouette Coefficient 29
Gambar 2. 8. Exploratory Data Analysis Heatmap 31
Gambar 3.1. Kerangka Kerja Penelitian..... 33
Gambar 3.2. Skenario Eksperimen Malicious URL 35
Gambar 3.3. Skenario riset malicious URL..... 36
Gambar 3.4. Diagram Pengolahan Dataset URL..... 40
Gambar 3.5. Flowchart Drop Fitur Timestamp 41
Gambar 3.6. Fitur berupa data timestamp 42
Gambar 3.7. Flowchart hapus URL sama 43
Gambar 3.8. Flowchart Clustering K-Means 43
Gambar 3.9. Pseudocode K – Means..... 44
Gambar 3.10. Data sebelum di clustering..... 45
Gambar 3.11. Data yang telah di clustering 45
Gambar 3.12. Pseudocode Silhouette Coefficient 46
Gambar 4.1. Hasil Ektraksi File PDF 49
Gambar 4.2. Fitur Obj 50
Gambar 4.3. Fitur Age..... 51
Gambar 4.4. Fitur Host..... 51
Gambar 4.5. Fitur TTL 51
Gambar 4.6. Fitur Connection Speed 52
Gambar 4.7. Fitur is_live..... 52
Gambar 4.8. Fitur Total_Updates..... 52
Gambar 4.9. Fitur Intended_Life_Span..... 53

Gambar 4. 10. Fitur life_remaining 53
Gambar 4.11. Fitur avg_updates_days 53
Gambar 4.12. Fitur reg_country 54
Gambar 4.13. Fitur days_since_last_seen 54
Gambar 4.14. Fitur days_since_first_seen 54
Gambar 4. 15. Fitur num_open_ports, isp, num_subdomains, open_ports 55
Gambar 4.16. Fitur registration_date 55
Gambar 4.17. Fitur expiration_date 56
Gambar 4.18. Fitur days_since_first_seen 57
Gambar 4.19. Fitur days_since_last_seen 57
Gambar 4.20. Fitur last_updates_dates 58
Gambar 4.21. Fitur first_seen 59
Gambar 4.22. Fitur last_seen 59
Gambar 4.23. Hasil pengecekan dengan VirusTotal Malicious URL 60
Gambar 4.24. Hasil pengecekan dengan VirusTotal Benign URL 61
Gambar 4.25. Isi fitur timestamp 61
Gambar 4.26. Hasil Seleksi Fitur Timestamp 62
Gambar 4.27. Hasil Ekstraksi file .pdf dengan url yang sama 62
Gambar 4.28. Hasil Penghapusan URL yang sama 63
Gambar 4.29. Penyebaran data sebelum clustering dengan mengseleksi fitur timestamp 64
Gambar 4.30. *Metode* Elbow *dengan Jumlah* Cluster 64
Gambar 4.31. Hasil Clustering dengan K – Means 65
Gambar 4.32. Hasil Pengukuran Euclidean Distance diantara 2 buah cluster 65
Gambar 4.33. Silhouette Plot dan Silhouette Score 66
Gambar 4. 34. Hasil Heatmap Korelasi Setiap Fitur 67
Gambar 4.35. Confusion Matrix 68

DAFTAR TABEL

Tabel 2. 1 Penelitian terdahulu yang menjadi rujukan 8
Tabel 2.2. Host - Based Features[28] 18
Tabel 3.1. Spesifikasi perangkat keras 34
Tabel 3.2. Spesifikasi Perangkat Lunak 34
Tabel 3.3. Host – Based Feature Extraction 37
Tabel 3. 4. Spesifikasi Paramater yang diujikan 47
Tabel 3. 5. Spesifikasi Parameter Grid Search Cross Validation 47
Tabel 4.1. Jumlah Dataset URL yang berhasil diekstraksi dari File PDF 50
Tabel 4. 2. Spesifikasi dan Nilai Score Grid Search Cross Validation 68
Tabel 4. 3. Hasil Validasi Benign dan Malicious URL 69

BAB I

PENDAHULUAN

1.1. Latar Belakang

Serangan atau ancaman yang dihadapi oleh pengguna internet saat ini memiliki tipe serangan yang berbagai macam. Serangan – serangan tersebut adalah serangan berupa *phising*, *malware*, *spyware*, *spam,* dan *ransomware* [1] [2] [3]. Salah satunya sarana serangan siber yang sangat efektif dilakukan oleh penyerang yaitu dengan menggunakan URL, dimana hacker membuat suatu URL baru dengan menggunakan nama domain yang hampir sama dengan domain yang resmi atau legal dari website yang ditiru. Dengan begitu pengguna dari internet yang tidak terlalu memperhatikan URL yang akan mereka klik dapat membuat mereka terjebak dan masuk kedalam URL yang salah dan bukan masuk ke halaman resmi dari websitenya.

URL (Uniform Resource Locator) adalah sebuah alamat yang digunakan untuk menemukan lokasi dari sebuah file yang berada di Internet. URL memiliki dua komponen utama yaitu protocol indentifier merupakan indefikasi dari protokol yang digunakan oleh website, kedua, resource name yaitu menunjukkan spesifik dari IP address atau domain name itu berada dari website [4]. Hal ini membuat URL digunakan sebagai salah satu metode untuk melakukan serangan siber disebut sebagai Malicious URL. Malicious URL atau situs berbahaya di internet memuat banyak konten berupa spam, phising, yang digunakan untuk memulai serangan. Pengguna yang tidak curiga mengunjungi situs web tersebut dan menjadi pasien dari berbagai jenis penipuan, termasuk kerugian ekonomi, pencurian informasi pribadi (identitas, kartu kredit, dll.). URL berbahaya juga dapat disembunyikan di tautan unduhan yang dianggap aman dan dapat menyebar dengan cepat melalui berbagi file dan pesan di jaringan bersama [5].

Hal ini mendorong perkembangan dari malicious URL ini begitu cepat dan bervariasi, dengan menggunakan teknik deteksi tradisional tidak dapat dilakukan dengan baik dalam melakukan pengenalan dan klasifikasi, dikarenakan kurangnya kemampuan dalam melakukan deteksi terhadap malicious URL yang baru. Untuk

meningkatkan dalam hal kemampuan dalam mendeteksi malicious URL, maka machine learning digunakan untuk membantu dalam mengidenfikasi dan mendeteksi malicious URL dalam beberapa tahun terakhir ini [6]. Hal ini menjadi ancaman yang menakutkan bagi pengguna internet dikarenakan URL yang merupakan malicious sulit untuk dibedakan dengan URL yang legal atau resmi sehingga bagi pengguna internet yang masih awam akan kesulitan dalam mengenalinya. Menurut penelitian dari [6] serangan berupa malicious URL sangat sulit untuk deteksi karena masih sedikitnya teknologi yang digunakan dalam mendeteksi URL berbahaya ini.

Dengan sulitnya dalam melakukan analisis, identifikasi, klasifikasi dan deteksi pada malicious URL dan benign URL, membuat Serangan melalui malicious URL ini menjadi serangan yang sering digunakan oleh para hacker dalam menyerang pengguna internet dan mencuri informasi dari pengguna tersebut tanpa dicurigai oleh pengguna yang terkena serangan dari malicious URL. Malicious URL menjadi ancaman yang serius yang menakutkan, karena serangan ini tidak terlihat oleh pengguna dan kemudian mencuri informasi dari penggunanya dan bahkan bisa mengambil alih perangkat penggunanya untuk membuka celah agar bisa masuk ke perangkat penggunanya. Berdasarkan data pada 5 tahun terakhir ini, malicious URL terus mengalami peningkatan dengan pertambahan sebanyak 90% dari pengguna yang terdampak dari serangan ini [7].

Pada penelitian yang menggunakan metode blacklisting [8] menunjukkan hasil yang baik dengan tingkat false negatif yang tinggi dibandingkan ekspetasi dari tingkat false positif dan dengan error 5%. Tetapi metode ini memiliki kelamahan dalam melakukan deteksi terhadap URL yang baru muncul diluar dari blacklist yang ada, selain itu blacklisting tidak terlalu mampu dalam mengatasi malicious URL dalam jumlah besar. Dan penyerang dapat dengan mudah dalam memanipulasi sistem dengan sedikit melakukan modifikasi dari satu atau lebih komponen string dari URL dan URL yang tidak ada dalam blacklist maka URL tersebut tidak dicurigai sebagai malicious [9].

Upaya untuk mengatasi kekurangan dari metode sebelumnya adalah dikembangkannya pendekatan dengan menggunakan algoritma *machine learning*. Ada beberapa metode machine learning yang digunakan dalam melakukan deteksi,

klasifikasi, dan clustering, yaitu *Random Forest*, *Support Vector Machine*, *Decission Tree*, *K-nearest Neighbor*, dan *K-means* [10]. Metode Machine Learning memberikan kemudahan dalam melakukan fungsi prediksi yang digunakan untuk mengelompokkan URL sebagai malicious dan benign. Dan dengan pendekatan machine learning, dapat digunakan untuk menganalisis informasi dari URL dengan melakukan ekstraksi fitur yaitu seperti lexical based feature extraction, host based-feature extraction dan content -based feature extraction yang menghasilkan fitur masukan yang akan digunakan dalam melakukan analisis, identifikasi, klasifikasi dan deteksi pada malicious URL dan benign URL, seperti URL length, domain name length, IP address, host-name URL. Berikut ini merupakan hasil penelitian yang menggunakan machine learning dan feature extraction yang digunakan untuk mengidentifikasi malicious URL yang menjadi acuan utama dalam penelitian ini.

Pada penelitian [10] menggunakan fitur ekstraksi yaitu lexical feature dan host-based feature extraction. Penelitian tersebut menggunakan dua metode machine learning yaitu Random Forest dan SVM. Penelitian tersebut dengan menggunakan 10 iterasi dan menggunakan tiga pembagian rasio 60:40, 70:30, dan 80:20. Hasil penelitian ini, memiliki tingkat akurasi pada SVM lebih kecil dibandingkan Random Forest dan dengan hasil plot dari Random Forest dengan variasi model yang dipakai 82 – 90%.

Pada penelitian [11] menggunakan pendekatan hybrid deep-learning yang dinamakan URLdeepDetect yang digunakan untuk melakukan analisis dan klasifikasi untuk mendeteksi malicious URL. URLdeepdetect menggunakan mekanisme supervised dan unsupervised yaitu LSTM (Long Short-Term Memory) dan K-Means Clustering untuk klasifikasi URL. Pada penelitian ini mendapatkan hasil yang baik yaitu dengan 98.3% akurasi untuk LSTM dan 99.7% untuk K-Means Clustering. Dalam penelitian ini, hasil yang didapatkan baik tetapi pada metode K-means Clustering masih perlu ditingkatan kembali pada bagian presisi dan recall yang masih rendah dibandingkan metode LSTM.

Pada penelitian [12] dengan judul Analyzing Malicious URLs using a Threat Intelligence System, penelitian ini menggunakan pendekatan berupa clustering digunakan untuk menganalis malicious URL. Dengan menerapkan pemprosesan data secara ekstensif yang digunakan untuk tokenizing, sanitizing, dan

vectorizing pada Dataset URL. Penelitian ini juga mendemonstrasikan pendekatan dengan k-means clustering untuk mengelompokkan malicious URL dalam dataset. Dan dengan menggunakan metode elbow dalam seleksi model dan dengan k = 35, menghasilkan pendekatan yang baik dengan koefisien 0.383 untuk dataset yang mengandung 11.000 malicious URL. Dengan menerapkan model ini, berhasil mengindentifikasi lebih dari 80% dari URL pada dataset yang merupakan malicious URL. Penelitian ini masih belum sempurna dan masih dapat untuk ditingkatkan kembali hasilnya.

Dari penelitian terdahulu yang telah dijelaskan pada pembahasan sebelumnya dengan performa dan hasil dari metode yang digunakan, maka peneliti mengangkat judul Analisis Malicious *Url* Pada *File* Menggunakan Metode *K-Means Clustering* Berbasis *Host-Based Feature Extraction* dengan menggunakan dataset dari file pdf garuda yang diparser untuk mengambil URL yang ada didalam file pdf.

1.2. Rumusan Masalah

Berdasarkan penjelasan dari latar belakang yang telah disampaikan diatas, pada penelitian ini akan melakukan analisis malicious url pada file menggunakan metode k-means clustering berbasis host-based feature extraction. Berikut rumusan masalah pada penelitian ini:

1. Proses atau langkah – langkah yang dilakukan dalam mengekstrak fitur dari URL menggunakan Host – Based Feature Extraction.
2. Menerapkan proses clustering terhadap URL menggunakan K – Means dalam mengelompokkan URL Benign dan URL Malicious.
3. Menerapkan uji performa terhadap K – Means dalam melakukan clustering terhadap URL Benign dan URL Malicious.

1.3. Batasan Masalah

Adapun batasan masalah pada penelitian ini, sebagai berikut :

1. Pada penelitian yang dilakukan ini berfokus untuk melakukan analisis terhadap malicious URL dan benign URL.
2. Menerapkan Ekstraksi Fitur URL dengan Host-Based Feature Extraction.

3. Analisis dilakukan dengan menerapkan teknik dari metode K-means Clustering dalam membuat kluster pengelompokan berdasarkan fitur yang digunakan.

1.4. Tujuan

Adapun tujuan dari penelitian ini, sebagai berikut :

1. Menerapkan ekstraksi fitur terhadap malicious URL dan Benign URL untuk mendapatkan informasi sesuai dengan fitur yang digunakan untuk mengoptimalkan dalam melakukan proses analisis terhadap malicious URL dan Benign URL.
2. Menerapkan metode K-means Clustering dalam pengklusteran dari malicious URL dan Benign URL.
3. Melakukan analisis dan visualisasi dari malicious URL dan benign URL berdasarkan dataset dan hasil dari kluster yang dibentuk dengan metode K-means Clustering.

1.5. Manfaat

Adapun manfaat dari penelitian ini, sebagai berikut :

1. Menerapkan metode K-means Clustering dalam melakukan analisis terhadap malicious URL dan Benign URL untuk penelitian selanjutnya.
2. Menganalisis dan memvisualkan hasil dari malicious URL dan Benign URL dengan metode K-means Clustering
3. Mengoptimalkan hasil dari K-means Clustering dengan Host-Based Feature Extraction.

1.6. Metodologi Penelitian

Metodologi penelitian terdapat beberapa tahapan dalam penelitian ini, sebagai berikut :

1. Metode Studi Pustakan dan Literature

   Pada bagian ini peneliti mengumpulkan data mengenai cara maupun proses ekstraksi fitur dengan Host-Based Feature Extraction dan proses

clustering dengan metode K-means Clustering dari berbagai sumber ilmu dalam membantu pembuatan Tugas Akhir ini.

2. Metode Konsultasi

   Pada bagian ini, mengacu pada bagian – bagian yang sudah memiliki pengetahuan dan pemahaman yang baik untuk mengatasi masalah yang dihadapi saat peneliti menulis.

3. Metode Pengumpulan Data

   Pada bagian ini, mengumpulkan data mengenai Malicious URL dan Benign URL, berdasarkan skema ekstraksi fitur dari URL dan proses analisis dengan bantuan pengklusteran URL.

4. Metode Pengujian

   Pada langkah ini, pembuatan rancang kluster – kluster berdasarkan fitur yang telah diektraksi dari URL untuk mendapatkan visual dari dataset untuk memudahkan dalam melakukan analisis.

5. Metode Analisis dan Kesimpulan

   Pada langkah terakhir ini, hasil dari ekstraksi fitur dan hasil visual dari kluster yang telah dibuat berdasarkan dataset di analisis selanjutnya untuk dilakukan penarikan kesimpulan dari penelitian ini.

1.7. Sistematika Penulisan

Sistematikan penulisan pada penelitian ini sebagai berikut :

BAB I. PENDAHULUAN

Dalam bab I, membahas latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat pada penelitian, metode penelitian dan sistematika penulisan.

BAB II. TINJAUAN PUSTAKA

Dalam bab II, mencakup pembahasan teori mendasar dari URL, Malicious URL, Host-Based Feature Extraction, K-means Clustering dan teori yang relevan lainnya untuk tugas akhir ini.

BAB III. METODOLOGI

Dalam bab III, meliputi tahap peneliti yang kerjakan dengan melakukan ekstraksi fitur dari URL dan menerapkan metodologi penelitian tugas akhir.

BAB IV. HASIL DAN ANALISIS

Dalam bab IV, mencakup tahap peneliti untuk melihat visual dari hasil clustering dan analisis terhadap malicious URL dan benign URL.

BAB V. KESIMPULAN DAN SARAN

Dalam bab V, terdapat kesimpulan dan saran di tarik dari pembahasan sebelumnya dan membagikan beberapa masukan sebagai referensi.

DAFTAR PUSTAKA

[1] S. Angadi and S. Shukla, "Malicious URL Detection Using Machine Learning Techniques," *Lect. Notes Networks Syst.*, vol. 458, no. 6, pp. 657–669, 2022, doi: 10.1007/978-981-19-2894-9_50.

[2] B. Cui, S. He, P. Shi, and X. Yao, "Malicious URL detection with feature extraction based on machine learning," *Int. J. High Perform. Comput. Netw.*, vol. 12, no. 2, pp. 166–178, 2018, doi: 10.1504/ijhpcn.2018.094367.

[3] G. Palaniappan, S. Sangeetha, B. Rajendran, Sanjay, S. Goyal, and B. S. Bindhumadhava, "Malicious Domain Detection Using Machine Learning on Domain Name Features, Host-Based Features and Web-Based Features," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 654–661, 2020, doi: 10.1016/j.procs.2020.04.071.

[4] D. Sahoo, C. Liu, and S. C. H. Hoi, "Malicious URL Detection using Machine Learning: A Survey," vol. 1, no. 1, pp. 1–37, 2017, [Online]. Available: http://arxiv.org/abs/1701.07179.

[5] C. Do Xuan, H. D. Nguyen, and T. V. Nikolaevich, "Malicious URL detection based on machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 148–153, 2020, doi: 10.14569/ijacsa.2020.0110119.

[6] Y. Peng, S. Tian, L. Yu, Y. Lv, and R. Wang, "A Joint Approach to Detect Malicious URL Based on Attention Mechanism," *Int. J. Comput. Intell. Appl.*, vol. 18, no. 3, pp. 1–14, 2019, doi: 10.1142/S1469026819500214.

[7] T. Manyumwa, P. F. Chapita, H. Wu, and S. Ji, "Towards Fighting Cybercrime: Malicious URL Attack Type Detection using Multiclass Classification," *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, pp. 1813–1822, 2020, doi: 10.1109/BigData50022.2020.9378029.

[8] S. Sinha, M. Bailey, and F. Jahanian, "Shades of Grey: On the effectiveness of reputation-based blacklists," *3rd Int. Conf. Malicious Unwanted Software, MALWARE 2008*, pp. 57–64, 2008, doi: 10.1109/MALWARE.2008.4690858.

[9] J. Ispahany and R. Islam, “Detecting malicious COVID-19 URLs using machine learning techniques,” *2021 IEEE Int. Conf. Pervasive Comput. Commun. Work. other Affil. Events, PerCom Work. 2021*, pp. 718–723, 2021, doi: 10.1109/PerComWorkshops51409.2021.9431064.

[10] R. Patgiri, H. Katari, R. Kumar, and D. Sharma, *Empirical study on malicious URL detection using machine learning*, vol. 11319 LNCS. Springer International Publishing, 2019.

[11] S. Afzal, M. Asim, A. R. Javed, M. O. Beg, and T. Baker, “URLdeepDetect: A Deep Learning Approach for Detecting Malicious URLs Using Semantic Vector Models,” *J. Netw. Syst. Manag.*, vol. 29, no. 3, 2021, doi: 10.1007/s10922-021-09587-8.

[12] S. Nayak, D. Nadig, and B. Ramamurthy, “Analyzing Malicious URLs using a Threat Intelligence System,” *Int. Symp. Adv. Networks Telecommun. Syst. ANTS*, vol. 2019-Decem, pp. 2–5, 2019, doi: 10.1109/ANTS47819.2019.9118051.

[13] N. V. Kishan and V. S. Teja, “Classification of Phishing Website Based on URL Features,” vol. 7, no. 5, pp. 9–11, 2019.

[14] S. Kumi, C. Lim, and S. G. Lee, “Malicious url detection based on associative classification,” *Entropy*, vol. 23, no. 2, pp. 1–12, 2021, doi: 10.3390/e23020182.

[15] M. Aljabri *et al.*, “An Assessment of Lexical, Network, and Content-Based Features for Detecting Malicious URLs Using Machine Learning and Deep Learning Models,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3241216.

[16] J. S. Ambata, J. Gaurana, D. Jacinto, and J. De Goma, “Malicious URL Classification Using Extracted Features, Feature Selection Algorithm, and Machine Learning Techniques,” *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, no. 2016, pp. 2421–2429, 2021.

[17] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: Learning to detect malicious web sites from suspicious URLs,” *Proc. ACM*

SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 1245–1253, 2009, doi: 10.1145/1557019.1557153.

[18] S. Rustam, "Analisa Clustering Phising Dengan K-Means Dalam Meningkatkan Keamanan Komputer," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 175–181, 2018, doi: 10.33096/ilkom.v10i2.309.175-181.

[19] K. S. Ray and R. Kusshwaha, "Detection of Malicious URLs Using Deep Learning Approach," *Lect. Notes Networks Syst.*, vol. 163, no. 03, pp. 189–212, 2021, doi: 10.1007/978-981-15-9317-8_8.

[20] Q. Wang, L. Li, B. Jiang, Z. Lu, J. Liu, and S. Jian, *Malicious domain detection based on k-means and smote*, vol. 12138 LNCS. Springer International Publishing, 2020.

[21] V. Vundavalli, F. Barsha, M. Masum, H. Shahriar, and H. Haddad, "Malicious URL Detection Using Supervised Machine Learning Techniques," *ACM Int. Conf. Proceeding Ser.*, 2020, doi: 10.1145/3433174.3433592.

[22] D. Vaishnavi, S. Suwetha, Y. B. Jinila, R. Subhashini, and S. P. Shyry, "A comparative analysis of machine learning algorithms on malicious URL prediction," *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, no. Iciccs, pp. 1398–1402, 2021, doi: 10.1109/ICICCS51141.2021.9432138.

[23] R. Verma and A. Das, "What's in a URL: Fast feature extraction and malicious URL detection," *IWSPA 2017 - Proc. 3rd ACM Int. Work. Secur. Priv. Anal. co-located with CODASPY 2017*, pp. 55–63, 2017, doi: 10.1145/3041008.3041016.

[24] C. Johnson, B. Khadka, R. B. Basnet, and T. Doleck, "Towards detecting and classifying malicious urls using deep learning," *J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl.*, vol. 11, no. 4, pp. 31–48, 2020, doi: 10.22667/JOWUA.2020.12.31.031.

[25] R. Naresh, A. Gupta, and S. Giri, "Malicious URL detection system using combined SVM and logistic regression model," *Int. J. Adv. Res. Eng. Technol.*, vol. 11, no. 4, pp. 63–73, 2020, doi:

10.34218/IJARET.11.4.2020.008.

[26] M. Aldwairi and R. Alsalman, “MALURLs: Malicious URLs Classification System,” 2011, doi: 10.5176/978-981-08-8113-9_ita29.

[27] R. Kumar, X. Zhang, H. A. Tariq, and R. U. Khan, “Malicious URL detection using multi-layer filtering model,” *2016 13th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. ICCWAMTIP 2017*, vol. 2018-February, pp. 97–100, 2017, doi: 10.1109/ICCWAMTIP.2017.8301457.

[28] H. M. J. Khan, Q. Niyaz, V. K. Devabhaktuni, S. Guo, and U. Shaikh, “Identifying Generic Features for Malicious URL Detection System,” *2019 IEEE 10th Annu. Ubiquitous Comput. Electron. Mob. Commun. Conf. UEMCON 2019*, pp. 0347–0352, 2019, doi: 10.1109/UEMCON47517.2019.8992930.

[29] C. A. Germain, “URLs: Uniform resource locators or unreliable resource locators,” *Coll. Res. Libr.*, vol. 61, no. 4, pp. 359–365, 2000, doi: 10.5860/crl.61.4.359.

[30] A. B.Sayamber and A. M. Dixit, “Malicious URL Detection and Identification,” *Int. J. Comput. Appl.*, vol. 99, no. 17, pp. 17–23, 2014, doi: 10.5120/17464-8247.

[31] K. D. Vara, V. S. Dimble, M. M. Yadav, and A. A. Thorat, “Based on URL Feature Extraction Identify Malicious Website Using Machine Learning Techniques,” vol. 6, no. 3, pp. 144–148, 2022.

[32] O. V. Lee *et al.*, “A malicious URLs detection system using optimization and machine learning classifiers,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 17, no. 3, pp. 1210–1214, 2020, doi: 10.11591/ijeecs.v17.i3.pp1210-1214.

[33] D. Chiba, K. Tobe, T. Mori, and S. Goto, “Detecting malicious websites by learning IP address features,” *Proc. - 2012 IEEE/IPSJ 12th Int. Symp. Appl. Internet, SAINT 2012*, pp. 29–39, 2012, doi: 10.1109/SAINT.2012.14.

[34] V. Hamon, “Malicious URI resolving in PDF documents,” *J. Comput. Virol.*, vol. 9, no. 2, pp. 65–76, 2013, doi: 10.1007/s11416-013-0179-2.

[35] A. K. Jain and B. B. Gupta, “A machine learning based approach for phishing

detection using hyperlinks information," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 5, pp. 2015–2028, 2019, doi: 10.1007/s12652-018-0798-z.

[36] E. Aghaei and G. Serpen, "Host-based anomaly detection using Eigentraces feature extraction and one-class classification on system call trace data," 2019, [Online]. Available: http://arxiv.org/abs/1911.11284.

[37] M. Alazab and S. Fellow, "Malicious URL Detection using Deep Learning," pp. 1–9, 2020.

[38] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.

[39] N. Wakhidah, "Clustering Menggunakan K-Means Algorithm," *J. Transform.*, vol. 8, no. 1, p. 33, 2010, doi: 10.26623/transformatika.v8i1.45.

[40] D. Fotakis, G. Piliouras, and S. Skoulakis, "Efficient Online Learning for Dynamic k-Clustering," pp. 1–27, 2021, [Online]. Available: http://arxiv.org/abs/2106.04336.

[41] A. Singh, A. Yadav, and A. Rana, "K-means with Three different Distance Metrics," *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, 2013, doi: 10.5120/11430-6785.

[42] E. Umargono, J. E. Suseno, and V. G. S. K., "K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median," no. Conrist 2019, pp. 234–240, 2020, doi: 10.5220/0009908402340240.

[43] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[44] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.

[45] S. Nayak, D. Nadig, and B. Ramamurthy, “Analyzing Malicious URLs using a Threat Intelligence System,” *Int. Symp. Adv. Networks Telecommun. Syst. ANTS*, vol. 2019-Decem, pp. 6–9, 2019, doi: 10.1109/ANTS47819.2019.9118051.

[46] I. J. Good, “Exploratory data analysis,” *J. Stat. Comput. Simul.*, vol. 37, no. 3–4, pp. 243–245, 1990, doi: 10.1080/00949659008811311.