DETEKSI MALICIOUS URL PADA FILE BERBASIS FITUR LEKSIKAL MENGGUNAKAN METODE *RANDOM FOREST*

OLEH:

RACHMAWATI DWINANTI PUTRI

09011281823064

JURUSAN SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA

2023

LEMBAR PENGESAHAN

Deteksi Malicious URL Pada File Berbasis Fitur Leksikal Menggunakan Metode *Random Forest*

PROPOSAL TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh

RACHMAWATI DWINANTI PUTRI
09011281823064

Indralaya, 22 Agustus 2023

Mengetahui,

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

AUTHENTICATION PAGE

Detection of Malicious URLs In Lexical Feature Based Using The Random Forest Method

FINAL TASK

Submitted To Fulfill One Of The Requirements
To Obtain A Bachelor's Degree in Computer Science

By

RACHMAWATI DWINANTI PUTRI
09011281823064

Indralaya, 22 August 2023

Acknowledge,

Head of Computer System Deparment	***Supervisor***
Dr. Ir. Sukemi, M.T.	**Ahmad Heryanto, S. Kom, M.T.**
NIP. 196612032006041001	**NIP. 198701222015041002**

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada :

Hari : Selasa

Tanggal : 18 Juli 2023

Tim Penguji :

1. **Ketua** : **Dr. Ahmad Zarkasi, M.T.**
2. **Sekretaris** : **Adi Hermansyah, M.T.**
3. **Penguji** : **Huda Ubaya, M.T.**
4. **Pembimbing I** : **Ahmad Heryanto, S.Kom, M.T.**

Mengetahui,

Ketua Jurusan Sistem Komputer

Dr.Ir. Sukemi, M.T.

NIP. 196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Rachmawati Dwinanti Putri

NIM : 09011281823064

Judul : Deteksi Malicious URL Pada File Berbasis Fitur Leksikal Menggunakan Metode Random Forest

Hasil Pengecekan Software *iThenticate/Turnitin* : 3%

Menyatakan bahwa laporan tugas akhir saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsur 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, Agustus 2023

Rachmawati Dwinanti Putri

NIM.09011281823064

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 "Deteksi Malicious URL Pada File Berbasis Fitur Leksikal Menggunakan Metode *Random Forest*".

Dalam melaksankan dan menyusun tugas akhir ini, tidak akan tercapai tanpa bantuan, bimbingan dan dukungan dari berbagai pihak yang telah membantu penulis yang berupa do'a dari mama, kakak, dan keluarga besar dari penulis serta bantuan seperti bimbingan dan nasihat kepada penulis agar penyusunan dan penulisan tugas akhir ini dapat diselesaikan dengan baik. Oleh karena itu, pada kesempatan ini penulis mengucapankan rasa syukur dan terima kasih yang sebesar – besarnya kepada:

1. Allah Subhanahu Wa Ta'ala yang telah memberikan berkah, nikmat dan hidayah-Nya yang tidak terhitung.
2. Mama dan kakak yang telah memberikan do'a, 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. 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 meluangkan waktunya untuk membimbing dan memberikan saran terbaik kepada penulis dalam menyelesaikan Tugas Akhir ini.
6. Kak Tri Wanda Septian, S.Kom, M.Sc., yang telah membantu dan memberikan arahan, masukan, serta saran kepada penulis.
7. Mbak Renny Virgasari selaku Admin Jurusan Sistem Komputer yang telah membantu penulis dalam mengurus seluruh berkas administrasi.

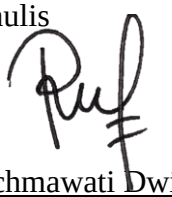
8. Teman – teman seperjuangan riset malicious URL, Muhammad Imam Rafi, Rizky Valen Mafaza, dan Muhammad Andiko Putra yang telah sama – sama berjuang dan membantu penulis dalam menyelesaikan tugas akhir ini.
9. Ades Harafi duri, Nia anita, dan Novi Yuningsih yang telah berkenan menjadi teman rantau dari awal masuk perkuliahan, berbagi suka duka, dan selalu ada untuk mendengarkan keluh kesah.
10. Dimas Aditya Kristianto yang telah membantu dan mengajari penulis dalam permasalahan codingan selama perkuliahan dan tugas akhir.
11. Teman – teman RTB, Muhammad Farhan Al Harits, Muhammad Realdi, dan Muhammad Furqon Rabbani yang selalu memberi semangat, saran dan kritik kepada penulis.
12. Teman – teman sejak SMP sampai saat ini, Arsyita Khairunnisa, Hera Marcelina, dan Regita Cahyani yang selalu ada di masa – masa sulit.
13. Teman – teman seperjuangan Sistem Komputer angkatan 2018.
14. Kepada semua pihak yang tidak dapat disebutkan satu persatu yang telah membantu selama penyelesaian tugas akhir ini.
15. Serta 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, Agustus 2023

Penulis



Rachmawati Dwinanti Putri

NIM. 09011281823064

DETEKSI MALICIOUS URL PADA FILE BERBASIS FITUR LEKSIKAL MENGGUNAKAN METODE RANDOM FOREST

Rachmawati Dwinanti Putri (09011281823064)
Department of Computer System, Faculty of Computer Science, University of Sriwijaya
Palembang, Indonesia
Email : rrahmaput@gmail.com

ABSTRAK

Dengan adanya bentuk model penyerangan seperti melakukan phising dan menyebarkan malware, tindakan tersebut diawali dengan mengakses *Uniform Resource Locator* (URL) atau file yang mengandung link berbahaya di dalamnya. *Uniform Resource Locator* (URL) adalah pengidentifikasi khusus yang digunakan untuk menemukan resource melalui internet. URL dapat menjadi ancaman terhadap ketersediaan, pengendalian kerahasiaan, dan integritas data yang salah satu ancamannya berupa malicious URL. Untuk membedakan malicious URL dan URL yang normal adalah menggunakan ektraksi fitur guna untuk mengidentifikasi karakteristik penting dari malicious URL. Fitur ekstraksi yang digunakan adalah fitur leksikal yang terdiri dari 18 fitur. Setelah diekstraksi, hasil dari dataset tidak seimbang maka diproses resampling menggunakan oversampling dengan SMOTE. Untuk mengklasifikasi dataset, penelitian ini menggunakan algoritma machine learning berupa random forest. Random Forest adalah algoritma yang membangun banyak decision tree atau pohon keputusan. Algoritma ini mampu mencapai akurasi klasifikasi yang tinggi serta memberikan hasil yang baik. Pada penelitian ini, memperoleh hasil evaluasi dengan nilai akurasi sebesar 90,97%, presisi 99,05%, recall 85,39% dan f1-score 91,71%.

Kata Kunci : URL, Malicious URL, Fitur Leksikal, Synthetic Minority Over-sampling Technique (SMOTE), Random Forest, Machine Learning.

DETECTION OF MALICIOUS URLs IN LEXICAL FEATURE BASED USING THE RANDOM FOREST METHOD

Rachmawati Dwinanti Putri (09011281823064)
Department of Computer System, Faculty of Computer Science, University of Sriwijaya
Palembang, Indonesia
Email : rrahmaput@gmail.com

ABSTRACT

With the existence of attack models such as phishing and malware distribution, these actions begin by accessing a Uniform Resource Locator (URL) or files containing harmful links within them. A Uniform Resource Locator (URL) is a specific identifier used to locate resources through the internet. URLs can pose threats to availability, control, confidentiality, and data integrity, with one of the threats being malicious URLs. To differentiate between malicious URLs and normal URLs, feature extraction is employed to identify important characteristics of malicious URLs. The extraction features used are lexical features consisting of 18 attributes. After extraction, due to the imbalanced dataset, resampling is performed using oversampling with SMOTE. To classify the dataset, this research utilizes a machine learning algorithm known as random forest. Random Forest is an algorithm that constructs multiple decision trees. This algorithm can achieve high classification accuracy and provide good results. In this study, the evaluation yields results with an accuracy value of 90.97%, precision of 99.05%, recall of 85.39%, and an f1-score of 91.71%.

Keyword : *Uniform Resource Locator, Malicious URL, Lexical Features, Synthetic Minority Over-sampling Technique (SMOTE), Random Forest, Machine Learning.*

DAFTAR ISI

LEMBAR PENGESAHAN……..…….i
AUTHENTICATION PAGE…….…..ii
HALAMAN PERSETUJUAN……iii
HALAMAN PERNYATAAN……. iv
KATA PENGANTAR……...……v
ABSTRAK……..…...……..vii
ABSTRACT……..…...….. viii
DAFTAR ISI……..…...….. ix
DAFTAR GAMBAR……..…….... xii
DAFTAR TABEL……...….... xiv
BAB I PENDAHULUAN……..….. 1
1.1 Latar Belakang…….. 1
1.2 Perumusan Masalah……. 3
1.3 Tujuan……... 4
1.4 Manfaat……. 4
1.5 Batasan Masalah……... 4
1.6 Metodologi Penelitian……...5
1.7 Sistematika Penulisan……5
BAB II TINJAUAN PUSTAKA……..……7
2.1 Penelitian Terdahulu……. 7
2.2 Uniform Resource Locator……..19
2.3 Malicious URL……21
2.4 Machine Learning……... 22
2.5 Random Forest…… 23
2.6 SMOTE……...….... 28
2.7 Fitur Leksikal…….. 29
2.8 Confusion Matrix…… 33
BAB III METODOLOGI PENELITIAN……. 36
3.1 Pendahuluan……36
3.2 Kerangka Kerja Penelitian…….. 36

3.3 Kerangka Kerja Metodologi Penelitian……..… 37
3.4 Kebutuhan Perangkat…….. 39
3.5 Skenario Eksperimen…….. 39
3.6 Skenario Riset……. 40
3.7 Persiapan Dataset…… 41
3.8 Ekstraksi Fitur Leksikal…….. 46
3.9 Pre-processing…….49
3.9.1 Synthetic Minority Oversampling Technique (SMOTE) ……..…49
3.9.2 Grid Search Cross Validation…….....51
3.10 Klasifikasi Random Forest……54
3.11 Validasi……. 58
BAB IV HASIL DAN ANALISA…….…… 59
4.1 Pendahuluan……59
4.2 Hasil Parser Data…… 59
4.3 Hasil Ekstraksi File PDF…… 60
4.4 Hasil Ekstraksi Fitur Leksikal…… 61
4.5 Pre-processing…… 65
4.5.1 Hasil SMOTE…….... 65
4.5.2 Hasil Grid Search Cross Validation……... 67
4.6 Hasil Kasifikasi Random Forest……. 68
4.7 Hasil Validasi Tanpa Fitur Leksikal…… 70
4.7.1 Hasil Validasi data Training 90% dan Data Testing 10%……70
4.7.2 Hasil Validasi data Training 80% dan Data Testing 20%……71
4.7.3 Hasil Validasi data Training 70% dan Data Testing 30%……71
4.7.4 Hasil Validasi data Training 60% dan Data Testing 40%……72
4.7.5 Hasil Validasi data Training 50% dan Data Testing 50%……73
4.8 Hasil Validasi Menggunakan Fitur Leksikal……...…74
4.8.1 Hasil Validasi data Training 90% dan Data Testing 10%…….74
4.8.2 Hasil Validasi data Training 80% dan Data Testing 20%…….74
4.8.3 Hasil Validasi data Training 70% dan Data Testing 30%…….75
4.8.4 Hasil Validasi data Training 60% dan Data Testing 40%…….76

4.8.5 Hasil Validasi data Training 50% dan Data Testing 50%…….77
4.9 Analisis Hasil Validasi…….77
BAB V KESIMPULAN DAN SARAN……... 79
5.1 Kesimpulan…….79
5.2 Saran ……...79
DAFTAR PUSTAKA……80

DAFTAR GAMBAR

Gambar 2.1 Struktur Uniform Resource Locator 19
Gambar 2.2 Arsitektur Random Forest untuk intruksi deteksi 24
Gambar 2.3 Visualisasi Random Forest 28
Gambar 2.4 Synthetic Minority Over-Sampling Technique.....29
Gambar 3.1 Kerangka Kerja Penelitian 37
Gambar 3.2 Kerangka Kerja Metode Penelitian.....38
Gambar 3.3 Skenario Eksperimen Malicious URL 40
Gambar 3.4 Skenario Riset Malicious URL 41
Gambar 3.5 File PDF.....41
Gambar 3.6 Alur Persiapan dataset 42
Gambar 3.7 Pengumpulan File PDF.....43
Gambar 3.8 Parser Data dengan code strings (namafilePDF) 43
Gambar 3.9 Parser Data dengan code pdf-parser -O namafile.pdf | grep http.....43
Gambar 3.10 Anlisis URL 44
Gambar 3.11 Atribut obj.....45
Gambar 3.12 Diagram Alir Proses Oversampling.....50
Gambar 3.13 Flowchart Proses Grid Search Cross Validation.....52
Gambar 3.14 Proses Random Forest 55
Gambar 4.1 Hasil Parser Data 59
Gambar 4.2 Jumlah Persentase Dataset.....60
Gambar 4.3 10 Daftar File PDF berisi Malicious URL.....60
Gambar 4.4 Hasil Ekstraksi File PDF.....61
Gambar 4.5 Hasil Ekstraksi Fitur Leksikal.....61
Gambar 4.6 Hasil Perubahan False dan True 62
Gambar 4.7 Fitur Host.....63
Gambar 4.8 Fitur tld 64
Gambar 4.9 Fitur Scheme.....64
Gambar 4.10 Grafik Nilai 18 Fitur Leksikal 65
Gambar 4.11 Data Imbalance 66
Gambar 4.12 Data balance.....67

Gambar 4.13 Visualisasi Random Forest 69
Gambar 4.14 Hasil Tanpa Fitur Leksikal (90:10).....70
Gambar 4.15 Hasil Tanpa Fitur Leksikal (80:20).....71
Gambar 4.16 Hasil Tanpa Fitur Leksikal (70:30).....72
Gambar 4.17 Hasil Tanpa Fitur Leksikal (60:40).....72
Gambar 4.18 Hasil Tanpa Fitur Leksikal (50:50).....73
Gambar 4.19 Hasil Validasi Data Training 90% dan Data Testing 10% 74
Gambar 4.20 Hasil Validasi Data Training 80% dan Data Testing 20% 75
Gambar 4.21 Hasil Validasi Data Training 70% dan Data Testing 30% 75
Gambar 4.22 Hasil Validasi Data Training 60% dan Data Testing 40% 76
Gambar 4.23 Hasil Validasi Data Training 50% dan Data Testing 50% 77

DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu.....7
Tabel 2.2 Daftar fitur leksikal.....31
Tabel 2.3 Confusion Matrix.....34
Tabel 3.1 Kebutuhan Perangkan Keras 39
Tabel 3.2 Kebutuhan Perangkat Lunak 39
Tabel 3.3 Detail Jumlah Data 45
Tabel 3.4 Fitur Leksikal.....46
Tabel 3.5 Spesifikasi Parameter SMOTE.....50
Tabel 3.6 Spesifikasi Parameter yang diujikan 53
Tabel 3.7 Spesifikasi Parameter Grid Search Cross Validation 53
Tabel 4.1 Jumlah Data file PDF dan URL.....61
Tabel 4.2 Detail Jumlah Fitur Host 62
Tabel 4.3 Detail Jumlah Fitur tld.....63
Tabel 4.4 Detail Jumlah Fitur Scheme 64
Tabel 4.5 Detail Jumlah Data Imbalance.....66
Tabel 4.6 Spesifikasi dan Score Grid Search Cross Validation 67
Tabel 4.7 Pembagian Data.....68
Tabel 4.8 Parameter Graphviz Random Forest 68
Tabel 4.9 Hasil Validasi Data Tanpa Fitur Leksikal (90:10) 70
Tabel 4.10 Hasil Validasi Data Tanpa Fitur Leksikal (80:20) 71
Tabel 4.11 Hasil Validasi Data Tanpa Fitur Leksikal (70:30) 72
Tabel 4.12 Hasil Validasi Data Tanpa Fitur Leksikal (60:40) 73
Tabel 4.13 Hasil Validasi Data Tanpa Fitur Leksikal (50:50) 73
Tabel 4.14 Hasil Validasi Data Training 90% dan data testing 10% 74
Tabel 4.15 Hasil Validasi Data Training 80% dan data testing 20% 75
Tabel 4.16 Hasil Validasi Data Training 70% dan data testing 30% 76
Tabel 4.17 Hasil Validasi Data Training 60% dan data testing 40% 76
Tabel 4.18 Hasil Validasi Data Training 50% dan data testing 50% 77
Tabel 4.19 Hasil Validasi Keseluruhan Tanpa Fitur 78
Tabel 4.20 Hasil Validasi Keseluruhan Dengan Fitur Leksikal 78

BAB I PENDAHULUAN

1.1 Latar Belakang

Dengan popularitas perkembangan internet yang kuat dan jaringan yang meluas, internet telah menjadi bagian tak tergantikan dari aktivitas masyarakat, memberikan kemudahan yang luar biasa. Namun, internet juga merupakan pedang bermata dua dengan dua sisi yang berbeda. Disatu sisi positif, internet memungkinkan pengguna untuk dengan mudah mencari informasi seperti data, gambar, dan pengetahuan. Namun, disisi negatifnya internet menjadi platform aktif bagi para penyerang [1]. Dengan adanya bentuk model penyerangan seperti melakukan phising dan menyebarkan malware. Semua tindakan tersebut diawali dengan mengakses *Uniform Resource Locator* (URL) atau file yang mengandung link berbahaya di dalamnya. Hal ini karena URL atau link tersebut berperan sebagai alamat atau vektor yang menggarahkan pengguna internet ke dalam jaringan tertentu [2].

Uniform Resource Locator (URL) adalah pengidentifikasi khusus yang digunakan untuk menemukan *resource* melalui internet. URL digunakan di World Wide Web untuk mengakses resource yang sah. Namun, ketika URL digunakan untuk tujuan lain, URL tersebut dapat menjadi ancaman terhadap ketersediaan, pengendalian, kerahasiaan, dan integritas data[3]. Salah satu ancaman tersebut adalah malicious URL. Malicious URL dan situs web yang telah terinfeksi adalah akar dari ancaman tersebut. Dampak dari URL yang telah terpapar malicious atau terpasang link berbahaya akan mengarah ke eksploitasi informasi dan aset pengguna[4]. Untuk membedakan malicious URL dari URL yang normal, solusi yang efektif adalah menggunakan ekstraksi fitur untuk mengidentifikasi karakteristik penting dari malicious URL. Salah satu fitur ekstraksi yang dapat digunakan adalah fitur leksikal, yang secara logis dapat mendeteksi dan menyimpulkan pola dalam malicious URL. Penggunaan fitur leksikal ini memberikan keuntungan berupa komputasi yang ringan, tingkat keamanan yang tinggi dan akurasi yang baik dalam mengklasifikasikan URL dengan menganalisis perbedaan antara benign URL dan malicious URL[5].

Berikut adalah hasil penelitian terdahulu yang membahas tentang malicious URL, menjadi referensi utama untuk topik penelitian ini.

Dalam penelitian [6] berjudul "Malicious URL Detection: A Comparative Study", peneliti mengusulkan penggunaan metode Random Forest. Penelitian ini menggunakan dataset yang diambil dari repositori Kaggle. Hasil penelitian menunjukkan performa yang baik, dengan mencapai akurasi 92,6%. Meskipun demikian, untuk meningkatkan keakuratan pengklasifikasi Random Forest, perlu dilakukan peningkatan dengan menggunakan data yang lebih seimbang, yaitu data yang berisi malicious dan benign dalam proporsi yang hampir sama.

Dalam Penelitian [7] berjudul "Towards Detecting and Classifying Malicious URLs Using Deep Learning", peneliti menggunakan metode Random Forest . Penelitian ini mencapai hasil performa akurasi sebesar 96,26% untuk klasifikasi multi-class. Dataset yang digunakan dalam penelitian ini adalah ISCX-URL-2016. Meskipun penelitian menghasilkan performa akurasi yang baik, namun terdapat masalah dalam pelatihan dan waktu pengujian, serta kelemahan tambahan berupa overfitting pada model yang mengurangi generalisasi dari hasil penelitian tersebut.

Dalam penelitian [8] berjudul "Malicious URL Detection Using Supervised Machine Learning Techniques", peneliti mengusulkan penggunaan metode Logistic Regression dengan menggunakan dataset dari Kaggle website. Hasil penelitian ini menunjukkan akurasi sebesar 86,25%, presisi sebesar 33,33%, recall sebesar 5,80%, dan F1-Score sebesar 9,88%. Meskipun metode Logistic Regression memiliki akurasi yang baik, namun presisi dan recall mendapatkan hasil yang rendah, sehingga diperlukan peningkatan dalam memprediksi malicious URL serta optimalisasi dalam menyesuaikan parameter.

Dalam menghadapi penyerangan yang menggunakan malicious URL, penggunaan metode Random Forest telah terbukti efektif. Random Forest dapat mengatasi masalah ini dengan menghitung berbagai ukuran variabel penting yang memberikan informasi untuk mengetahui kontribusi masing – masing variabel terhadap akurasi model[5]. Kelebihan Random Forest terletak pada kemampuannya dalam mengambil rata – rata, sehingga mencapai keseimbangan yang proposional. Selain itu, Random Forest memiliki sedikit jumlah parameter yang dapat diatur dan

dapat digunakan secara langsung dengan pengaturan default, menjadikannya alat yang sederhana untuk digunakan tanpa mengorbankan kecepatan dan efisiensi dalam meghasilkan model yang masuk akal[9].

Random Forest adalah algoritma yang membangun banyak decision tree atau pohon keputusan. Selama proses klasifikasi, algoritma ini mengembalikan kelas yang paling sering muncul dari himpunan pohon individu. Semua pohon dibangun menggunakan subset acak dari fitur – fitur yang ada dalam dataset[10]. Dalam konteks pengklasifikasi, Random Forest digunakan untuk analisis klasifikasi pada data deteksi intruksi. Dengan membangun berbagai pohon keputusan selama tahap pelatihan, Random Forest menghasilkan label kelas yang paling sering muncul. Algoritma ini mampu mencapai akurasi klasifikasi yang tinggi, mengatasi outlier dan noise dalam data, serta memberikan hasil yang baik dalam klasifikasi data dibeberapa kasus dengan kemungkinan terjadinya *overfitting* yang lebih [11].

Berdasarkan penelitian terdahulu yang telah dijelaskan sebelumnya, peneliti memilih metode Random Forest karena selain mampu menghasilkan akurasi tinggi, metode ini juga menunjukkan kompleksitas yang lebih rendah dan kinerja yang tinggi dalam masalah klasifikasi. Oleh karena itu, peneliti memilih judul "Deteksi malicious URL pada file berbasis fitur leksikal menggunakan metode Random Forest".

1.2 Perumusan Masalah

Berdasarkan penjelasan latar belakang yang telah diuraikan sebelumnya, penelitian ini bertujuan mendeteksi serangan malicious URL menggunakan metode *Random Forest*. Berbeda dengan penelitian sebelumnya[6] yang hanya menerapkan 14 fitur ekstraksi, penelitian ini akan menerapkan 18 fitur ekstraksi berdasarkan fitur leksikal. Selain itu, implementasi penelitian ini juga akan melakukan pengklasifikasi secara binary untuk mengklasifikasi URL dan menentukan statusnya. Penelitian ini juga akan melakukan uji validasi untuk mengukur akurasi, presisi, recall, dan F1-Score dari model yang dikembangkan.

1.3 Tujuan

Tujuan yang akan dicapai dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan metode *Random Forest* dalam klasifikasi Malicious dan Benign URL pada file PDF
2. Meningkatkan kinerja metode *Random Forest* dengan menerapkan ekstraksi Fitur Leksikal untuk menghasilkan model yang lebih optimal dibandingkan penelitian sebelumnya.
3. Mengevaluasi performa ekstraksi fitur dan metode berdasarkan akurasi, presisi, recall, dan F1-Score untuk memahami kualitasnya.

1.4 Manfaat

Manfaat untuk penelitian ini adalah sebagai berikut:

1. Mampu mengolah data dengan memanfaatkan sistem operasi Kali Linux untuk memparser kumpulan file PDF dan menghasilkan dataset baru.
2. Mengoptimalkan proses ekstraksi dengan memanfaatkan fitur leksikal guna meningkatkan efisiensi dalam ekstraksi.
3. Mampu menganalisis hasil ekstraksi file PDF yang telah diproses untuk memperoleh pemahaman mengenai akurasi dataset menggunakan metode Random Forest.
4. Membandingkan peforma akhir dari tahap klasifikasi malicious URL dengan menggunakan confusion matrix guna mendapatkan gambaran yang jelas mengenai hasil yang dicapai.

1.5 Batasan Masalah

Berikut batasan masalah pada Tugas Akhir ini, yaitu:

1. Penelitian ini menggunakan dataset yang diperoleh dari file PDF GARUDA.
2. Penelitian ini dilakukan untuk mengklasifikasi Malicious dan Benign URL.
3. Menerapkan prosedur klasifikasi malicious URL dengan mengaplikasikan metode Random Forest.

1.6 Metodologi Penelitian

Metodologi penelitian yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Metode Studi Pustaka dan Literature

   Pada metode ini peneliti mengumpulkan dan menganalisis data yang akan diolah menggunakan metode Random Forest dari berbagai macam referensi dalam membantu pembuatan Tugas Akhir ini.

2. Metode Konsultasi

   Pada metode ini peneliti memilih bahan – bahan yang sudah mempunyai pengetahuan serta pemahaman yang baik dalam mengatasi masalah yang ditemui pada saat peneliti menulis.

3. Metode Pengumpulan Data

   Pada metode ini, dilakukan pengumpulan data dengan cara melabeli URL yang telah didapatkan dari file PDF menggunakan virus total.

4. Metode Pengujian

   Pada metode ini, melakukan pengekstraksi dengan cara mengekstraksi fitur menggunakan fitur leksikal yang selanjutnya akan dilakukan pengujian dari hasil ekstraksi URL dengan metode Random Forest

5. Metode Analisis dan Kesimpulan

   Pada langkah terakhir untuk penelitian ini diperoleh hasil validasi performa model dari pengujian terhadap metode Random Forest yang akan dianalisis dan ditarik beberapa kesimpulan.

1.7 Sistematika Penulisan

Sistematika penulisan yang akan digunakan dalam penulisan tugas akhir adalah sebagai berikut:

BAB I PENDAHULUAN

Bab pertama akan menjelaskan tentang latar belakang, tujuan penelitian, rumusan masalah, serta bentuk sistematika penulisan penelitian.

BAB II TINJAUAN PUSTAKA

Bab kedua akan menjelaskan tentang teori-teori dasar yang berupa model sistematis yang berkaitan dengan penelitian ini. Dasar teori yang akan dibahas adalah literatur

mengenai URL, Malicious URL, Fitur Leksikal, *Random Forest* dan performa validasi.

BAB III METODOLOGI PENELITIAN

Bab ketiga akan menjelaskan secara bertahap tentang proses dan langkah – langkah kegiatan dalam penelitian. Penelitian akan dimulai dari persiapan data, pra pengolahan dataset PDF, fitur ekstraksi dan klasifikasi.

BAB IV HASIL DAN USAHA

Bab keempat akan menunjukkan hasil pengujian dari pengujian yang telah dilakukan dan menjelaskan analisa terhadap hasil penelitian yang telah dilakukan.

BAB V KESIMPULAN DAN SARAN

Bab keenam akan membuat kesimpulan yang telah didapatkan dari hasil keseluruhan peneltian dan tujuan yang telah dicapai terhadap penelitian yang telah dilakukan.

DAFTAR PUSTAKA

[1] Y. Chen, Y. Zhou, Q. Dong, and Q. Li, "A Malicious URL Detection Method Based on CNN," *2020 IEEE Conf. Telecommun. Opt. Comput. Sci. TOCS 2020*, pp. 23–28, 2020, doi: 10.1109/TOCS50858.2020.9339761.

[2] M. Yunus, D. Widiastuti, H. Rasjid, and ..., "… Klasifikasi Untuk Deteksi Uniform Resource Locator (URL) Berdasarkan Jenis Serangan Menggunakan Algoritma Naive Bayes, C4. 5 dan K-Nearest Neighbor," *Pros. ...*, vol. 3, 2019, [Online]. Available: http://sunuy165.staff.gunadarma.ac.id/Publications/files/6057/Prosiding+Moh+Yunus+268-Article+Text-786-1-10-20200926.pdf

[3] 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.

[4] P. Varaprasada Rao, S. Govinda Rao, P. Chandrasekhar Reddy, B. S. Anil Kumar, and G. Anil Kumar, "Detection of malicious uniform resource locator," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2, pp. 41–47, 2019, doi: 10.35940/ijrte.A1265.078219.

[5] G. S. B, C. Tang, W. Gao, and Y. Yin, "MD- VC M atrix : An Efficient Scheme for Publicly Verifiable Computation," vol. 1, pp. 349–362, 2016, doi: 10.1007/978-3-319-46298-1.

[6] Shantanu, B. Janet, and R. Joshua Arul Kumar, "Malicious URL Detection: A Comparative Study," *Proc. - Int. Conf. Artif. Intell. Smart Syst. ICAIS 2021*, pp. 1147–1151, 2021, doi: 10.1109/ICAIS50930.2021.9396014.

[7] 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.

[8] 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.

[9] A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine leaning," *RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc.*, vol. 2018-Janua, pp. 1432–1436, 2017, doi: 10.1109/RTEICT.2017.8256834.

[10] C. Liu, L. Wang, B. Lang, and Y. Zhou, "Finding effective classifier for malicious URL detection," *ACM Int. Conf. Proceeding Ser.*, pp. 240–244, 2018, doi: 10.1145/3180374.3181352.

[11] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, vol. 6, no. c, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[12] A. Saleem Raja, R. Vinodini, and A. Kavitha, "Lexical features based malicious URL detection using machine learning techniques," *Mater. Today Proc.*, vol. 47, no. xxxx, pp. 163–166, 2021, doi: 10.1016/j.matpr.2021.04.041.

[13] J. Yuan, G. Chen, S. Tian, and X. Pei, "Malicious URL detection based on a parallel neural joint model," *IEEE Access*, vol. 9, pp. 9464–9472, 2021, doi: 10.1109/ACCESS.2021.3049625.

[14] S. Sindhu, S. P. Patil, A. Sreevalsan, F. Rahman, and A. N. Saritha, "Phishing detection using random forest, SVM and neural network with backpropagation," *Proc. Int. Conf. Smart Technol. Comput. Electr. Electron. ICSTCEE 2020*, pp. 391–394, 2020, doi: 10.1109/ICSTCEE49637.2020.9277256.

[15] 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.

[16] 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.

[17] A. Joshi, L. Lloyd, P. Westin, and S. Seethapathy, “Using lexical features for malicious URL Detection - A machine learning approach,” *arXiv*, 2019.

[18] C. D. Morales-Molina, D. Santamaria-Guerrero, G. Sanchez-Perez, H. Perez-Meana, and A. Hernandez-Suarez, “Methodology for malware classification using a random forest classifier,” *2018 IEEE Int. Autumn Meet. Power, Electron. Comput. ROPEC 2018*, no. Ropec, pp. 1–6, 2019, doi: 10.1109/ROPEC.2018.8661441.

[19] C. WU, M. LI, L. YE, X. ZOU, and B. QIANG, “Malicious Website Detection Based on URLs Static Features,” *DEStech Trans. Comput. Sci. Eng.*, no. mso, pp. 0–6, 2018, doi: 10.12783/dtcse/mso2018/20499.

[20] 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.

[21] 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.

[22] M. Weedon, D. Tsaptsinos, and J. Denholm-Price, “Random forest explorations for URL classification,” *2017 Int. Conf. Cyber Situational Awareness, Data Anal. Assessment, Cyber SA 2017*, pp. 3–6, 2017, doi: 10.1109/CyberSA.2017.8073403.

[23] 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-Febru, pp. 97–100, 2017, doi: 10.1109/ICCWAMTIP.2017.8301457.

[24] L. MacHado and J. Gadge, “Phishing Sites Detection Based on C4.5 Decision Tree Algorithm,” *2017 Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2017*, pp. 1–5, 2018, doi: 10.1109/ICCUBEA.2017.8463818.

[25] 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.

[26] P. Iyappan, R. Muthaiya Ram, R. Barani, and S. Kumarakrishnan, “Enhanced shortened uniform resource locator based on phishing and malware detection algorithm-a secure model,” *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, pp. 1–6, 2019, doi: 10.1109/ICSCAN.2019.8878684.

[27] J. Puchýř and M. Holeňa, “Random-forest-based analysis of URL paths,” *CEUR Workshop Proc.*, vol. 1885, pp. 129–135, 2017.

[28] H. Zhao, Z. Chang, W. Wang, and X. Zeng, “Malicious Domain Names Detection Algorithm Based on Lexical Analysis and Feature Quantification,” *IEEE Access*, vol. 7, pp. 128990–128999, 2019, doi: 10.1109/ACCESS.2019.2940554.

[29] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Learning to detect malicious URLs,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, doi: 10.1145/1961189.1961202.

[30] G. Tan, P. Zhang, Q. Liu, X. Liu, C. Zhu, and F. Dou, “Adaptive Malicious URL Detection: Learning in the Presence of Concept Drifts,” *Proc. - 17th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. 12th IEEE Int. Conf. Big Data Sci. Eng. Trust. 2018*, pp. 737–743, 2018, doi: 10.1109/TrustCom/BigDataSE.2018.00107.

[31] S. V. Mahadevkar *et al.*, “A Review on Machine Learning Styles in Computer Vision - Techniques and Future Directions,” *IEEE Access*, vol. 10, no. August, pp. 107293–107329, 2022, doi:

10.1109/ACCESS.2022.3209825.

[32] P. Zhao and S. C. H. Hoi, “Cost-sensitive online active learning with application to malicious URL detection,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. Part F1288, pp. 919–927, 2013, doi: 10.1145/2487575.2487647.

[33] A. Primajaya and B. N. Sari, “Random Forest Algorithm for Prediction of Precipitation,” *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[34] A. Cutler, D. R. Cutler, and J. R. Stevens, “Ensemble Machine Learning,” *Ensemble Mach. Learn.*, 2012, doi: 10.1007/978-1-4419-9326-7.

[35] V. Y. Kulkarni and P. K. Sinha, “Effective Learning and Classification using Random Forest Algorithm,” *Int. J. Eng. Innov. Technolgy*, vol. 3, no. 11, pp. 267–273, 2014.

[36] S. Ciss, “Random Uniform Forests,” pp. 1–19, 2015.

[37] N. Sapountzoglou, J. Lago, and B. Raison, “Fault diagnosis in low voltage smart distribution grids using gradient boosting trees,” *Electr. Power Syst. Res.*, vol. 182, no. February, p. 106254, 2020, doi: 10.1016/j.epsr.2020.106254.

[38] J. Yang, J. Gong, W. Tang, Y. Shen, C. Liu, and J. Gao, “Delineation of urban growth boundaries using a patch-based cellular automata model under multiple spatial and socio-economic scenarios,” *Sustain.*, vol. 11, no. 21, 2019, doi: 10.3390/su11216159.

[39] A. S. Manjeri, R. Kaushik, A. Mnv, and P. C. Nair, “A Machine Learning Approach for Detecting Malicious Websites using URL Features,” *Proc. 3rd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2019*, pp. 555–561, 2019, doi: 10.1109/ICECA.2019.8821879.

[40] K. R. Mahmudah, B. Purnama, F. Indriani, and K. Satou, “Machine learning algorithms for predicting chronic obstructive pulmonary disease from gene expression data with class imbalance,” *Bioinforma. 2021 - 12th Int. Conf.*

Bioinforma. Model. Methods Algorithms; Part 14th Int. Jt. Conf. Biomed. Eng. Syst. Technol. BIOSTEC 2021, vol. 3, no. Biostec, pp. 148–153, 2021, doi: 10.5220/0010316501480153.

[41] S. S. Shastri, P. C. Nair, D. Gupta, R. C. Nayar, R. Rao, and A. Ram, "Breast cancer diagnosis and prognosis using machine learning techniques," *Adv. Intell. Syst. Comput.*, vol. 683, pp. 327–344, 2018, doi: 10.1007/978-3-319-68385-0_28.

[42] X. T. Dang *et al.*, "A novel over-sampling method and its application to miRNA prediction," *J. Biomed. Sci. Eng.*, vol. 06, no. 02, pp. 236–248, 2013, doi: 10.4236/jbise.2013.62a029.

[43] H. Kumar, P. Gupta, and R. P. Mahapatra, "Protocol based ensemble classifier for malicious URL detection," *Proc. 3rd Int. Conf. Contemp. Comput. Informatics, IC3I 2018*, pp. 331–336, 2018, doi: 10.1109/IC3I44769.2018.9007255.

[44] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL names say it all," *Proc. - IEEE INFOCOM*, pp. 191–195, 2011, doi: 10.1109/INFCOM.2011.5934995.

[45] 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.

[46] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious URLs: An application of large-scale online learning," *ACM Int. Conf. Proceeding Ser.*, vol. 382, 2009, doi: 10.1145/1553374.1553462.