DETEKSI ANCAMAN *URL* BERBAHAYA BERBASIS *EMBEDDING FEATURE EXTRACTION* MENGGUNAKAN METODE *ARTIFICIAL NEURAL NETWORK*

SKRIPSI

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

FARHAN RADHI ZUHRI

09011282126079

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2025

DETEKSI ANCAMAN *URL* BERBAHAYA BERBASIS *EMBEDDING FEATURE EXTRACTION* MENGGUNAKAN METODE *ARTIFICIAL NEURAL NETWORK*

SKRIPSI

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

OLEH:

FARHAN RADHI ZUHRI

09011282126070

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2025

HALAMAN PENGESAHAN

SKRIPSI

Deteksi Ancaman URL Berbahaya Berbasis Embedding Feature Extraction Menggunakan Metode Artificial Neural Network

Sebagai salah satu syarat untuk penyelesaian studi di
Program Studi S1 Sistem Komputer

Oleh:

FARHAN RADHI ZUHRI

09011282126070

Pembimbing 1 : **Aditya Putra Perdana Prasetyo, S.Kom., M.T.**
NIP 198810202023211018

Pembimbing 2 : **Abdurahman, S.Kom., M.Han**
NIP. 199410222024211018

Mengetahui

Ketua Jurusan Sistem Komputer

Dr. Ir. Sukemi, M.T

196612032006041001

AUTHENTIFICATION PAGE

FINAL TASK

Malicious URL Threat Detection Based on Embedding Feature Extraction Using Artificial Neural Network Method

As one of the requirements for completing the Bachelor's
Degree Program in Computer Systems

By:

FARHAN RADHI ZUHRI

09011282126070

Supervisor 1 : **Aditya Putra Perdana Prasetyo, S.Kom., M.T.**

NIP 198810202023211018

Supervisor 2 : **Abdurahman, S.Kom., M.Han**

NIP. 199410222024211018

Approved by

Head of Computer System Departement

Dr. Ir. Sukemi, M.T

196612032006041001

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada:

Hari : Jum'at

Tanggal : 25 Juli 2025

Tim Penguji:

1. Ketua : Dr. Ahmad Zarkasi, M.T.
2. Penguji : Huda Ubaya, M.T.
3. Pembimbing I : Aditya Putra Perdana Prasetyo, M.T.
4. Pembimbing II : Abdurahman, S.Kom., M.Han.

Mengetahui, 19/8/25
Ketua Jurusan Sistem Komputer

Dr. Ir. Sukemi, M.T
NIP. 196612032006041001

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : Farhan Radhi Zuhri

NIM : 09011282126070

Judul : Deteksi Ancaman URL Berbahaya Berbasis Embedding Feature Extraction Menggunakan Metode Artificial Neural Network

Hasil Pengecekan Software Turnitin: 1%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsur penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya saya bersedia menerima sanksi akademik dari Universitas Sriwijaya.

Demikian pernyataan ini saya buat dalam keadaan sadar dan tidak dipaksakan.

Indralaya, Agustus 2025

Farhan Radhi Zuhri
09011282126070

KATA PENGANTAR

Segala puji dan syukur atas kehadirat Allah SWT yang telah memberikan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan penulisan skripsi berjudul "**Deteksi Ancaman URL Berbahaya Berbasis *Embedding Feature Extraction* Menggunakan Metode *Artificial Neural Network***". Salawat dan salam kepada baginda Nabi Muhammad SAW yang telah menjadi guru terbaik dan suri tauladan bagi seluruh umat islam. Tidak lupa penulis menyampaikan ucapan terima kasih kepada seluruh pihak yang telah membantu, membimbing serta terus mendukung sehingga penulis dapat menyelesaikan laporan akhir magang & studi independen bersertifikat dengan baik, di antaranya:

1. Allah SWT yang telah memberi rahmat kesehatan, kenikmatan, kemudahan dengan kelancaran sehingga hamba dapat menyelesaikan laporan ini dengan baik.
2. Nabi Muhammad SAW yang telah membimbing umat islam dari alam kebodohan ke alam yang lebih berilmu pengetahuan yang telah dirasakan saat ini.
3. Bapak Lukman dan ibu Irdayanti selaku kedua orang tua penulis yang selalu mendoakan, mendukung serta memotivasi penulis dalam penyusunan skripsi ini.
4. Bapak Prof. Dr. Erwin, S.Si, M.Si selaku dekan fakultas Ilmu Komputer Universitas Sriwijaya.
5. Bapak Dr. Ir. Sukemi, M.T. selaku ketua jurusan Sistem Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.
6. Bapak Huda Ubaya, M.T selaku sekretaris jurusan Sistem Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Bapak Aditya Putra Perdana Prasetyo, S.Kom., MT selaku dosen pembimbing I yang telah bersedia untuk membimbing, mengarahkan, serta memberikan kritik dan saran dalam pembuatan skripsi.

8. Bapak Abdurahman, S.Kom., M.Han selaku dosen pembimbing akademik dan dosen pembimbing II yang telah bersedia untuk memberikan masukan, kritik dan saran dalam pembuatan skripsi.
9. Seluruh dosen jurusan Sistem Komputer yang telah memberikan ilmu dan nasehat kepada penulis.
10. Kak M. Angga Pratama selaku admin jurusan Sistem Komputer Universitas Sriwijaya.
11. Kak Dendi Renaldo Permana yang telah mendukung, membantu dan membimbing dalam proses penyusunan laporan skripsi.
12. Teman-teman yang tergabung pada grup botanisme Fakhrul, Reihan, Adam, Ade, Azriel, Quddus yang selalu memberikan semangat dan mendukung penulis.
13. Teman-teman Sistem Komputer 2021 yang telah mendukung penulis untuk menyelesaikan skripsi ini.

Penulisan menyadari bahwa kesempurnaan hanya milik Allah SWT. Sehingga sangat besar kemungkinan apabila skripsi ini masih terdapat banyak kekurangan. Penulis mengharapkan kritik dan saran dari berbagai pihak untuk memperbaiki laporan skripsi ini agar lebih baik. Akhir kata penulis pengucapkan terima kasih banyak kepada semua pihak yang terlibat maupun membantu penulis dalam proses penyusunan laporan skripsi ini.

Indralaya, Agustus 2025
Penulis,

Farhan Radhi Zuhri
NIM.09011282126070

Deteksi Ancaman *URL* Berbahaya Berbasis *Embedding Feature Extraction* Menggunakan Metode *Artificial Neural Network*

FARHAN RADHI ZUHRI (09011282126070)
Jurusan Sistem Komputer, Fakultas Ilmu Komputer
Universitas Sriwijaya
Email: farhanzuhri10@gmail.com

ABSTRAK

URL yang berbahaya sering digunakan dalam serangan siber, khususnya dalam upaya untuk mencuri infomasi sensitif, menyebarkan malware atau melakukan penipuan daring. Pelaku kejahatan siber biasanya akan menyebarkan serangan ini melalui iklan palsu, spam email dan berbagai cara lainnya untuk menarik perhatian korban. Berbagai cara telah dilakukan untuk mencegah serangan ini terus terjadi. Penelitian ini mengusulkan pendekatan inovatif dengan mengintergrasikan teknik *subword embedding* untuk ekstraksi fitur. SMOTE digunakan untuk mengatasi ketidakseimbangan antar kelas. Model klasifikasi *artificial neural network* (ANN) diusulkan dengan mempertimbangkan keunggulan model tersebut dalam menangkap informasi yang lebih kompleks.

Kata Kunci : URL berbahaya, subword embedding, SMOTE, artificial neural network.

Malicious URL Threat Detection Based on Embedding Feature Extraction Using Artificial Neural Network Method

FARHAN RADHI ZUHRI (09011282126070)
Department of Computer Systems, Faculty of Computer Science
Sriwijaya University
Email: farhanzuhri10@gmail.com

ABSTRACT

Malicious URLs are often used in cyber attacks, especially in an attempt to steal sensitive information, spread malware or commit online fraud. Cybercriminals will usually spread these attacks through fake advertisements, email spam and various other ways to attract victims' attention. There are various ways to prevent these attacks from continuing to occur. This research proposes an innovative approach by integrating subword embedding technique for feature extraction. SMOTE is used to overcome the imbalance between classes. An artificial neural network (ANN) classification model is proposed considering its superiority in capturing more complex information.

Keywords : malicious URL, subword embedding, SMOTE, artificial neural network

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi yang berkembang pesat telah membawa banyak kemudahan dalam kehidupan sehari-hari manusia. Internet kini menjadi bagian penting dalam kehidupan manusia, memfasilitasi akses informasi, komunikasi, dan berbagai layanan daring yang diberikan dengan mudah, namun dibalik kemudahan itu, ada ancaman serius yang dapat mencuri informasi–informasi sensitif pengguna internet, salah satunya adalah ancaman *cyber threat intelligence* (CTI). CTI adalah informasi tentang potensi ancaman siber yang dikumpulkan, dianalisis dan dibagikan untuk membantu organisasi atau individu dalam meningkatkan keamanan data [1]. CTI mencakup berbagai jenis ancaman, seperti serangan *malware*, *phising*, *ransomware* dan lain-lain. Informasi ini sangat penting untuk dipahami sebagai upaya dalam meningkatkan keamanan dan kewaspadaan pengguna internet terhadap ancaman-ancaman siber yang terus berkembang hingga saat ini [2].

Salah satu ancaman siber yang paling umum terjadi adalah *uniform resource relocator* (URL) berbahaya, URL adalah alamat dari sebuah web yang biasa diketik di dalam *browser* saat ingin mengunjungi suatu situs. URL berbahaya dirancang untuk mencuri informasi pribadi, menyebarkan virus, atau melakukan tindakan kejahatan lainnya. Cara kerja serangan ini yang tersembunyi sehingga sulit untuk dideteksi, hal ini menjadikan ancaman serius bagi pengguna internet baik individu maupun organisasi [3]. Pelaku akan menyebarkan informasi palsu atau iklan palsu untuk menarik perhatian pengguna internet untuk mengklik URL berbahaya yang sudah disiapkan. Pelaku akan melakukan berbagai cara untuk menginfeksi perangkat pengguna dengan *payload* berbahaya atau menipu korban dengan cara berinteraksi dengan pelaku untuk melakukan penipuan atau serangan lainnya [4].

Kesulitan saat mendeteksi serangan URL berbahaya mengakibatkan meningkatnya jumlah kerugian dan jumlah *website phising* yang ada pada internet.

Akibat dari kejadian tersebut, para pihak keamanan siber melakukan berbagai cara untuk mencegah serangan ini terjadi, terkhususnya serangan yang mengarah pada perusahaan. Penggunaan *machine learning* dirasa mampu untuk mendeteksi URL berbahaya [5].

Deteksi dan mitigasi terhadap URL berbahaya sangat penting dilakukan untuk mencegah kerugian yang lebih besar, baik dari segi finansial ataupun reputasi, oleh karena itu perlunya pengembangan metode yang efektif diperlukan untuk mendeteksi URL berbahaya sangat dibutuhkan [6]. Penggunaan *machine learning* untuk melakukan deteksi, klasifikasi dan analisis sangat diperlukan untuk mengidentifikasi pola – pola yang ada pada URL yang berpotensi berbahaya.

Pada penelitian [3] melakukan deteksi URL berbahaya menggunakan model *bidirectional encoder representations from transformers* (BERT) untuk melakukan tokenisasi dari URL untuk ekstraksi fitur berbasis kontekstual yang menggunakan mekanisme *self-attention* untuk memahami korelasi antar token dalam pendeteksian URL berbahaya. Akurasi yang didapatkan oleh model BERT terhadap tiga *dataset* yang berbeda, yaitu sebesar 98,78%, 96,71% dan 99,98%. Model ini menunjukan fleksibilitas dengan mengklasifikasikan URL dari berbagai *domain* secara efektif. Keunggulan model ini memampu menangangi berbagai format URL. Kekurangan model BERT memiliki komputasi yang lebih kompleks dan memerlukan lebih banyak sumber daya komputasi untuk pelatihan dan inferensi model.

Pada penelitian [7] melakukan identifikasi dan klasifikasi pada sekumpulan *dataset* URL berbahaya dengan 5 (lima) kelas yaitu *benign*, *spam, phising*, *malware* dan *defacement*. Metode yang digunakan adalah *ensemble learning* yang dapat menggabungkan beberapa metode *ensemble* seperti *ensemble of bagging trees (en_bag), ensemble of k-nearest neighbor, ensemble of boosted decision trees (en_bos),* dan *ensemble of subspace discriminator*. Hasil terbaik yang diraih oleh model *en_bag* dengan akurasi sebesar 99,3% pada klasifikasi 2 kelas dan 97,92% pada klasifikasi 5 kelas. Hasil menunjukan bahwa kelas *phising* mengalami tingkat kesalahan terbesar (*false positive* dan *false negative*) di antara kelas–kelas lainnya. Hal ini menunjukan bahwa model tersebut kesulitan dalam mengklasifikasikan URL *phising* dengan tingkat keakuratan yang rendah dibanding dengan kelas lain.

Pada penelitian [8] melakukan deteksi URL *phising* dan *spoofing* dari dua *dataset* berbeda menggunakan metode *hybrid convolutional neural network-long short term memory* (CNN-LSTM). Penelitian ini berfokus pada penggabungan metode *convolutional neural network* (CNN) dan *Long short term memory* LSTM untuk mendapatkan hasil akurasi yang baik terhadap deteksi *phising* dan *spoofing website*. Hasil akurasi yang di dapatkan dari hybrid CNN-LSTM yaitu 98,9% pada dataset UCL dan 96,8% pada dataset phistank. Kelebihan model ini adalah memiliki akurasi yang lebih tinggi dibandingkan model *deep learning* biasa dan mampu melakukan deteksi URL dengan penerapan waktu nyata. Kekurangan model hybrid CNN-LSTM memiliki komputasi yang cukup besar dibandingkan dengan metode deep learning konvensional sehingga tidak cocok diimplementasikan ke dalam perangkat yang ringan.

Pada penelitian [9] melakukan deteksi URL berbahaya menggunakan model *classification based on associations* (CBA) untuk mendeteksi *phising, malware* dan *drive-by-download* dengan memanfaatkan fitur leksikal URL dan fitur konten halaman web. Algoritma *apriori* digunakan untuk menghasilkan *class association rules* (CARs) yang dikemudian digunakan oleh pengklasifikasi CBA untuk memberi label URL sebagai berbahaya atau tidak berbahaya. Model ini mendapatkan hasil *accuracy* 95,8%, *recall* 97,67%, dan *precision* 91,3% dengan tingkat *false negative* sebesar 1,35%. Metode ini mempunyai interpretabilitas yang tinggi karena aturan yang dihasilkannya mudah dipahami dan kemampuannya untuk mendeteksi ancaman dengan menggabungkan fitur URL dan konten halaman web. Kekurangan dari model ini yaitu memiliki kompleksitas komputasi yang tinggi dan ketergantungan pada data pelatihan yang telah dilabel.

Pada penelitian [10] melakukan deteksi URL berbahaya dengan mengusulkan model *hybrid deep learning* yang menggabungkan *capsule network* (CapsNet) dan *independent recurrent neural network* (IndRNN) untuk meningkatkan deteksi URL berbahaya dengan memanfaatkan fitur semantik dan visual. Metode ini mengubah URL menjadi visual gambar dalam format *grayscale* untuk mengekstraksi pola tekstur lalu dilakukan pemprosesan menggunakan *character* dan *word embedding* untuk mengekstraksi pola semantik. Model ini mendapatkan hasil *accuracy* sebesar 99,78%, *recall* sebesar 99,98% dan *f1-score* sebesar 99,78%. Keunggulan model

ini meliputi akurasi yang tinggi, kemampuan untuk mendeteksi URL berbahaya yang tidak dikenal, dan pengurangan ketergantungan pada fitur manual. Kekurangan dari model ini yaitu memiliki komputasi yang tinggi dan potensi *overfitting* pada pola URL tertentu.

Penelitian ini diharapkan dapat menghasilkan model deteksi URL berbahaya berbasis *embedding feature extraction* dan *artificial neural network* (ANN) dengan parameter metrik yang optimal, seperti *accuracy*, *precision, recall,* dan *f1-score* yang sangat baik. Teknik *embedding* yang disematkan pada model ANN diharapkan dapat mengekstraksi fitur URL secara lebih informatif dan efisien, sehingga model ANN mampu mengidentifikasi pola URL berbahaya dengan akurat. Penelitian ini juga diharapkan mencapai kecepatan komputasi yang baik dan efisien dalam proses deteksi. Efisiensi komputasi ini diharapkan dicapai melalui optimisasi arsitektur ANN dengan penggunaan *embedding layer* yang mengurangi dimensi fitur tanpa kehilangan informasi. kombinasi akurasi tinggi dan kecepatan deteksi yang cepat, model ini diharapkan dapat diimplementasi secara praktis dalam sistem keamanan siber.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan diatas, penelitian ini akan melakukan deteksi pada URL berbahaya menggunakan metode ANN berbasis *embedding feature extraction*. Berikut adalah rumusan masalah pada penelitian ini.

1. Bagaimana teknik *embedding feature extraction* dapat merepresentasikan karakteristik URL berbahaya secara efektif untuk deteksi ancaman?
2. Bagaimana metode ANN dapat mendeteksi dan mengklasifikasikan URL berbahaya berdasarkan fitur *embedding* yang diekstraksi?
3. Bagaimana performa model ANN berbasis *embedding feature extraction* dalam melakukan deteksi URL *benign* dan URL *malicious* menggunakan metrik *accuracy, precision, recall,* dan *f1-score*?

1.3 Tujuan

Berdasarkan rumusan masalah yang sudah dipaparkan sebelumnya, tujuan penelitian ini adalah sebagai berikut.

1. Merancang teknik *embedding feature extraction* yang efektif untuk merepresentasikan karakteristik URL berbahaya.
2. Mengembangkan model deteksi dan klasifikasi menggunakan ANN terhadap URL berbahaya berdasarkan fitur *embedding* yang diekstraksi.
3. Mengevaluasi performa model ANN dalam melakukan deteksi URL yang terdeteksi *benign* dan *malicious* menggunakan metrik *accuracy, precision, recall,* dan *f1-score*.

1.4 Manfaat

Adapun manfaat dari penelitian ini adalah sebagai berikut.

1. Peningkatan keamanan jaringan dengan sistem yang dapat mendeteksi URL berbahaya.
2. Mengurangi risiko keamanan terhadap serangan URL berbahaya terhadap pengguna atau organisasi yang berpotensi merusak sistem atau mencuri data sensitif.
3. Memberi kontribusi terhadap perkembangan siber, khususnya dalam meningkatkan sistem deteksi dan pencegahan serangan berbasis URL.

1.5 Batasan Masalah

Penelitian ini memiliki batasan masalah yaitu sebagai berikut.

1. Penelitian ini berfokus pada deteksi URL berbahaya seperti *phising, malware* dan *defacement* menggunakan *embedding feature extraction* dan ANN.
2. Penelitian ini berfokus pada teknik *embedding* karena kemampuan merepresentasikan data teks URL ke dalam vektor numerik yang lebih bermakna.
3. Penelitian ini berfokus pada penggunaan metode ANN yang dapat menangani pola kompleks dan non-linier pada data URL.
4. Penelitian ini melakukan analisis pada URL yang dikumpulkan dari *dataset* publik.
5. Penelitian diterapkan pada konteks keamanan siber untuk proteksi pengguna internet umum.

6. Hasil penelitian ini menargetkan pengguna *internet* ataupun pengembang sistem keamanan.

1.6 Metodologi Penelitian

Metodologi penelitian yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Metode Studi Pustaka dan Literatur

   Penelitian ini menggunakan metode deksriptif eksploratif dengan pendekatan kuantitatif. Fokus utama adalah untuk menganalisis performa dari ANN dalam mendeteksi URL benign dan URL berbahaya yang berasal dari berbagai macam referensi dalam membantu pembuatan penelitian ini.
2. Metode Konsultasi

   Pada metode ini memilih bahan – bahan yang sesuai untuk mendukung penelitian penulis dalam mengatasi masalah yang akan ditemui.
3. Metode Pengumpulan Data

   Pada metode ini, peneliti mengumpulkan data melalui data publik yang telah tersedia untuk dilakukan penelitian.
4. Metode Pengujian

   Pada metode ini akan melakukan ekstraksi fitur berbasis *embedding* yang akan digunakan untuk pengujian deteksi menggunakan metode ANN.
5. Metode Analisis

   Langkah terakhir dalam penelitian ini untuk memperoleh akurasi dari model prediksi dari pengujian metode ANN yang akan dianalisis dan ditarik beberapa kesimpulan.

1.7 Sistematika Penulisan

Sistematika penulisan diperlukan untuk mempermudah melihat dan mengetahui isi pembahasan yang ada dalam penelitian ini secara menyeluruh. Penelitian ini terbagi dalam beberapa bab yaitu.

BAB I PENDAHULUAN

Bab ini meliputi latar belakang, perumusan masalah, tujuan, manfaat dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini berisi tentang pembahasan dari pengertian URL, feature extraction, data preprocessing, machine learning dan berbagai macam penjelasan.

BAB III METODOLOGI PENELITIAN

Pada bab ini membahas tentang metode yang digunakan dalam penelitian yang dilakukan oleh penulis untuk deteksi URL benign dan URL berbahaya.

BAB IV HASIL DAN PEMBAHASAN

Bab ini terdiri dari gambaran hasil dan analisa penelitian yang telah dilakukan peneliti dengan mengemukakan hasil tersebut dalam literatur.

BAB V PENUTUP

Bab ini berisi tentang kesimpulan dan saran dari penelitian yang sudah dilakukan. Kesimpulan dikemukakan dari hasil penelitian yang disampaikan secara objektif. Sedangkan saran berisi tentang solusi untuk mengatasi masalah dan kelemahan yang ada.

DAFTAR PUSTAKA

[1] N. Sun *et al.*, "Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives," *IEEE Commun. Surv. Tutorials*, vol. 25, no. 3, pp. 1748–1774, 2023, doi: 10.1109/COMST.2023.3273282.

[2] J. Liu *et al.*, "TriCTI: an actionable cyber threat intelligence discovery system via trigger-enhanced neural network," *Cybersecurity*, vol. 5, no. 1, Dec. 2022, doi: 10.1186/s42400-022-00110-3.

[3] M. Y. Su and K. L. Su, "BERT-Based Approaches to Identifying Malicious URLs," *Sensors (Basel).*, vol. 23, no. 20, 2023, doi: 10.3390/s23208499.

[4] N. Reyes-Dorta, P. Caballero-Gil, and C. Rosa-Remedios, "Detection of malicious URLs using machine learning," *Wirel. Networks*, vol. 30, no. 9, pp. 7543–7560, 2024, doi: 10.1007/s11276-024-03700-w.

[5] E. S. Shombot, G. Dusserre, R. Bestak, and N. B. Ahmed, "An application for predicting phishing attacks: A case of implementing a support vector machine learning model," *Cyber Secur. Appl.*, vol. 2, Jan. 2024, doi: 10.1016/j.csa.2024.100036.

[6] 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.

[7] Q. Abu Al-Haija and M. Al-Fayoumi, "An intelligent identification and classification system for malicious uniform resource locators (URLs)," *Neural Comput. Appl.*, vol. 35, no. 23, pp. 16995–17011, 2023, doi: 10.1007/s00521-023-08592-z.

[8] B. C. Ujah-Ogbuagu, O. N. Akande, and E. Ogbuju, "A hybrid deep learning technique for spoofing website URL detection in real-time applications," *J. Electr. Syst. Inf. Technol.*, vol. 11, no. 1, 2024, doi: 10.1186/s43067-023-00128-8.

[9] 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.

[10] 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.

[11] Z. Chen, Y. Liu, C. Chen, M. Lu, and X. Zhang, “Malicious URL Detection Based on Improved Multilayer Recurrent Convolutional Neural Network Model,” *Secur. Commun. Networks*, vol. 2021, 2021, doi: 10.1155/2021/9994127.

[12] Y. C. Chen, Y. W. Ma, and J. L. Chen, “Intelligent Malicious URL Detection with Feature Analysis,” *Proc. - IEEE Symp. Comput. Commun.*, vol. 2020-July, 2020, doi: 10.1109/ISCC50000.2020.9219637.

[13] A. Das, A. Das, A. Datta, S. Si, and S. Barman, “Deep Approaches on Malicious URL Classification,” *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, 2020, doi: 10.1109/ICCCNT49239.2020.9225338.

[14] P. L. Indrasiri, M. N. Halgamuge, and A. Mohammad, “Robust Ensemble Machine Learning Model for Filtering Phishing URLs: Expandable Random Gradient Stacked Voting Classifier (ERG-SVC),” *IEEE Access*, vol. 9, pp. 150142–150161, 2021, doi: 10.1109/ACCESS.2021.3124628.

[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] S. R. Abdul Samad *et al.*, “Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection,” *Electron.*, vol. 12, no. 7, 2023, doi: 10.3390/electronics12071642.

[17] A. Aljofey *et al.*, “An effective detection approach for phishing websites using URL and HTML features,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–19, 2022, doi: 10.1038/s41598-022-10841-5.

[18] S. Li and O. Dib, “Enhancing Online Security: A Novel Machine Learning Framework for Robust Detection of Known and Unknown Malicious URLs,” *J. Theor. Appl. Electron. Commer. Res.* , vol. 19, no. 4, pp. 2919–2960, 2024, doi: 10.3390/jtaer19040141.

[19] S. Remya, M. J. Pillai, K. K. Nair, S. Rama Subbareddy, and Y. Y. Cho, “An Effective Detection Approach for Phishing URL Using ResMLP,” *IEEE*

Access, vol. 12, no. May, pp. 79367–79382, 2024, doi: 10.1109/ACCESS.2024.3409049.

[20] M. Aljabri *et al.*, "Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions," *IEEE Access*, vol. 10, no. November, pp. 121395–121417, 2022, doi: 10.1109/ACCESS.2022.3222307.

[21] A. E. Belfedhal and M. A. Belfedhal, "A Lightweight Phishing Detection System Based on Machine Learning and URL Features," *Int. Conf. Manag. Bus. Through Web Anal.*, pp. 307–319, 2022, doi: 10.1007/978-3-031-06971-0_22.

[22] F. A. Ghaleb, M. Alsaedi, F. Saeed, J. Ahmad, and M. Alasli, "Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning," *Sensors*, vol. 22, no. 9, pp. 1–19, 2022, doi: 10.3390/s22093373.

[23] S. Singh and A. Mahmood, "The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures," *IEEE Access*, vol. 9, pp. 68675–68702, 2021, doi: 10.1109/ACCESS.2021.3077350.

[24] M. Omar, S. Choi, D. Nyang, and D. Mohaisen, "Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions," *IEEE Access*, vol. 10, no. June, pp. 86038–86056, 2022, doi: 10.1109/ACCESS.2022.3197769.

[25] D.-D. Science, "A Method of Short Text Representation Based on the," 2021.

[26] D. Wang, J. Su, and H. Yu, "Feature extraction and analysis of natural language processing for deep learning english language," *IEEE Access*, vol. 8, pp. 46335–46345, 2020, doi: 10.1109/ACCESS.2020.2974101.

[27] S. R. C. Liew and N. F. Law, "Use of subword tokenization for domain generation algorithm classification," *Cybersecurity*, vol. 6, no. 1, 2023, doi: 10.1186/s42400-023-00183-8.

[28] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[29] Y. Bao and S. Yang, "Two Novel SMOTE Methods for Solving Imbalanced

Classification Problems," *IEEE Access*, vol. 11, no. December 2022, pp. 5816–5823, 2023, doi: 10.1109/ACCESS.2023.3236794.

[30] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable Machine Learning for Scientific Insights and Discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020, doi: 10.1109/ACCESS.2020.2976199.

[31] C. S. Wickramasinghe, K. Amarasinghe, D. L. Marino, C. Rieger, and M. Manic, "Explainable Unsupervised Machine Learning for Cyber-Physical Systems," *IEEE Access*, vol. 9, pp. 131824–131843, 2021, doi: 10.1109/ACCESS.2021.3112397.

[32] R. A. Nazib and S. Moh, "Reinforcement Learning-Based Routing Protocols for Vehicular Ad Hoc Networks: A Comparative Survey," *IEEE Access*, vol. 9, pp. 27552–27587, 2021, doi: 10.1109/ACCESS.2021.3058388.

[33] T. Guillod, P. Papamanolis, and J. W. Kolar, "Artificial neural network (ann) based fast and accurate inductor modeling and design," *IEEE Open J. Power Electron.*, vol. 1, no. June, pp. 284–299, 2020, doi: 10.1109/OJPEL.2020.3012777.

[34] A. Al Bataineh, D. Kaur, and S. M. J. Jalali, "Multi-Layer Perceptron Training Optimization Using Nature Inspired Computing," *IEEE Access*, vol. 10, pp. 36963–36977, 2022, doi: 10.1109/ACCESS.2022.3164669.

[35] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[36] S. Poudel, "A Study of Disease Diagnosis Using Machine Learning," p. 8, 2023, doi: 10.3390/iech2022-12311.

[37] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis : A," *Healthcare*, pp. 1–30, 2022.

[38] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," pp. 37–63, 2020, [Online]. Available: http://arxiv.org/abs/2010.16061

[39] G. M. Foody, "Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient," *PLoS One*, vol. 18, no. 10 October, pp. 1–27, 2023, doi:

10.1371/journal.pone.0291908.

[40] R. Yacouby and D. Axman, “Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models,” pp. 79–91, 2020, doi: 10.18653/v1/2020.eval4nlp-1.9.

[41] S. Albahra *et al.*, “Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts,” *Semin. Diagn. Pathol.*, vol. 40, no. 2, pp. 71–87, 2023, doi: 10.1053/j.semdp.2023.02.002.