Zaprimljen: 10.6.2025.
Odobren za objavu: 23.7.2025.
Pregledni rad

STROJNO UČENJE ZA DETEKCIJU MREŽNE KRAĐE IDENTITETA ANALIZOM URL ADRESA

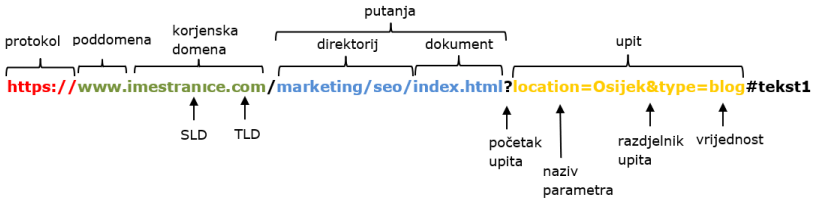
Ivana Hartmann Tolić[1], Mirta Vujnovac[2]
[1] Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek, Kneza Trpimira 2b, 31000 Osijek, Hrvatska
[2] III. gimnazija Osijek, Kamila Firingera 14, 31000 Osijek, Hrvatska
ePošta: ivana.hartmann@ferit.hr, mirta.vujnovac@gmail.com

Sažetak: U posljednje vrijeme phishing napadi i mrežne krađe identiteta predstavljaju značajnu prijetnju kibernetičke sigurnosti koristeći lažne web stranice kako bi prevarili korisnike u otkrivanju osjetljivih podataka. Phishing je oblik društvenog inženjeringa u kojem napadači daju pogrešne informacije putem lažnih web stranica kako bi prevarili žrtvu da ustupi osobne podatke radi dobivanja dodatnih informacija ili ostvarivanja financijske koristi. Zbog brzog razvoja tehnologije i taktika krađe identiteta te sve češće razmjene informacija putem interneta, potrebne su učinkovite metode za otkrivanje lažnih URL-ova. Cilj ovog rada bio je procijeniti učinkovitost različitih modela strojnog i dubokog učenja u klasifikaciji zlonamjernih i sigurnih web adresa bez analize sadržaja stranica. Eksperimentalni rezultati pokazuju da konvolucijske neuronske mreže (CNN) mogu postići točnost do 98,7 %, dok ensemble modeli poput Random Foresta i XGBoosta također bilježe visoku točnost iznad 96 %, čime se značajno nadmašuju tradicionalni pristupi poput logističke regresije.
Kako se strategije krađe identiteta nastavljaju razvijati, tako će adaptivni modeli poput ensemble tehnika učenja i arhitektura dubokog učenja biti ključni za zaštitu online sigurnosti te za razumijevanje učinkovitog suzbijanja novonastalih kibernetičkih prijetnji.

Ključne riječi: društveni inženjering, ensemble modeli, kibernetički napadi, klasifikacija URL adresa, SMOTE

1. Uvod

Phishing napadi ubrajaju se među najčešće prijetnje u kibernetičkoj sigurnosti jer iskorištavaju povjerenje korisnika i ranjivosti internetskih protokola. Prema *Anti-Phishing Working Group* (APWG) u četvrtom tromjesečju 2024. godine zabilježeno je 989.123 *phishing* napada, što predstavlja znatan porast u odnosu na prethodna tromjesečja te ukazuje na kontinuirani rast ove prijetnje.
Ovi napadi rabe zavaravajuće URL-ove kako bi žrtve preusmjerili na lažne stranice koje oponašaju legitimne servise. Tradicionalni pristupi obrani, kao što su crne liste i heuristike, ne mogu pratiti dinamične taktike napadača. Primjerice, autori u (Karim et al., 2023) pokazali su učinkovitost hibridnih modela temeljenih na kombinaciji klasičnih klasifikatora i dubokih mreža, dok Khan i suradnici u (Khan et al., 2024) ističu prednost dubokih modela u prepoznavanju skrivenih obrazaca u URL-ovima. Stoga se sve više primjenjuju metode strojnog i dubokog učenja koje na temelju karakteristika URL-a mogu predvidjeti radi li se o legitimnoj ili *phishing* adresi, te svojom učinkovitošću nadilaze tradicionalne metode.



Slika 1: Struktura URL-a s označenim elementima

2. Metodologija

Razvoj pouzdane metode za detekciju *phishing web* stranica zahtijeva pažljivo planiran eksperimentalni okvir koji obuhvaća prikupljanje, obradu, selekciju i modeliranje podataka. Ova faza od ključnog je značaja jer kvaliteta i relevantnost ulaznih značajki izravno utječu na točnost predviđanja. Cilj je metodologije identificirati reprezentativne značajke URL-ova, pravilno pripremiti podatke za algoritme te primijeniti optimalne modele uz preciznu evaluaciju. Analiza se temelji na javno dostupnim skupovima podataka koji sadrže tisuće označenih URL-ova (Slika 1). Svaki je URL predstavljen kao vektor značajki koje uključuju strukturne, simboličke i domenske informacije. Prije treniranja modela provodi se obrada podataka koja uključuje čišćenje i dekodiranje URL znakova, kodiranje kategoriziranih značajki, standardizaciju numeričkih vrijednosti (npr. duljina URL-a), eliminaciju redundantnih značajki korištenjem analiza korelacije i RFE-a (*Recursive Feature Elimination*). Odabir značajki dodatno se provodi pomoću testova značajnosti poput *chi-squared* i ANOVA testa te izračuna informacijske mjere kao što je uzajamna informacija (engl. *mutual information*), čime se osigurava da u model ulaze samo one varijable koje najviše doprinose klasifikaciji (Hajizada & Jahan, 2023). Standardizacija odabira značajki uključuje transformaciju podataka kako bi svi atributi bili u istom rasponu (npr. 0–1) i jednako ponderirani. Time se izbjegava dominacija značajki s većim numeričkim rasponima i omogućuje stabilnije treniranje modela. Zbog neravnoteže između broja *phishing* i legitimnih URL-ova koristi se SMOTE (*Synthetic Minority Over-sampling Technique*) metoda koja generira sintetičke uzorke manjinske klase i tako sprječava pristranost modela (Omari, Taoussi, & Oukhatar, 2025).

2.1. Značajke URL-ova i njihova uloga u detekciji

U analizi *phishing web* stranica značajke URL-ova igraju ključnu ulogu jer zlonamjerne adrese često slijede specifične obrasce dizajnirane kako bi zavarale korisnike. Analizom tih obrazaca moguće je identificirati sumnjive URL-ove bez potrebe za otvaranjem stranice što znatno ubrzava i osigurava proces detekcije.

Jedna od najčešće korištenih skupina značajki strukturne su značajke pri čijoj se analizi ispituju elementi kao što su ukupna duljina URL-a, broj točaka i specijalnih znakova te prisutnost IP adrese umjesto naziva domene. *Phishing* URL-ovi često imaju iznimno dugu strukturu ili sadrže nasumične nizove znakova kako bi prikrili pravi identitet dok je korištenje IP adrese umjesto domene pokušaj da se zaobiđu domenski filtri (Karim et al., 2023).

Druga su važna skupina značajke povezane s domenom uključujući duljinu same domene, broj poddomena, vršnu domenu (TLD – *top-level domain)* i prisutnost HTTPS protokola. *Phishing* adrese često rabe dugačke domene s višestrukim poddomenama kako bi oponašale legitimne *web* stranice.

Također, iako HTTPS protokol implicira sigurnost, zlonamjerne ga stranice često rabe kako bi stekle povjerenje korisnika što znači da prisutnost HTTPS-a sama po sebi nije jamstvo sigurnosti (Hajizada & Jahan, 2023).

Treća su kategorija semantičke značajke koje analiziraju sadržaj samog URL-a, poput prisutnosti riječi kao što su *login*, *verify*, *secure* ili *update*. Ove se riječi često rabe kako bi potaknule korisnika na djelovanje (npr. prijavu ili ažuriranje podataka) i stoga su česte u *phishing* kampanjama. Također se promatra broj ponavljanja znakova jer zlonamjerni URL-ovi ponekad rabe ponavljanje slova ili znakova kako bi vizualno zavarali korisnika (Almomani et al., 2022).

Značajke koje se temelje na WHOIS podatcima, a posebno starost domene i trajanje registracije, dodatno pomažu u prepoznavanju sumnjivih URL-ova. *Phishing* stranice često rabe novoosnovane domene koje su registrirane na kratak period što ih razlikuje od legitimnih *web* stranica koje obično imaju dužu povijest i stabilnu registraciju.

Modeli strojnog učenja s visokom preciznošću temeljem analize značajki URL-ova razlikuju legitimne od zlonamjernih bez potrebe za ulaskom u samu stranicu čime se povećava učinkovitost sustava za detekciju *phishing* napada.

2.2. Klasični modeli

Klasični modeli strojnog učenja već su dugi niz godina temelj pristupa u detekciji phishing napada. Njihova popularnost proizlazi iz jednostavne implementacije, visoke interpretabilnosti rezultata te relativno niskih računalnih zahtjeva što ih čini osobito prikladnima za sustave s ograničenim resursima ili kao polazište u fazama prototipiranja i istraživanja. Kada su podatci uravnoteženi, a značajke kvalitetno odabrane i dobro pripremljene, ovi modeli mogu imati visoku preciznost.

Klasični modeli strojnog učenja, poput logističke regresije, Naivnog Bayesovog klasifikatora i stabla odlučivanja, široko su rabljeni za binarnu klasifikaciju phishing i legitimnih URL-ova. Njihova točnost u recentnim istraživanjima najčešće doseže od 92 % do 96 %, ovisno o izboru značajki i kvaliteti obrade podataka (Alam et al., 2020; Almomani et al., 2022; Omari et al., 2023). Logistička regresija temelji se na linearnoj kombinaciji značajki za procjenu vjerojatnosti pripadnosti klasi, Naivni Bayes pretpostavlja neovisnost značajki dok stabla odlučivanja grade hijerarhijsku strukturu odluka. Iako su ovi modeli interpretabilni i brzi za treniranje, ensemble pristupi poput Random Foresta i Gradient Boostinga u više su radova postigli najvišu točnost, često iznad 96 %, zahvaljujući robusnosti i boljem prepoznavanju složenih obrazaca (Alam et al., 2020; Almomani et al., 2022; Omari et al., 2023). Primjerice, Random Forest je u navedenim istraživanjima redovito ostvarivao točnost od 96 % do 97,7 % dok su Gradient Boosting modeli postizali slične ili nešto više vrijednosti što ih čini najpouzdanijim izborom za detekciju phishing web stranica (Almomani et al., 2022; Omari et al., 2023).

2.3 Modeli temeljeni na dubokom učenju

S obzirom na sve veću složenost obrazaca *phishing* napada i kontinuirani rast količine dostupnih podataka, duboko učenje afirmiralo se kao izuzetno učinkovit pristup za klasifikaciju zlonamjernih URL-ova. Za razliku od tradicionalnih algoritama koji se oslanjaju na ručno definirane značajke, duboki modeli omogućuju automatsko učenje reprezentacija iz podataka čime se otkrivaju kompleksni i nelinearni odnosi među atributima što je osobito važno za detekciju prikrivenih obrazaca karakterističnih za phishing (Khan et al., 2024; Sahingoz et al., 2024).

Ključna prednost dubokih modela leži u njihovoj višeslojnoj arhitekturi koja omogućuje hijerarhijsko učenje značajki. Višeslojni perceptron (MLP) tretira sve

značajke URL-a kao ulazni vektor dok konvolucijske neuronske mreže (*CNN*) prepoznaju lokalne obrasce u nizovima znakova poput neuobičajenih domena ili zamjene slova brojevima što su česti obrasci u phishing adresama (Khan et al., 2024; Sahingoz et al., 2024). Rekurentne neuronske mreže, uključujući LSTM i GRU varijante, posebno su učinkovite u analizi sekvencijalnih podataka omogućujući prepoznavanje semantičkih i sintaktičkih obrazaca duž cijelog URL-a. Autoenkoderi omogućuju nenadzirano učenje komprimiranih reprezentacija URL-ova što može dodatno unaprijediti performanse kasnijih klasifikatora osobito u uvjetima ograničenih označenih podataka.

Empirijska istraživanja dosljedno potvrđuju da duboki modeli, osobito *CNN*, nadmašuju klasične pristupe s točnostima koje dosežu ili premašuju 98 % (Sahingoz et al., 2024; Khan et al., 2024). Unatoč povećanim zahtjevima za količinom podataka i računalnim resursima, njihova sposobnost generalizacije i prilagodbe novim prijetnjama čini ih temeljem suvremenih sustava za detekciju *phishing* napada.

2.4. Ensemble i hibridni modeli

Ensemble i hibridni modeli sve se češće rabe za poboljšanje točnosti, robusnosti i otpornosti sustava za detekciju *phishing* URL-ova. Osnovna je ideja da kombinacija više klasifikatora iskorištava njihove komplementarne prednosti i smanjuje slabosti čime se povećava ukupna učinkovitost modela (Karim et al., 2023).

Ensemble modeli, poput *Random Foresta*, *Gradient Boostinga* i *XGBoosta*, integriraju predikcije više osnovnih klasifikatora u jedinstvenu odluku. *Random Forest* kombinira predviđanja brojnih stabala odlučivanja dok *XGBoost* rabi *boosting* za sekvencijalno ispravljanje pogrešaka prethodnih modela. *Stacking* dodatno rabi predikcije različitih modela kao ulaze za meta-klasifikator čime se postiže dodatna robusnost. Hibridni modeli integriraju različite faze obrade, primjerice koristeći duboke modele (CNN) za automatsku ekstrakciju značajki, a klasične klasifikatore (*Random Forest*, MLP) za završnu odluku (Karim et al., 2023). Iako zahtijevaju više resursa i složeniju validaciju, istraživanja pokazuju da ensemble i hibridni modeli dosljedno nadmašuju pojedinačne pristupe osobito u okruženjima s neuravnoteženim klasama.

3. Rezultati i rasprava

Empirijska analiza potvrđuje da duboki modeli, značajno nadmašuju klasične pristupe u detekciji *phishing* URL-ova. Prema rezultatima Sahingoz i sur. (2024), *CNN* arhitektura postigla je najvišu točnost od 98,74 % na velikom, uravnoteženom skupu od više od pet milijuna URL-ova. RNN i druge duboke arhitekture također su pokazale visoku učinkovitost, ali s nižom točnošću i znatno duljim vremenima treniranja.

Klasični modeli poput logističke regresije i *Random Foresta* i dalje su relevantni zbog nižih računalnih zahtjeva i jednostavnosti implementacije, ali postižu znatno nižu točnost – primjerice, logistička regresija doseže 93,8 %, a *Random Forest* 87,7 % na istom skupu podataka (Sahingoz et al., 2024). Ensemble pristupi poput *XGBoosta* u drugim studijama također pokazuju visoku stabilnost i otpornost na neuravnotežene podatke s točnostima koje često prelaze 97 % (Karim et al., 2023; Omari et al., 2025). Hibridni modeli, koji kombiniraju duboke mreže za ekstrakciju značajki i klasične klasifikatore za završnu odluku, postižu optimalnu ravnotežu između točnosti i učinkovitosti (Karim et al., 2023). Primjena tehnike SMOTE za balansiranje klasa dodatno poboljšava performanse modela osobito u uvjetima neuravnoteženih podataka (Omari et al., 2025).

Analiza eksperimentalnih rezultata pokazuje da ne postoji univerzalno najbolje rješenje – izbor optimalnog

modela ovisi o karakteristikama podatkovnog skupa, dostupnim računalnim resursima i zahtjevima za interpretabilnošću i brzinom izvođenja. Buduća istraživanja trebala bi se usmjeriti na daljnje kombiniranje različitih pristupa i razvoj interpretabilnih sustava temeljenih na umjetnoj inteligenciji (Sahingoz et al., 2024). Usporedba točnosti različitih modela iz relevantne literature prikazana je u (Tablici 1).

Tablica 1. Usporedba točnosti modela prema izvorima iz literature

Model/Metoda	**Točnost**	**Autori**
CNN	98,7	Tang (2021)
Decision Tree	91,9-96,3	Omari (2023) Alam (2020) Karim (2023)
Logistic regression	87,1-98,89	Hajizada (2023) Sahingoz (2024) Tang (2021)
Naive Bayes	54,9-70,34	Khan (2024) Karim (2023) Omari (2025)
Random Forest	87,7-97,1	Alam (2020) Tang (2021) Khan (2024) Karim (2023) Omari (2023) Yang(2021)
Support Vector Machine	60,78-98,85	Hajizada (2023) Omari (2025) Tang (2021)

4. Zaključak

Analiza URL adresa pokazala se kao brz i pouzdan način za rano otkrivanje *phishing web* stranica, osobito kada analiza sadržaja nije moguća ili je preskupa. U radu su uspoređeni klasični i duboki modeli strojnog učenja za detekciju *phishing* URL-ova. Klasični modeli, poput logističke regresije i stabala odlučivanja, prikladni su u okruženjima s ograničenim resursima zbog jednostavnosti, interpretabilnosti i brze implementacije, dok duboke arhitekture poput CNN-a i RNN-a omogućuju prepoznavanje kompleksnih obrazaca i postižu točnost veću od 98 %, što ih čini pogodnim za sustave kojima je prioritet visoka preciznost. Ensemble i hibridni pristupi povećavaju robusnost kombiniranjem prednosti različitih modela. Ove se metode već koriste u komercijalnim sigurnosnim rješenjima kao što su filtri za *e-poštu*, sigurnosni dodaci za preglednike i zaštita korisnika tijekom pretraživanja. Implementacija modela omogućuje automatsku detekciju *phishing* pokušaja u stvarnom vremenu. Ipak, pristupi temeljeni na strojnom učenju suočavaju se s određenim ograničenjima, uključujući pristranost podataka koja može smanjiti učinkovitost na stvarnim, raznolikim URL-ovima, kao i izazove generalizacije na nepoznate vrste napada koji nisu bili zastupljeni u treniranju modela. Stoga je ključno kontinuirano ažurirati skupove podataka i nadzirati performanse modela. Buduća istraživanja trebaju razvijati interpretabilne duboke modele i multimodalne pristupe koji kombiniraju

URL analizu s dodatnim podacima poput DNS zapisa i WHOIS informacija, čime bi se izgradili otporniji i skalabilniji sustavi za prevenciju *phishing* napada u stvarnom vremenu. Posebno je važno ugraditi etička načela u dizajn i korištenje sustava strojnog učenja kako bi se osigurala transparentnost, pravednost i zaštita privatnosti korisnika, čime se jača povjerenje u pouzdanost rješenja za kibernetičku sigurnost.

5. Literatura

Alam, M. N., Sarma, D., Lima, F. F., Saha, I., Ulfath, R.-E., & Hossain, S. (2020). Phishing attacks detection using machine learning approach. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1173–1179). IEEE. https://doi.org/10.1109/ICSSIT48917.2020.9214225

Almomani, A., Alauthman, M., Shatnawi, M. T., Alweshah, M., Alrosan, A., Alomoush, W., & Gupta, B. B. (2022). Phishing website detection with semantic features based on machine learning classifiers: a comparative study. International Journal on Semantic Web and Information Systems (IJSWIS), 18(1), 1-24.

Hajizada, A., & Jahan, S. (2023, February). Feature selections for phishing urls detection using combination of multiple feature selection methods. In Proceedings of the 2023 15th International Conference on Machine Learning and Computing (pp. 444-450).

Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S. B., & Joga, S. R. K. (2023). Phishing detection system through hybrid machine learning based on URL. IEEE Access, 11, 36805-36822.

Khan, M. A., et al. (2024). Phishing website detection using deep learning models. Information Security Journal, 33(1), 12–28.

Omari, K. (2023). Comparative study of machine learning algorithms for phishing website detection. International Journal of Advanced Computer Science and Applications, 14(9).

Omari, K., Taoussi, C., & Oukhatar, A. (2025). Comparative Analysis of Undersampling, Oversampling, and SMOTE Techniques for Addressing Class Imbalance in Phishing Website Detection. International Journal of Advanced Computer Science & Applications, 16(2).

Sahingoz, O. K., BUBE, E., & Kugu, E. (2024). Dephides: Deep learning based phishing detection system. IEEE Access, 12, 8052-8070.

Tang, L., & Mahmoud, Q. H. (2021). A deep learning-based framework for phishing website detection. IEEe Access, 10, 1509-1521.

Yang, R., Zheng, K., Wu, B., Wu, C., & Wang, X. (2021). Phishing website detection based on deep convolutional neural network and random forest ensemble learning. Sensors, 21(24), 8281.

MACHINE LEARNING APPROACHES FOR PHISHING DETECTION BASED ON URL ANALYSIS

Abstract: Phishing attacks have posed a significant threat to cybersecurity in recent years. Phishing is a form of social engineering in which attackers provide misleading information via fake websites in order to trick the victim into disclosing private information to obtain further information or gain a financial advantage. With the rapid development of technology and phishing tactics, access to information and the frequent exchange of information, effective methods for detecting fake URLs are needed. The goal is to evaluate the effectiveness of different models in classifying malicious and legitimate web addresses without analyzing the content of the page. This study aimed to evaluate the effectiveness of various machine learning and deep learning models in classifying malicious and legitimate web addresses without analyzing page content. Experimental results show that convolutional neural networks (CNNs) can achieve accuracy rates of up to 98.7%, while ensemble models such as Random Forest and XGBoost also demonstrate high accuracy, exceeding 96%, significantly outperforming traditional approaches like logistic regression.
As phishing strategies continue to evolve, adaptive models such as ensemble learning techniques, deep learning architectures will be fundamental to securing online security and crucial to understanding how to effectively counter emerging cybersecurity threats.

Keywords: Cyber attacks, Ensemble models, SMOTE, Social engineering, URL classification