VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

FUNDAMENTINIŲ MOKSLŲ FAKULTETAS

INFORMACINIŲ SISTEMŲ KATEDRA

Laimonas Janutėnas

VAIZDO KLASTOJIMO GILIOJO MOKYMOSI METODAI

DEEP-LEARNING METHODS TO DETECT IMAGE FALSIFICATION

Baigiamasis magistro darbas

Duomenų gavybos technologijų studijų programa, valstybinis kodas 6211BX016

Studijų kryptis: Informatikos inžinerija

Vilnius, 2023

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

FUNDAMENTINIŲ MOKSLŲ FAKULTETAS

INFORMACINIŲ SISTEMŲ KATEDRA

TVIRTINU
Katedros vedėjas

_____
(Parašas)

_____
(Vardas, pavardė)

_____
(Data)

Laimonas Janutėnas

VAIZDO KLASTOJIMO GILIOJO MOKYMOSI METODAI

DEEP-LEARNING METHODS TO DETECT IMAGE FALSIFICATION

Baigiamasis magistro darbas

Duomenų gavybos technologijų studijų programa, valstybinis kodas 6211BX016

Studijų kryptis: Informatikos inžinerija

Vadovas: Prof. dr. Dmitrij Šešok _____ _____

(Pedag. vardas, vardas, pavardė) (Parašas) (Data)

Lietuvių kalbos konsultantas: _____ _____ _____

(Pedag. vardas, vardas, pavardė) (Parašas) (Data)

Vilnius, 2023

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
INFORMACINIŲ TECHNOLOGIJŲ KATEDRA

Studijų kryptis: Informatikos inžinerija
Studijų programa: Informacinės technologijos , valstybinis kodas 6211BX016
Specializacija: Duomenų gavybos technologijos

TVIRTINU
Katedros vedėjas
Dmitrij Šešok
2023-05-18

MAGISTRO BAIGIAMOJO DARBO UŽDUOTIS

Nr. DGTfm-21–3442

Vilnius

Studentas (-ė): Laimonas Janutėnas

Baigiamojo darbo tema: Vaizdo klastojimo giliojo mokymosi metodai

Baigiamojo darbo užbaigimo terminas pagal numatytą studijų kalendorinį grafiką.

BAIGIAMOJO DARBO UŽDUOTIS:

Tikslas: Ištirti mokslinėje literatūroje aprašytą „Deepfake“ aptikimo metodą ir jį patobulinti.

Uždaviniai:
1. Išanalizuoti „Deepfake“ giliojo mašininio mokymosi metodus, skirtus atpažinti suklastotus vaizdus.
2. Pasirinkti „Deepfake“ aptikimo būdą, ir identifikuoti atitinkamus duomenis reikalingus mašininiam mokymui.
3. Atlikti modelio tyrimą, pateikti rezultatus ir išvadas.
4. Pasiūlyti patobulintą suklastotų „Deepfake“ vaizdų atpažinimo metodą.

Planuojami rezultatai: Pasiūlyti pasirinkto mašininio mokymo modelio patobulinimą

Vadovas profesorius Dmitrij Šešok

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Informacinių technologijų katedra

ISBN | ISSN
Egz. sk. .....
Data .....-.....-.....

Antrosios pakopos studijų **Informacinių technologijų** programos magistro baigiamasis darbas 4

Pavadinimas	**Vaizdo klastojimo giliojo mokymosi metodai**
Autorius	**Laimonas Janutėnas**
Vadovas	**Dmitrij Šešok**

Kalba: lietuvių

Anotacija

Baigiamajame darbe nagrinėjamos „Deepfake“ giliojo mokymosi pagrindu sukuriamos vaizdo klastotės bei jų atpažinimo ir aptikimo metodai. Šiame darbe išnagrinėtas ir patobulintas mokslinėje literatūroje aprašytas „Deepfake“ atpažinimo metodas. Darbe atlikta išsami „Deepfake“ veikimo ir atpažinimo metodų mokslinės literatūros analizė, taip pat buvo atliktas pasirinkto mašininio mokymo modelio tyrimas, keičiant įvairius modelio parametrus, kurio metu gauti 135 skirtingi rezultatai. Šie rezultatai buvo analizuojami siekiant rasti optimalius modelio parametrus. Remiantis tyrimo rezultatais, buvo sukurtas naujas patobulintas modelis, kuris pasižymi 1,08 % didesniu tikslumu tam tikrais atvejais, palyginti su mokslinėje literatūroje aprašytu modeliu.

Baigiamasis darbas apima 63 puslapius, neįskaitant priedų, ir pateikia 26 iliustracijas, 9 lentelės bei 25 bibliografinių šaltinių.

Prasminiai žodžiai: Mašininis mokymas, vaizdo klastotės, gilusis mokymas

Vilnius Gediminas Technical University
Faculty of Fundamental Sciences
Department of Information Technologies

ISBN | ISSN
Copies No. .....
Date .....-.....-.....

Master Degree Studies **Information Technologies** study programme Master Graduation Thesis

Title	**Deep-Learning Methods for Image Falsification**
Author	**Laimonas Janutėnas**
Academic supervisor	**Dmitrij Šešok**

Thesis language: Lithuanian

Annotation

The master's thesis analyzes how deep learning based image forgeries such as "Deepfakes" are created and how they could be detected. In this paper, we analyze a "Deepfake" recognition method that is described in scientific literature and make improvements. This paper provides a comprehensive analysis of scientific literature on "Deepfake" generation and recognition methods, then a selected machine learning model's investigation was conducted, by changing various model parameters, during which 135 different results were obtained. These results were analyzed to find the most optimal model parameters. Based on the research results, a new improved model was developed, in which the accuracy is higher by 1,08 % in certain cases compared to the model described in scientific literature.

The final work consists of 63 pages, presents 26 illustrations, 9 tables, and 25 bibliographic sources.

Keywords: Machine learning, Deepfakes, Deep learning

Turinys

Įvadas 5
1. „Deepfake“ aptikimo mokslinių darbų analizė 8
    1.1. „Deepfake“ aptikimo metodai analizuojant vaizdo pasikeitimą laike 11
    1.2. „Deepfake“ aptikimo metodai analizuojant veido sritį 14
    1.3. Kiti „Deepfake“ aptikimo metodai 18
    1.4. „Deepfake“ duomenų rinkiniai 21
    1.5. „Deepfake“ aptikimo metodų apžvalga 22
2. „Deepfake“ atpažinimo modelio tyrimas ir tobulinimas 24
    2.1. Mašininio modelio parametrų tyrimas 27
3. „LRNet“ modelio tobulinimas ir rezultatai 45
Išvados 53
Šaltiniai 54

Paveikslų sąrašas

1 pav. „Deepfake“ pavyzdys (Castillo Camacho & Wang, 2021) 8
2 pav. „Deepfake“ generavimo principas (Afchar et al., 2018) 9
3 pav. GAN modelio struktūra (Google, 2023) 10
4 pav. Pasiūlytas modelis, kuriame tikrinamos paklaidos veide per tam tikrą laikotarpį (Lee et al., 2021) 11
5 pav. tikrų ir netikrų vaizdo įrašų neatitikimas laike (Gu et al., 2021) 12
6 pav. vaizdo ir garso sinchronizavimo neatitikimas „Deepfake“ vaizduose (Agarwal et al., 2020) 13
7 pav. „LRNet“ modelio architektūra (Sun et al., 2021) 14
8 pav. ADD „Deepfake“ aptikimo sistema (Khormali et al., 2021) 15
9 pav. biologinių ženklu aptikimas „Deepfake“ vaizduose (Ciftci et al., 2020) 16
10 pav. trijų srautų „Deepfake“ aptikimo būdas (Yavuzkilic et al., 2021) 17
11 pav. Dėmesiu pagrįsta klastotės aptikimo architektūra (Ganguly et al., 2022) 18
12 pav. suklastotų vaizdų aptikimas naudojant maišos reikšmes (Tanaka et al., 2021) 19
13 pav. „Deepfake“ aptikimas pagal dažnių bangų koeficientus (Wolter et al., 2022) 20
14 pav. „Deepfake“ aptikimas pagal dažnių srities analizę (Durall et al., 2020) 21
15 pav. „LRNet“ modelio struktūra 26
16 pav. Mokymosi greičio ir išmetimo procento įtaka AUC. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 31
17 pav. Mokymosi greičio ir išmetimo procento įtaka tikslumui. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 32
18 pav. Mokymosi greičio ir išmetimo procento įtaka F1. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 33
19 pav. Mokymosi greičio ir įrašo ilgio įtaka AUC. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 34
20 pav. Mokymosi greičio ir įrašo ilgio įtaka tikslumui. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 35
21 pav. Mokymosi greičio ir įrašo ilgio įtaka F1. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 36
22 pav. Išmetimo procento ir įrašo ilgio įtaka AUC. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 37
23 pav. Išmetimo procento ir įrašo ilgio įtaka tikslumui. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 38
24 pav. Išmetimo procento ir įrašo ilgio įtaka F1. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 39
25 pav. Modelio treniravimo laikas priklausomai nuo įrašo ilgio. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius 40
26 pav. a) patobulintas modelis, b) tyrimo metu naudojamas modelis – „LRNet“. 46

Lentelių sąrašas

1 lentelė. „Deepfake“ aptikimo metodų palyginimas 22
2 lentelė. „LRNet“ modelio rezultatai 27
3 lentelė. Atkartoti „LRNet“ modelio rezultatai 27
4 lentelė. Rezultatų palyginimas pagal optimizatorių 29
5 lentelė. Palyginimas straipsnio, atkartoto ir geriausių „Adam“ ir „RMSprop“ tyrimo rezultatais, su „c23“ duomenų lygiu 40
6 lentelė. Palyginimas straipsnio, atkartoto ir geriausių „Adam“ parametrų rezultatais atlikus 5 dalių kryžminę patikrą, su visais duomenų lygiais. 42
7 lentelė. Geriausių parametrų modelio rezultatų patikimumas 44
8 lentelė. Palyginimas straipsnio, atkartoto ir geriausių „Adam“ parametrų rezultatais ir patobulinto modelio, su visais duomenų lygiais, atliekant 5 dalių kryžminę patikrą. 48
9 lentelė. Patobulinto modelio rezultatų patikimumas 50

Santrumpos

ReLU – padalyta tiesinė funkcija, kuri išves įvestį, jei ji yra teigiama, priešingu atveju ji išves nulį.

GAN – yra mašininio mokymosi modelis, kuriame du neuroniniai tinklai konkuruoja tarpusavyje, siekdami tikslesnių prognozių.

AE – yra dirbtinio neuroninio tinklo tipas, naudojamas efektyviam nepažymėtų duomenų kodavimui išmokti.

„Deepfake“ – vaizdas apie asmenį, kuriame jo veidas ar kūnas buvo skaitmeniniu būdu pakeistas taip, kad jis atrodo kaip kažkas kitas.

Artefaktas – bet kokios informacijos supratimo ar pateikimo klaida, kurią sukėlė kompiuterinė įranga ar technika.

Įvadas

Vaizdo redagavimo priemonės leidžia pakeisti skaitmeninius vaizdus kaip norima ar sukurti visiškai naujus ir tikroviškus vaizdus, net neturint realaus pagrindo. Taip pat, kompiuterinė įranga tapo pakankamai pažangi ir prieinama visiems, kad iš esmės tai gali padaryti bet koks žmogus. Dėl šių priežasčių atsirado naujos kūrybos formos ir net naujos profesijos, kurios specializuojasi į vaizdų koregavimą, kuriančios naujus ir kūrybiškus vaizdus, kad galėtų geriau perteikti emocijas ar naujas idėjas. Tai papildant internetu, atsirado galimybė šias idėjas paskleisti greitai bei leisti tobulinti šias technologijas sparčiau nei bet kada.

Tačiau yra ir kita šių technologijų pusė. Kadangi modifikuoti vaizdus tapo prieinama ir efektyvu, atsirado ir daug kenkėjiškų galimybių jomis pasinaudoti. Vaizdus galima klastoti ir tai galima realizuoti labai tikroviškai, kad žmogaus akys niekada negalės atskirti ar tai padirbtas vaizdas. Suklastoti vaizdai gali būti naudojami skleisti dezinformaciją, pateikti vaizdus kurie pakreiptų žmonių nuomonę viena ar kita linkme. O internetu informacija pasklinda greičiau nei galima įrodyti vaizdo autentiškumą. Anksčiau, kai nebuvo šių technologijų, nuotraukos ar filmuota medžiaga galėjo būti svarus faktų įrodymas. Dabar joks internete ar žiniasklaidoje esantis vaizdas nebėra patikimas, šių vaizdų naudojimas teismuose nebėra tinkamas. Atsiranda galimybė pateikti suklastotus įrodymus arba paneigti visus įrodymus susijusius su vaizdais. Dėl šių priežasčių svarbu turėti įrankius, kurie gali padėti mums suprasti tiesą. Atsirado poreikis kurti įrankius, kurie galėtų patikrinti vaizdų autentiškumą.

Viena iš tokių naujausių vaizdo klastočių yra „Deepfake“. Ši technologija naudojantis mašininio mokymo algoritmais gali pakeisti žmonių veidus kitais. Sistema yra apmokoma iš žmogaus veido nuotraukų, o vėliau tikroviškai pakeičia šiuo veidu, kitą veidą. Nors ši technologija dar turi savo trūkumų išgaunant labai kokybiškas klastotes, bet turint gerą duomenų rinkinį ir tinkamas sąlygas, gali būti sukurta įtikinama klastotė. Šios klastotės yra ypač pavojingos, nes jos gali paveikti žmones asmeniškai ar apkaltinti kitą žmogų teisme, todėl reikia turėti įrankius, galinčius įrodyti šių vaizdų autentiškumą.

Tyrimo objektas – suklastotų vaizdų aptikimo būdai

Darbo tikslas – ištirti mokslinėje literatūroje aprašytą „Deepfake“ aptikimo metodą ir jį patobulinti.

Darbo uždaviniai:

1. Išanalizuoti „Deepfake“ giliojo mašininio mokymosi metodus, skirtus atpažinti suklastotus vaizdus.

2. Pasirinkti „Deepfake“ aptikimo būdą, ir identifikuoti atitinkamus duomenis reikalingus mašininiam mokymui.
3. Atlikti modelio tyrimą, pateikti rezultatus ir išvadas.
4. Pasiūlyti patobulintą suklastotų „Deepfake“ vaizdų atpažinimo metodą.

Temos naujumas – mašininis mokymas šiuo metu yra labai populiarus visose srityse. Nors vaizdų klastojimas yra pakankamai sena tema, klastojimas naudojantis mašininiu mokymu yra pakankamai nauja sritis. Suklastotų vaizdų kiekis internete per pastaruosius metus sparčiai padidėjo ir tikėtina tik augs, todėl svarbu surasti būdą kaip atskirti netikrus vaizdus nuo tikrų.

Temos aktualumas – šiuo metu dar nėra labai gero sprendimo kaip atpažinti suklastotus vaizdus, nors vaizdų klastojimas kol kas nėra masinė problema, tikėtina, kad netolimoje ateityje gali atsirasti daug atvejų kur reikės įsitikinti vaizdo autentiškumu. Teismuose gali tekti patikrinti kiekvieną vaizdinį faktą, kol jį bus galima laikyti įrodymu. Taip pat internete gali tekti įrengti automatines sistemas, kur patikrintų nuotraukos autentiškumą, tam kad išvengti masinės dezinformacijos.

Tyrimo metodika – analitinei daliai atlikti buvo naudojama su tema susijusių mokslinių straipsnių literatūros analizė.

Mokslinė darbo vertė – pasiūlytas ir įgyvendintas naujas metodas suklastotų vaizdų aptikimui, kuris pagerintų „Deepfake“ atpažinimą.

Darbo rezultatai:

1. Atlikta literatūros šaltinių, susijusių su „Deepfake“ veikimo ir aptikimo metodais, analizė.

2. Atliktas pasirinkto mašininio mokymo modelio tyrimas, keičiant įvairius modelio parametrus.

3. Pasiūlytas pasirinkto mašininio mokymo modelio patobulinimas.

Darbo struktūra:

1. Įvadas, jame pateikiamas darbo aktualumas, tikslas, uždaviniai, temos naujumas, aktualumas, darbo metodika, vertė, rezultatai ir struktūra.

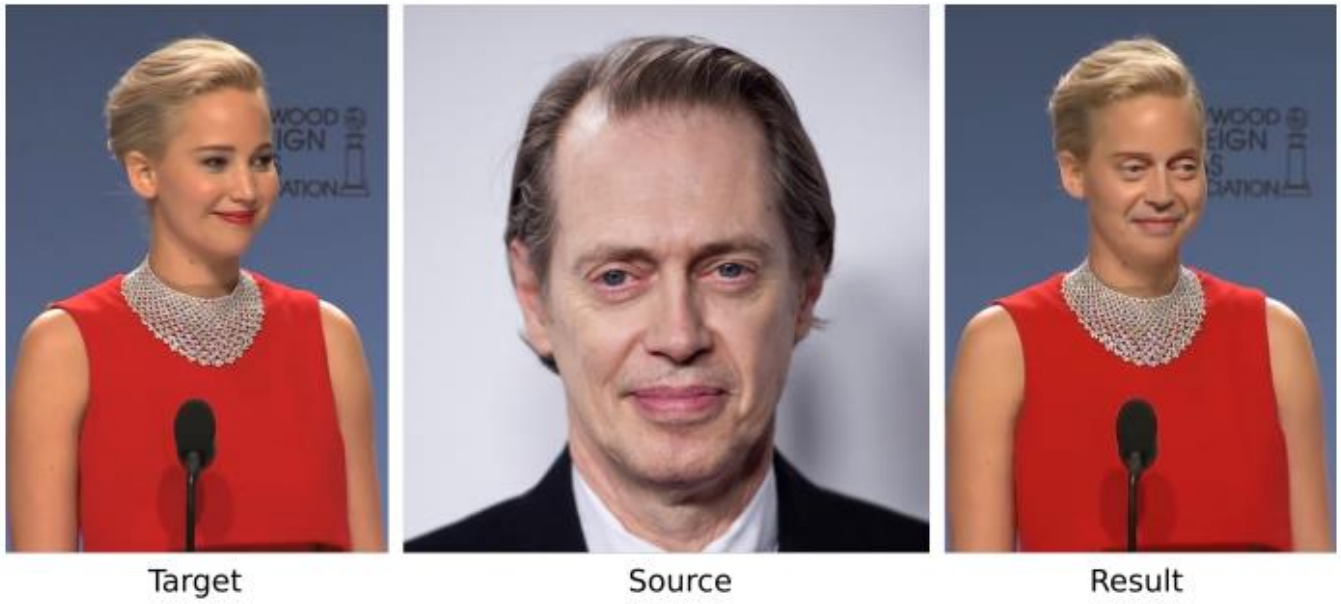
2. Pirmajame skyriuje analizuojamas „Deepfake“ vaizdų generavimo metodas. Taip pat, analizuojama naujausia „Deepfake“ vaizdų atpažinimo metodų mokslinė literatūra. Galiausiai palyginami šie aptikimo metodai.

3. Antrame skyriuje pateikiamas mašininio mokymo modelio tyrimas, kuriame modelis buvo treniruojamas keičiant įvairius modelio parametrus. Viso ištirti ir palyginti 135 skirtingi šių parametrų kombinacijų rezultatai.

4. Trečiajame skyriuje pasiūlytas patobulintas mašininio mokymo modelis ir palygintas tiek su pasirinkto modelio rezultatais, tiek su tyrimo rezultatais.

1. „Deepfake“ aptikimo mokslinių darbų analizė

Pastaraisiais metais atsirado daugybė gilaus mokymosi metodų, galinčių sukurti netikrą vaizdinį turinį. Šie metodai, bendrai vadinami „Deepfake“, jie apima automatinius kodavimo įrenginius (angl. *autoencoder*) AE ir GAN (angl. *generative adversarial network*) (P. Yang et al., 2020). GAN modeliai yra naudojami įvairiose srityse ir pakeitė daug laiko reikalaujančias užduotis, skirtas aukštos kvalifikacijos ekspertams, į paprastas ir greitas operacijas. Daugeliu atvejų, ši programa yra naudojama pakeisti asmenį (arba asmens veidą) esamame vaizde ar vaizdo įraše su kitu asmeniu (ar kito asmens veidu). 1 paveiksle pateikiamas pavyzdys, kaip pakeičiamas veidas (vaizdas kairėje), naudojant šaltinį (vaizdas viduryje) ir gaunamas naujas suklastotas kadras (vaizdas dešinėje). Nors ši nuotrauka yra rodoma tik mokomiesiems tikslams, tokia technika gali turėti rimtesnį poveikį kitose situacijose, pavyzdžiui, nuotraukos gali būti panaudotos politiniams tikslams. Pastaruoju metu yra skiriamas didelis dėmesys tokio tipo turinio aptikimui, dėl to kad „Deepfake“ yra pakankamai nesudėtinga sukurti, o gali turėti didelį neigiamą poveikį (Castillo Camacho & Wang, 2021).

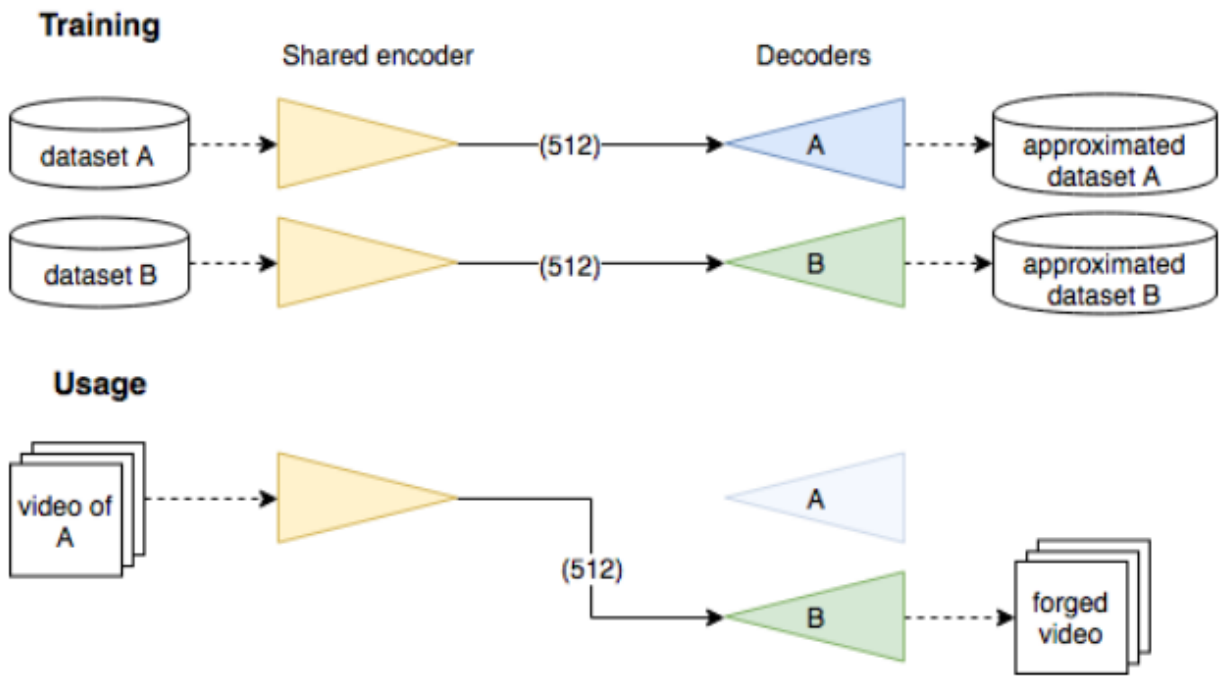


1 pav. „Deepfake“ pavyzdys (Castillo Camacho & Wang, 2021)

„Deepfake“ pirmą kartą pasirodė 2017 m., o vėliau šią techniką patobulino nedidelė bendruomenė, kad sukurtų programą „FakeApp“. Pagrindinė idėja yra lygiagretus dviejų automatinių kodavimo įrenginių mokymas. Jų architektūra gali skirtis priklausomai nuo failų dydžio, pageidaujamo mokymo laiko, numatomos kokybės ir turimų išteklių. Užkodavimo paskirtis yra sumažinti duomenų dimensijas iš įvesties sluoksnio, sumažinant kintamųjų skaičių. Atkodavimo tikslas yra naudoti tuos gautus

kintamuosius ir išgauti apytikslį įvesties vaizdą. Optimizavimo fazėje atliekamas palyginimas tarp įvesties ir sukurto apytikslio vaizdo, ir bandoma sumažinti skirtumą tarp šių vaizdų (Afchar et al., 2018).

„Deepfake“ vaizdų generavimo procesas yra dviejų skirtingų žmonių A ir B, veidų nuotraukų surinkimas, tada automatinio užkoduotojo EA apmokymas atkurti A veidus iš A veido vaizdų duomenų rinkinio, o automatinis užkoduotojas EB apmokomas atkurti B veidus iš B veido vaizdų duomenų rinkinio. Pagrindinis dalykas yra dviejų automatinių užkodavimo įrenginių EA ir EB užkodavimo svorių pasidalijimas, tačiau jų atkodavimas yra atskirtas. Atlikus optimizavimą, bet koks vaizdas, kuriame yra A veidas, gali būti užkoduotas naudojant šį bendrą užkodavimo įrenginį, bet atkoduotas naudojant EB atkoduotoją. Šis principas parodytas 2 paveiksle (Afchar et al., 2018).

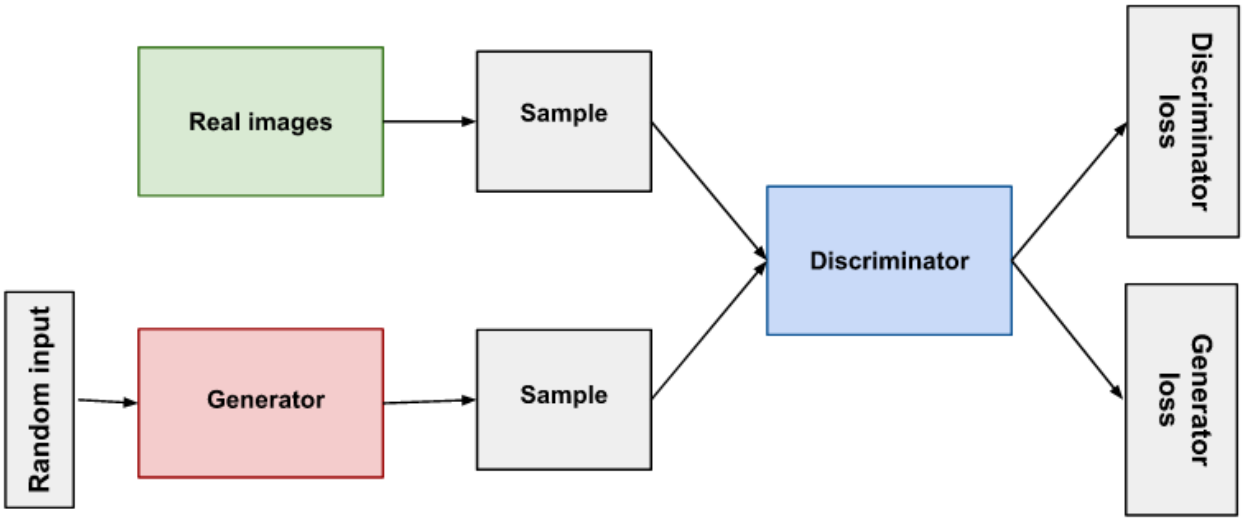


2 pav. „Deepfake“ generavimo principas (Afchar et al., 2018)

Bendra šio modelio idėja yra turėti vieną užkodavimo įrenginį, suteikiantį galimybę koduoti bendrą informaciją apie apšvietimą, veido padėtį ir išraišką, ir tada turėti atskirus užkodavimo įrenginius kiekvienam asmeniui, kad būtų galima atkurti veido charakteristikas, asmens veido formą ir detales. Tai atskiria kontekstinę informaciją ir morfologinę informaciją. Praktiniai šio modelio rezultatai yra įspūdingi, o tai paaiškina technologijos populiarumą. Paskutinis žingsnis, kad būtų galima gauti vaizdo įrašą, tai ištraukti ir sulygiuoti veidus iš kiekvieno kadro, sugeneruoti kitą veidą su tuo pačiu apšvietimu ir išraišką, ir tada vėl sujungti į vieną vaizdo įrašą (Afchar et al., 2018).

Nepaisant privalumų ši technika nėra be klaidų. Veidų ištraukimas ir jų reintegravimas gali nepavykti, ypač veido okliuzijos atveju: kai kurie kadrai gali neturėti veido išraiškos, būti neryškūs arba turėti dvigubą veido kontūrą (angl. *doubled facial contour*). Tačiau šių klaidų galima lengvai išvengti naudojant pažangesnius tinklus. Dar daugiau, ir tai pasakytina apie kitas programas, automatiniai užkoduotojai yra linkę prastai atkurti smulkias detales, nes nuotraukos dažniausiai būna suspaustos įvairių failų formatų, todėl rezultatas dažnai atrodo šiek tiek neryškus. Nors kai kurie sprendimai gali gerai atkurti žmogaus veido detales, gautas veidas vis tiek praranda tikroviškumą, nes linkęs panašėti į pradinį veidą. (Afchar et al., 2018).

GAN taip pat yra viena iš technologijų naudojama sukurti „Deepfake“. Generatyviniai priešpriešiniai tinklai (GAN) yra galinga dirbinio intelekto technologija, kuria siekiama sukurti naujus, kuo panašesnius duomenis į mokymosi imtį. GAN sąvoką pirmą kartą pristatė 2014 metais, o šiandien ši technologija naudojama įvairiose srityse, tokiose kaip vaizdų generavimas, vaizdo stiliaus atkartojimas, teksto generavimas ir t.t. GAN struktūra yra iš esmės dviejų neuroninių tinklų priešprieša, iš kurių vienas yra generatorius (G), o kitas - diskriminatorius (D). Generatorius siekia sukurti kuo panašesnius duomenis, o diskriminatorius siekia atskirti tikruosius nuo suklastotų duomenų. Šie du tinklai mokosi vienas iš kito, tobulindami savo gebėjimus. Modelio architektūra pavaizduota 3 paveiksle.



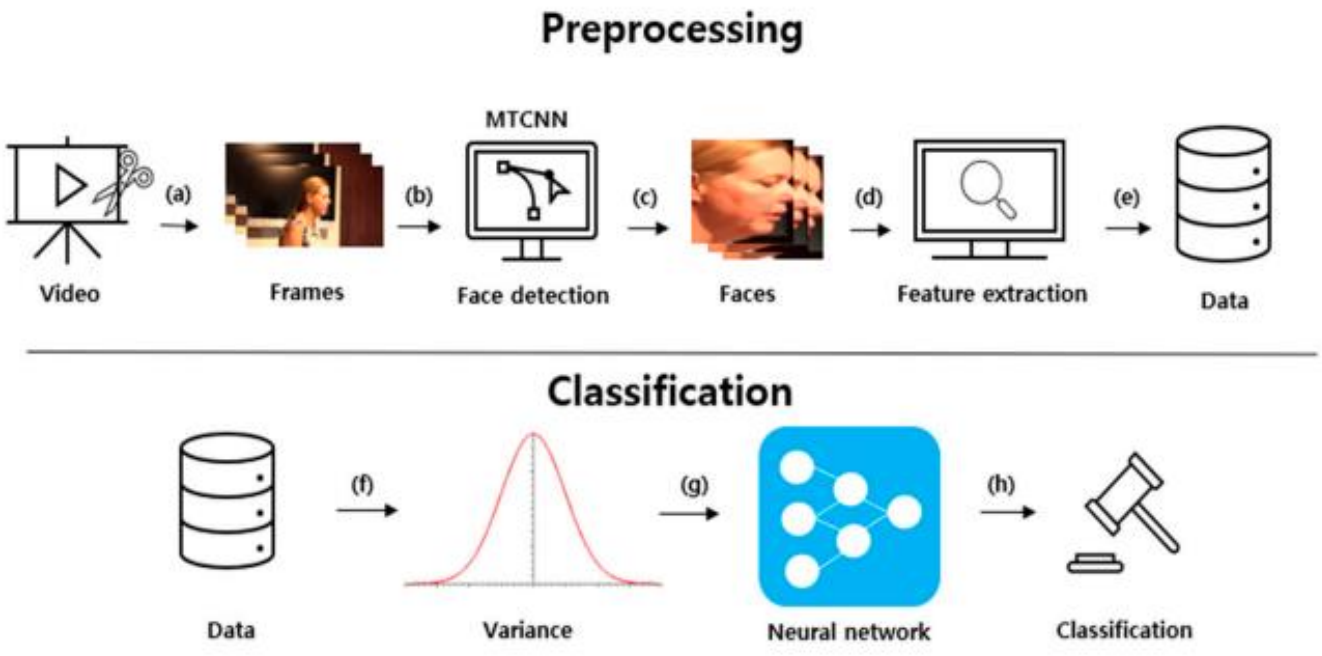
3 pav. GAN modelio struktūra (Google, 2023)

Generatorius pradeda nuo triukšmo (atsitiktinių skaičių), kurį pamažu transformuoja į norimą duomenų formą. Tuo tarpu diskriminatorius lygina generuotus duomenis su tikraisiais, stengdamasis kuo

tiksliau juos atskirti. Procesas vyksta žingsnis po žingsnio, kuomet generatorius vis gerina savo duomenų kokybę, o diskriminatorius tampa vis gudresnis atskirdamas nuo tikrųjų duomenų. Šis procesas vadinamas priešpriešiniu mokymusi ir tęsiasi, kol pasiekiamas pusiausvyros taškas, kai diskriminatorius nebegali patikimai atskirti generuotų duomenų nuo tikrųjų. Tokiu būdu GAN sugeba pasiekti aukšto lygio generavimo rezultatus, kurie dažnai yra sunkiai atpažįstami žmogaus akimi. Vis dėlto, nors GAN technologija pasižymi dideliu potencialu, ji kelia keletą iššūkių, tokių kaip modelių nestabilumas, sudėtingas mokymasis ir kontrolės stoka (Google, 2023). O „Deepfake“ sugeneruotų vaizdų atveju, GAN sukurtuose vaizduose taip pat lieka aptinkamų artefaktų ir klaidų, kurios nors ir nėra matomos žmogaus akiai, bet gali būti pastebėtos kompiuterio.

1.1. „Deepfake“ aptikimo metodai analizuojant vaizdo pasikeitimą laike

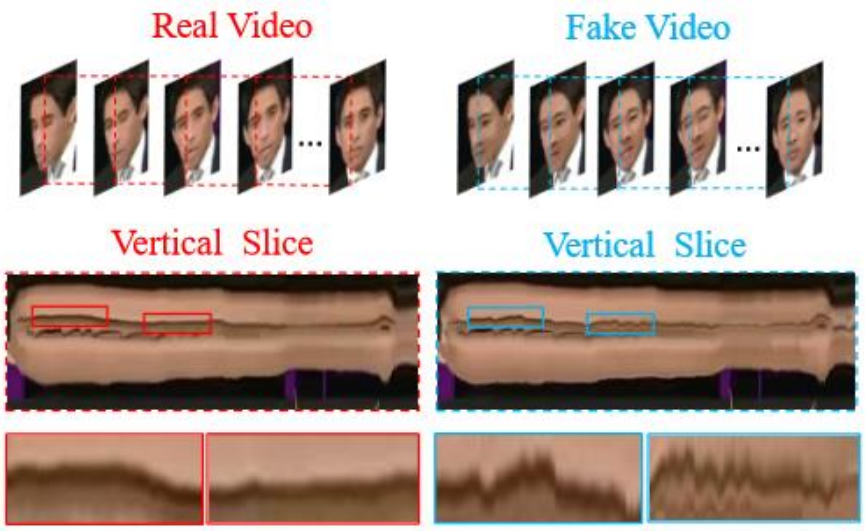
Lee et al. (2021) siūlo stebėti kaip keičiasi veidas vaizdo įrašuose ir ar jis ne per daug pasikeičia, per tam tikrą laikotarpį. 4 paveiksle parodyta siūloma sistemos struktūra: metodas yra padalintas į pirminio apdorojimo ir klasifikavimo procesus; išankstinio apdorojimo procesas išgauna veidą iš kadro išskiria veido ypatybes ir tada nustato skirtumą tarp kadrų. Klasifikavimo procesas aptinka klastotes, tikrindamas paklaidas atsirandančias per tam tikrą kadrų skaičių iš apdorotų duomenų.



4 pav. Pasiūlytas modelis, kuriame tikrinamos paklaidos veide per tam tikrą laikotarpį (Lee et al., 2021)

Pirmiausia, vaizdo įrašas suskirstomas į kadrus. Tada kiekviename kadre aptinkama ir iškerpama veido dalis naudojant MTCNN (angl. *Multi-Task Cascaded Convolutional Neural Networks*) įrankį. Iškirpti veidai sutvarkomi ir toliau išskiriami įvairūs veido požymiai. Šie požymiai atrandami naudojant skaičiavimus, klasterizavimą ir filtravimą. Paimamas tam tikras skaičius apskaičiuotų požymių ir paklaidų tarp šių požymių. Tada šie duomenys yra naudojami neuroninio tinklo mokymui, kuris pagal tai klasifikuoja tikrus ir netikrus vaizdo įrašus.

Gu et al. (2021) teigia, kad „Deepfake“ vaizdo įrašai yra sukuriami manipuliuojant kiekvieną kadrą atskirai ir tik tada šie kadrai yra sudedami į vaizdo įrašą, todėl „Deepfake“ vaizdo įrašų veido sintezė turi ne tik nenuoseklią informaciją apie erdvę (kurioje yra veidas), bet ir apie laiką (5 paveikslas).

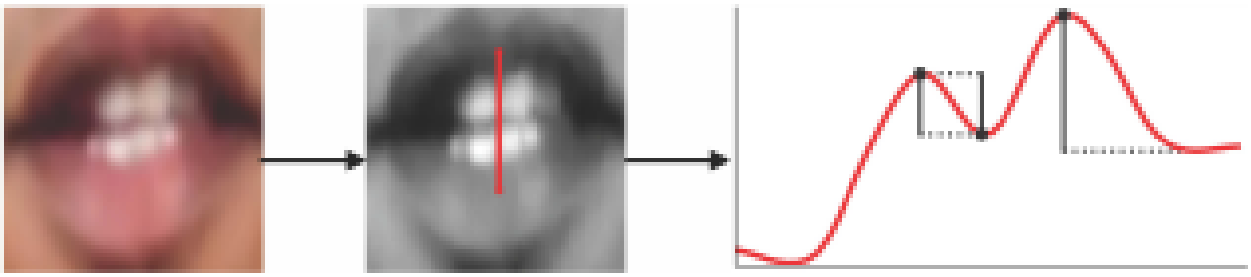


5 pav. tikrų ir netikrų vaizdo įrašų neatitikimas laike (Gu et al., 2021)

Kiekvienas kadras vaizdo įraše skaidomas vertikaliai fiksuotoje horizontalioje padėtyje ir tada jungiamas, kad suformuotų žemėlapį, kuriame tikrųjų vaizdų perėjimai yra tolydūs, o suklastotųjų – pasižymi šiurkštumu. Norint atskirti „Deepfake“ ir originalius vaizdo įrašus yra siūlomas metodas pagrįstas 3-D CNN, kuris mokosi tiek iš erdvinės, tiek laiko informacijos, gautų iš nuoseklios kadrų sekos. Pritaikytas 3-D konvoliucinis tinklas, kuris iš trumpos nuoseklios kadrų sekos išskiria erdvės ir laiko ypatybes. 3-D konvoliucija atliekama padauginant 3-D branduolį su kubu, kuris yra sukrautas iš eilės kadrų. Taigi šis modelis gali užfiksuoti erdvės ir laiko informaciją apie veidus iš kadrų sekos.

Agarwal et al. (2020) bando analizuoti tiek garso, tiek vaizdinę medžiagą, norint aptikti klastotes. Jų metode išnaudojamas faktas, kad nors lūpų sinchronizavimo klastotės yra pakankamai įtikinamos, burnos formos dinamika kartais vis tiek nesuderinama su šnekamąja fonema. Pavyzdžiui, sakant žodį, prasidedantį M ar B – mama, brolis – pastebima, kad lūpos visiškai užsidaro. Jei žmogus nėra pilvakalbis,

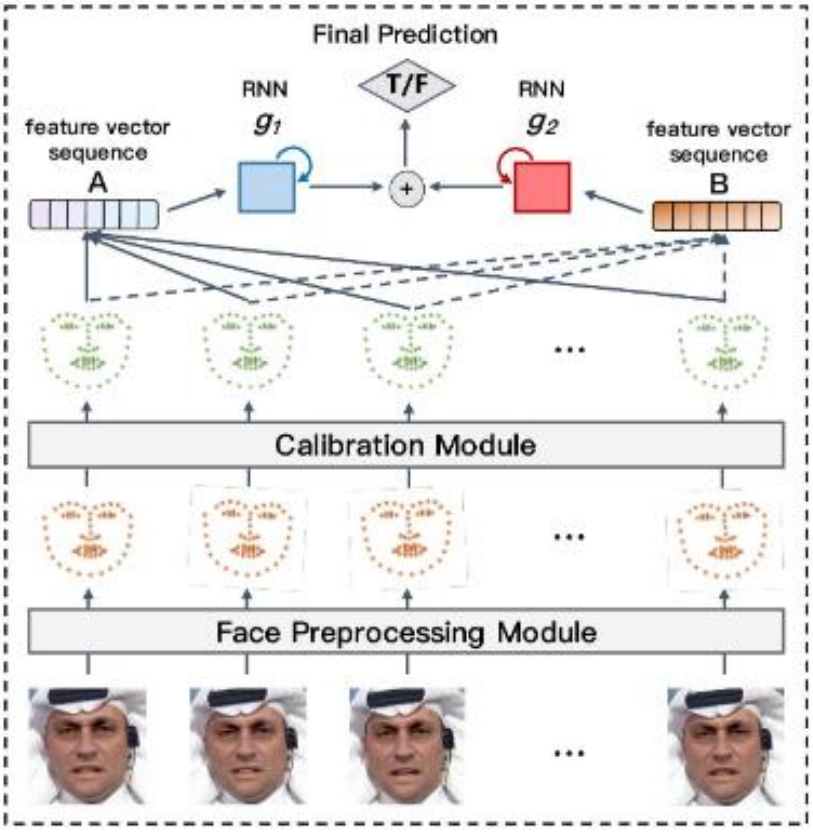
bus sunku tinkamai ištarti žodį „mama“, neuždarant lūpų. Pastebima, kad šio tipo fonemos klastotėse dažnai yra pažeidžiamos, net jei iš pirmo žvilgsnio tai nėra akivaizdu. Šiuos garso ir teksto lūpų sinchronizavimo neatitikimus galima panaudoti norint aptikti ar vaizdas yra suklastotas ar ne. Lūpos analizuojamos naudojant tris generavimo metodus: garso į vaizdo įrašą (A2V), teksto į vaizdo įrašą, kuriuose generuojami tik trumpi posakiai (T2V-S) ir teksto į vaizdo įrašą, kuriame generuojamos ilgesnės ištaros frazės (T2V-L). Taikant A2V generavimo techniką, kaip įvestis imamas kalbančio asmens vaizdo įrašas ir naujas garso įrašas, ir generuojamas naujas vaizdo įrašas, kuriame asmens burna sinchronizuojama su nauju garsu. T2V generavimo metodai kaip įvestį ima vaizdo įrašą, kuriame matomas kalbantis asmuo ir norimas pasakyti tekstas, ir generuoja naują vaizdo įrašą, kuriame asmens burna sinchronizuojama su naujais žodžiais. Analizuojant iš burnos vidurio ištraukiamas vertikalus intensyvumo profilis (6 paveikslas).



6 pav. vaizdo ir garso sinchronizavimo neatitikimas „Deepfake“ vaizduose (Agarwal et al., 2020)

Tikimasi, kad šis intensyvumo profilis kokybiškai skirsis, kai burna yra atvira arba uždaryta. Pavyzdžiui, burnos vertikalus intensyvumo profilis keisis iš odos tono į šviesų (dantys), į tamsų (apatinė burnos dalis), į šviesią (dantys), ir tada grįš į odos atspalvį. Kitu atveju, bus aptikta, kad burna yra tiktai atvira, bet intensyvumo profilis rodys priešingus duomenis.

Sun et al. (2021) siūlo „Deepfake“ vaizdo įrašų aptikimą naudojant modeliavimą laike pagal tikslias geometrines ypatybes. Naudojamas kalibravimo modulis sukurtas siekiant pagerinti geometrinių ypatybių tikslumą ir labiau juos išskiriant, o dviejų srautų rekurentinis neuronų tinklas (RNN) sukurtas taip, kad būtų išnaudotos požymių laike pasikeitimo ypatybės. Jis atskleidžia manipuliuojamus veidus, aptikdamas nenormalius veido judesius ir nesutapimus laike. „Deepfake“ aptikimo sistemą „LRNet“ architektūra pavaizduota 7 paveiksle.



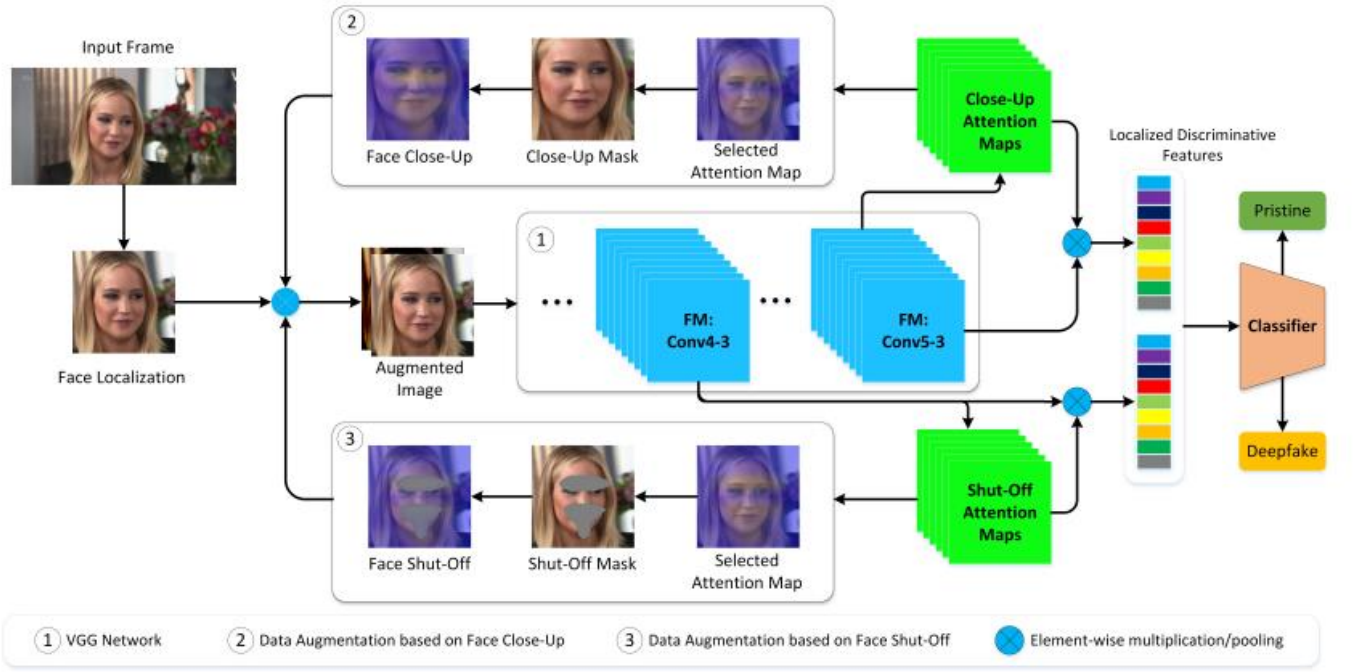
7 pav. „LRNet“ modelio architektūra (Sun et al., 2021)

„LRNet“ yra suskirstytas į keturis komponentus, t.y. veido apdorojimo modulį, kalibravimo modulį, veido ypatybių įdėjimo modulį ir RNN klasifikavimo modelį. Veido apdorojimo modulis atlieka geometrinių ypatybių ištraukimą iš veido nuotraukų, o jį sudaro veidų aptikimas, geometrinių ypatybių aptikimas ir jų sulygiavimas. Po to veido požymiai kalibruojami, kad būtų sumažintas triukšmas ir pagerintas tikslumas. Tada įterpiami veido ypatybių duomenys yra naudojami dviejų srautų RNN modelyje. Šių abiejų srautų rezultato vidurkis nusako galutinį rezultatą, t.y. ar vaizdo įrašas yra suklastotas ar ne. Vaizdo įrašas suskirstomas į fiksuoto ilgio trumpus įrašus, o iš tų trumpų vaizdo įrašų rezultatų vidurkio gaunamas vienas bendras rezultatas.

1.2. „Deepfake“ aptikimo metodai analizuojant veido sritį

Khormali et. al. (2021) pristato dėmesiu į detales pagrįstą skaitmeninę vaizdo autentifikavimo sistemą ADD. Bendras pateikto dėmesiu pagrįsto „Deepfake“ aptikimo metodas parodytas 8 paveiksle. ADD sudaro trys pagrindiniai komponentai: lokalizuotas diskriminacinių savybių ištraukimas, dėmesiu į detales pagrįstas duomenų papildymas ir klasifikatorius, skirtas atskirti originalius kadrus nuo „Deepfake“. „Deepfake“ generavimo algoritmai daugiausia manipuliuoja veido sritimis ir palieka

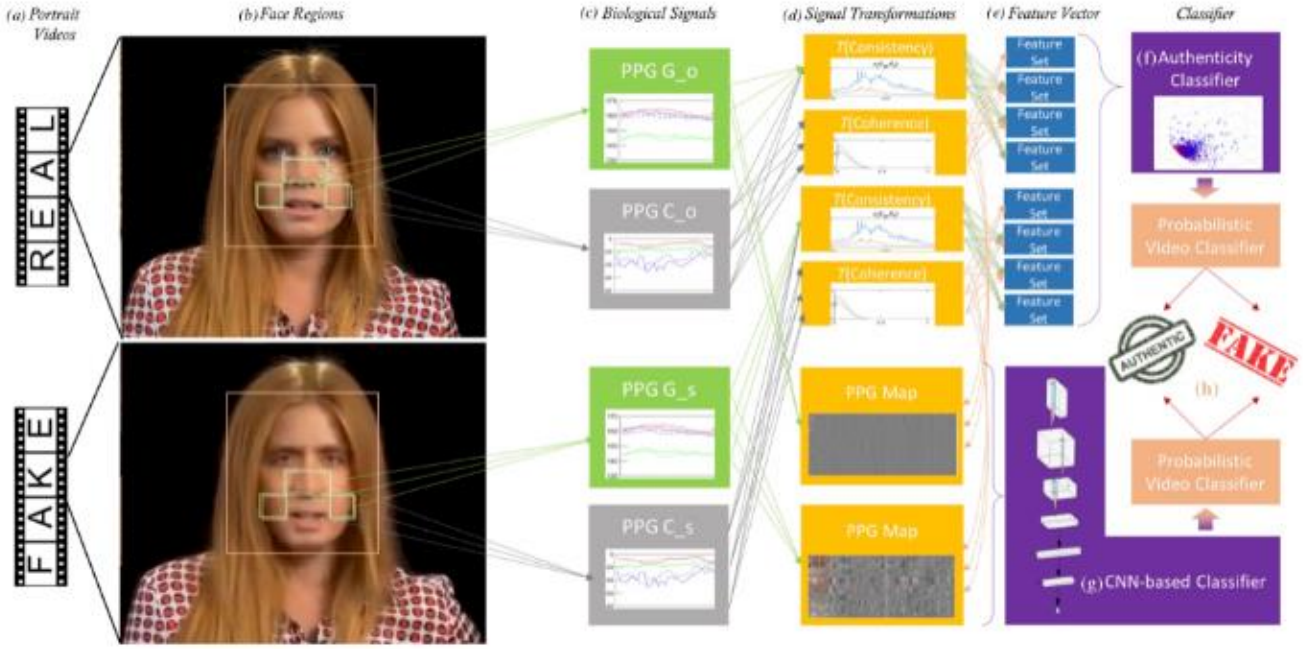
nepažeistą fono dalį. Todėl fokusavimas į veido sritį, užuot analizuojant visą kadrą, ne tik pagerina aptikimo našumą, nes sumažina foninį triukšmą, bet ir sumažina skaičiavimo laiką, sumažinant įvesties dydį. Naudojantis veido atpažinimo technologijomis surenkamos veido sritys. Gauti veido orientyrai naudojami apkarpyti, sulygiuoti ir pakeisti veidus į standartinę konfigūraciją. Šie apkarpyti kadrai, kuriuose yra tik veido sritys, toliau naudojami dėmesiu į detales pagrįstam vaizdų pagerinimui ir veido savybių ištraukimui.



8 pav. ADD „Deepfake“ aptikimo sistema (Khormali et al., 2021)

Šiame darbe naudojami du dėmesiu į detales pagrįsti duomenų gerinimo metodai, veido iš arti analizavimas ir veido ypatybių išėmimas. Pirmoji atidžiau žiūri į konkrečią veido sritį, o antroji nepaiso atitinkamos veido srities ir mato kitas veido sritis. Pagrindinis veido iš arti metodo tikslas yra atidžiau pažvelgti į konkrečias veido sritis, pvz., akis, kaktą, lūpas ir t.t., ir suteikti modeliui ryškesnių vietinių bruožų, kad būtų galima geriau interpretuoti vietinį vaizdą. Šis metodas sutelkia dėmesį į klastojamą sritį ir smulkesnes suklastojimo ypatybes. Nors veido iš arti metodas leidžia atidžiau pažvelgti į konkrečias veido sritis, tai ne visada pasiteisina. Kai kuriais atvejais modelis neišmoks naujų savybių. Siekiant išvengti šios problemos ir išskirti daugiau savybių iš kitų vaizdo regionų, naudojamas veido ypatybių išėmimo metodas. Šiuo metodu išimamos kai kurios veido dalys (akys ir burna) ir tai savo ruožtu priverčia modelį matyti kitas vaizdo dalis ir pamatyti kitas klastotės savybes.

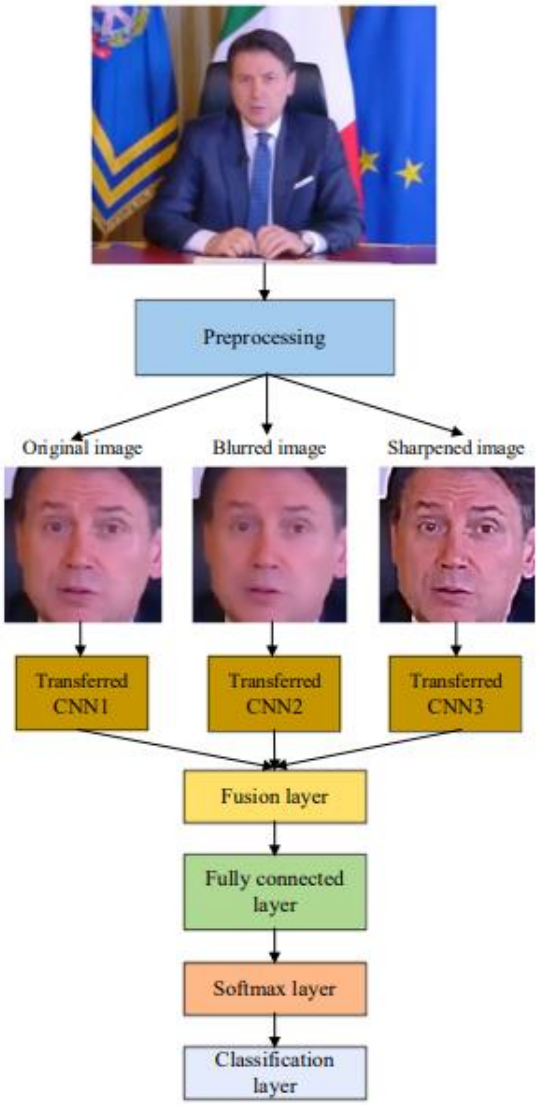
Ciftci et al. (2020) pastebi, kad nors GAN mokosi ir generuoja fotorealistinius vaizdus, kurių žmogaus akis negali atskirti, gamtos paslėpti biologiniai signalai (9 paveikslas) vis tiek nėra lengvai atkartojami. Biologiniai signalai yra vaizdo įrašą papildančios detalės, kurios taip pat įeina į „Deepfake“ klastotes. Be to, vaizdo įrašai, priešingai nei vaizdai, turi dar vieną sudėtingumo sluoksnį, jie turi sukurti tinkamas klastotes visame vaizdo įraše ir būti įtikinami visą laiką. Biologiniai signalai ir laiko nuoseklumas yra raktas aptinkant autentišką turinį.



9 pav. biologinių ženklu aptikimas „Deepfake“ vaizduose (Ciftci et al., 2020)

Nors naujausi gilaus mokymosi metodai, skirti aptikti netikrą turinį yra pakankamai geri, juos riboja specifinis generacinis modelis, duomenų rinkinys, žmonės arba rankomis sukurtos funkcijos. Priešingai nei kitose sistemose, šiame darbe pasirenkama ieškoti kai kurių natūralių, biologinių ženklų, užuot darant tam tikras suklastoto turinio prielaidas. Sistemos procesas vaizduojamas 9 paveiksle. Sistema apdoroja įvesties vaizdo įrašus rinkdama fiksuoto ilgio vaizdo įrašų segmentus su veido dalimis, apibrėždama dominančius regionus kiekviename veide ir išgaunant biologinius signalus iš tų regionų, tuose laiko segmentuose. Sprendimas gaunamas nagrinėjant išgautus biologinius signalus, jų transformacijas į skirtingas sritis (laikas, dažnis, laikas-dažnis) ir jų sąsajas. Antroje dalyje apjungiama informacija, kad būtų sukurtas apibendrintas autentiškumo klasifikatorius, veikiantis didelės dimensijos ypatybių erdvėje. Trečioje dalyje signalai paverčiami fiksuotos trukmės segmentų signalų žemėlapiais, skirtais išmokyti paprastą konvoliucinį tinklą. Tada duomenys apibendrinami ir pateikiamas rezultatas ar vaizdas tikras ar ne.

Yavuzkilic et al. (2021) pristato naują didelio masto duomenų rinkinį WPDD, skirtą gerinti „Deepfake“ aptikimo algoritmus. Duomenų rinkinio sukūrimui buvo gauti 20 skirtingų politikų vaizdo įrašai iš interneto. Atsisiųsti vaizdo įrašai buvo padalinti į mažas dalis ir gauta daugiau nei 320 000 kadrų. Šie kadrai buvo sugrupuoti ir jiems buvo pritaikytos septynios skirtingos manipuliacijų klasės. Taip pat, pasiūlyta nauja trijų srautų architektūra, kuri pavaizduota 10 paveiksle.

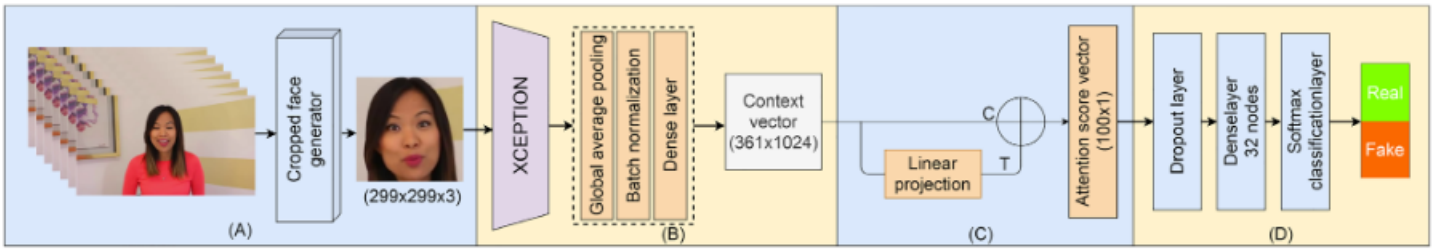


10 pav. trijų srautų „Deepfake“ aptikimo būdas (Yavuzkilic et al., 2021)

Kaip matyti 10 paveiksle, siūloma architektūra susideda iš trijų srautų, kurie yra sujungiami į vieną. Po sujungimo sluoksnio eina pilnai sujungtas sluoksnis ir „Softmax“ klasifikacijos sluoksnis skirtas identifikuoti ar nuotrauka suklastota ar ne. Pirmajame sraute naudojamas pradinis vaizdas, kad sistema išmoktų bendrų dalykų apie tokias savybes kaip galvos forma ir plaukų spalva. Antrasis srautas naudojamas odos spalvos analizavimui, tam naudojamas išblukintas vaizdas. Paskutinis srautas susijęs su

veido bruožais, tam naudojamas išryškintas įvesties vaizdas. Iš nuotraukų aptinkamos veido sritys, jos iškerpamos, o tada pakeičiamos iki 224 × 224 failo dydžio. Mokymuisi naudojami iš anksto apmokytų CNN modelių svoriai naujai klasifikacijos problemai mokyti. Pirmajam srautui naudojamas iš anksto apmokytas VGG16 modelis. Antrame sraute naudojamas iš anksto apmokytas VGG19 modelis, o trečiame - iš anksto apmokytas ResNet18 modelis. Kiekvienas iš srautų mokosi su atitinkamomis nuotraukomis ir galiausiai šie srautai yra sujungiami į vieną bendrą, iš kurio vėliau gaunamas klasifikavimo rezultatas.

Ganguly et. al. (2022) pastebi kad dirbtinai sugeneruotose nuotraukose yra matomos anomalijos. Todėl giliuoju mokymu paremtas bruožų ištraukimo modelis, kuris nustato likusių artefaktų buvimą ar nebuvimą, gali būti tinkama technika aptikti suklastotus vaizdus. Daugelis gilaus mokymosi metodų, siekiantys aptikti svarbius veido bruožus, naudoja įvairius dėmesio į detales pagrįstus mechanizmus, bet jie atlieka daug skaičiavimų. Atsižvelgdami į šiuos faktus, šiame darbe autoriai siūlo ne tokį skaičiavimams intensyvų sprendimą, kuris yra paremtas „Xception“ modeliu. Siūlomas modelis pavaizduotas 11 paveiksle.



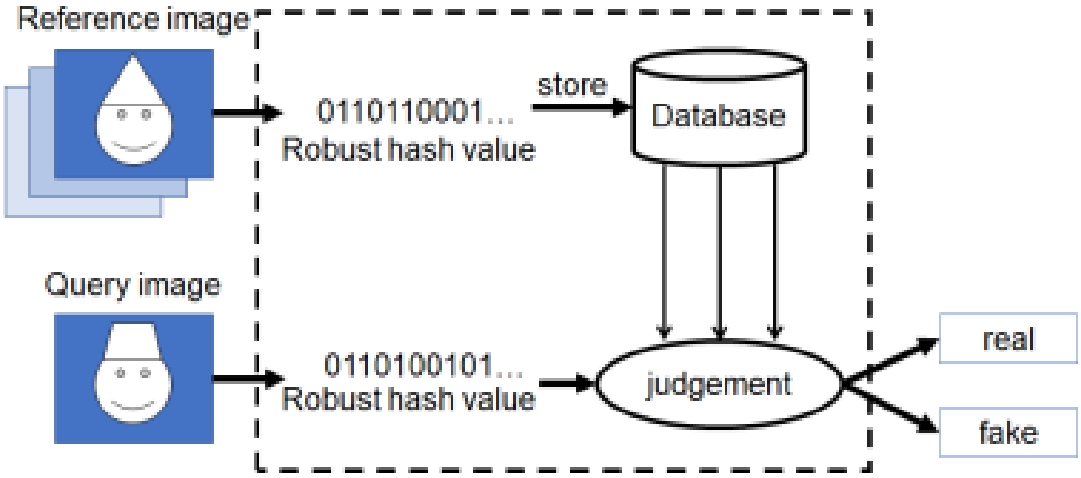
11 pav. Dėmesiu pagrįsta klastotės aptikimo architektūra (Ganguly et al., 2022)

Iš vaizdo įrašų kadrų išskiriama veido sritis naudojantis MTCNN algoritmu, o tada ji perduodama per iš anksto paruoštą „Xception“ modelį, kuris buvo iš anksto apmokytas ant „ImageNet“ duomenų rinkinio, kad būtų gaunami požymių žemėlapiai. Toliau, naudojant vizualinio dėmesio mechanizmą, sutelkiamas dėmesys į vaizdo manipuliavimo likusius artefaktus, kurie gali padėti nustatyti vaizdų ar vaizdo įrašų autentiškumą.

1.3. Kiti „Deepfake“ aptikimo metodai

Tanaka et al. (2021) tiria galimybę sukurti netikrų vaizdų aptikimo metodą su šifravimo maišos funkcijomis. Suklastotų vaizdų maišos reikšmių skaičiavimas gali būti pakankamai patikimas, net ir kai vaizdai yra suspausti ar jų dydis pakeistas. Todėl sutelkiamas dėmesys į atsparios maišos funkcijos metodo naudojimą, kuris siekia atrasti nuotraukas, panašias į užklausos nuotraukas. Generuojamos maišos

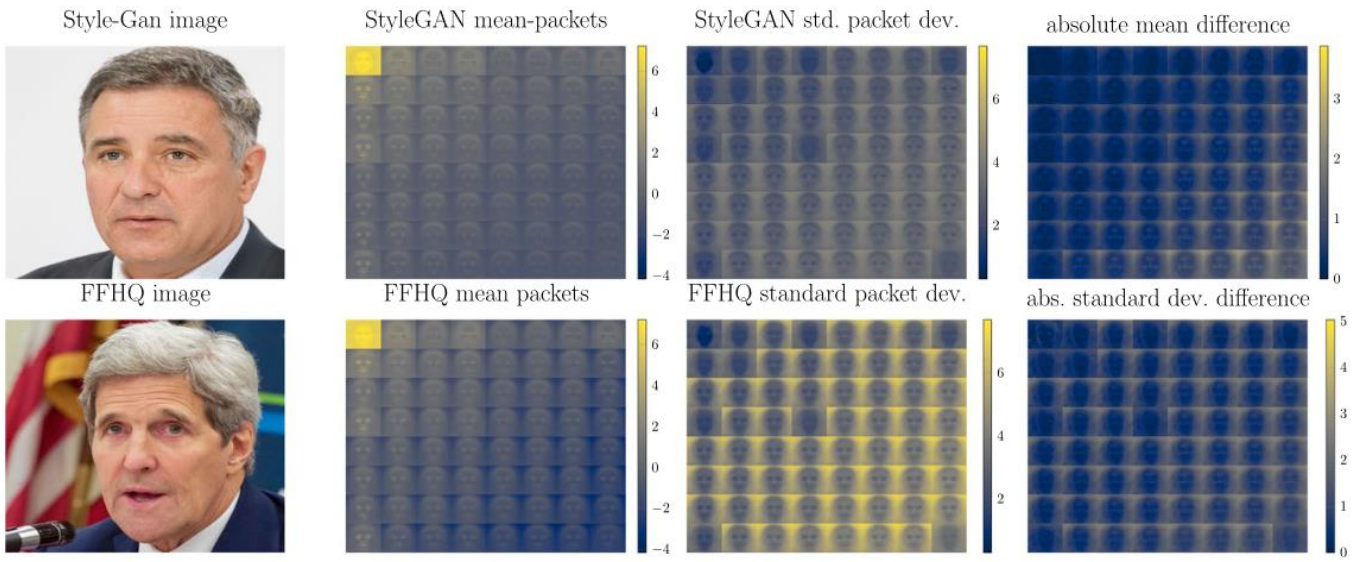
reikšmės turi būti jautrios pasikeitimams, kurie yra atliekami klastojant vaizdus. 12 paveiksle parodyta siūlomo metodo apžvalga.



12 pav. suklastotų vaizdų aptikimas naudojant maišos reikšmes (Tanaka et al., 2021)

Sistemoje maišos reikšmės yra apskaičiuojamos rinkiniui vaizdų (iš anksto surinktų duomenų) ir yra saugomos duomenų bazėje. Kai norima patikrinti ar vaizdas buvo suklastotas, vaizdas yra paduodamas į sistemą, gaunama maišos reikšmė ir ji yra lyginama su esamomis duomenų bazėje. Jei palyginimui nėra originalaus maišos kodo, siūlomas netikrų vaizdų aptikimo metodas negalimas. Priešingai, jei palyginimui yra originalus maišos kodas, jis gali patikimiau aptikti netikrus vaizdus nei įprasti metodai. Be to, siūlomas metodas leidžia naudoti bet kokius vaizdus, kuriems negali būti taikomi kiti aptikimo metodai. Užklausto vaizdo maišos reikšmė lyginama su duomenų bazėje saugomomis vertėmis, o užklausos vaizdas yra tikras arba netikras, atsižvelgiant į atstumą tarp maišos reikšmių.

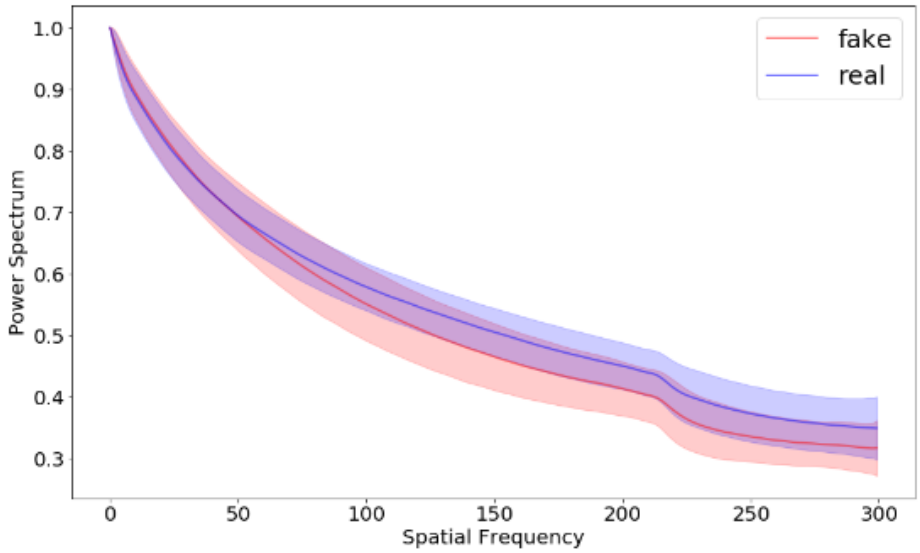
Wolter et. al (2022) lygina nuotraukų bangų dažnių koeficientus. Taip nustatomi reikšmingi skirtumai tarp tikrų ir netikrų vaizdų. Palyginti su kitais darbais analizuojančius dažnius, pirmą kartą nagrinėjamos GAN erdvines ir dažninės savybes. Randami skirtumai tarp tikrų ir netikrų vaizdų tiek pagal dažnių vidurkį, tiek pagal standartinį nuokrypį. Dažnių vidurkio ir standartinio nuokrypio skirtumai taip pat yra matomi tarp skirtingų GAN modelių sukurtų vaizdų (13 paveikslas).



13 pav. „Deepfake“ aptikimas pagal dažnių bangų koeficientus (Wolter et al., 2022)

Tikrų ir netikrų vaizdų bangų dažnių koeficientų palyginimas leidžia pamatyti skirtumus tarp šių vaizdų. Taip pat, tokio klasifikatoriaus našumas yra konkurencingas arba geresnis už kitus modelius, netgi esant mažam tinklo dydžiui. 13 paveiksle viduriniai stulpeliai rodo bangų dažnių paketų vidurkius ir standartinį nuokrypį. Dešiniajame stulpelyje parodytas absoliutus paketų dažnių vidurkių ir standartinio nuokrypio skirtumus. Pastebima, kad GAN sugeneruotų vaizdų dažnių vidurkių paketai dažnai šviesesni ir tai leidžia sukurti mechanizmą sugeneruotų vaizdų atpažinimui.

Durall et. al (2020) požiūris remiasi klasikine vaizdų dažnių analize, atskleidžiančia skirtingus elgesius aukštuose dažniuose. 14 paveiksle parodyta, kaip tam tikras dažnių komponentų diapazonas elgiasi, kai vaizdai yra dirbtinai generuojami.



14 pav. „Deepfake“ aptikimas pagal dažnių srities analizę (Durall et al., 2020)

Palyginti su kitais modeliais, kuriuose reikia tiekti didelius kiekius pažymėtų duomenų, šis metodas parodo labai gerus rezultatus naudojant tik kelis pažymėtus mokymo pavyzdžius ir netgi pasiekia gerą tikslumą visiškai neprižiūrimuose scenarijuose. Vaizdų charakteristikas galima analizuoti pagal Furjė transformaciją. Tokia transformacija nurodo, kaip signalo energija paskirsto dažnių diapazone. Pritaikius Furjė transformaciją, informacija pateikiama išvestyje vis tiek yra 2D informacija. Todėl apskaičiuojamas azimutalinis vidurkis, kuris duomenis paverčia į 1D FFT galios spektro vaizdą. Turint šiuos duomenis, naudojantis paprastu klasifikatoriumi galima lengvai atskirti tikrus ir netikrus vaizdus.

1.4. „Deepfake“ duomenų rinkiniai

Tolosana et al. (2022) įvardina tris pagrindinius „Deepfake“ duomenų rinkinius: „FaceForensics++“, „Celeb-DF“ ir „DFDC“.

- **„FaceForensics++“**. 2019 m. buvo pristatyta „FaceForensics++“ (Rössler et al., 2019). duomenų bazė kaip pradinio „FaceForensics“ pratęsimas (Rössler et al., 2018), kuris buvo orientuotas tik į ekspresijos keitimo manipuliacijas. „FaceForensics++“ turi 1000 tikrų vaizdo įrašų, ištrauktų iš „Youtube“. Netikri vaizdo įrašai buvo generuojami naudojant tiek kompiuterinę grafiką, tiek GAN metodus (1000 netikrų vaizdo įrašų kiekvienam metodui). Duomenų rinkinyje yra 3 skirtingų lygių priklausomai nuo vaizdo įrašo suspaudimo lygio. “Raw” – aukščiausios raiškos vaizdo įrašai, “c23” – vidutiniškai suspausti vaizdo įrašai, “c40” – prasčiausios kokybės vaizdo įrašai.

- **Celeb-DF v2**. „Celeb-DF v2“ duomenų bazė (Li et al., 2020) siekia pateikti geresnės vaizdo kokybės suklastotus vaizdo įrašus, panašius į populiarius vaizdo įrašus, kurie yra internete. Šią duomenų bazę sudaro 890 tikrų vaizdo įrašų, ištrauktų iš „Youtube“, atitinkančius 59 įžymybių interviu, įvairaus pasiskirstymo pagal lytį, amžių ir etninę grupę. Be to, šiuose vaizdo įrašuose yra daug įvairių aspektų kaip veido dydžiai (pikseliais), orientacijos, apšvietimo sąlygos ir foninis vaizdas. Iš viso buvo sukurti 5639 vaizdo įrašai naudojant patobulintą viešai prieinamo „DeepFakes“ generavimo algoritmo versija. Patobulinti aspektai, pvz., maža susintetintų veidų skiriamoji geba ir spalvų neatitikimai.
- **DFDC preview**. DFDC duomenų bazė (Dolhansky ir kt., 2019) kurią „Facebook“ išleido bendradarbiaudama su kitomis įmonėmis ir akademinėmis institucijomis, kaip „Microsoft“, „Amazon“ ir MIT. Svarbu pastebėti, kad nebuvo naudojami jokie viešai prieinami duomenys ar duomenys iš socialinės žiniasklaidos šiam duomenų rinkiniui sukurti, skirtingai nuo kitų populiarių duomenų bazių. Iš viso buvo sukurta 4119 netikrų vaizdo įrašų naudojant du skirtingus nežinomus metodus. Buvo sukurti suklastoti vaizdo įrašai sukeičiant panašios išvaizdos subjektus, t. y. su panašiu veidu, odos tonu, plaukais, akiniais ir kt. Svarbu pabrėžti, kad DFDC duomenų bazėje atsižvelgiama į skirtingus scenarijus (t. y. patalpose ir lauke), apšvietimo sąlygas (t. y. diena, naktis ir pan.), atstumai nuo asmens iki fotoaparato ir pozų variantai, ir t.t.

1.5. „Deepfake“ aptikimo metodų apžvalga

Visi metodai siūlo skirtingus sprendimus ir pasiekia skirtingus efektyvumo lygius. Šie metodai yra palyginami 1 lentelėje, kad būtų galima suprasti kokie metodai jau išbandyti ir kokius metodus dar būtų galima patobulinti.

1 lentelė. „Deepfake“ aptikimo metodų palyginimas

Autorius	Grupė	Aptikimo metodas	Efektyvumas
Lee et al. (2021)	Analizuojamas vaizdo pasikeitimas laike	Veido pasikeitimas laike	97,39 %
Gu et al. (2021)	Analizuojamas vaizdo pasikeitimas laike	Trimatis veido lyginimas laike	99,78 %
Agarwal et al.	Analizuojamas vaizdo	Vaizdo ir garso lyginimas laike	93,90 %

Autorius	**Grupė**	**Aptikimo metodas**	**Efektyvumas**
(2020)	pasikeitimas laike		
Sun et al. (2021)	Analizuojamas vaizdo pasikeitimas laike	Veido požymių pasikeitimas laike	99,90 %
Khormali et al. (2021)	Analizuojama veido sritis vaizduose	Dėmesių į detales pagrįstas vaizdo autentifikavimas	98,37 %
Ciftci et al. (2020)	Analizuojama veido sritis vaizduose	Analizuojami biologiniai veido signalai	94,65 %
Yavuzkilic et al. (2021)	Analizuojama veido sritis vaizduose	Modelis apmokomas iš 3 skirtingų vaizdų tipų	99,80 %
Ganguly et al. (2022)	Analizuojama veido sritis vaizduose	Dėmesiu pagrįstas aptikimas	70,10 %
Tanaka et al. (2021)	Kita	Vaizdai lyginami pagal maišos reikšmes	88,01 %
Durall et al. (2020)	Kita	Dažnių srities analizė	90,00 %
Wolter et al. (2022)	Kita	Nagrinėjamos GAN erdvines ir dažninės savybės	98,02 %

Be abejo, kadangi yra analizuojami „Deepfake“ vaizdai, tai visi metodai telkia pagrindinį dėmesį į veido sritį, kurioje yra atliekami visi klastojimo veiksmai. Pagrindinė šių metodų prielaida yra tai, kad „Deepfake“ vaizdai nėra visiškai tikslūs, nepilnai atvaizduoja pagrindinius veido požymius kaip: akys, burna ir nosis. O ypatingai tai matosi vaizdo įrašuose, nes veido išraiškos kečiasi laike, todėl atsiranda daugiau klaidų. Taip pat, kai kurie metodai stebi ar vaizduose yra įvairiausi artefaktai, kurie dažniausiai yra sukuriami suspaudžiant vaizdo failus ar kitaip programiškai apdorojant juos.

Šiuos metodus sudėtinga vertinti pagal jų efektyvumą, nes jie yra patikrinti su skirtingais duomenų rinkiniais, todėl sunku teigti, kuris metodas yra geriausias, bet galima tikrai išskirti kai kuriuos, labai gerus metodus, turinčius didesnį nei 95 % tikslumą. Visi metodai yra vienaip ar kitaip patikimi, todėl jų

patobulinimai ar kombinacija galia duoti dar geresnius rezultatus. Bet Sun et al. (2021) metodas išsiskiria didžiausiu AUC rezultatu, todėl jis pasirinktas tolimesniam nagrinėjimui.

2. „Deepfake“ atpažinimo modelio tyrimas ir tobulinimas

Pasirinktas Sun et al. (2021) metodas, kuriame naudojamas RNN tinklas, skirtas aptikti geometrinių veido ypatybių skirtumus laike. Kiti „Deepfake“ vaizdo įrašų aptikimo metodai daugiausia orientuojasi į veido išvaizdos ypatybes, bet sudėtingas vaizdų manipuliavimas šias aptikimo sistemas gali apeiti bei šie aptikimo metodai yra labai paveikūs vaizdo triukšmui. Autorių sukurta sistema „LRNet“ atpažįsta „Deepfake“ vaizdo įrašus naudodama naują kalibravimo modulį, kuris pagerina geometrinių ypatybių nustatymą bei naudoja dviejų srautų rekurentinius neuroninius tinklus (RNN), sukurtus atskirti šias ypatybes laike. Palyginti su kitais modeliais, siūlomas modelis taip pat yra lengviau ir greičiau treniruojamas.

Sun et al. (2021) pastebi, kad suklastoti vaizdo įrašai yra labai tikslūs ir raiškūs atskiruose kadruose, bet jie vis tiek turi nenatūralias išraiškas ar veido judesius. Tai yra būdingas „Deepfake“ trūkumas, nes suklastoti vaizdo įrašai generuojami kadras po kadro ir neatsižvelgia į veido išraiškų elgesį, bei neatsižvelgia į vaizdų klastojimo tęstinumą laike. Autoriai, norėdami geriau užfiksuoti šias klaidas laike, pasirinko duomenims naudoti tik geometrines veidų ypatybes, o ne visą nuotrauką, nes šios ypatybės yra tinkamesnės modeliuojant veido išraiškų kitimą. Šios geometrines veido ypatybės yra taškų rinkinys, apibrėžiantis pagrindinių veido dalių kontūrus.

Šiame darbe naudojamas dviejų srautų rekurentinis neuroninis tinklas, kuris gali atrasti šias klaidas laike. Be to, įvertinta, kad netiksliai gautos veido ypatybes ir orientyrai gali pakenkti norint gauti geresnius atpažinimo rezultatus. Todėl šiame darbe sukurtas kalibravimo modulis, skirtas sumažinti geometrinių ypatybių kitimą laike. Šio modelio pavadinimas – „LRNet“.

Veido vaizdų pakeitimas į geometrines ypatybes yra veiksmingas duomenų dimensijos sumažinimas ir palyginti su kitais gilaus mokymo modeliais, tai turi mažiau įtakos, kai vaizdo įrašai yra prastos kokybes. Modelis sudaro keturi komponentai: veido apdorojimo modulis, kalibravimo modulis, veido ypatybių įdėjimo (angl. *face embedding*) modulis ir RNN klasifikavimo modelis. Skirtingai nuo daugelio siūlomų metodų, juos reikia mokyti nuo pradžios iki galo, o „LRNet“ mokymą reikia atlikti tik RNN modelio dalyje.

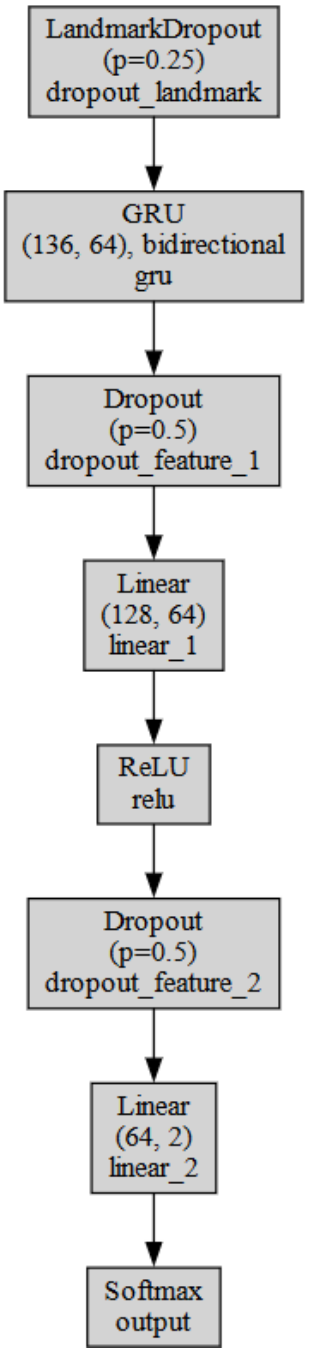
Veido apdorojimo modulis ištraukia geometrines ypatybes iš veido nuotraukų. Jį sudaro veidų aptikimas, veido geometrinių ypatybių aptikimas ir jų sulygiavimas. Veidų aptikimas nuotraukose

atliekamas kiekviename vaizdo įrašo kadre ir iš nuotraukos iškerpama tik reikalinga veido dalis. Iškirpus veidus iš nuotraukų, surandami 68 veido požymiai, kurie apibūdina veidą ir galiausiai požymiai yra sulygiuojami su iš anksto nustatyta padėtimi.

Išgauti veidų požymiai dar nėra visai tinkami naudoti modeliui. Aptikti požymiai turi akivaizdų virpėjimą (angl. *jittering*) laike, net jei veidas beveik nejuda. Todėl šiai problemai išspręsti naudojamas veido požymių kalibravimo modulis. Prognozuojamos požymių pozicijos naudojant nuosekliai einančius kadrus, remiantis Lucas-Kanade optinio srauto skaičiavimo algoritmu. Tinkamos prognozės sujungiamos su atitinkamais Kalmano filtro rezultatais, kad būtų sumažintas triukšmas ir kalibruoti požymiai būtų dar tikslesni.

Sutvarkyti veido ypatybių duomenys yra įterpiami į dviejų tipų vektorių seką ir tada naudojami dviejų srautų RNN modelyje. *g1* srautas yra naudojamas aptikti veido formos kitimą, o *g2* naudojamas rasti veido ypatybių skirtumus. Po RNN išvesties yra pridedamas visiškai sujungtas sluoksnis, kuriuo naudojantis bus gautas atsakymas ar vaizdo įrašas suklastotas ar ne. Šių abiejų srautų rezultato vidurkis nusako galutinį rezultatą. Suklastotiems vaizdo įrašams aptikti, kiekvienas vaizdo įrašo pavyzdys yra suskirstytas į fiksuoto ilgio trumpus įrašus, o tada iš tų trumpų vaizdo įrašų rezultato gaunamas vienas bendras rezultatas, kuris nusako, ar vaizdo įrašas suklastotas ar ne.

Dviejų srautų tinklas susideda iš GRU (angl. *Gated Recurrent Unit*) sluoksnio su 64 išvestimis ir dviem 64 neuronų visiškai sujungtų sluoksnių ir galiausiai 2 išvestys nusakančios ar vaizdo įrašas suklastotas ar ne. Neuroninio tinklo modelis pavaizduotas 15 paveiksle.



15 pav. „LRNet“ modelio struktūra (Sun et al., 2021)

Tarp įvesties ir GRU sluoksnio įterptas „Dropout“ sluoksnis su išmetimo tikimybe *dr1* = 0,25, kuris išmeta dalį veido požymių, o kitur papildomai įterpiami ir dar 2 „Dropout“ sluoksniai, kurių išmetimo tikimybė *dr2* = 0,5, kurie yra naudojami tarp likusių sluoksnių. Be to, duomenų rinkinys padalijimas į dvi dalis: 80 % mokymui ir 20 % testavimui. Kiekvienas vaizdo įrašas yra segmentuojamas į klipus, kurių fiksuota trukmė yra 60 kadrų, tai sumuojasi į 2 sekundžių trukmės vaizdo įrašus, kai kadrų

per sekundę greitis yra 30. Taip pat, naudojamas „Adam“ optimizatorius su mokymosi greičiu *lr* = 0,001, o „batch size“ – 1024. Šis klasifikavimo modelis bus apmokytas iki 500 ciklų (angl. *epoch*).

2.1. Mašininio modelio parametrų tyrimas

Sun et. al (2021) savo modeliui kurti naudojo „FaceForensics++“ duomenų rinkinį ir testavo jį skirtinguose duomenų rinkinio lygiuose (2 lentelė):

2 lentelė. „LRNet“ modelio rezultatai

Duomenų rinkinio lygis	Tikslumas	AUC	g1 tikslumas	g2 tikslumas
“raw”	99,7	99,9	83,4	98,3
“c23”	-	97,3	-	-
“c40”	91,2	95,7	80,4	85,2

Atliktame tyrime naudojamas tikslumas ir AUC metrika. AUC yra metrika, apibūdinanti modelio tikslumą, bet ši metrika atsižvelgia į skirtingų klasių duomenų disbalansą (Shiksha, 2023). Su „raw“ suspaudimo lygiu buvo gautas tikslumas 99,7 %, o AUC – 99,9 %, su „c23“ tikslumas nenurodytas, o AUC – 97,3 %, o su „c40” tikslumas buvo 91,2 %, AUC – 95,7 %.

Kad būtų galima geriau palyginti tyrimo rezultatus, šio modelio treniravimas buvo atkartotas naudojantis kita kompiuterine įranga, kurios parametrai yra:

- Vaizdo ploktštė – NVIDIA „GeForce“ RTX 3070 „Laptop“ GPU 8GB GDDR6
- Procesorius – AMD „Ryzen“ 9 5900HX
- Operatyvioji atmintis – 24GB 3200MHz
- Kietasis diskas – SSD CT1000P2SSD8 1TB

Atkartojus šio modelio mokymą, buvo gauti rezultatai, kurie pateikti 3 lentelėje -

3 lentelė. Atkartoti „LRNet“ modelio rezultatai

Duomenų rinkinio lygis	g1 tikslumas	g2 tikslumas	Bendras tikslumas	AUC	Precision	Recall	F1	Laikas sekundėmis

Duomenų rinkinio lygis	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**	**Laikas sekundėmis**
„raw”	86,3806	91,7724	95,7143	97,8705	96,9965	98,0357	97,5133	898,1953
„c23“	63,5757	87,1847	92,9333	96,8073	96,7629	93,3333	95,0172	887,5831
„c40“	81,5299	82,444	91,2857	93,7111	93,5315	95,5357	94,523	1095,184

Modelio mokymas buvo atliktas 3 kartus su kiekvienu duomenų rinkinio lygiu ir paimti geriausi rezultatai. Buvo naudojami tie patys mokymo ir testavimo duomenys kaip ir straipsnyje. Skirtingai nuo straipsnyje pateiktų duomenų, papildomai paskaičiuota „Precision“ – naudojama įvertinti modelio tikslumą, kai prognozuojama teigiama klasė (Iguazio, 2023), „Recall“ – naudojama įvertinti modelio gebėjimą atpažinti teigiamus pavyzdžius (c3.ai, 2023), „F1“ – yra viena iš klasifikavimo metrikų, kuri apjungia „Precision“ ir „Recall“ metrikas į vieną metriką, siekiant gauti išsamų modelio veiklos vertinimą (v7labs, 2023), ir laikas sekundėmis.

Su „raw“ tikslumas buvo gautas 95,71%, o AUC – 97,87%, „g1“ – 86,38 % ir „g2“ – 91,77 % kas atitinkamai skiriasi 3,99 %, 2,03 %, 2,98 % ir 6,53 % nuo straipsnyje pateiktų rezultatų. „Precision“ – 96,99 %, „Recall“ – 98,03 %, „F1“ – 97,51 % ir mokymo laikas 898 sekundės.

„c23“ duomenų rinkinio lygiu, tikslumas buvo gautas 92,93 %, o AUC – 96,81 %, „g1“ – 63,57 % ir „g2“ – 87,18 %. Palyginti su straipsnyje pateiktais duomenimis tik AUC skiriasi 0,49 %, kiti parametrai nebuvo pateikti. „Precision“ – 96,7629 %, „Recall“ – 93,3333 %, „F1“ – 95,0172 % ir mokymo laikas 887 sekundės.

Su „c40“ tikslumas buvo gautas 91,28 %, o AUC – 93,71 %, „g1“ – 81,52 % ir „g2“ – 82,44 %, tai atitinkamai skiriasi 0,09 %, 1,99 %, 1,13 % ir 2,78 % nuo straipsnyje pateiktų rezultatų. Bendras „Precision“ – 93,53 %, „Recall“ – 95,53 %, „F1“ – 94,52 % ir mokymo laikas 1095 sekundės.

Atkartoto modelio rezultatai yra vidutiniškai keliais procentais prastesni. Tai gali būti dėl įvairių priežasčių, pirmiausia dėl to, kad skiriasi kompiuterinės sistemos, kuriose buvo atlikti mokymai, antra – paprasčiausiai straipsnio autoriams pavyko gauti geresnį rezultatą kuriant modelį. Šio tyrimo metu bus atsižvelgta tiek į straipsnyje pateiktus rezultatus, tiek gautus rezultatus atkartojus mokymą.

Tyrimo metu atliktas eksperimentas, kuriame buvo keičiami šio modelio parametrai:

- „Block size“ – vaizdo įrašo ilgis kadrais. Galimi variantai: 30, 60, 120. Pradiniame modelyje buvo naudojama 60 kadrų vaizdo įrašai.
- „Dropout rate“ – išmetimo procentas. Galimi variantai: 0,3, 0,4, 0,5, 0,6, 0,7. Pradiniame modelyje išmetimo tikimybė buvo 0,5.
- Mokymosi greitis. Galimi variantai: 0,001, 0,005, 0,01. Pradiniame modelyje naudojama 0,001.
- Optimizatorius. Galimi variantai: „Adam“, „RMSprop“, „SGD“. Pradiniame modelyje naudojama „Adam“.

Eksperimentas buvo atliktas su „c23“ duomenų rinkinio lygių, su visomis galimomis šių parametrų kombinacijomis, kurių iš viso gavosi 135.

Gauti rezultatai padalinti pagal optimizatorių, paskaičiuotas šių rezultatų vidurkis, standartinis nuokrypis ir variacijos koeficientas (4 lentelė).

4 lentelė. Rezultatų palyginimas pagal optimizatorių

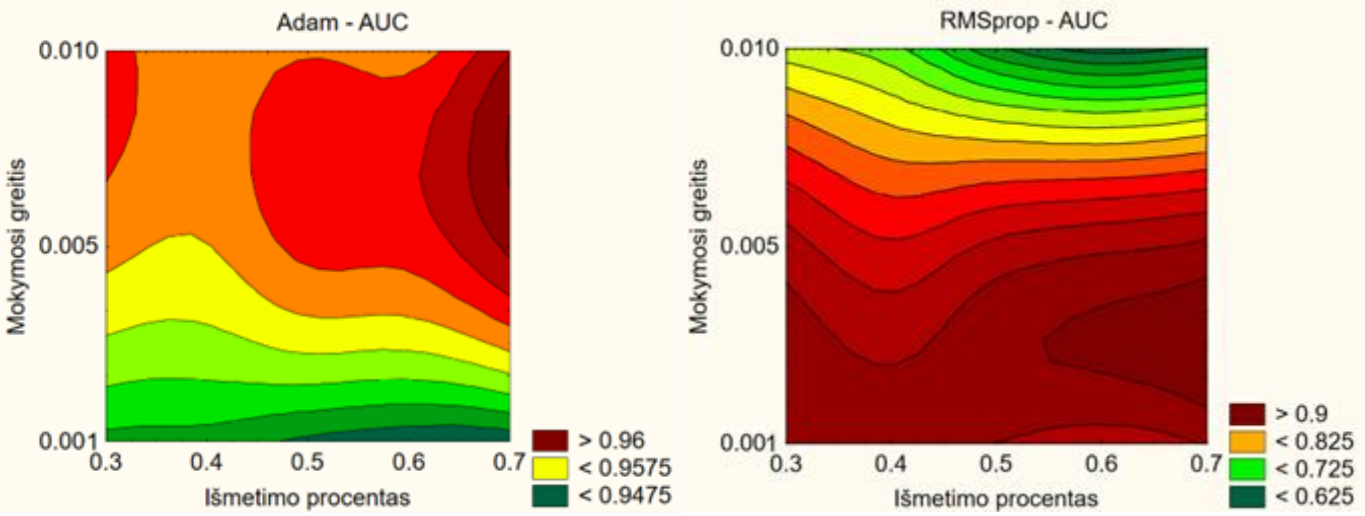
Optimizatorius - parametras	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
„SGD“ - vidurkis	0,3358	0,3849	0,2696	0,5009	0,1629	0,1003	0,1051
„SGD“ - standartinis nuokrypis	0,1799	0,2487	0,1531	0,0054	0,3310	0,2615	0,2580
„SGD“ - variacijos koeficientas	53,5699	64,6357	56,7945	1,0917	203,0915	260,5321	245,4264
„Adam“ - vidurkis	0,5778	0,8688	0,9143	0,9572	0,9698	0,9065	0,9368
„Adam“ - standartinis nuokrypis	0,1088	0,0214	0,0172	0,0137	0,0084	0,0269	0,0140
„Adam“ - variacijos koeficientas	18,8349	2,4706	1,8836	1,4382	0,8703	2,9699	1,5017
„RMSprop“ - vidurkis	0,6084	0,7499	0,7972	0,8427	0,8358	0,8025	0,8069

Optimizatorius - parametras	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
„RMSprop“ standartinis nuokrypis	0,0898	0,2504	0,2328	0,1734	0,2684	0,3124	0,3010
„RMSprop“ variacijos koeficientas	14,7599	33,3940	29,2101	20,5766	32,1170	38,9303	37,3123

„SGD“ optimizatorius pasižymi mažiausiu tikslumu, AUC, „Precision“, „Recall“ ir F1 reikšmėmis, palyginti su kitais dviem optimizatoriais. Be to, variacijos koeficientai yra gana dideli, rodydami didesnį rezultatų pasiskirstymą. „Adam“ optimizatorius pasižymi geriausiais rezultatais visuose matavimo rodikliuose: didžiausiu tikslumu, AUC, „Precision“, „Recall“ ir F1. Be to, šio optimizatoriaus variacijos koeficientai yra mažesni, rodydami mažesnį rezultatų pasiskirstymą, kas yra pageidautina. „RMSprop“ optimizatorius rodo geresnius rezultatus nei „SGD“, tačiau prastesnius nei „Adam“. Jo tikslumo, AUC, „Precision“, „Recall“ ir F1 rodikliai yra geri, bet variacijos koeficientai yra didesni nei „Adam“ optimizatoriaus, bet mažesni nei „SGD“.

Remiantis šia analize, „Adam“ optimizatorius yra geriausias pasirinkimas tolimesnei analizei, nes jis pasižymi didžiausiu tikslumu, AUC, „Precision“, „Recall“ ir F1, o jo variacijos koeficientai yra mažesni, rodydami mažesnį rezultatų pasiskirstymą. „RMSprop“ antroje vietoje, o „SGD“ optimizatorius šioje situacijoje yra netinkamas dėl prastesnių rezultatų ir didesnio rezultatų pasiskirstymo. Todėl toliau bus analizuojami rezultatai gauti naudojant „Adam“ ir „RMSprop“ optimizatorius.

Pagal gautus rezultatus bus lyginama, kaip kelių modelio parametrų kombinacijos turi įtakos AUC, tikslumui ir F1, pagal skirtingą optimizatorių. 16, 17 ir 18 paveiksluose lyginama pagal išmetimo procentą ir mokymosi greitį.

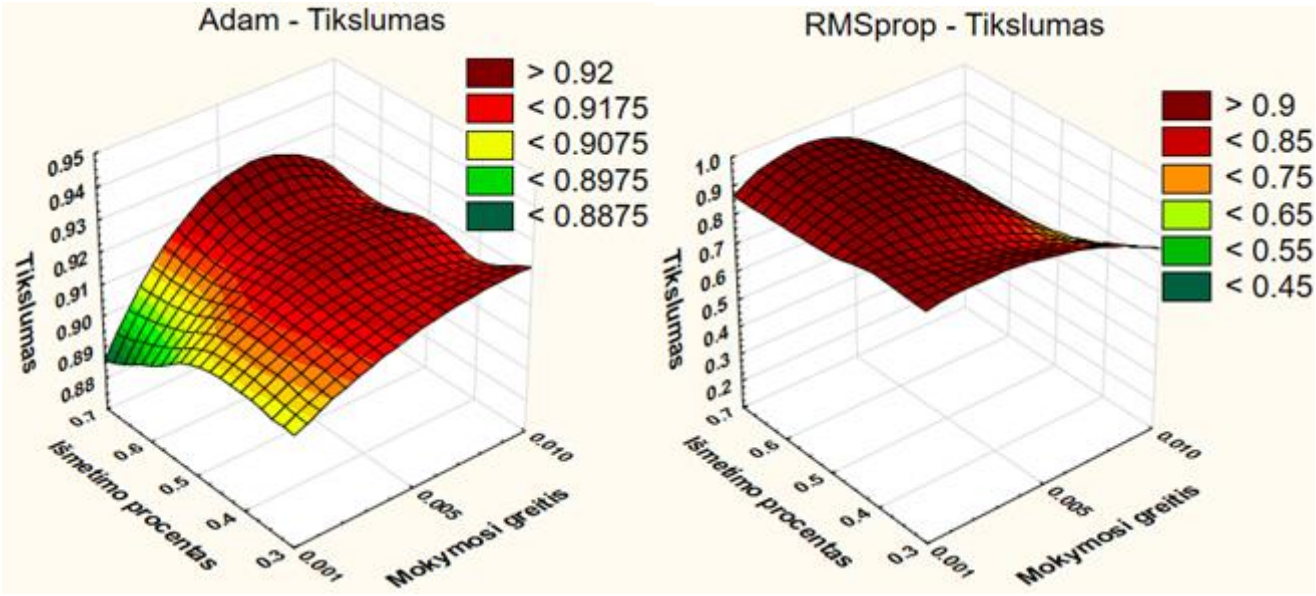


a) b)

16 pav. Mokymosi greičio ir išmetimo procento įtaka AUC. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Su „Adam“ matyti, kad AUC reikšmės yra palyginti didelės ir yra tarp 92,79 % ir 97,45 % ir iš duomenų galima matyti, kad AUC reikšmės šiek tiek kinta priklausomai nuo mokymosi greičio ir išmetimo koeficientų. Taip pat, galima daryti išvadą, kad modelis, kuris naudoja didesnį mokymosi tempą ir didesnį išmetimo koeficientą (pvz., mokymosi tempas 0,005 ir išmetimo koeficientas 0,7), gali pasiekti geresnę AUC reikšmę. O su „RMSprop“ matyti, kad mažesni išmetimo procento (0,3–0,4) ir mažesni mokymosi greičio (0,001) deriniai tendencijos rodo didesnes AUC reikšmes. Tuo tarpu didesni išmetimo procento (0,6–0,7) ir didesni mokymosi greičio (0,01) deriniai tendencijos rodo mažesnes AUC reikšmes, kas gali reikšti, kad šis derinys gali sukelti per didelį svyravimą mokymo metu ir trukdyti modeliui tinkamai mokytis. Vertinant šiuos duomenis, galima daryti išvadą, kad tinkamas mokymosi greičio ir išmetimo procento derinys turi svarbią įtaką AUC kitimui. Tinkamai parinkus šiuos parametrus, galima pagerinti modelio veikimą ir padidinti AUC reikšmes.

Su „Adam“ optimizatoriumi didžiausias AUC – 97,45 %, su mokymosi greičiu 0,005 ir išmetimo procentu 0,7, o prasčiausias AUC – 92,79 %, su mokymosi greičiu 0,01 ir išmetimo procentu 0,6. Su „RMSprop“ optimizatorius didžiausias AUC – 96,82 %, su mokymosi greičiu 0,005 ir išmetimo procentu 0,4, o prasčiausias AUC – 50 %, su mokymosi greičiu 0,01 ir išmetimo procentu 0,4. „Adam“ optimizatorius vidutiniškai rodo geresnius rezultatus.

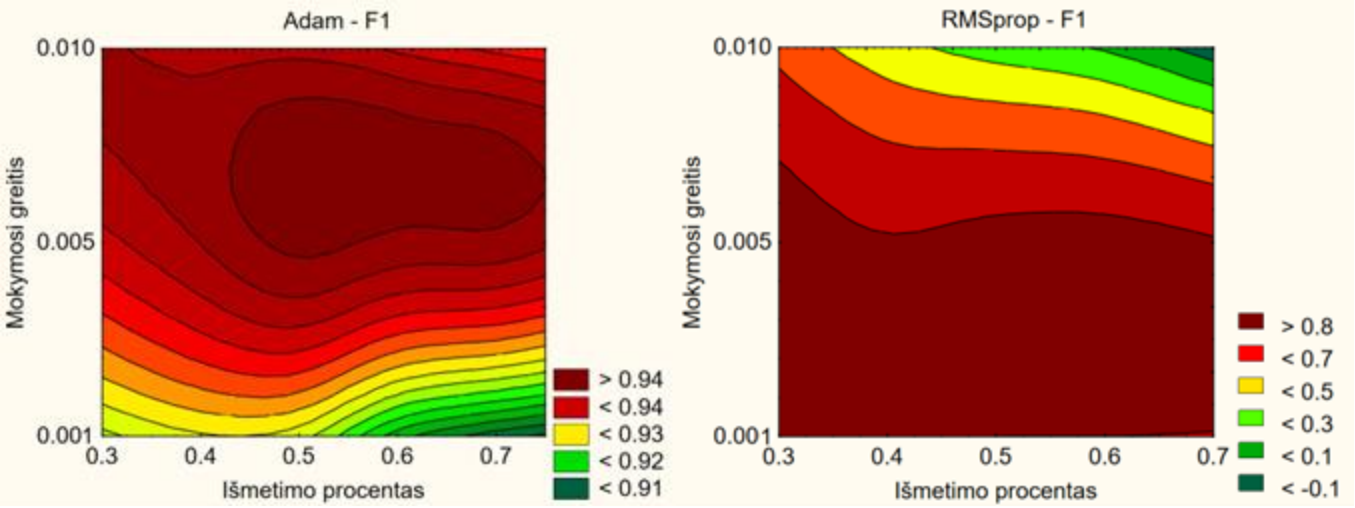


a) b)

17 pav. Mokymosi greičio ir išmetimo procento įtaka tikslumui. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Iš duomenų matyti, kad naudojant „Adam“, kai mokymosi greitis yra 0,005 ir išmetimo procentas 0,7 modelis pasiekia aukščiausią tikslumo rezultatą (94,6 %), o mažiausias tikslumas (87,33 %) su mokymosi greičiu 0,001 ir išmetimo procentu 0,3. Su „RMSprop“ didžiausias tikslumas (93,8 %) gautas naudojant mokymosi greitį 0,005 ir išmetimo procentą 0,6, o mažiausias tikslumas (21 %) su mokymosi greičiu 0,01 ir išmetimo procentu 0,7. Bendrai visi „Adam“ optimizatoriaus rezultatai duoda gerus rezultatus, o „RMSprop“ rezultatai labai skirtingi.

Iš „Adam“ duomenų matyti, kad su mažesniu išmetimo procentu vidutiniškai geresni tikslumo rezultatai, o su didesnėmis išmetimo procento reikšmėmis (0,7) ir mažesniu mokymosi greičiu (0,001) gali sumažinti tikslumą. Tiek mokymosi greitis, tiek išmetimo procentas turi įtakos modelio tikslumo rodikliams. O su „RMSprop“ analizuojant pateiktus duomenis, taip pat galime pastebėti, kad skirtingi mokymosi greičiai ir išmetimo procentas turi įvairią įtaką modelio tikslumui. Pavyzdžiui, tinklas su 0,01 mokymosi greičiu turi vidutiniškai prastesnius tikslumo rezultatus (netgi iki 21 %), o su mažesnėmis išmetimo procento reikšmėmis geresnis tikslumas.

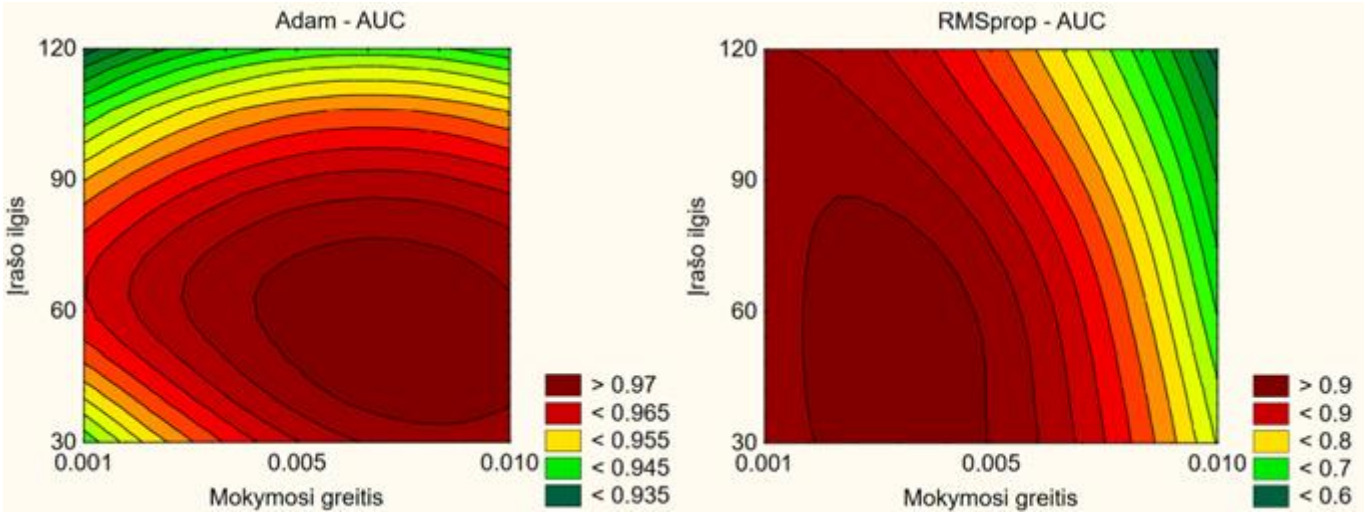


a) b)

18 pav. Mokymosi greičio ir išmetimo procento įtaka F1. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Tyrimo duomenys rodo, kad skirtingos mokymosi greičio ir išmetimo procento reikšmės turi mažą poveikį F1 vertei. Pavyzdžiui su „Adam“ optimizatoriumi naudojant 0,7 išmetimo procentą ir 0,005 mokymosi greitį F1 vertė yra didžiausia – 96,45 %. Tuo tarpu esant išmetimo procentui 0.6 ir mokymosi greičiui 0,01, F1 vertė sumažėja iki 90,74 %. Bendrai „Adam“ optimizatoriaus F1 reikšmės yra pakankamai didelės, bet duomenys parodo, kad sunku pasakyti ar mokymosi greitis ir išmetimo procentas turi tam tikros įtakos F1 vertei. Tačiau galima pastebėti, kad esant mažesnėms mokymosi greičio reikšmėms (0,001), F1 rezultatas yra mažesnis. Su „RMSprop“ rezultatai yra labai skirtingi. Naudojant 0,3 išmetimo procentą ir 0,005 mokymosi greitį F1 vertė yra didžiausia - 96,14 %. Tuo tarpu esant išmetimo procentui 0,7 ir mokymosi greičiui 0,01, F1 vertė sumažėja iki 0 %. Bendrai „RMSprop“ optimizatoriaus F1 reikšmės yra labai pasklidusios, bet duomenys parodo, kad mokymosi greitis ir išmetimo procentas turi tam tikros įtakos F1 vertei. Esant mažesniam išmetimo procentui ir mažesniam mokymosi greičiui rezultatai geresni. Lyginant „Adam“ ir „RMSprop“, tiek viename, tiek kitame optimizatoriuje didesnės mokymosi greičio reikšmės lėmė blogesnius F1 rezultatus.

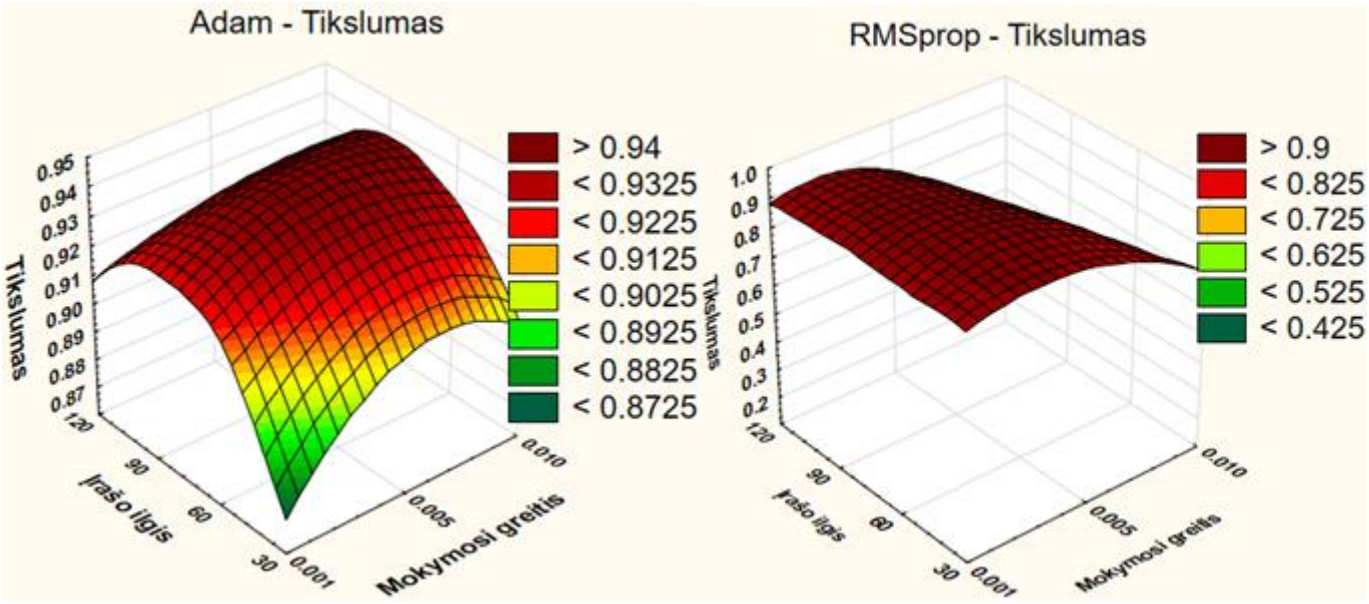
19, 20 ir 21 paveiksluose lyginami AUC, tikslumo ir F1 rezultatai pagal įrašo ilgį ir mokymosi greitį.



a) b)

19 pav. Mokymosi greičio ir įrašo ilgio įtaka AUC. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Iš „Adam“ duomenų galima matyti, kad AUC reikšmės kinta priklausomai nuo mokymosi greičio ir įrašo ilgio. Taip pat, galima daryti išvadą, kad modelis, kuris naudoja mažesnį įrašo ilgį ir didesnį mokymosi tempą (pvz., įrašo ilgis 30 – 60 ir mokymosi tempą 0,005 – 0,01), gali pasiekti geresnę AUC reikšmę. Pavyzdžiui, kai bloko dydis yra 30, mokymosi greitis 0,01, AUC yra 96,81%, o kai bloko dydis yra 120, AUC sumažėja iki 92,79 %, beveik 4 %. O su „RMSprop“ matome, kad mažesni įrašo ilgiai (30) ir mažesni mokymosi greičio (0,001) deriniai rodo didesnes AUC reikšmes. Tuo tarpu didesni įrašo ilgiai (120) ir didesni mokymosi greičio (0,01) deriniai rodo mažesnes AUC reikšmes. Pavyzdžiui, kai bloko dydis yra 30, mokymosi greitis 0,001, AUC yra 96,69 %, o kai bloko dydis yra 120, AUC sumažėja iki 93,67 %. Vertinant šiuos duomenis, galima daryti išvadą, kad tinkamas mokymosi greičio ir įrašo ilgio derinys turi svarbią įtaką AUC kitimui.

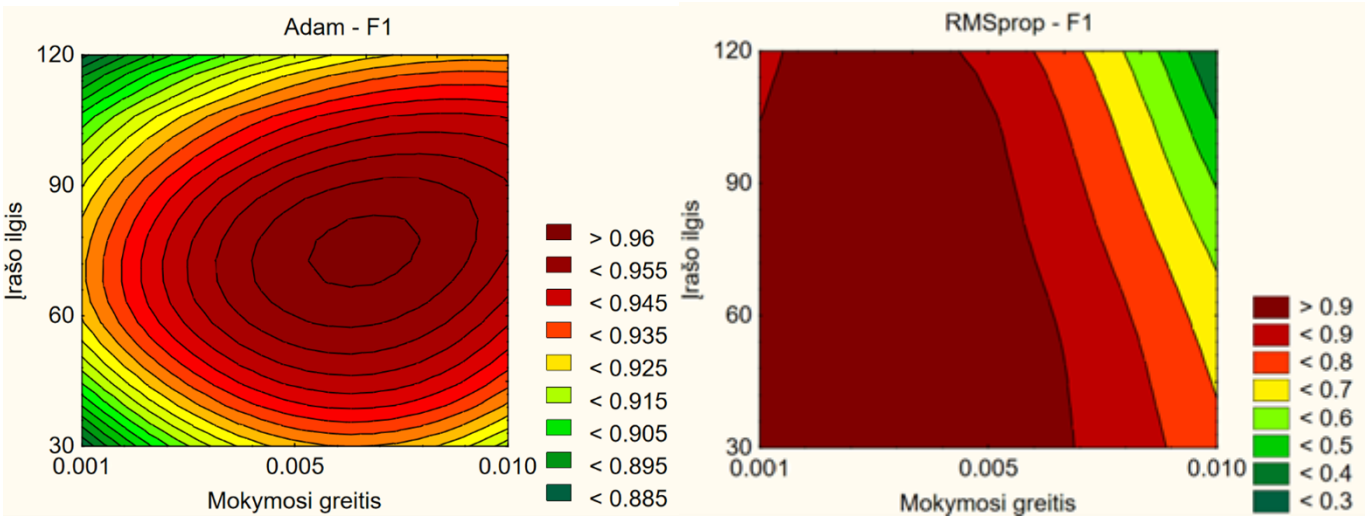


a) b)

20 pav. Mokymosi greičio ir įrašo ilgio įtaka tikslumui. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Iš duomenų matome, kad naudojant „RMSprop“ didžiausias tikslumas (93,8 %) gautas naudojant mokymosi greitį 0,005 ir įrašo ilgį 60, o mažiausias tikslumas (21 %) su mokymosi greičiu 0,01 ir visais įrašo ilgiais (30, 60 ir 120). O su „Adam“, kai mokymosi greitis yra 0,005 ir įrašo ilgis 60 modelis pasiekia aukščiausią tikslumo rezultatą (94,6 %), o mažiausias tikslumas (87,33 %) su mokymosi greičiu 0,001 ir įrašo ilgiu 30.

„RMSprop“ optimizatoriaus rezultatams įrašo ilgis neturi didelės įtakos. Bendrai visi įrašo ilgio variantai vienodai pasiskirstę, bet su 120 įrašo ilgiu, truputį prastesni rezultatai. Pavyzdžiui, tinklas su 0.001 mokymosi greičiu ir įrašo ilgiu 120 turi tikslumą 90,4 %, su 30 įrašo ilgiu 91 %, o galiausiai su 60 įrašo ilgiu 92 % tikslumas. Mokymosi greitis tikrai turi įtakos, naudojant mokymosi greitį 0,01 gaunami prastesni rezultatai. Su „Adam“ duomenimis matoma, kad mokymosi greitis turi mažesnę įtaką rezultatams. Visi rezultatai su mokymosi greičiu yra vienodai pasiskirstę. Bet įrašo ilgio rezultatai turi didesnę įtaką. Su įrašo ilgiu 60, gaunami geriausi rezultatai, o su 30 prasčiausi. 120 įrašo ilgio rezultatai yra maždaug per vidurį. Pavyzdžiui, tinklas su įrašo ilgiu 60, turi tikslumą 92,4 %, su 30 įrašo ilgiu 89,46 %, o su 120 įrašo ilgiu 91,86 %.

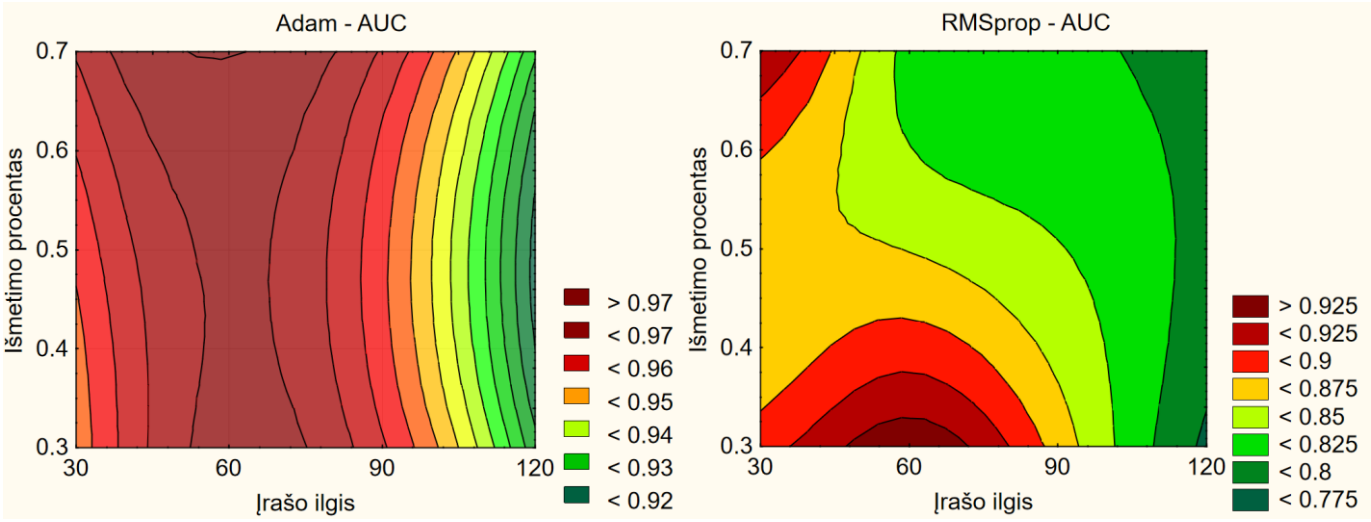


a) b)

21 pav. Mokymosi greičio ir įrašo ilgio įtaka F1. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Su „Adam“ galima pastebėti, kad įrašo ilgio ir mokymo greičio pasirinkimas turi įtakos F1 vertei. Didinant įrašo ilgį iš 30 iki 60, F1 vertė didėja, tačiau tolesnis didinimas iki 120 ne visada teigiamai veikia F1 vertę. Tai gali būti susiję su tuo, kad didesnis bloko dydis padeda modeliui geriau suprasti ilgesnius kontekstus, tačiau per didelis bloko dydis gali pakenkti mokymui dėl didesnio sudėtingumo ir padidintų atminties reikalavimų. Kalbant apie mokymo greitį, matoma, kad F1 vertės pasiskirstymas yra gana įvairus, tai rodo, kad mokymo greitis nelabai turi įtakos. O su „RMSprop“ galima matyti, kad mokymosi greitis turi daugiau įtakos negu įrašo ilgis. Esant mokymosi greičiui 0,01 gaunami prasčiausi rezultatai, bet tiek su 0,005, tiek 0,001 gaunami geri rezultatai. O geriausi rezultatai pagal įrašo ilgį yra su 60, bet taip pat ir su 30 bei 120, bet su 120 vidutiniškai prastesni rezultatai.

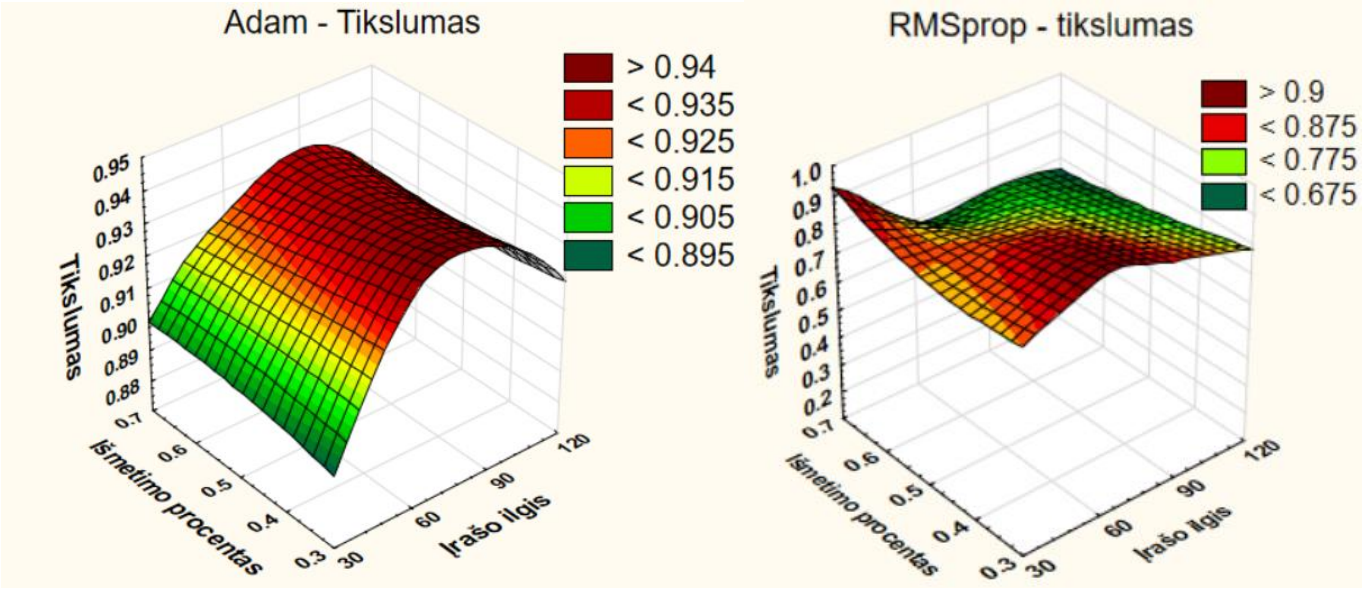
22, 23 ir 24 paveiksluose lyginami AUC, tikslumo ir F1 rezultatai pagal išmetimo procentą ir įrašo ilgį.



a) b)

22 pav. Išmetimo procento ir įrašo ilgio įtaka AUC. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

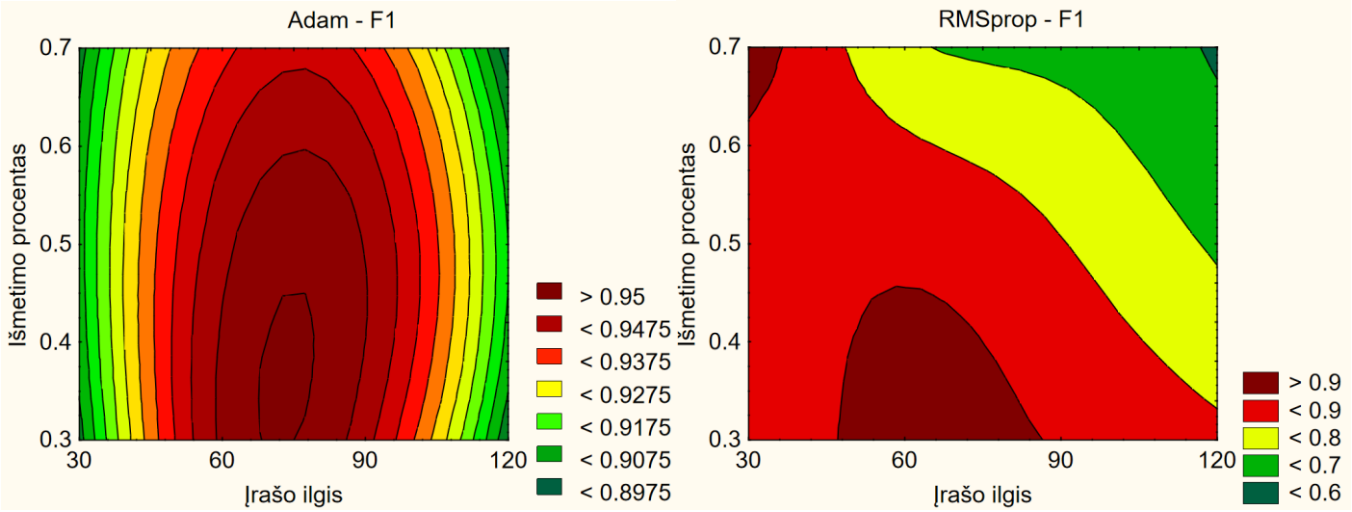
Nagrinėjant „Adam“ duomenis galima pastebėti, kad didesnės išmetimo procento reikšmės (0,6–0,7) su įrašo ilgiu 60 rezultatai dažniausiai pasiekia aukštesnes AUC reikšmes. Pavyzdžiui, su išmetimo procentu 0,7 ir įrašo ilgiu 60, AUC siekia 97,45 %, o su išmetimo procentu 0,6 – 97,29 %. Bet su didesniu įrašu ilgiu gaunamos prastesnės AUC reikšmės. Pavyzdžiui, su išmetimo procentu 0,7 ir įrašo ilgiu 120, AUC siekia 95,51 %, o su išmetimo procentu 0.6 – 93,91 %. O kai išmetimo procentas yra mažesnis (0,3–0,4) gaunami prastesni rezultatai. Pavyzdžiui esant įrašo ilgiui 30 ir išmetimo procentui 0,7 – 97,11 %, o su išmetimo procentu 0,3 – 95,93 %. O su „RMSprop“ matoma, esant įrašo ilgiui 120 gaunamos mažesnės AUC reikšmes, o su kitomis įrašo ilgio kombinacijomis priklauso nuo išmetimo procento. Pavyzdžiui, su išmetimo procentu 0,3 ir įrašo ilgiu 60, AUC siekia 96,38 %, o su išmetimo procentu 0,7 ir įrašo ilgiu 30 – 95,39 %. Didžiausios reikšmės yra su įrašo ilgiu 60 ir mažu išmetimo procentu arba su mažesniu 30 įrašo ilgiu ir didesnius išmetimo procentu (0,6–0,7).



a) b)

23 pav. Išmetimo procento ir įrašo ilgio įtaka tikslumui. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Nagrinėjant „Adam“ duomenis galima pastebėti, kad išmetimo procento reikšmės turėjo mažiau įtakos negu įrašo ilgis. Su įrašo ilgiu 60 rezultatai dažniausiai pasiekia aukštesnes tikslumo reikšmes. Pavyzdžiui, su išmetimo procentu 0,7 ir įrašo ilgiu 60, tikslumas siekia 94,6 %, o su išmetimo procentu 0,3 – 93,06 %, skirtumas yra, bet nedidelis. O su kitokiu įrašu ilgiu gaunamos prastesnės tikslumo reikšmės. Pavyzdžiui, su išmetimo procentu 0,7 ir įrašo ilgiu 120, tikslumas siekia 92,2 %, o su išmetimo procentu 0,3 ir įrašo ilgiu 30 – 89,13 %. Su „RMSprop“ matoma, kad išmetimo procentas taip pat turi mažesnę įtaką negu įrašo ilgis. Bet su mažesnėmis įrašo ilgio (30–60) reikšmėmis, gaunamas geresnis tikslumas. Pavyzdžiui, su išmetimo procentu 0,3 ir įrašo ilgiu 60, tikslumas siekia 93,73 %, o su išmetimo procentu 0,3 ir įrašo ilgiu 120 – 86,46 %, bet esant išmetimo procentui 0,7 – 90,6 %.

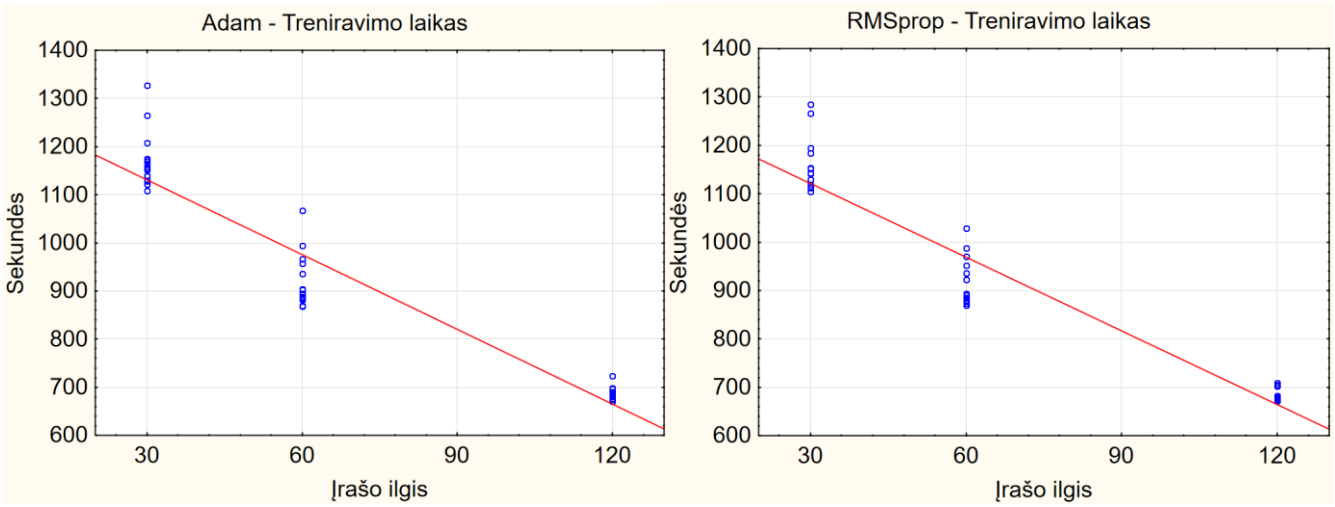


a) b)

24 pav. Išmetimo procento ir įrašo ilgio įtaka F1. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Tyrimo duomenys rodo, kad skirtingos įrašo ilgio ir išmetimo procento reikšmės turi poveikį F1 vertei. Su „Adam“ mažesnis išmetimo procentas (pvz., 0,3 ir 0,4) lemia didesnį F1 vertę palyginti su didesniu išmetimo procentu (pvz., 0,6 ir 0,7), o tiek 30 ir 120 įrašo ilgis lemia prastesnius rezultatus. Pvz., naudojant 0,7 išmetimo procentą ir 30 įrašo ilgį F1 vertė yra – 92,14 %, ir esant 120 įrašo ilgiu – 90,81 % Tuo tarpu esant išmetimo procentui 0,3 ir įrašo ilgiui 60 F1 vertė yra - 95,51 %. Bendrai „Adam“ duomenys parodo, kad išmetimo procentas ir įrašo ilgis turi įtakos F1 vertei. Su „RMSprop“ rezultatai yra truputi kitokie. Ryškiai prastesni rezultatai matomi tik su 120 įrašo ilgiu, o su kitais parametrais rezultatai panašus. Geriausias rezultatas gautas naudojant 0,3 išmetimo procentą ir 60 įrašo ilgį – 96,14 %, o su išmetimo procentu 0,7 ir įrašo ilgiu 30 gauta F1 vertė yra - 95,44 %. Tuo tarpu esant išmetimo procentui 0,7 ir įrašo ilgiui 120, F1 vertė sumažėja iki 90,08 %. Bendrai „RMSprop“ duomenys parodo, kad išmetimo procentas ir įrašo ilgis turi įtakos F1 vertei. Lyginant „Adam“ ir „RMSprop“, F1 rezultatai labai panašūs, bet „Adam“ optimizatoriaus reikšmės šiek tiek didesnės.

25 paveiksle lyginama modelio mokymo trukmė priklausomai nuo įrašo ilgio.



a) b)

25 pav. Modelio treniravimo laikas priklausomai nuo įrašo ilgio. a) „Adam“ optimizatorius; b) „RMSprop“ optimizatorius

Pirmiausia, didesnis bloko dydis – 120, vidutiniškai lemia mažesnius mokymo laikus, palyginti su mažesniais bloko dydžiais 30 ir 60. Tai galima aiškinti tuo, kad didesnis bloko dydis suteikia efektyvesnį duomenų panaudojimą, mažinant atnaujinimų skaičių per epochą. Siekiant geriau suprasti laiko ir bloko dydžio sąsają, verta apžvelgti kai kuriuos vidutinius laikus pagal bloko dydį:

- Bloko dydis 30: apie 1100 sekundžių
- Bloko dydis 60: apie 900 sekundžių
- Bloko dydis 120: apie 700 sekundžių

Iš šių rezultatų matome, kad didinant įrašo ilgį nuo 30 iki 120, vidutinis mokymo laikas sumažėja apie 37 %. Tačiau svarbu atkreipti dėmesį į tai, kad nors didesnis bloko dydis gali padėti sumažinti mokymo laiką, jis turi įtakos modelio tikslumui.

Apibendrinti tyrimo rezultatų duomenys yra pavaizduoti 5 lentelėje.

5 lentelė. Palyginimas straipsnio, atkartoto ir geriausių „Adam“ ir „RMSprop“ tyrimo rezultatais, su „c23“ duomenų lygiu

Rezultatas	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
Atkartotas straipsnio modelis	63,5756	87,1847	92,9333	96,8073	96,7629	93,3333	95,0171
Straipsnio duomenys	-	-	-	97,3	-	-	-
„Adam“ – geriausi tyrimo rezultatai	51,8731	88,6591	94,6	97,453	97,6644	95,2743	96,4545
„RMSprop“ – geriausi tyrimo rezultatai	57,3184	86,6276	92,8667	96,8271	94,7631	96,2025	95,4774

Pradiniai atkartoti straipsnio duomenys pasižymi gana aukštais rezultatais, ypač „g1“ tikslumas – 63,58 %, aukščiausias iš visų kitų modelių, o kiti rezultatai: „g2“ tikslumas - 87,18 %, bendras tikslumas - 92,93 %, AUC – 96,80 %, „Precision“ – 96,76 %, „Recall“ – 93,33 % ir F1 – 95,02 %. Straipsnyje su „c23“ duomenų lygiu buvo suteikia ribotą informaciją, tačiau AUC yra labai aukštas – 97,3 %. Tyrimo rezultatai su „Adam“ pasiekia aukščiausius rezultatus tarp visų modelių, beveik visuose kriterijuose, išskyrus „g1“ – 51,87 % ir „Recall“ – 95,27 %, toliau „g2“ tikslumas - 88,66 %, bendras tikslumas – 97,45 %, AUC - 97,45 %, „Precision“ - 97,66 %, „Recall“ - 95,27 % ir F1 - 96,45%. „RMSprop“ su geriausiais parametrais taip pat pasižymi aukštais rodikliais, bet bendrai turi prastesnius rezultatus negu „Adam“ tyrimo rezultatų: „g1“ tikslumas – 57,31 %, „g2“ tikslumas – 86,62 %, bendras tikslumas – 92,86 %, AUC – 96,82 %, „Precision“ – 94,76 %, „Recall“ – 96,20 % ir F1 – 95,47 %.

Pagrindiniai parametrai yra bendras tikslumas, AUC ir F1, tyrimo metu galima pamatyti, kad geriausi rezultatai su „Adam“ optimizatoriumi pralenkia straipsnyje nurodytus AUC rezultatus 0,1 %, o palyginti su atkartotais straipsnio modelio rezultatais AUC pagerintas 0,6 %. Taip pat, palyginti su atkartotais straipsnio modelio rezultatais, „g2“ pagerintas 1,5 %, bendras tikslumas 1,7 %, F1 – 1,4 %. O lyginant atkartoto straipsnio modelio rezultatus su geriausiais „RMSprop“ tyrimo rezultatai, „RMSprop“ rezultatai yra truputį prastesni arba beveik vienodi, išskyrus F1 rezultatą, jis yra geresnis 0,4 %.

Remiantis šia analize, geriausi „Adam“ tyrimo rezultatai parodo, kad šiam modeliui yra parametrų, su kuriais gaunamas geresnis rezultatas ir šie parametrai yra pasirenkami tolimesnei analizei, nes jie pasižymi aukščiausiais rezultatais beveik visuose matavimo rodikliuose. Šie geriausi parametrai yra:

- Mokymosi greitis – 0,005
- Išmetimo procentas – 0,7
- Įrašo ilgis – 60
- Optimizatorius – „Adam“

Taigi, pasirinkus šiuos naujus parametrus, modelis yra treniruojamas naudojant ankstesnius treniravimo duomenis, bet su jais atliekant 5 dalių kryžminę patikrą, norint dar papildomai patikrinti mokymo paklaidas ar modelis pastoviai pasiekia geresnius rezultatus negu straipsnyje nurodyti rezultatai. Testavimo duomenys yra palikti tokie patys kaip ir straipsnyje nurodyti. Kryžminės patikros modelio mokymo rezultatai pavaizduoti 6 lentelėje.

6 lentelė. Palyginimas straipsnio, atkartoto ir geriausių „Adam“ parametrų rezultatais atlikus 5 dalių kryžminę patikrą, su visais duomenų lygiais.

Duomenų rinkinio lygis	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
Atkartotas - „raw”	86,3806	91,7724	95,7143	97,8705	96,9965	98,0357	97,5133
Atkartotas - „c23“	63,5757	87,1847	92,9333	96,8073	96,7629	93,3333	95,0172
Atkartotas - „c40“	81,5299	82,444	91,2857	93,7111	93,5315	95,5357	94,523
Straipsnio - „raw”	83,4	98,3	99,7	99,9	-	-	-
Straipsnio - „c23“	-	-	-	97,3	-	-	-
Straipsnio -	80,4	85,2	91,2	95,7	-	-	-

Duomenų rinkinio lygis	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
„c40“							
Geriausi parametrai - „raw”	21,7164	90,5671	95,3142	97,4330	96,7902	97,9285	97,3553
Geriausi parametrai - „c23“	21,7164	91,5485	96,4285	98,0241	97,4188	98,1428	97,7770
Geriausi parametrai - „c40“	32,6412	82,9144	89,6571	90,6779	90,8831	97,25	93,9395

Šioje lentelėje pateikti atkartoti straipsnio modelio rezultatai, straipsnyje pateikti rezultatai ir rezultatų vidurkis atlikus kryžminę patikrą su geriausiais tyrimo metu rastais parametrais. Geriausių parametrų „raw“ rezultatai rodo g1 tikslumą 21,71 %, g2 tikslumą 90,56 %, bendrą tikslumą 95,31 %, AUC - 97,43 %, „Precision“ – 96,79 %, „Recall“ – 97,93 % ir F1 – 97,36 %. Rezultatai rodo, kad esant šiems parametrams su „raw“ duomenų rinkinių, nėra pagerinamas nei vienas parametrus tiek su atkartotai duomenimis, tiek su straipsnio duomenimis. Lyginant su straipsnio duomenimis tikslumas yra mažesnis apie 5 %, o AUC apie 2 %, o g1 rezultatas visiškai prastas, praktiškai veikiantis nekorektiškai. Su „c23“ duomenų rinkiniu rezultatai rodo g1 tikslumą 21,71 %, g2 tikslumą 91,54 %, bendrą tikslumą 96,43 %, AUC – 98,02%, „Precision“ – 97,42 %, „Recall“ – 98,14% ir F1 – 97,78%. Lyginant su straipsnyje pateiktu AUC rezultatu, tai jis buvo pagerintas 0,7 %, kiti rezultatai straipsnyje nebuvo pateikti šiam duomenų rinkinio lygiui. O lyginant su atkartotais modelio rezultatais, pagerinti visi rezultatai išskyrus g1. Tikslumas pagerintas 3,5 %, AUC 1,2 % o F1 pagerintas 2,7 %. Galiausiai su „c40“ rodo g1 tikslumą – 32,64 %, g2 tikslumą – 82,91 %, bendrą tikslumą – 89,66 %, AUC – 90,68 %, „Precision“ – 90,88 %, „Recall“ – 97,25 % ir F1 reitingą – 93,94 %. Lyginant su straipsnio pateiktais „c40“ rezultatais, visi rezultatai yra keliais procentais prastesni, pvz. tikslumas 1,7 % prastesnis, o AUC 5 % prastesnis, o lyginant su atkartotais modelio rezultatais, tikslumo ir AUC rezultatai vis tiek prastesni, bet g1 tikslumas

0,5 % geresnis ir „Recall“ yra 1,7 % geresnis. Apibendrinant, su šiais naujais modelio parametrais, buvo šiek tiek pagerintas „c23“ AUC rezultatas.

Tam, kad įsitikinti ar gauti rezultatai su naujais parametrais yra pastovūs, 7 lentelėje pateikti kryžminės patikros rezultatų standartiniai nuokrypiai ir variacijos koeficientai.

7 lentelė. Geriausių parametrų modelio rezultatų patikimumas

Duomenų rinkinys - parametras	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
„raw“ - standartinis nuokrypis	0	0,0058	0,0064	0,0042	0,0052	0,0027	0,0029
„raw“ - variacijos koeficientas	0	0,6510	0,6819	0,4331	0,5469	0,2765	0,2996
„c23“ - standartinis nuokrypis	0	0,0015	0,0050	0,0029	0,0076	0,0044	0,0030
„c23“ - variacijos koeficientas	0	0,1698	0,5237	0,2966	0,7877	0,4566	0,3078
„c40“ standartinis nuokrypis	0,2442	0,0182	0,0254	0,0338	0,0241	0,0130	0,0129
„c40“ variacijos koeficientas	74,8396	2,2038	2,8382	3,7282	2,6524	1,3405	1,3780

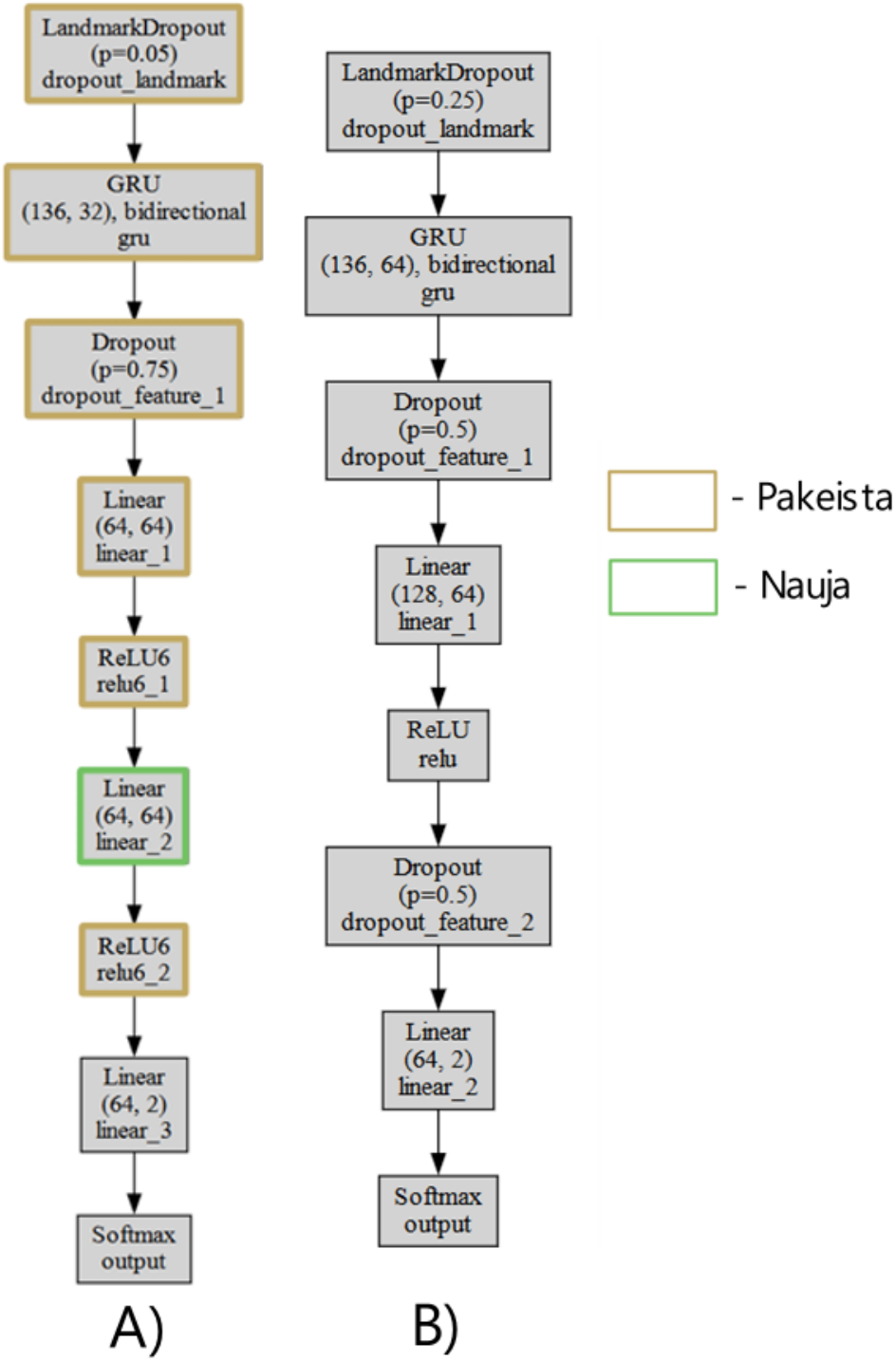
Lyginant skirtingų duomenų rinkinių standartinį nuokrypį, „raw“ duomenų rinkinys turi didesnius standartinio nuokrypio dydžius g2 tikslumui, bendram tikslumui ir AUC, palyginti su „c23“ duomenų rinkiniu. „c40“ duomenų rinkinys turi didesnius standartinio nuokrypio dydžius visiems parametrams, palyginti su „raw“ ir „c23“ duomenų rinkiniais, išskyrus g1 tikslumą, kurio standartinis nuokrypis yra lygus 0. Bet apibendrinus su visais duomenų rinkiniais standartinis nuokrypis yra mažas.

Lyginant skirtingų duomenų rinkinių variacijos koeficientą, „raw“ duomenų rinkinys turi didesnius variacijos koeficiento dydžius g2 tikslumui, bendram tikslumui, AUC, „Precision“ ir „Recall“, palyginti su „c23“ duomenų rinkiniu. „c40“ duomenų rinkinys turi žymiai didesnį variacijos koeficientą g1 tikslumui, palyginti su kitais modeliais, taip pat didesnius variacijos koeficiento dydžius visiems kitiems parametrams, palyginti su „raw“ ir „c23“ modeliais.

„c23“ duomenų rinkinio rezultatų AUC standartinis nuokrypis (0,0050) yra mažesnis nei „raw“ (0,0064) ir „c40“ (0,0254) modelių standartiniai nuokrypiai. Mažesnis standartinis nuokrypis rodo, kad „c23“ AUC rezultatai yra labiau sutelkti aplink vidurkį, todėl modelis yra stabiliau veikiantis ir turi mažesnę rezultatų sklaidą. Taip pat, „c23“ duomenų rinkinio AUC variacijos koeficientas (0,5237) yra mažesnis nei „raw“ (0.6819) ir „c40“ (2,8382) variacijos koeficientai. Mažesnis variacijos koeficientas rodo, kad „c23“ AUC rezultatai yra mažiau išsklaidyti palyginti su kitais duomenų rinkinių rezultatais. Iš šių palyginimų galime daryti išvadą, kad su „c23“ duomenų rinkiniu, modelis yra efektyvesnis ir stabiliau veikiantis nei su kitais duomenų rinkiniais, kai kalbama apie AUC rezultatus ir galima teigti šie rezultatai efektyviai pagerina straipsnyje nurodytus AUC rezultatus.

3. „LRNet“ modelio tobulinimas ir rezultatai

Remiantis praeitame skyiuje gautais rezultatais 26 paveiksle pateikiamas naujas patobulintas modelio variantas. Kuriant šį modelį, buvo bandoma pagerinti „c23“ duomenų lygio AUC rezultatą ir tikslumą, negu nurodyta straipsnyje ar tyrimo metu gautų pagerintų rezultatų. „c23“ duomenų lygis taip pat yra svarbus, nes šiems duomenis yra pritaikyta duomenų kompresija, kuri yra dažnai sutinkama internete, todėl šie duomenys galima sakyti yra svarbiausi atsižvelgiant į modelio pritaikymo aktualumą. O AUC bei tikslumas labiausiai nusako modelio efektyvumą atskirti tikrus vaizdus nuo netikrų.



26 pav. a) patobulintas modelis, b) tyrimo metu naudojamas modelis – „LRNet“.

Pagrindiniai skirtumai tarp dviejų modelių yra pirmojo „Dropout“ koeficientas, GRU paslėptų neuronų skaičius, „Dropout“ sluoksnių skaičius ir tiesinių sluoksnių bei aktyvacijos funkcijų konfigūracija. Patobulintas modelis turi mažesnį GRU paslėptų neuronų skaičių ir papildomus tiesinius

bei ReLU6 sluoksnius, tuo tarpu „LRNet“ turi didesnį pirmąjį „Dropout“ koeficientą ir vienu daugiau „Dropout“ sluoksniu.

Lyginant du modelius, patobulintą modelį ir „LRNet“, pastebimi keli pagrindiniai skirtumai jų struktūroje. Pirmiausia, patobulintame modelyje naudojamas mažesnis pirmasis „Dropout“ koeficientas - 0,05, palyginti su „LRNet“ modeliu - 0,25. Antra, GRU sluoksnio paslėptų neuronų skaičius yra mažesnis patobulintame modelyje - 32, nei „LRNet“ modelyje - 64. Be to, patobulintas modelis turi vieną „Dropout“ sluoksnį su 0,75 išmetimo procentu, tuo tarpu „LRNet“ turi du „Dropout“ sluoksnius su 0,5 išmetimo procentu. Tiesinių sluoksnių ir aktyvacijos funkcijų konfigūracija taip pat yra skirtinga. Patobulintas modelis turi tris tiesinius sluoksnius su ReLU6 aktyvacijos funkcijomis, o „LRNet“ turi du tiesinius sluoksnius su ReLU aktyvacijos funkcija. Taip pat, patobulintame modelyje tiesinių sluoksnių neuronų skaičius yra mažesnis, nes GRU sluoksnio išvestis yra mažesnė. Patobulinto modelio tiesiniai sluoksniai yra: 64x64, 64x64, 64x2, o „LRNet“: 128x64, 64x2. Galiausiai, abiejų modelių išvesties sluoksnis yra „Softmax“ su viena išvestimi. Patobulintas modelis yra sudėtingesnis, su papildomais tiesiniais sluoksniais ir ReLU6 aktyvacijos funkcijomis, tuo tarpu „LRNet“ turi didesnį pirmojo „Dropout“ sluoksnio koeficientą ir vienu daugiau „Dropout“ sluoksniu. Šių modelių palyginimas punktais:

- Pirmojo „Dropout“ koeficientas
    - Patobulintas modelis: 0,05.
    - „LRNet“: 0,25.
- GRU sluoksnis:
    - Patobulintas modelis: paslėptų neuronų skaičius 32.
    - „LRNet“: paslėptų neuronų skaičius 64.
- „Dropout“ sluoksniai:
    - Patobulintas modelis: vienas sluoksnis su 0,75 išmetimo procentu.
    - „LRNet“: du sluoksniai su 0,5 išmetimo procentu.
- Tiesiniai sluoksniai ir aktyvacijos funkcijos:
    - Patobulintas modelis:
        - Pirmasis tiesinis sluoksnis su įvesties dydžiu 64 ir išvesties dydžiu 64.
        - ReLU6 aktyvacijos funkcija po pirmojo tiesinio sluoksnio.
        - Antrasis tiesinis sluoksnis su įvesties dydžiu 64 ir išvesties dydžiu 64.
        - ReLU6 aktyvacijos funkcija po antrojo tiesinio sluoksnio.

        - Trečiasis tiesinis sluoksnis su įvesties dydžiu 64 ir išvesties dydžiu 2.
    - „LRNet“:
        - Pirmasis tiesinis sluoksnis su įvesties dydžiu 128 ir išvesties dydžiu 64.
        - ReLU aktyvacijos funkcija po pirmojo tiesinio sluoksnio.
        - Antrasis tiesinis sluoksnis su įvesties dydžiu 64 ir išvesties dydžiu 2.
- Išvesties sluoksnis:
    - Patobulintas modelis: „Softmax“ su viena išvestimi.
    - „LRNet“: „Softmax“ su viena išvestimi.

Taigi, kaip ir su tyrimo metu rastais naujais parametrais, modelis yra treniruojamas naudojant ankstesnius treniravimo duomenis, bet su jais atliekant 5 dalių kryžminę patikrą, norint dar papildomai patikrinti mokymo paklaidas ar modelis pastoviai pasiekia geresnius rezultatus. Bet testavimo duomenys yra palikti tokie patys kaip nurodyti straipsnyje. Kryžminės patikros patobulinto modelio rezultatai pavaizduoti 8 lentelėje.

8 lentelė. Palyginimas straipsnio, atkartoto ir geriausių „Adam“ parametrų rezultatais ir patobulinto modelio, su visais duomenų lygiais, atliekant 5 dalių kryžminę patikrą.

Duomenų rinkinio lygis	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
Atkartotas - „raw”	86,3806	91,7724	95,7143	97,8705	96,9965	98,0357	97,5133
Atkartotas - „c23“	63,5757	87,1847	92,9333	96,8073	96,7629	93,3333	95,0172
Atkartotas - „c40“	81,5299	82,444	91,2857	93,7111	93,5315	95,5357	94,523
Straipsnio - „raw”	83,4	98,3	99,7	99,9	-	-	-
Straipsnio - „c23“	-	-	-	97,3	-	-	-
Straipsnio -	80,4	85,2	91,2	95,7	-	-	-

Duomenų rinkinio lygis	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
„c40“							
Geriausi parametrais - „raw”	21,7164	90,5671	95,3142	97,4330	96,7902	97,9285	97,3553
Geriausi parametrais - „c23“	21,7164	91,5485	96,4285	98,0241	97,4188	98,1428	97,7770
Geriausi parametrais - „c40“	32,6412	82,9144	89,6571	90,6779	90,8831	97,25	93,9395
Patobulintas modelis - „raw”	79,7798	90,7014	94,5428	96,9191	95,9633	98,1071	97,0193
Patobulintas modelis - „c23“	77,7686	92,4029	96,1714	98,3896	96,7534	98,7142	97,7213
Patobulintas modelis - „c40“	76,7089	83,4552	88,5428	89,1908	89,142	98,2857	93,4827

Šioje lentelėje pateikti atkartoti straipsnio modelio rezultatai, straipsnyje pateikti rezultatai, rezultatų vidurkis atlikus kryžminę patikrą su geriausiais tyrimo metu rastais parametrais ir patobulinto modelio rezultatų vidurkis atlikus kryžminę patikrą.

Patobulinto modelio „raw“ variantas pasiekė labai panašius rezultatus į atkartoto modelio „raw“ ir geriausių parametrų modelio „raw“ rezultatus. Tai rodo, kad patobulintas modelis su „raw“ duomenų rinkiniu turi truputį prastesnius rodiklius, tokius kaip g1 tikslumas (79,78 %), g2 tikslumas (90,70 %), bendras tikslumas (94,54 %), AUC (96,92 %), „Precision“ (95,96 %), „Recall“ (98,11 %) ir F1 (97,02 %). Lyginant su atkartotu modeliu AUC yra 1 % mažesnis, su straipsnio modeliu AUC 3 % mažesnis, o su geriausiu parametru modeliu 0,5 % mažesnis.

Antra, patobulinto modelio „c23“ variantas pasiekė geresnius rezultatus nei atkartotas modelis beveik visuose kriterijuose. g1 tikslumas padidėjo nuo 63,58 % iki 77,77 %, g2 tikslumas iš 87,18 % išaugo iki 92,40 %, bendras tikslumas nuo 92,93 % iki 96,17 %, AUC nuo 96,80 % iki 98,39 %, „Precision“ nepakito, „Recall“ pakilo nuo 93,33 % iki 98,71 % ir F1 nuo 95,01 % iki 97,72 %. Toliau lyginant su straipsnio pateiktais duomenimis AUC buvo pagerintas 1,08 %, lyginant su geriausių parametrų modeliu 0,36 %, o su atkartotu modeliu 1,58 % geresnis.

Galiausiai, patobulintas modelis „c40“ variante parodė panašius rezultatus į atkartotą modelį. g1 tikslumas sumažėjo nuo 81,53 % iki 76,71 %, tačiau g2 tikslumas padidėjo nuo 82,44 % iki 83,46 %, bendras tikslumas sumažėjo nuo 91,28 % iki 88,54 %, AUC nuo 93,71 % iki 89,19 %, „Precision“ nuo 93,53 % iki 89,14 %, „Recall“ padidėjo nuo 95,53 % iki 98,29 % ir F1 sumažėjo nuo 94,52 % iki 93,48 %. Lyginant su geriausių parametrų modeliu, beveik visi rezultatai yra apie 1 % mažesni

Apibendrinant, patobulintas modelis rodo geriausius AUC rezultatus su „c23“ duomenų rinkiniu iš visų modelių ir geriausius kitų parametrų rezultatus, bet su kitais duomenų rinkiniais rodo šiek tiek prastesnius rezultatus. Tam, kad įsitikinti ar gauti rezultatai su patobulintu modeliu yra pastovūs, 9 lentelėje pateikti kryžminės patikros rezultatų standartiniai nuokrypiai ir variacijos koeficientai.

9 lentelė. Patobulinto modelio rezultatų patikimumas

Duomenų rinkinys - parametras	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
„raw“ - standartinis nuokrypis	0,0113	0,0023	0,0077	0,0102	0,0129	0,0009	0,0062
„raw“ - variacijos koeficientas	1,4258	0,2596	0,823	1,0610	1,3523	0,099	0,6465
„c23“ - standartinis nuokrypis	0,0291	0,0029	0,0079	0,0049	0,0093	0,0029	0,0040
„c23“ - variacijos koeficientas	3,7488	0,3164	0,8230	0,5056	0,9692	0,2972	0,4178
„c40“ standartinis	0,0483	0,0073	0,0156	0,0316	0,0148	0,0081	0,0076

Duomenų rinkinys - parametras	**g1 tikslumas**	**g2 tikslumas**	**Bendras tikslumas**	**AUC**	**Precision**	**Recall**	**F1**
nuokrypis							
„c40“ variacijos koeficientas	6,3069	0,8753	1,772	3,5487	1,6709	0,8286	0,8144

Lyginant skirtingų duomenų rinkinių standartinį nuokrypį, matome, kad „raw“ duomenų rinkinys pasižymi mažesniais standartinio nuokrypio dydžiais g2 tikslumui, bendram tikslumui, AUC ir F1, palyginti su „c40“ duomenų rinkiniu, o lyginant su „c23“ turi mažesni tiktai g2 ir bendro tikslumo, bet labai nedideliu skirtumu. Kita vertus, „c40“ duomenų rinkinys turi didesnius standartinio nuokrypio dydžius visiems parametrams, palyginti su „raw“ ir „c23“ duomenų rinkiniais. Apibendrinant, su visais duomenų rinkiniais standartinis nuokrypis yra mažas, bet svarbiausia, kad su „c23“ duomenų rinkiniu AUC ir F1 standartinis nuokrypis mažiausias.

Kalbant apie skirtingų duomenų rinkinių variacijos koeficientą, pastebime, kad „c40“ duomenų rinkinys turi didesnius variacijos koeficiento dydžius g1 ir g2 tikslumui, bendram tikslumui, AUC, „Precision“ ir „Recall“ ir F1, palyginti su „c23“ ir „raw“ duomenų rinkiniais. „raw“ duomenų rinkinys pasižymi mažiausiais variacijos koeficientais išskyrus AUC, „Precision“ ir F1 rezultatais. O geriausiais AUC, „Precision“ ir F1 rezultatais pasižymi „c23“ duomenų rinkinys.

„c23“ duomenų rinkinio AUC standartinis nuokrypis (0,0049) yra mažesnis nei „raw“ (0,0102) ir „c40“ (0,0316) modelių standartiniai nuokrypiai. Mažesnis standartinis nuokrypis rodo, kad „c23“ AUC rezultatai yra labiau sutelkti aplink vidurkį, todėl modelis veikia stabiliau ir turi mažesnę rezultatų sklaidą. Be to, „c23“ duomenų rinkinio AUC variacijos koeficientas (0,5056) yra mažesnis nei „raw“ (1,0610) ir „c40“ (3,5487) variacijos koeficientai. Mažesnis variacijos koeficientas parodo, kad „c23“ AUC rezultatai yra mažiau išsklaidyti palyginti su kitų duomenų rinkinių rezultatais.

Iš šių palyginimų galime daryti išvadą, kad su „c23“ duomenų rinkiniu, modelis yra efektyvesnis ir stabiliau veikiantis nei su kitais duomenų rinkiniais, kai kalbama apie AUC ir tai parodo, kad šis patobulintas modeliais stabiliai pagerina straipsnio modelio rezultatus ir sugeba geriau atpažinti suklastotus vaiz-

dus nuo tikrų vaizdų. Taip pat, buvo pagerintas bendras tikslumas, lyginant su atkartoto modelio rezultatais. Dėja su kitais duomenų rinkiniais ir kitais modelio parametrais, modelis nepasiekė geresnių rezultatų.

Išvados

1. Atlikta „Deepfake“ vaizdų generavimo ir jų klastočių atpažinimo metodų mokslinių straipsnių analizė. Išanalizuota 11 būdų aptikti „Deepfake“ vaizdus, kuriuos galima suskirstyti į 3 grupes: pagal analizuojamą vaizdo pasikeitimą laike, pagal analizuojamą veido sritį vaizduose ir kitokie aptikimo būdai. Dauguma šių aptikimo būdų pasižymi bent 90 % suklastotų vaizdų atpažinimo tikslumu.
2. Identifikuoti 3 duomenų rinkiniai, kurie yra naudojami treniruojant „Deepfake“ vaizdų atpažinimo modelius. Tolimesniam tyrimui pasirinktas Sun et al. (2021) „Deepfake“ vaizdų aptikimo metodas, kuris analizuoja vaizdo pasikeitimą laike ir taip identifikuoja vaizdo įrašus. Šis metodas pasirinktas todėl, nes palyginti su kitais modeliais jis turėjo didžiausia AUC tikslumą.
3. Atliktas modelio tyrimas, kuriame buvo keičiami modelio parametrai, siekiant pagerinti mokslinėje literatūroje pateiktus rezultatus. Keičiami parametrai buvo: įrašo ilgis, išmetimo procentas, mokymosi greitis ir optimizatorius. Tyrimo metu buvo gauti 135 rezultatai ir nustatyta, kad esant: mokymosi greičiui – 0,005, išmetimo procentui – 0,7, įrašo ilgiui – 60 ir optimizatoriui – „Adam“, esant „c23“ duomenų rinkiniui gaunamas geresnis AUC rezultatas.
4. Pasiūlytas patobulintas suklastotų „Deepfake“ vaizdų atpažinimo metodas, kuriame pakeistas pirmojo sluoksnio „Dropout“ koeficientas, GRU paslėptų neuronų skaičius, „Dropout“ sluoksnių skaičius ir tiesinių sluoksnių bei aktyvacijos funkcijos pakeistos į ReLU6. Šis patobulintas modelis turi geresnius „c23“ duomenų rinkinio AUC rezultatus negu Sun et al. (2021) straipsnyje nurodyti rezultatai, bet taip pat ir geresnis AUC rezultatas negu su tyrimo metu rastais geriausiais parametrais.

Šaltiniai

Analytics Vidhya, 2023. *Guide to AUC ROC Curve in Machine Learning : What Is Specificity?* [žiūrėta 2023-03-25] Prieiga per internetą: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: a Compact Facial Video Forgery Detection Network. *10th IEEE International Workshop on Information Forensics and Security, WIFS 2018*. https://doi.org/10.1109/WIFS.2018.8630761

Agarwal, S., Farid, H., Fried, O., & Agrawala, M. (2020). Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. *Undefined*, *2020-June*, 2814–2822. https://doi.org/10.1109/CVPRW50498.2020.00338

C3.ai, 2023. *Precision* [žiūrėta 2023-03-25] Prieiga per internetą: https://c3.ai/glossary/machine-learning/precision/

Castillo Camacho, I., & Wang, K. (2021). A Comprehensive Review of Deep-Learning-Based Methods for Image Forensics. *Journal of Imaging 2021, Vol. 7, Page 69*, *7*(4), 69. https://doi.org/10.3390/JIMAGING7040069

Ciftci, U. A., & Demir, I. (2019). FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *X*, 1–1. https://doi.org/10.1109/TPAMI.2020.3009287

Durall, R., Keuper M., Pfreundt F., Keuper J. (2020). *Unmasking DeepFake with simple* Features. https://arxiv.org/abs/1911.00686

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Canton, C., Ai, F., Team, R., & Ai, F. (2019). The Deepfake Detection Challenge (DFDC) Preview Dataset. *Nem-Initiative.Org*. https://arxiv.org/abs/1910.08854v2

Ganguly, S., Mohiuddin, S., Malakar, S. *et al.* Visual attention-based deepfake video forgery detection. *Pattern Anal Applic* **25**, 981–992 (2022). https://doi.org/10.1007/s10044-022-01083-2

Google, 2023. *Overview of GAN Structure* [žiūrėta 2023-03-25] Prieiga per internetą: https://developers.google.com/machine-learning/gan/gan_structure

Iguazio, 2023. *What is Recall in Machine Learning?* [žiūrėta 2023-03-25] Prieiga per internetą:

https://www.iguazio.com/glossary/recall/

Khormali, A., Yuan, J.-S., Khormali, A. ;, & Yuan, J.-S. (2021). ADD: Attention-Based DeepFake Detection Approach. *Big Data and Cognitive Computing 2021, Vol. 5, Page 49*, *5*(4), 49. https://doi.org/10.3390/BDCC5040049

Lee, G., Kim, M., Zhang, Y., Tak Wu Sam, K., Long, X., & Zhao, T. (2021). Deepfake Detection Using the Rate of Change between Frames Based on Computer Vision. *Sensors 2021, Vol. 21, Page 7367*, *21*(21), 7367. https://doi.org/10.3390/S21217367

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2019). Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3204–3213. https://doi.org/10.1109/CVPR42600.2020.00327

Gu, Z., Chen, Y., Youtu Lab, T., Yao, T., Ding, S., Li Tencent Youtu Lab, J., Huang Tencent Youtu Lab, F., & Ma, L. (2021). Spatiotemporal Inconsistency Learning for DeepFake Video Detection. https://doi.org/10.1145/3474085.3475508

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. https://arxiv.org/abs/1803.09179v1

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-October*, 1–11. https://doi.org/10.1109/ICCV.2019.00009

Shiksha, 2023. ROC-AUC vs Accuracy: Which Metric Is More Important? [žiūrėta 2023-03-25] Prieiga per internetą: https://www.shiksha.com/online-courses/articles/roc-auc-vs-accuracy/

Sun, Z., Han, Y., Hua, Z., Ruan, N., & Jia, W. (2021). *Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features*. http://arxiv.org/abs/2104.04480

Tanaka, M., Shiota, S., & Kiya, H. (2021). A Detection Method of Operated Fake-Images Using Robust Hashing. *Journal of Imaging 2021, Vol. 7, Page 134*, *7*(8), 134. https://doi.org/10.3390/JIMAGING7080134

Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Gonzalez-Sosa, E., & Fierrez, J. (2022). DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. *Engineering Applications of Artificial Intelligence*, *110*, 104673. https://doi.org/10.1016/J.ENGAPPAI.2022.104673

V7Labs, 2023. *F1 Score in Machine Learning: Intro & Calculation* [žiūrėta 2023-03-25] Prieiga per internetą: https://www.v7labs.com/blog/f1-score-guide

Yang, P., Baracchi, D., Ni, R., Zhao, Y., Argenti, F., & Piva, A. (2020). A Survey of Deep Learning-Based Source Image Forensics. *Journal of Imaging 2020, Vol. 6, Page 9*, *6*(3), 9. https://doi.org/10.3390/JIMAGING6030009

Yavuzkilic, S., Sengur, A., Akhtar, Z., & Siddique, K. (2021). Spotting Deepfakes and Face Manipulations by Fusing Features from Multi-Stream CNNs Models. *Symmetry 2021, Vol. 13, Page 1352*, *13*(8), 1352. https://doi.org/10.3390/SYM13081352

Wolter, M., Blanke, F., Heese, R. *et al.* Wavelet-packets for deepfake image analysis and detection. *Mach Learn* **111**, 4295–4327 (2022). https://doi.org/10.1007/s10994-022-06225-5