ELEKTROTEHNIŠKI VESTNIK 93(3): 142–148, 2026
IZVIRNI ZNANSTVENI ČLANEK

Globoki ponaredki na slovenskem političnem parketu: zaznavanje v realnih pogojih družbenih omrežij

Marija Ivanovska

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana, Slovenija
E-pošta: marija.ivanovska@fe.uni-lj.si

Povzetek. V tem prispevku obravnavamo problem zaznavanja globokih ponaredkov slovenskih politikov v realnem okolju družbenih omrežij. Z namenom empiričnega ovrednotenja metod za detekcijo ponarejenih slik smo sestavili podatkovno zbirko, ki vsebuje 8000 obraznih slik osmih slovenskih politikov, pri čemer 4000 slik izhaja iz globokih ponaredkov javno objavljenih na Facebooku in Instagramu. Preostalih 4000 pa izhaja iz avtentičnih videoposnetkov slovenskih medijskih hiš. Na zbrani zbirki smo ovrednotili sedem uveljavljenih metod za zaznavanje globokih ponaredkov: CapsuleNet, CORE, FFD, MesoNet, RECCE, SRM in Xception. Rezultati kažejo, da je uspešnost metod izrazito slabša v primerjavi z uspešnostjo, ki jo te metode dosegajo kadar jih ovrednotimo na standardiziranih podatkovnih zbirkah, ki so nastale v laboratorijskih pogojih. Pri tem smo odkrili, da je več metod sistematično zamenjavalo razreda pristnih in ponarejenih slik. Kvalitativna analiza je pokazala, da so artefakti v sodobnih političnih ponaredkih pogosto lokalni, kratkotrajni in subtilni, zaradi česar jih je težko zaznati med predvajanjem ponarejenih posnetkov z običajno hitrostjo. Rezultati potrjujejo pomembno vrzel med laboratorijskim vrednotenjem detektorjev in njihovo uporabnostjo pri realnih političnih vsebinah z družbenih omrežij.

Ključne besede: globoki ponaredki v političnem prostoru, zaznavanje globokih ponaredkov, umetna inteligenca, računalniški vid, razpoznavanje vzorcev

Deepfakes in Slovene Politics: Detection Performance on Real-World Social Media Content

This paper tackles the problem of detecting deepfakes of Slovene politicians in the real-world environment of social media. For the purpose of empirically evaluating deepfake detection methods, we constructed a dataset containing 8000 facial images of eight Slovene politicians. The dataset consists of 4000 images extracted from publicly available deepfake videos posted on Facebook and Instagram, and 4000 images extracted from authentic videos published by different national medias. Using this dataset, we evaluated seven established deepfake detection methods: CapsuleNet, CORE, FFD, MesoNet, RECCE, SRM, and Xception. The results show that the performance of the evaluated methods significantly degrades when applied to real-world deepfakes instead of standardized benchmark datasets. Furthermore, several methods systematically confused the authentic and manipulated image classes. Qualitative analysis revealed that artefacts in modern political deepfakes are often local, short-lived, and subtle. The obtained results confirm a substantial gap between laboratory evaluation of deepfake detectors and their practical applicability to real political content distributed through social media.

Keywords: political deepfakes, detection of deepfakes, artificial intelligence, computer vision, pattern recognition

Prejet 11. maj, 2026
Odobren 22. maj, 2026

1 UVOD

Sodobni generativni modeli so vedno bolj dostopni in omogočajo relativno hitro in enostavno ustvarjanje umetnih video posnetkov. Med ostalim omogočajo tudi ustvarjanje prepričljivih posnetkov oseb, ki govorijo ali počnejo nekaj, kar se v resnici ni zgodilo [21], [14]. Tovrstni globoki ponaredki (angl. deepfakes) izgledajo precej resnično in imajo zato moč spodkopati integriteto lažno uprizorjenih oseb. Obenem pa zmanjšujejo tudi zaupanje javnosti v digitalno multimedijsko vsebino in povečajo možnost, da se pristni posnetki zavračajo kot domnevni ponaredki [26], [6]. Globoki ponaredki so še posebej problematični v politiki, saj se lahko uporabijo kot orodje za zavajanje državljanov in celo vplivajo na mnenje volivcev v času predvolilnih kampanj [9], [29], [11]. Iz tega razloga je njihova učinkovita in pravočasna zaznava ključna.

Kljub hitremu napredku na področju avtomatiziranega zaznavanja globokih ponaredkov je njihovo zanesljivo prepoznavanje v realnih razmerah izredno zahtevno. Obstoječe metode so večinoma razvite in ovrednotene z uporabo podatkovnih zbirk, kjer je vrsta slikovnih manipulacij jasno definirana, napake, ki manipulacije razkrijejo, pa so natančno označene [21], [14]. Sodobni generativni modeli v praksi pogosto ustvarjajo vse bolj realistične ponarejene obraze, z manj očitnimi napakami, prepričljivo mimiko in natančno sinhronizacijo govora.

Vse to dodatno otežuje razlikovanje med pristnimi in umetno ustvarjenimi posnetki [10]. V političnem kontekstu problem zaostri tudi dejstvo, da se globoki ponaredki najpogosteje širijo prek družbenih omrežij, kot sta Facebook in Instagram. Takšni posnetki so običajno kratki, močno kompresirani in pogosto dodatno obogateni z napisi ali drugimi vizualnimi elementi [23]. Posledično se številni nizkonivojski forenzični kazalci, na katere se detektorji ponaredkov zanašajo, oslabijo ali celo izgubijo, kar zmanjšuje uspešnost zaznave [2].

Opisani izzivi niso zgolj teoretični, temveč se pojavljajo tudi v slovenskem medijskem prostoru. Na družbenih omrežjih so se na primer nedavno pojavili videoposnetki in slike, ki so prikazovali politične akterje v zavajajočih ali umetno ustvarjenih situacijah. Del teh vsebin je bil v javnosti že obravnavan kot umetno ustvarjen ali tehnološko manipuliran material, kar kaže, da vprašanje političnih globokih ponaredkov ni več zgolj hipotetično, temveč tudi aktualen problem slovenskega medijskega prostora [22].

V tem prispevku predstavimo študijo *slovenskih političnih globokih ponaredkov*, zbranih iz javno dostopnih profilov in strani na družbenih omrežjih. Osredotočamo se na dve ravni analize: 1) kvalitativno analizo značilnih slikovnih artefaktov, ki se pojavljajo v zbranih ponarejenih posnetkih, 2) empirično ovrednotenje obstoječih orodij za zaznavanje teh ponaredkov. Cilj naše sistematične analize je preiskati v kolikšni meri obstoječa orodja in forenzični kazalci zadoščajo za zaznavo globokih ponaredkov, ki niso bili ustvarjeni v laboratorijskih pogojih.

2 SORODNA DELA

2.1 Zaznavanje globokih ponaredkov

Algoritmi za zaznavanje globokih ponaredkov se običajno osredotočajo na iskanje vizualnih nepravilnosti, ki nastanejo med ustvarjanjem umetne vsebine. Primeri slikovnih napak vključujejo: nenaravno zlivanje meje obraza z okolico, nekonsistentno osvetlitev in sence, nenavadno teksturo kože, podvojene obrvi, popačeno zobovje itd. [12] Tipični detektorji ponaredkov predstavljajo globoke konvolucijske mreže, ki lokalne slikovne napake iščejo s primerjavo pristnih in ponarejenih primerov slik [1]. Novejše metode uporabljajo tudi bolj napredne arhitekture, kot so transformerji, ki omogočajo modeliranje globalnih semantičnih nekonsistentnosti [7]. Poleg prostorskih napak detektorji pogosto analizirajo tudi časovno konsistentnost posnetkov ali nepravilnosti v frekvenčni domeni [17].

Ker so lastnosti slikovnih nepravilnosti ponaredkov močno odvisne od generativne metode, ki jih je ustvarila, so se v preteklosti pojavili tudi detektorji, ki so naučeni na simuliranih ponaredkih in zato lažje zaznavajo širši nabor napak [24]. Še vedno pa so tovrstni algoritmi omejeni na zaznavo videza napak, ki ga modelirajo med učenjem. Kot rešitev za to težavo so raziskovalci razvili detektorje, ki zaznavajo odstopanja v obrazni mimiki in drži glave. Li idr. so tako predlagali metodo za zaznavanje nenaravnega mežikanja, saj se pri manipuliranju obrazov pogosto nepravilno reproducira odpiranje in zapiranje oči [13]. Yang idr. pa so pokazali, da je nekonsistentnost položaja glave ravno tako dober kazalec nepristnosti posnetkov [28].

Detektorji za zaznavanje globokih ponaredkov običajno dosegajo zelo dobre rezultate na standardnih podatkovnih zbirkah kot so FaceForensics++ [21], Celeb-DF [14] in DFDC [10]. Pri tem je treba poudariti, da vse te zbirke vsebujejo nadzorovano ustvarjene visokokakovostne manipulacije, ki so natančno opisane in označene. Kadar se detektorji ovrednotijo na podatkih iz realnega sveta, ki so ponavadi ustvarjeni z neznano metodo, pa pride do degradacije njihove uspešnosti in zanesljivosti [27], [16], [5]. Še posebej je problematično zaznavanje ponaredkov z družbenih omrežij, saj so ti kompresirani, pomanjšani in pogosto dodatno opremljeni z vizualnimi elementi, kar zakrije ali celo odstrani forenzične sledi [18].

Namen našega prispevka je v tem kontekstu ovrednotiti uspešnost delovanja uveljavljenih detektorjev globokih ponaredkov na primeru nedavnih posnetkov slovenskih politikov iz družbenih omrežij Facebook in Instagram.

2.2 Politično motivirani globoki ponaredki

V zadnjem času smo priča širjenju vedno bolj prepričljivih globokih ponaredkov političnih akterjev. Študije sicer kažejo, da imajo tudi tehnično nedovršeni politični ponaredki nezanemarljiv vpliv na javno mnenje in integriteto lažno prikazanih politikov [11]. Z napredkom generativne tehnologije pa so politično motivirani ponaredki postali tudi orodje za zavajanje državljanov v času predvolilnih kampanj [29]. Trokielewicz idr. so z analizo lažnih političnih oglasov, pokazali, da sodobni ponaredki ohranijo biometrično prepoznavnost politika, hkrati pa prepričljivo spremenijo semantično vsebino političnega sporočila [25].

Med iskanjem rešitev za omenjeno težavo so Sankaranarayanan idr. pokazali, da splošni detektorji ponaredkov ne zagotavljajo zanesljive zaznave lažnih posnetkov ameriških politikov [23]. Lin idr. pa so ugotovili, da je uspešnost in zanesljivost detektorjev še slabša pri zaznavanju ponaredkov, objavljenih na družbenih omrežjih [15]. Boháček idr. so v svoji raziskavi predlagali identitetno usmerjen forenzični pristop, ki modelira značilne obrazne izraze in gibalne vzorce svetovnih voditeljev [3]. Ta pristop sloni na ideji, da ima javna oseba prepoznaven način govora in gibanja, ki ga algoritmi za ustvarjanje ponaredkov pogosto ne reproducirajo pravilno. Boháček idr. so to idejo nadgradili s pristopom modeliranja obrazne mimike, gestikulacij in glasovnih lastnosti politikov, ki so tarča napadov [2]. Vendar implementacija takšnega pristopa v praksi zahteva kompleksne multimodalne modele, ki se učijo z uporabo

velikih količin reprezentativnih podatkov o posameznem politiku, kar omejuje njegovo praktično uporabnost.

V tem prispevku se osredotočamo na analizo javno dostopnih umetnih posnetkov slovenskih politikov, ki so v medijskem prostoru dosegli veliko gledanost. Naš pristop vključuje tako kvalitativno analizo slikovnih artefaktov, kot tudi empirično vrednotenje obstoječih metod za detekcijo umetne vsebine. S tem prispevamo k razumevanju, kako se metode za zaznavanje globokih ponaredkov obnašajo v manjšem jezikovnem in političnem prostoru, kjer je manj standardiziranih referenčnih posnetkov, manj javnih podatkovnih zbirk in večja odvisnost od kontekstualnega preverjanja.

3 METODOLOGIJA IN EKSPERIMENTALNI PROTOKOL

3.1 Metode za zaznavanje globokih ponaredkov

V eksperimentalnem delu smo uporabili sedem uveljavljenih metod za zaznavanje globokih ponaredkov: CapsuleNet [19], CORE [20], FFD [8], MesoNet [1], RECCE [4], SRM [17] in Xception [21]. Uporabljene metode pokrivajo več različnih pristopov k zaznavanju ponaredkov. Xception in MesoNet predstavljata diskriminativne konvolucijske pristope, ki se učijo razlikovati med pristnimi in ponarejenimi obraznimi slikami neposredno v slikovnem prostoru. CapsuleNet uporablja kapsulne mreže, katerih namen je bolj eksplicitno modeliranje prostorskih odnosov med značilnicami obraza, zato lahko zaznava tudi nekatere geometrijske in strukturne nepravilnosti. CORE in FFD predstavljata pristopa, ki poskušata izboljšati generalizacijo detektorjev z implementacijo mehanizmov, ki preprečujejo prekomerno prilagajanje detektorja na specifične artefakte posamezne generativne metode. RECCE se nauči kompaktne predstavitve pristnih obrazov, nato pa odstopanja med rekonstruirano in vhodno sliko uporabi kot dodaten vir informacij za zaznavanje ponaredkov. SRM pa poleg običajnih slikovnih značilnic uporablja tudi informacije iz visoko-frekvenčnega oziroma rezidualnega prostora. Vse uporabljene metode so bile prednaučene in javno dostopne v okviru orodja DeepfakeBench*, ki je standardizirano orodje za ovrednotenje in primerjavo metod za zaznavanje globokih ponaredkov. Ker je cilj našega prispevka ocena uspešnosti delovanja obstoječih metod v realnem okolju, detektorjev nismo dodatno učili na slovenskih podatkih. S tem smo ovrednotili sposobnost prenosa metod iz standardnih učnih zbirk na realne politične posnetke z družbenih omrežij.

3.2 Podatkovna zbirka

Za potrebe raziskave smo sestavili lastno podatkovno zbirko obraznih slik slovenskih politikov. Ponarejeni del zbirke temelji na javno objavljenih videoposnetkih s Facebook strani *Politična satira*† in Instagram profila *Slomemija*‡. Ponarejene posnetke smo kot globoke ponaredke označili na podlagi javnega konteksta objave, vizualne analize in dejstva, da so bili objavljeni kot umetno ustvarjene oziroma satirične manipulacije političnih oseb. V zbirko nismo vključili primerov, pri katerih ni bilo mogoče zanesljivo sklepati, da gre za umetno ustvarjeno ali tehnološko manipulirano obrazno vsebino. Skupno smo pridobili 10 videoposnetkov globokih ponaredkov, iz katerih smo izvozili 4000 ponarejenih obraznih slik. Zbirka vključuje 8 različnih slovenskih politikov, pri čemer je število slik po politikih uravnoteženo: za vsakega politika smo uporabili 500 ponarejenih obraznih slik. Za primerjavo smo v zbirko dodali tudi enako število pristnih obraznih slik. Te smo pridobili iz 8 avtentičnih video posnetkov, javno objavljenih na platformi YouTube na kanalih medijskih hiš RTV Slovenija§, POP TV¶ in SVET24‖. Za vsakega izmed 8 politikov smo izvozili 500 pristnih obraznih slik, kar skupaj predstavlja 4000 pristnih slik. Celotna podatkovna zbirka tako vsebuje 8000 obraznih slik, od tega 4000 ponarejenih in 4000 pristnih. Takšna uravnotežena zasnova podatkovne zbirke omogoča neposredno primerjavo uspešnosti detektorjev brez vpliva neuravnoteženosti razredov. Pri izvozu slik iz videoposnetkov smo za vsako sekundo posnetka izbrali štiri časovno enakomerno razporejene slike. S tem smo zmanjšali možnost, da bi v zbirko vključili veliko število skoraj identičnih zaporednih slik. Na vsaki izvoženi sliki smo z detektorjem obrazov DLib** zaznali lokacijo obraza politika. Zaznani obraz smo nato obrezali s 30-odstotno margino okrog zaznanega obraznega območja. Obrezana slika tako poleg osrednjega dela obraza vključuje tudi robove obraza, lase, del vratu in neposredno okolico obraza, kjer se pogosto pojavijo artefakti zlivanja med generiranim obrazom in ozadjem. Po obrezovanju smo obraze tudi poravnali z namenom geometrijske normalizacije slike na podlagi značilnih obraznih točk, kot so položaji oči, nosu in ust. Cilj poravnave je, da so obrazi v vseh vhodnih slikah predstavljeni v podobni legi, merilu in orientaciji. S tem zmanjšamo vpliv različnih položajev glave in izreza ter zagotovimo, da detektorji primerjajo primerljive obrazne regije. Tako pripravljene slike smo nato uporabili kot vhod v vse izbrane metode za zaznavanje globokih ponaredkov. Primeri slik so podani na Sliki 1.

3.3 Metrike

Uspešnost metod za zaznavanje ponaredkov smo ovrednotili z metrikami AUC, inverzno vrednostjo AUC_{inv},

*https://github.com/sclbd/deepfakebench

†https://www.facebook.com/people/Politi%C4%8Dna-satira/61586948309543/

‡https://www.instagram.com/slomemija/

§https://www.youtube.com/@rtvsloinfo

¶https://www.youtube.com/@POPolnkanal

‖https://www.youtube.com/@svet24si

**https://github.com/davisking/dlib-models

Slika 1 Izbrani primeri pristnih (prva vrstica) in ponarejenih slik slovenskih politikov (druga in tretja vrstica) izvoženih iz javno objavljenih videoposnetkov na družbenih omrežjih Facebook in Instagram. Opazimo lahko, da so sodobni politični ponaredki pogosto vizualno zelo prepričljivi in brez večjih očitnih napak.

natančnostjo N (angl. Precision), priklicem P (angl. Recall) in ROC krivuljo. Naj bo $y_i \in \{0, 1\}$ prava oznaka i-te slike, kjer $y_i = 1$ označuje ponarejeno sliko, $y_i = 0$ pa pristno sliko. Naj bo $s_i \in [0, 1]$ predikcija detektorja, ki predstavlja verjetnost oziroma stopnjo zaupanja, da je i-ta slika ponarejena. Za izbrani prag τ sliko razvrstimo kot ponarejeno, če velja $s_i \geq \tau$, sicer pa kot pristno.

Pri danem pragu $\tau = 0,5$ definiramo število pravilno zaznanih ponaredkov kot TP (angl. True Positive), število napačno zaznanih ponaredkov kot FP (angl. False Positive), število pravilno zaznanih pristnih slik kot TN (angl. True Negative) in število spregledanih ponaredkov kot FN (angl. False Negative). Natančnost N pove, kolikšen delež slik, ki jih je detektor označil kot ponaredke, je dejansko ponarejen:

$$\mathrm{N} = \frac{TP}{TP + FP}. \tag{1}$$

Priklic P pa pove, kolikšen delež vseh dejanskih ponaredkov je detektor pravilno zaznal:

$$\mathrm{P} = \frac{TP}{TP + FN}. \tag{2}$$

Za izris krivulje ROC uporabimo stopnjo pravilno pozitivnih zaznav TPR (angl. True Positive Rate) in stopnjo napačno pozitivnih zaznav oziroma FPR (angl. False Positive Rate). Definirani sta kot:

$$\mathrm{TPR} = \frac{TP}{TP + FN}, \tag{3}$$

$$\mathrm{FPR} = \frac{FP}{FP + TN}. \tag{4}$$

Krivulja ROC prikazuje razmerje med TPR in FPR pri različnih vrednostih praga τ. Ker se prag spreminja od zelo strogega do zelo ohlapnega, ROC krivulja ne meri uspešnosti detektorja samo pri eni izbrani nastavitvi, temveč pokaže njegovo obnašanje skozi celoten razpon možnih pragov. To je posebej pomembno pri zaznavanju globokih ponaredkov, saj optimalen prag v realni uporabi ni nujno enak za vse detektorje, vse podatkovne zbirke ali vse vrste ponaredkov.

Površina pod ROC krivuljo oziroma AUC predstavlja splošno sposobnost detektorja, da ponarejenim slikam pripiše višje ocene kot pristnim slikam. Formalno jo lahko zapišemo kot:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d\mathrm{FPR}. \tag{5}$$

Vrednost $\mathrm{AUC} = 1$ pomeni popolno ločevanje med pristnimi in ponarejenimi slikami, vrednost $\mathrm{AUC} = 0,5$ pa ustreza naključnemu razvrščanju. Poleg AUC v tem prispevku poročamo tudi inverzno vrednost AUC, označeno kot AUC_{inv}:

$$\mathrm{AUC}_{inv} = 1 - \mathrm{AUC}. \tag{6}$$

Za razliko od mere AUC, AUC_{inv} predpostavi, da detektor ponarejenim slikam pripisuje nižje verjetnosti kot pristnim slikam. AUC_{inv} je torej še posebej uporabna v primerih, ko detektor sistematično zamenjuje razreda pristnih in ponarejenih slik.

4 Rezultati

4.1 Kvantitativna analiza

Tabela 1 prikazuje rezultate kvantitativnega vrednotenja uporabljenih metod za zaznavanje globokih ponaredkov na zbrani zbirki slovenskih političnih posnetkov. Rezultati kažejo, da nobena izmed uporabljenih metod ni dosegla visoke uspešnosti zaznavanja. Vrednosti AUC vseh metod so tudi bistveno nižje od rezultatov, ki

Tabela 1 Kvantitativni rezultati ovrednotenja uveljavljenih metod za zaznavanje globokih ponaredkov kažejo, da je njihova uspešnost bistveno slabša kot uspešnost, ki jo dosegajo na standardiziranih posnetkih. Visoke vrednosti $\mathbf{AUC_{inv}}$ razkrivajo, da metode pogosto označujejo pristne slike kot ponarejene in obratno.

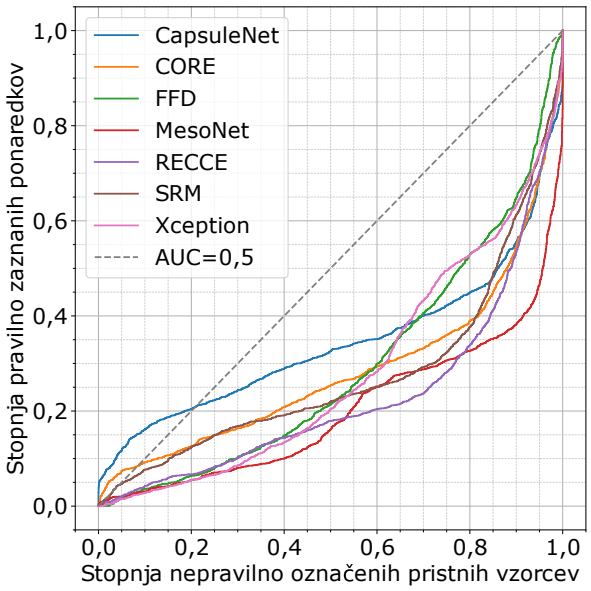
Model	**AUC** ↑	$\mathbf{AUC_{inv}}$ ↑	**N** ↑	**P** ↑
CapsuleNet [19]	$\mathbf{0,342}$	$0,658$	$0,295$	$0,304$
CORE [20]	$0,284$	$0,716$	$0,241$	$0,214$
FFD [8]	$0,291$	$0,709$	$0,226$	$0,272$
MesoNet [1]	$0,200$	$\mathbf{0,800}$	$\mathbf{0,380}$	$\mathbf{1,000}$
RECCE [4]	$0,228$	$0,772$	$0,172$	$0,197$
SRM [17]	$0,274$	$0,726$	$0,226$	$0,382$
Xception [21]	$0,281$	$0,719$	$0,218$	$0,260$
Povprečje	$0,271$	$0,729$	$0,251$	$0,376$

jih iste metode običajno dosegajo na standardnih podatkovnih zbirkah, kot so FaceForensics++, Celeb-DF ali DFDC. To potrjuje ugotovitve sorodnih raziskav, da se uspešnost detektorjev v realnih pogojih družbenih omrežij močno poslabša.

Najvišjo vrednost AUC je dosegla metoda CapsuleNet z vrednostjo $0,342$, kar pomeni, da je ta metoda najbolje ločevala med pristnimi in ponarejenimi slikami. Kljub temu je dosežena uspešnost še vedno nizka in pod pragom naključnega razvrščanja. AUC_{inv} vrednosti so po drugi strani relativno visoke, saj najboljši detektor MesoNet dosega AUC_{inv} z vrednostjo $0,8$, drugi najboljši (RECCE) pa $0,772$. Tako visoke vrednosti razkrivajo, da izbrani detektorji pristnim slikam pogosto pripisujejo visoke verjetnosti ponaredka, ponarejenim slikam pa nizke verjetnosti, kar je ravno obratno od pričakovanega načina obratovanja metod. Detektorji torej pristne slike pogosto označijo kot ponarejene z visoko stopnjo zaupanja, medtem ko ponarejenim slikam pripisujejo visoko verjetnost pristnosti. To lahko kaže na prisotnost uporabne, vendar napačno interpretirane informacije, lahko pa tudi na izrazit domenski zamik med učnimi zbirkami detektorjev in realnimi političnimi posnetki z družbenih omrežij.

Pri interpretaciji rezultatov je pomembno upoštevati tudi vrednosti natančnosti N in priklica P. MesoNet je na primer dosegel popoln priklic ($\mathrm{P} = 1$), vendar hkrati razmeroma nizko natančnost ($\mathrm{N} = 0,38$). To pomeni, da metoda ni spregledala ponarejenih slik, vendar je kot ponaredke označila tudi veliko pristnih slik, kar kaže na visoko stopnjo napačno pozitivnih zaznav.

Dodatni vpogled v obnašanje detektorjev podajajo ROC krivulje, prikazane na Sliki 2. ROC krivulje vseh metod so v večini primerov pod diagonalo, ki predstavlja naključno razvrščanje ($AUC = 0,5$). Le pri metodi CapsuleNet opazimo, da je njena ROC krivulja nad diagonalo, ko je stopnja nepravilno označenih pristnih vzorcev majhna. Stopnja pravilno zaznanih ponaredkov v tem območju vseeno ne preseže vrednosti $0,2$.



Slika 2 ROC krivulje ovrednotenih metod za zaznavanje globokih ponaredkov slovenskih politikov potekajo pod diagonalo, ki predstavlja naključno razvrščanje ($AUC = 0,5$), kar kaže na sistematično zamenjavo pristnih in ponarejenih slik.

Opazimo lahko tudi, da se večina krivulj začne izraziteje dvigovati šele pri zelo visokih vrednostih stopnje napačno pozitivnih zaznav, kar pomeni, da morajo detektorji za doseganje višjega priklica sprejeti veliko število napačnih zaznav pristnih slik.

Kvantitativni rezultati eksperimentov vseh ovrednotenih metod potrjujejo glavno hipotezo tega prispevka, da so politični globoki ponaredki iz družbenih omrežij bistveno zahtevnejši problem od standardnih laboratorijskih primerov.

4.2 Kvalitativna analiza

Poleg kvantitativnega vrednotenja detektorjev smo izvedli tudi kvalitativno analizo značilnih artefaktov, prisotnih v zbranih političnih globokih ponaredkih. Posnetki, uporabljeni za sestavo podatkovne zbirke, so bili pogosto kratki, večkrat kompresirani in dodatno opremljeni z vizualnimi elementi, kot so napisi in logotipi. Poleg tega so bili analizirani ponaredki relativno kakovostni in časovno konsistentni, brez večjih očitnih napak, kar nakazuje uporabo sodobnejših generativnih modelov. Takšne lastnosti otežujejo generalizacijo detektorjev, naučenih na nadzorovanih podatkovnih zbirkah, kjer so manipulacije pogosto ustvarjene v bolj kontroliranih pogojih.

Kljub relativno visoki kakovosti posnetkov smo v posameznih primerih zaznali več različnih vrst artefaktov. Primeri značilnih napak so prikazani na Sliki 3. Med najbolj pogostimi artefakti je prisotnost znamenj na koži, ki jih prikazana oseba v resnici nima. Takšna znamenja se včasih pojavijo le v posameznih časovnih intervalih videoposnetka, nato pa izginejo ali spremenijo

Slika 3 Primeri značilnih artefaktov v analiziranih političnih globokih ponaredkih. Z rdečo barvo so označene subtilne prostorske nepravilnosti, kot so nekonsistentne sence, umetno ustvarjena znamenja na koži, popačeno zobovje in nepravilnosti v očesnem predelu kot je popačena šarenica. Poleg prostorskih artefaktov se pojavljajo tudi časovne nekonsistentnosti, na primer spreminjanje gostote ali oblike brade in obraznih dlak med zaporednimi slikami videoposnetka.

obliko, kar kaže na časovno nekonsistentnost generiranega obraza. V določenih primerih smo opazili tudi nekonsistentne sence na obrazu ali nenaravne prehode med osvetljenimi in neosvetljenimi deli kože.

Pogoste so bile tudi manj izrazite morfološke nepravilnosti. Na določenih posnetkih se je pojavilo nepravilno oblikovano zobovje, pri katerem so bili posamezni zobje deformirani, zamegljeni ali geometrijsko nekonsistentni. Občasno smo zaznali tudi popačeno strukturo šarenice oziroma očesnega predela, kjer je generativni model ustvaril nenaravno obliko zenice ali nepravilne odseve svetlobe. Takšne napake so pogosto subtilne in jih je težko zaznati pri običajnem predvajanju videoposnetka, vendar postanejo opaznejše pri analizi posameznih slik.

Poleg prostorskih artefaktov smo zasledili tudi časovne nekonsistentnosti. Pri nekaterih moških obrazih se je na primer skozi čas spreminjala gostota ali oblika brade in obraznih dlak. Tovrstne nepravilnosti nakazujejo, da generativni model med ustvarjanjem posameznih slik ne ohranja popolne časovne stabilnosti obraznih značilnic.

Kvalitativna analiza tako potrjuje, da sodobni politični globoki ponaredki pogosto vsebujejo bistveno manj očitnih napak kot starejši primeri iz standardnih podatkovnih zbirk. Artefakti so običajno lokalni, kratkotrajni in subtilni, zato jih je težko zaznati zgolj z vizualnim pregledom videa. To dodatno pojasnjuje nizko uspešnost uporabljenih detektorjev, saj so številni naučeni predvsem na izrazitejših in bolj sistematičnih napakah starejših generativnih metod. Naši rezultati zato nakazujejo, da zaznavanje sodobnih političnih globokih ponaredkov v realnem okolju družbenih omrežij ostaja odprt raziskovalni problem.

5 Zaključek

V tem prispevku smo obravnavali problem zaznavanja globokih ponaredkov politikov javno objavljenih na družbenih omrežjih. Ovrednotili smo sedem uveljavljenih metod, pri čemer smo pokazali, da je njihova uspešnost delovanja na teh posnetkih občutno slabša v primerjavi z uspešnostjo, ki jo dosegajo na standardnih laboratorijskih zbirkah. Pri tem smo ugotovili, da večina metod zelo pogosto razvrsti pristne slike kot ponaredke in obratno. Kvalitativna analiza je dodatno pokazala, da sodobni politični globoki ponaredki pogosto ne vsebujejo velikih in očitnih napak, značilnih za starejše generativne metode. Opazni artefakti so bili večinoma lokalni, kratkotrajni in subtilni. Takšne napake so pri običajnem ogledu videoposnetka težko zaznavne, hkrati pa jih lahko kompresija in dodatna obdelava na družbenih omrežjih še dodatno prikrijeta.

Pomembna omejitev naše raziskave je relativno majhno število izvornih videoposnetkov, saj je 8000 slik vzorčenih iz omejenega števila javno dostopnih posnetkov. Posamezne slike zato niso popolnoma neodvisni vzorci. Poleg tega se lahko pristni in ponarejeni posnetki razlikujejo tudi po viru, kompresiji, ločljivosti in produkcijskih pogojih, kar lahko vpliva na odziv detektorjev. Opravljeno raziskavo zato razumemo kot študijo primera delovanja detektorjev v realnem okolju, ne kot splošno primerjavo njihove absolutne uspešnosti.

Rezultati tega prispevka potrjujejo pomembno vrzel med laboratorijskim vrednotenjem metod za zaznavanje globokih ponaredkov in njihovo dejansko uporabnostjo v realnem medijskem okolju. Za zanesljivo obravnavo političnih globokih ponaredkov zato ni dovolj zgolj uporaba prednaučenih detektorjev, temveč je potrebna kombinacija avtomatske detekcije, ročne vizualne analize in kontekstualnega preverjanja.

Zahvala

Raziskava predstavljena v tem prispevku je bila delno financirana z ARIS temeljnim programom P2-0250(B) ”Metrologija in biometrični sistemi”in ARIS raziskovalnim projektom J2-50065 ”Odkrivanje globokih ponaredkov z metodami zaznave anomalij (DeepFake DAD)”.

Literatura

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. MesoNet: a Compact Facial Video Forgery Detection Network. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, page 1–7, 2018.

[2] M. Bohacek and H. Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48), 2022.

[3] M. Bohacek and H. Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48), 2022.

[4] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022.

[5] B. Cavia, E. Horwitz, T. Reiss, and Y. Hoshen. Real-time deepfake detection in the real-world, 2024.

[6] R. Chesney and D. K. Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN Electronic Journal*, 2018.

[7] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva. Raising the Bar of AI-generated Image Detection with CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 4356–4366, 2024.

[8] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[9] T. Dobber, N. Metoui, D. Trilling, N. Helberger, and C. de Vreese. Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *The International Journal of Press/Politics*, 26(1):69–91, 2020.

[10] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton-Ferrer. The DeepFake Detection Challenge Dataset. *arXiv preprint*, abs/2006.07397, 2020.

[11] M. Hameleers. Cheap Versus Deep Manipulation: The Effects of Cheapfakes Versus Deepfakes in a Political Setting. *International Journal of Public Opinion Research*, 36(1), 2024.

[12] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face X-Ray for More General Face Forgery Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5000–5009, 2020.

[13] Y. Li, M.-C. Chang, and S. Lyu. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, page 1–7, 2018.

[14] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3204–3213, 2020.

[15] G. Lin, L. Lin, C. P. Walker, D. S. Schiff, and S. Hu. Fit for purpose? deepfake detection in the real world, 2025.

[16] L. Lin, N. Gupta, Y. Zhang, H. Ren, C.-H. Liu, F. Ding, X. Wang, X. Li, L. Verdoliva, and S. Hu. Detecting Multimedia Generated by Large AI Models: A Survey. 2024.

[17] Y. Luo, Y. Zhang, J. Yan, and W. Liu. Generalizing Face Forgery Detection With High-Frequency Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16317–16326, 2021.

[18] A. Montibeller, D. Shullani, D. Baracchi, A. Piva, and G. Boato. Bridging the Gap: A Framework for Real-World Video Deepfake Detection via Social Network Compression Emulation. In *Proceedings of the 1st on Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media*, page 29–36, 2025.

[19] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019.

[20] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao. CORE: Consistent Representation Learning for Face Forgery Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 12–21, 2022.

[21] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 1–11, 2019.

[22] RTV Slovenija. Asta Vrečko zaradi videa, ustvarjenega z umetno inteligenco, razmišlja o kazenskem pregonu, 2026. Dostopano: 4.5.2026.

[23] A. Sankaranarayanan, M. Groh, R. Picard, and A. Lippman. The Presidential Deepfakes Dataset. In *CEUR Workshop Proceedings*, 2021.

[24] K. Shiohara and T. Yamasaki. Detecting Deepfakes with Self-Blended Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 18699–18708, 2022.

[25] E. B. Trokielewicz, K. Roszczewska, and A. Martinek. Faces of Deception - a multi-level analysis of lip-sync-based fake ads involving politicians and their impact on content semantics and recognition systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 946–955, 2026.

[26] C. Vaccari and A. Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1), 2020.

[27] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang. Deepfake Detection: A Comprehensive Survey from the Reliability Perspective. *ACM Computing Surveys*, 57(3):1–35, 2024.

[28] X. Yang, Y. Li, and S. Lyu. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 8261–8265, 2019.

[29] M. Łabuz and C. Nehring. On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science*, 23(4):454–473, 2024.

Marija Ivanovska je asistenka na Fakulteti za elektrotehniko, Univerze v Ljubljani (FE UL), kjer na 1. in 2. bolonjski stopnji poučuje različne predmete na področju analize signalov, predvsem v povezavi z računalniškim vidom in razpoznavanjem vzorcev. Leta 2019 je na FE UL magistrirala s področja elektrotehnike in procesne avtomatizacije, leta 2026 pa doktorirala na področju umetne inteligence in računalniškega vida. Njeno raziskovalno delo obsega razvoj algoritmov za zaznavanje anomalij v slikah s poudarkom na zaznavanju ponarejenih obraznih slik. Leta 2022 je bila gostujoča raziskovalka na Johns Hopkins University v ZDA, leta 2024 pa gostujoča asistentka na Queensland University of Technology v Avstraliji. Marija je avtorica več člankov objavljenih v vplivnih revijah in se s svojimi prispevki redno udeležuje uglednih konferenc. Aktivna je tudi v organizacijskih odborih vodilnih konferenc s področij njenega raziskovanja, kot so BMVC, WACV, IJCB. Marija je članica mednarodnega tehničnega odbora IEEE Information Forensics and Security Technical Committee, ki deluje znotraj organizacije IEEE Signal Processing Society z namenom spodbujanja znanstvenega in tehnološkega razvoja na področju digitalne forenzike in varnosti inteligentnih sistemov.