

Videti še ne pomeni verjeti: zaznava ponarejenih osebnih dokumentov z velikimi jezikovnimi modeli

Marija Ivanovska

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

Povzetek. V dobi digitalizacije družbe je oddaljena identifikacija uporabnikov prek zajema fotografije osebnega dokumenta ključna za učinkovito vodenje storitev. Kakorkoli, sistemi, ki to omogočajo, so še vedno ranljivi in podvrženi napadom, ki med drugim vključujejo lažno identifikacijo s ponarejenimi osebnimi dokumenti. Običajne rešitve te težave uporabljajo algoritme, ki analizirajo pogoste napake, ki jih je mogoče zaslediti na ponarejenih dokumentih. Težava nastane, ko napake zaradi vedno bolj učinkovitih pristopov digitalnega manipuliranja slikovnih vsebin spremenijo svoj videz. Poleg tega obstoječi algoritmi za zaznavanje ponaredkov ne znajo jezikovno razložiti, zakaj je bil določen dokument označen kot ponaredek. V tej študiji se omenjenih težav lotevamo z raziskovanjem sodobnih večmodalnih velikih jezikovnih modelov (VVJM) v kontekstu zaznavanja ponarejenih osebnih dokumentov. Predstavimo prvo sistematično ovrednotenje šestih odprtokodnih prednaučenih modelov (DeepSeek-VL2-Tiny, LLaVA-1.6-Mistral-7B, Qwen2.5-VL-7B, Gemma-3-12B, InternVL3.5-4B, Pixtral-12B) z uporabo največje standardizirane podatkovne zbirke osebnih (avtentičnih in ponarejenih) dokumentov IDNet. Naša raziskava pokaže, da so VVJM zelo učinkoviti pri zaznavanju različnih vrst digitalnih manipulacij osebnih dokumentov. Podrobnejša analiza po državah, ki so dokumente izdale, pa razkrije, da je uspešnost zaznave močno odvisna od vizualnega sloga dokumenta.

Ključne besede: zaznava ponarejenih osebnih dokumentov, večmodalni veliki jezikovni modeli, računalniški vid, razpoznavanje vzorcev, razložljiva umetna inteligenca

Seeing Is Not Believing: Detecting Fake ID Documents With Multimodal Large Language Models

The rapid digitalisation of society has enabled remote identity verification based on photographs of identity documents, supporting a wide range of online services from banking to public administration. However, such systems remain vulnerable to document forgery and identity misuse, especially as modern image manipulation techniques become increasingly sophisticated and visually convincing. Traditional detection approaches rely on supervised models trained to recognize known artifacts and common forgery patterns. These methods struggle to generalize to novel manipulation techniques and typically provide only binary decisions without meaningful explanations, limiting trust in high-stakes applications. In this work, we investigate whether multimodal large language models (MLLMs) can address these limitations in a zero-shot setting. We present a large-scale systematic evaluation of six open-source models — DeepSeek-VL2-Tiny, LLaVA-1.6-Mistral-7B, Qwen2.5-VL-7B, Gemma-3-12B, InternVL3.5-4B, and Pixtral-12B—on the IDNet benchmark, the largest standardised dataset of authentic and forged identity documents. Without task-specific training, we analyse their ability to detect forged documents and examine performance across different fraud categories and document templates. Our results indicate that MLLMs achieve strong performance in detecting diverse digital manipulations of identity documents. Nevertheless, a

finer-grained analysis across issuing countries demonstrates that detection accuracy is significantly affected by the visual design and structural characteristics of individual documents.

Keywords: Detection of forged identity documents, Multimodal large language models, Computer vision, Pattern recognition, Explainable artificial intelligence

1 UVOD

Digitalizacija družbe zagotavlja učinkovitost in konkurenčnost gospodarstva ter podpira razvoj digitalnih storitev, ki jih lahko uporabniki opravljajo na daljavo – od odpiranja računov in sklepanja pogodb do dostopa do e-uprave. Slovenija se na tem področju umešča med države, ki sistematično spodbujajo digitalno preobrazbo. V ta namen je bila ustvarjena Strategija digitalnih javnih storitev 2030 [19], ki postavlja ambiciozne cilje, kot je digitalna dostopnost ključnih javnih storitev, in poudarja vlogo enotne digitalne identitete za varno, uporabniku prijazno in široko uporabo e-storitev. Pri tem postaja ključna tudi identifikacija osebe na daljavo, kjer uporabnik kot dokaz identitete posreduje sliko osebnega dokumenta za izvedbo storitve brez fizičnega obiska.

Hkrati digitalizacija povečuje izpostavljenost sistemov zlorabam. Ko identifikacija osebe temelji na sliki osebnega dokumenta, se odpre prostor za ponarejanje ali zlorabo identitet, kar lahko omogoča finančne prevare,

Prejet 31. marec, 2026
Odobren 22. junij, 2026



Avtorske pravice: © 2026
Creative Commons Attribution 4.0
International License

nezakonite migracije in prikrivanje identitete pri hujših kaznivih dejanjih. Na Škotskem je bil na primer nedavno aretiran moški, ki je s ponarejenimi dokumenti pridobil več kot milijon funtov [15]. V širši regiji Balkana številne analize poudarjajo, da se ponarejeni dokumenti uporabljajo za lažje prehajanje meja in podporo kriminalnim mrežam [7]. Tovrstne goljufije so torej globalen problem, ki ga pospešujejo široko-dostopni računalniški algoritmi za ustvarjanje in urejanje slikovnih vsebin.

Večina obstoječih rešitev za samodejno zaznavanje ponarejenih dokumentov v slikovni obliki temelji na algoritmih, ki se učijo razlikovati med pristnimi in ponarejenimi primeri podatkov [13], [17]. Ker se tehnike ponarejanja nenehno izboljšujejo, takšni algoritmi pogosto vidijo le omejen nabor manipulacij, ki ne odraža najnovejših napadov. Zato se v zadnjem času vse pogostejše uveljavlja pristop, pri katerem se algoritmi učijo predvsem videza pristnih dokumentov in nato morebitne manipulacije obravnavajo kot odstopanja [11]. Še vedno pa algoritmi ne znajo jezikovno utemeljiti, zakaj so določen dokument označili kot avtentičen ali ponarejen.

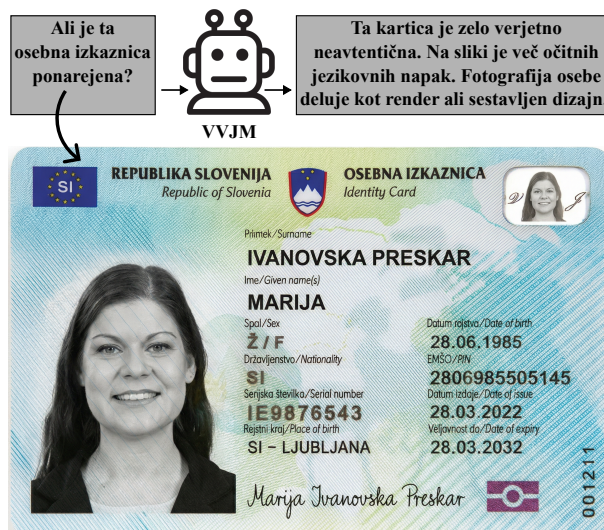
Naša študija se zgoraj omenjenih izzivov loteva z raziskovanjem večmodalnih velikih jezikovnih modelov (VVJM), ki so bili naučeni na ogromni količini slikovnih podatkov. Zato se domneva, da VVJM poznajo videz avtentičnih slikovnih vsebin, hkrati pa lahko svoje analize in odločitve podkrepijo z jezikovno razlago, kar je uporabno za forenziko in pomembno za zaupanje v sisteme (slika 1). V tem članku predstavljamo preliminarno študijo uporabe VVJM za zaznavo ponarejenih osebnih dokumentov. Naši ključni prispevki so: i) prva podrobna analiza uspešnosti šest široko-razširjenih VVJM pri zaznavanju slik ponaredkov; ii) analiza uspešnosti zaznave v odvisnosti od metode ponarejanja, ki pokaže, pri katerih vrstah manipulacij modeli uspešno zaznavajo strukturna neskladja ali odpovejo; iii) analiza uspešnosti modelov pri obdelavi izkaznic iz različnih držav, s čimer razkrijemo odvisnost uspešnosti modelov od vizualnega sloga dokumenta; iv) napotki za bodoče izboljšave sistemov, ki temeljijo na VVJM.

2 SORODNA DELA

Obstoječi algoritmi za vizualno zaznavo ponarejenih osebnih dokumentov predstavljajo konvolucijske mreže, naučene z različnimi kriterijskimi funkcijami [13].

2.1 Klasični nadzorovani algoritmi

Klasični algoritmi za zaznavanje ponaredkov največkrat izvajajo nadzorovano binarno razvrščanje slik osebnih dokumentov na avtentične in ponarejene. Pri tem se pogosto uporabljajo uveljavljene konvolucijske mreže, kot je MobileNet [5]. Z namenom izboljšanja uspešnosti zaznave Albiero idr. predlagajo dodatno primerjavo fotografije iz dokumenta z ločeno referenčno fotografijo osebe [1]. Ker so takšni algoritmi močno občutljivi za slab zajem slik, Gonzalez in Tapia



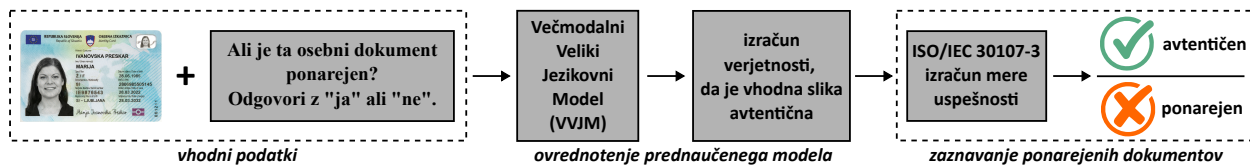
Slika 1 V tej študiji ovrednotimo uspešnost večmodalnih velikih jezikovnih modelov (VVJM) pri zaznavi ponarejenih osebnih dokumentov. V nasprotju z obstoječimi algoritmi, VVJM vrnejo oznako dokumenta skupaj z utemeljitvijo o avtentičnosti. Osebna izkaznica, prikazana na tej sliki, je bila ustvarjena z generativnim modelom Nano Banana Pro.

predlagata eksplicitno upoštevanje kakovosti slikovnih podatkov pri odločanju [6]. Kot dodatno orodje za preverjanje avtentičnosti dokumentov se vse pogostejše uporablja tudi frekvenčna analiza, saj so značilne sledi pri fizičnih napadih pogosto bolj izražene v frekvenčni kot v slikovni domeni. Li idr. tako predlagajo združevanje barvnih in frekvenčnih značilk za izboljšanje posploševanja med različnimi napravami in pogoji zajema podatkov. Chen idr. pa predstavijo popolnoma spektralno usmerjen pristop, ki manipulacije dokumentov zaznava s prilagojenim učenjem v frekvenčni domeni [3]. V scenarijih, kjer je ponarejen le manjši del dokumenta, Mudgalgundurao idr. predlagajo nadgradnjo klasičnega globalnega razvrščanja slik z razvrščanjem na ravni slikovnih elementov, kjer model poleg končne oznake poda tudi lokacijo najdenih manipulacij [9].

Čeprav so klasični nadzorovani algoritmi zelo učinkoviti v primeru obdelave nizkokakovostnih ponaredkov, njihova učinkovitost upada pri analizi sodobnejših napadov in pri obdelavi dokumentov neznanih oseb.

2.2 Šibko nadzorovani algoritmi in temeljni modeli

Razvoj nadzorovanih rešitev za slikovno zaznavo ponarejenih osebnih dokumentov zahteva dostop do velikega števila podatkov, ki so pogosto občutljivi iz vidika varnosti in zasebnosti. Sanchez idr. to težavo rešujejo s šibko nadzorovanim algoritmom, ki v učni fazi med matematičnim modeliranjem distribucije osebnih dokumentov uporabi le zelo majhno število vzorcev [14]. Markham idr. pa so razvili pristop, ki popolnoma od-



Slika 2 Visokonivojska vizualizacija postopka zaznavanja ponarejenih dokumentov z večmodalnimi velikimi jezikovnimi modeli.

pravi potrebo po resničnih podatkih, saj uporabljajo le sintetične slike pri čemer prave primere manipulacij nadomestijo s simuliranjem tipičnih napak ponaredkov [8]. Čeprav so poročani rezultati zelo dobri, avtorji opazajo, da simulacija forenzičnih sledi še ne zagotavlja robustnih detektorjev, saj se lahko modeli prekomerno prilagodijo umetno ustvarjenim napakam namesto dejanskim značilnostim napadov.

Najnovejše raziskave na področju zaznavanja biometričnih manipulacij vse pogostejše gradijo na temeljnih modelih [11]. Temeljni modeli so naučeni na ogromnih količinah slikovnih podatkov in zato že v osnovi zajemajo bogate, splošno uporabne vizualne značilke. Chettaoui idr. so pokazali, da so temeljni modeli zelo uspešni pri razpoznavanju obrazov [4]. Avtorji študij [12] in [2] pa so uspešno prilagodili tovrstne modele na področju zaznavanja obraznih manipulacij. Zaznavanje ponarejenih dokumentov s temeljnimi modeli je bilo do zdaj predstavljeno samo v študiji [16]. Preliminarni eksperimenti so v tem primeru pokazali izjemen potencial posploševanja detekcije na nove vrste slikovnih manipulacij, vendar so temeljni algoritmi še vedno premalo raziskani in obstaja prostor za znatne izboljšave. Večmodalni veliki jezikovni modeli so bili kot posebna skupina temeljnih modelov delno preizkušeni pri zaznavanju različnih obraznih manipulacij [10], a podobnih študij, ki obravnavajo zaznavanje slik ponarejenih dokumentov, do danes še ni bilo objavljenih.

Povzetek literature izpostavi tri pomembne pomanjkljivosti obstoječih študij: i) klasične metode za zaznavanje ponarejenih dokumentov so naučene na omejenih zbirkah vnaprej definiranih tipov manipulacij in pogosto ne zaznajo novih vrst ponaredkov; ii) raziskave temeljnih modelov v tem kontekstu so še vedno redke in metodološko neenotne; ter iii) obstoječi sistemi ne zagotavljajo jezikovno razumljive razlage zaznanih nepravilnosti v dokumentih, kar omejuje njihovo uporabnost v forenzičnih okoljih. V tej študiji obravnavamo našete pomanjkljivosti s sistematično analizo zmogljivosti sodobnih večmodalnih velikih jezikovnih modelov.

3 METODOLOGIJA IN EKSPERIMENTALNI PROTOKOL

V tem poglavju opišemo metodologijo dela, uporabljene eksperimentalne podatke ter mere vrednotenja uspešnosti zaznave slik ponarejenih dokumentov.

3.1 Metodologija

V študiji nalogo zaznavanja ponarejenih osebni dokumentov obravnavamo kot problem vizualnega sklepanja brez dodatnega učenja, kjer na podlagi ene same vhodne slike dokumenta večmodalni veliki jezikovni model (VVJM) presodi o njegovi avtentičnosti. Cilj raziskave je preveriti, ali VVJM, ki so prednaučeni na velikih količinah podatkov, omogočajo zaznavo strukturnih in semantičnih nepravilnosti ponarejenih dokumentov.

Naj bo p tekstovni poziv, ki od dotičnega VVJM zahteva oceno avtentičnosti slike brez dodatnih razlag:

Opravljaš forenzično analizo slike osebnega dokumenta. Ugotovi, ali slika vsebuje sledove digitalne manipulacije. Preveri znake, ki kažejo na zlivanje obrazov, zamenjavo obraza, podvojene ali kopirane dele dokumentov, spremenjeno ali prepisano besedilo, retuširanje, nenaravne meje ali napake zlivanja. Odgovori z ja ali ne. Ja = manipulacija, ne = avtentičen dokument. Odgovor:

Naj bo $\mathcal{X}$ množica slik osebni dokumentov, $\mathcal{Y} = \{ja, ne\}$ pa množica binarnih oznak, kjer se oznaka ja uporablja za označevanje ponaredkov, oznaka ne pa za označevanje avtentičnih vzorcev. Možnost, da določen VVJM označi dotično sliko x z oznako y , definiramo v obliki pogojne verjetnosti

$$P_{\theta}(y | x, p), \quad (1)$$

kjer so θ označeni parametri VVJM, s katerimi analiziramo vhodno sliko. Končna odločitev modela je definirana s pravilom:

$$\hat{y} = \arg \max_{y \in \mathcal{Y}} P_{\theta}(y | x, p), \quad (2)$$

kar pomeni, da se sliki x priredi tista oznaka y , ki ima največjo pogojno verjetnost glede na poziv p .

Poziv p , ki ga uporabljamo v tej študiji, je zasnovan kot forenzično usmerjen poziv, ki model eksplicitno usmeri k preverjanju značilnih vrst manipulacij osebni dokumentov, kot so zlivanje ali zamenjava obraza, podvajanje oziroma kopiranje delov dokumenta, spreminjanje besedila, retuširanje ter nenaravne meje in napake zlivanja. Poziv pri tem ne predpisuje konkretnih nizkonivojskih vizualnih značilk ali determinističnih pravil za zaznavanje posamezne vrste manipulacije, temveč modelu prepusti, da na podlagi svojega vizualno-jezikovnega sklepanja oceni prisotnost navedenih forenzičnih znakov in poda binarno odločitev o

avtentičnosti dokumenta. Visokonivojski pregled predlaganega postopka zaznavanja ponarejenih dokumentov je prikazan na sliki 2.

V naši raziskavi ovrednotimo šest VVJM, in sicer: InternVL (InternVL3.5-4B), DeepSeek (DeepSeek-VL2-Tiny), Qwen (Qwen2.5-VL-7B), Gemma (Gemma3-12B), LLaVA (LLaVA1.6-Mistral-7B), ter Pixtral (Pixtral-12B). Izbrani modeli predstavljajo različne nevronske arhitekture, ki so bile naučene z različnimi kriterijskimi funkcijami. Vsi modeli in njihovi prednaučeni parametri so bili pridobljeni iz repozitorija Hugging Face Model Hub*. Modeli so bili ovrednoteni v angleščini, brez dodatnega učenja ali prilagajanja algoritmov. Deterministično delovanje modelov in ponovljivost rezultatov sta bila zagotovljena z ovrednotenjem brez stohastičnega vzorčenja med ustvarjanjem odgovorov VVJM. Eksperimenti so bili izvedeni na eni grafični kartici NVIDIA A100-SXM4 z 80 GB pomnilnika. Vrednotenje modelov je bilo avtomatizirano s programskim jezikom Python 3.10 in knjižnicami PyTorch 2.0 in CUDA 11.8.

3.2 Podatkovna baza

Vsi eksperimenti so bili izvedeni z uporabo baze IDNet [18], ki je trenutno najboljše javno dostopna zbirka osebnih dokumentov za zaznavanje goljufij. Zbirka skupno vsebuje 837 060 slik. Polovica vzorcev predstavlja osebne izkaznice in vozniška dovoljenja iz 10 zveznih držav ZDA, in sicer iz Arizone (AZ), Kalifornije (CA), Nevade (NV), Severne Karoline (NC), Pensilvanije (PA), Južne Dakote (SD), Utah (UT), Washingtona D. C. (DC), Zahodne Virginije (WV) in Wisconsin (WI). Druga polovica vzorcev predstavlja potne liste in osebne izkaznice iz 10 evropskih držav, in sicer iz Albanije (ALB), Azerbajdžana (AZE), Estonije (EST), Finske (FIN), Grčije (GRC), Latvije (LVA), Rusije (RUS), Srbije (SRB), Slovaške (SVK) in Španije (ESP). Število slik je uravnoteženo tako, da je za vsako od 20 držav na voljo 5 979 avtentičnih dokumentov. Z vsakim avtentičnim dokumentom pa je bilo z uporabo naslednjih metod vizualne manipulacije generiranih šest ponarejenih različic:

- *kopiraj in prestavi (KP)*: naključno izbran izsek iz enega dokumenta se prenese na drug dokument;
- *zlivanje obrazov (ZO)*: obraza dveh oseb se združita v enega, kar simulira napade z zlivanjem;
- *zamenjava obraza (Z)*: zamenjava originalne osebne fotografije s fotografijo druge osebe;
- *izbriši in napiši (IN)*: izbrana beseda se odstrani z rekonstrukcijo ozadja z generativnim modelom, nato pa se na njenem mestu napiše nova beseda;
- *izreži in zamenjaj (IZ)*: izbrana beseda se izreže, na njeno mesto pa se vstavi nova beseda ali oblikovno spremenjenja različica stare (spremeni se njena pisava, velikost ali kontrast slikovnih elementov);
- *kombinacija (K)*: več različnih manipulacij.

Izbrani vzorci ponarejenih osebnih dokumentov iz baze IDNet so prikazani na sliki 3.

3.3 Mere vrednotenja

Uspešnost zaznavanja ponarejenih osebnih dokumentov z izbranim naborom VVJM je bila ovrednotena v skladu z mednarodnim standardom ISO/IEC 30107-3[†]. Kot glavno mero kvantitativnega ovrednotenja smo uporabili EER (angl. Equal Error Rate), ki predstavlja napako detekcije, pri kateri je stopnja napačne klasifikacije ponaredkov enaka stopnji napačne klasifikacije avtentičnih dokumentov. Pri izračunu mere EER smo uporabljali odločitveno verjetnost $P_{\theta}(y \in \mathcal{S} \mid x, p)$ (Enačba 1).

4 REZULTATI

V tem poglavju najprej predstavimo splošno uspešnost zaznavanja ponarejenih izkaznic ter uspešnost v odvisnosti od države, ki je dokument izdala. V nadaljevanju naredimo analizo uspešnosti glede na vrsto digitalne manipulacije. Poglavje se zaključi s kvalitativno analizo, kjer raziskujemo interpretabilnost odločitev modelov.

4.1 Splošna uspešnost modelov

Tabela 1 prikazuje povprečno napako EER (%) zaznavanja ponarejenih osebnih dokumentov po modelih in po posameznih državah. Skupna povprečna napaka EER (11.98%) kaže, da imajo prednaučeni VVJM zelo dober potencial za detekcijo ponarejenih osebnih dokumentov. Podrobnejši rezultati pa razkrijejo, da je uspešnost zaznavanja močno odvisna tako od arhitekture uporabljenega modela kot tudi od videza dokumenta. Glede na povprečno napako EER v vseh državah je najuspešnejši model DeepSeek z 5.68%, tesno za njim je Qwen z 6.13%, sledita Gemma z 6.34% in Pixtral z 7.21%. Model InternVL dosega nekoliko slabše rezultate (10.59%), medtem ko LLaVA izrazito odstopa z bistveno višjo povprečno napako EER (35.92%), kar nakazuje, da je najmanj primeren za zaznavanje ponaredkov z našim predlaganim pristopom. Relativno majhne razlike v napaki med najboljšimi tremi modeli, DeepSeek (z 1 milijardo parametrov), Qwen (s 7 milijardi parametrov) in Gemma (z 12 milijardi parametrov), kažejo, da velikost modela (število parametrov) sama po sebi ni edini odločilni dejavnik. Pomembno vlogo imajo njegova arhitektura, vizualni kodirnik, način učenja algoritma, ter predvsem sestava učne množice podatkov, ki je različna pri posameznih modelih.

4.2 Analiza uspešnosti po državah

Rezultati po posameznih državah razkrivajo izrazito variabilnost uspešnosti. V nasprotju z Gemmo in LLaVo, ki približno enako uspešno zaznavata ponarejene evropske in ameriške dokumente, sta Qwen in Pixtral precej uspešnejša pri zaznavanju evropskih ponaredkov, pri katerih dosežeta približno 60% nižjo povprečno napako

* <https://huggingface.co/docs/hub/en/models-the-hub>

[†] <https://www.iso.org/standard/67381.html>



Slika 3 Primeri avtentične in ponarejenih osebnih izkaznic iz baze IDNet [18]. Manipulacije so označene z rdečim okvirjem.

EER. Prav nasprotno pa DeepSeek in InternVL dosežeta vsaj 30% nižjo povprečno napako EER pri zaznavanju ameriških dokumentov v primerjavi z evropskimi.

Pri dokumentih nekaterih držav, kot so Albanija (ALB), Azerbajdžan (AZE) in Finska (FIN), ter večine držav iz ZDA dosejajo najboljši modeli skoraj ničelno napako EER, kar pomeni, da so manipulacije na teh dokumentih za modele razmeroma enostavno zaznavne. Zlasti izstopa model Pixtral, ki uspešno zazna vse ponarejene dokumente iz Albanije (ALB), Azerbajdžana (AZE) in Rusije (RUS), ter model DeepSeek, ki z napako EER 0% zazna vse ponaredke iz Kalifornije (CA). Vseeno pa opazimo, da so napake EER relativno visoke pri določenih državah, kot je Wisconsin (WI). To nakazuje, da imajo vizualne značilnosti dokumentov posameznih držav (razporeditev besedilnih polj, kompleksnost ozadja, tipografija, barvne sheme) pomemben vpliv na zaznavo manipulacij.

4.3 Analiza uspešnosti glede na vrsto manipulacije

Uspešnost zaznavanja različnih vrst digitalnih manipulacij osebnih dokumentov smo analizirali s pomočjo relativne napake EER, ki jo izračunamo tako, da iz napake EER zaznavanja določenega tipa manipulacije odštejemo napako EER, povprečeno po vseh tipih manipulacij. Relativna napaka EER posameznih VVJM je prikazana na sliki 4. Relativna napaka torej predstavlja odstopanje napake EER posameznega modela pri določenih vrsti napada glede na njegovo skupno povprečje. Negativne vrednosti pomenijo, da model to vrsto manipulacije zaznava bolje kot v povprečju, pozitivne vrednosti pa, da je zaznava slabša od povprečja. Iz grafa je razvidno, da modeli v povprečju najbolje zaznavajo kombinacijo več vrst manipulacij (K) ter manipulacije z zamenjavo obraza (Z). Nasprotno pa so modeli nekoliko manj uspešni pri manipulacijah tipa izbriši in napiši

(IN) ter izreži in zamenjaj (IZ). Tovrstne manipulacije pogosto vključujejo digitalno obdelavo dokumentov s sodobnimi generativnimi modeli, ki naredijo lokalne tekstovne ali kompozicijske spremembe, pri čemer za sabo pustijo le minimalne, na prvi pogled neopazne sledi.

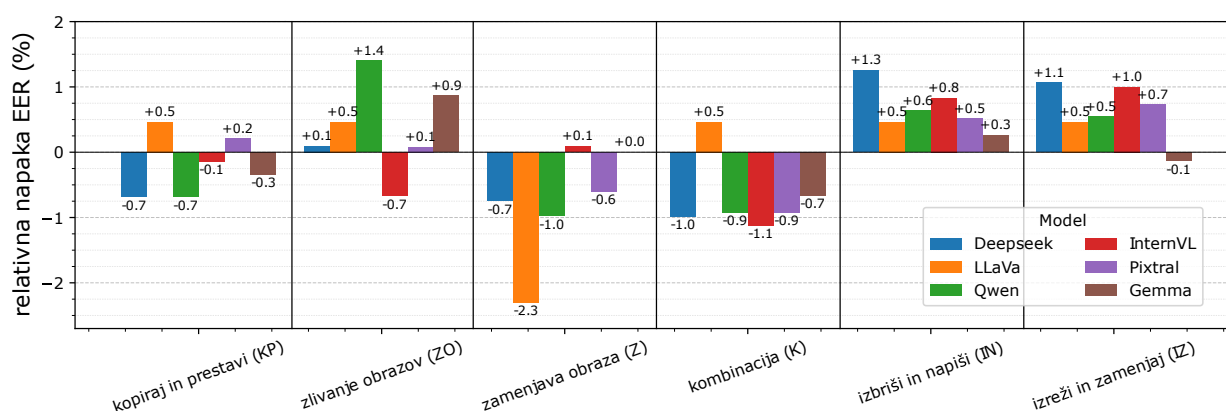
V vsakem primeru je pomembno poudariti, da je variabilnost napake EER med različnimi vrstami manipulacij razmeroma majhna, saj je v večini primerov znotraj približno $\pm 1\%$ relativne EER. To nakazuje, da modeli približno enako dobro in neodvisno od konteksta manipulacije zaznavajo splošne slikovne nepravilnosti.

4.4 Kvalitativna analiza rezultatov

Za boljše razumevanje delovanja VVJM smo izvedli kvalitativno analizo izbranih primerov osebnih dokumentov, ki so bili s strani VVJM pravilno označeni kot avtentični ali ponarejeni (slika 5). Analiza je bila izvedena z modelom, ki doseže najmanjšo povprečno napako EER za vse države (DeepSeek). Pri tem smo uporabili tekstovni poziv, ki je poleg oznake dokumenta od modela zahteval tudi utemeljitev. Ugotovili smo, da pri analizi avtentičnih dokumentov model pravilno izpostavi značilnosti, ki so tipične za uradne dokumente: pravilno poravnano besedilo, usklajen slog pisave ter poenoteno barvno shemo. Pri zaznavanju ponarejenih dokumentov pa je najpogostejše izpostavil splošne artefakte, ki jih postopek manipulacije pusti za sabo. Model po drugi strani ne izpostavi semantične nekonsistentnosti. Model na primer ne izpostavi, da je na vozniškem dovoljenju (1. dokument v 2. vrsti slike 5) fotografija otroka, ki bi glede na naveden datum rojstva moral biti ob izdaji dokumenta star 28 let. Iz razlag je torej razvidno, da model ocenjuje dokument na osnovi globalne vizualne konsistentnosti in tipografskega ujemanja.

Tabela 1 Povprečna napaka EER (%) zaznavanja ponarejenih osebnih izkaznic po državah. (najboljši rezultat, drugi najboljši)

Država		DeepSeek	Qwen	Gemma	Pixtral	InternVL	LLaVA	Povprečje
Evropa (E)	ALB	0.93	0.67	7.63	0.00	10.20	59.60	13.17
	AZE	0.77	0.17	18.43	0.00	4.13	87.20	18.45
	ESP	4.53	0.80	6.23	19.17	34.00	7.80	12.09
	EST	0.17	1.43	3.07	2.07	5.80	13.20	4.29
	FIN	5.53	0.80	1.60	0.17	15.00	3.20	4.38
	GRC	12.87	2.60	9.10	0.40	20.67	7.40	8.84
	LVA	34.83	5.60	4.37	0.33	2.13	43.20	15.08
	RUS	2.90	10.27	7.43	0.00	15.70	46.20	13.75
	SRB	0.27	2.40	7.32	1.37	0.80	53.60	10.96
	SVK	14.83	3.70	1.20	7.70	13.20	33.40	12.34
Združene države (ZDA)	AZ	24.90	1.13	0.83	1.40	1.27	56.60	14.36
	CA	0.00	8.40	7.40	2.00	22.80	33.40	12.33
	DC	0.30	19.60	2.77	21.80	15.80	21.80	13.68
	NC	0.50	10.80	8.87	12.50	5.67	9.40	7.96
	NV	1.53	10.93	8.43	35.20	5.13	70.00	21.87
	PA	0.80	24.73	7.50	20.80	15.50	67.40	22.79
	SD	0.97	5.63	1.47	3.93	6.80	2.40	3.53
	UT	0.20	2.77	7.13	2.20	13.87	67.00	15.53
	WI	6.50	5.20	12.90	6.60	3.10	8.20	7.08
	WV	0.17	4.90	3.07	6.47	0.20	27.40	7.04
E povprečje		7.76	2.84	6.64	3.12	12.16	35.48	11.34
ZDA povprečje		3.59	9.41	6.04	11.29	9.01	36.36	12.62
Skupno povprečje		5.68	6.13	6.34	7.21	10.59	35.92	11.98



Slika 4 Prikaz relativne EER napake (%) posameznih VVJM glede na vrsto digitalne manipulacije, kjer negativne vrednosti pomenijo boljše zaznavanje ponaredkov od povprečne uspešnosti modela, pozitivne pa slabše.

5 ZAKLJUČEK

V tej študiji smo predstavili prvo sistematično analizo večmodalnih velikih jezikovnih modelov (VVJM) z namenom zaznavanja ponarejenih osebnih dokumentov brez dodatnega učenja algoritmov. Rezultati na podatkovni bazi IDNet kažejo, da VVJM dosegajo prese- netljivo dobre rezultate zaznave ponaredkov, pri čemer so razlike med najboljšimi modeli Deepseek, Qwen in Gemma relativno majhne. To potrjuje, da velikost modela sama po sebi ni odločilni dejavnik uspešnosti,

temveč pomembno vlogo igrajo arhitektura, vizualni ko- dirnik ter način učenja modela in sestava učne množice podatkov.

Podrobnejša analiza po državah, ki so osebne doku- mente izdale, razkrije, da je uspešnost zaznave močno odvisna od vizualnega sloga dokumenta. Analiza rezul- tatov v odvisnosti od vrste digitalnih manipulacij pa pokaže, da modeli nekoliko boljše zaznavajo ponarejene dokumente, ki vsebujejo več vrst manipulacij ali zame- njavo obraza. Zaznava subtilnih sprememb napisov in lokalnih kompozicijskih sprememb je zahtevnejša.



Slika 5 Prikaz izbranih primerov, ki so bili z modelom Deepseek pravilno označeni kot avtentičen ali ponarejen osebni dokument. Pravilna oznaka oz. vrsta vsebovane manipulacije je nad sliko dokumenta. Pod vsako sliko pa je utemeljitev modela.

Kvalitativna analiza razkrije, da modeli svoje odločitve večinoma utemeljujejo z zaznavo globalne vizualne (ne)konsistentnosti, tipografskih neskladij in nenaravnega zliivanja elementov. Hkrati pa pogosto ne zaznajo semantičnih nekonistentnosti (npr. neskladja med starostjo osebe na sliki in starostjo izračunano na podlagi datuma rojstva). Kljub temu je zmožnost generiranja jezikovne razlage pomembna prednost VVJM v primerjavi s klasičnimi algoritmi, saj omogoča večjo transparentnost in potencialno večje zaupanje v odločitve sistema.

Na podlagi predstavljenih rezultatov lahko zaključimo, da imajo VVJM v ničelnem učnem režimu velik potencial za zaznavo digitalnih manipulacij osebnih dokumentov, vendar še ne dosegajo stopnje robustnosti, ki bi omogočala njihovo neposredno uporabo v varnostno kritičnih okoljih. Prihodnje raziskave bodo usmerjene v dodatno učenje VVJM z namenom izboljšanja občutljivosti za možne digitalne manipulacije slik osebnih dokumentov.

ZAHVALA

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

LITERATURA

- [1] V. Albiero, N. Srinivas, E. Villalobos, J. Perez-Facuse, R. Rosenthal, D. Mery, K. Ricanek, and K. W. Bowyer. Identity Document to Selfie Face Matching Across Adolescence. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2020.
- [2] E. Caldeira, G. Ozgur, T. Chettaoui, M. Ivanovska, P. Peer, F. Boutros, V. Struc, and N. Damer. MADation: Face Morphing Attack Detection with Foundation Models. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1565–1575, 2025.
- [3] C. Chen, Y. Li, B. Li, W. Yu, B. Chen, B. Li, and J. Huang. Moiré Spectral Augmentation and Masked Frequency Modeling for Document Presentation Attack Detection. *IEEE Transactions on Dependable and Secure Computing*, 22(05):5366–5381, 2025.
- [4] T. Chettaoui, N. Damer, and F. Boutros. FRoundation: Are foundation models ready for face recognition? *Image and Vision Computing*, 156:105453, 2025.
- [5] S. Gonzalez and J. E. Tapia. Forged presentation attack detection for ID cards on remote verification systems. *Pattern Recognition*, 162:111352, 2025.
- [6] S. González and J. Tapia. Towards Refining ID Cards Presentation Attack Detection Systems using Face Quality Index. In *30th European Signal Processing Conference (EUSIPCO)*, pages 1027–1031, 2022.
- [7] T. G. Initiative. Fraudulent or illegally obtained Balkans documents are permitting ease of travel for criminals worldwide. https://riskbulletins.globalinitiative.net/see-obs-014/05-fraudulent-or-illegally-obtained-balkans-documents.html?utm_source=chatgpt.com, 2023. Dostopano: 21.2.2026.
- [8] R. P. Markham, J. M. E. López, M. Nieto-Hidalgo, and J. E. Tapia. Open-Set: ID Card Presentation Attack Detection Using Neural Style Transfer. *IEEE Access*, 12:68573–68585, 2024.
- [9] R. Mudgalgundurao, P. Schuch, K. Raja, R. Ramachandra, and N. Damer. Pixel-wise supervision for presentation attack detection on identity document cards. *IET Biometrics*, 11:383–395, 2022.

- 2022.
- [10] K. Narayan, V. Vibashan, and V. M. Patel. FaceXBench: Evaluating Multimodal LLMs on Face Understanding. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2026.
 - [11] H. Otrosi Shahreza and S. Marcel. Foundation models and biometrics: A survey and outlook. *IEEE Transactions on Information Forensics and Security*, 20:9113–9138, 2025.
 - [12] G. Ozgur, E. Caldeira, T. Chettaoui, F. Boutros, R. Ramachandra, and N. Damer. FoundPAD: Foundation Models Reloaded for Face Presentation Attack Detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 697–707, 2025.
 - [13] E. M. Ruiz, J. E. Tapia, R. T. Soto, and C. Busch. Identity card presentation attack detection: A systematic review, 2025.
 - [14] A. Sanchez, J. M. Espín, and J. E. Tapia. Few-Shot Learning: Expanding ID Cards Presentation Attack Detection to Unknown ID Countries. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2024.
 - [15] T. S. Sun. Crook who raked in £1M from fake ID scheme while on run from cops for 14 years snared during police stop in Scots town. https://www.thescottishsun.co.uk/news/15888595/man-jailed-1-million-identity-fraud/?utm_source=chatgpt.com, 2026. Dostopano: 21.2.2026.
 - [16] J. E. Tapia and C. Busch. Can foundation models generalise the presentation attack detection capabilities on id cards?, 2025.
 - [17] J. E. Tapia, N. Damer, C. Busch, J. M. Espin, J. Barrachina, A. S. Rocamora, K. Ocvirk, L. Alessio, B. Batagelj, S. Patwardhan, R. Ramachandra, R. Mudgalgundurao, K. Raja, D. Schulz, and C. Aravena. First competition on presentation attack detection on id card. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2024.
 - [18] L. Xie, Y. Wang, H. Guan, S. Nag, R. Goel, N. Swamy, Y. Yang, C. Xiao, J. Prisby, R. Maciejewski, and J. Zou. IDNet: A Novel Identity Document Dataset via Few-Shot and Quality-Driven Synthetic Data Generation. In *IEEE International Conference on Big Data (BigData)*, pages 2244–2253, 2024.
 - [19] M. za javno upravo. Strategija digitalnih javnih storitev 2030. https://www.gov.si/assets/ministrstva/MDP/Dokumenti/Strategija_digitalnih_javnih_storitev_2030-1.pdf, 2023. Dostopano: 21.2.2026.

Marija Ivanovska je asistentka 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 univerzi Johns Hopkins University v ZDA, leta 2024 pa gostujoča asistentka na univerzi 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, IEEE WACV, IAPR/IEEE IJCB, IEEE FG. 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.