Sveučilište u Rijeci
University of Rijeka
https://www.uniri.hr

Polytechnic: Journal of Technology Education, Volume 9, Number 2 (2025)
Politehnika: Časopis za tehnički odgoj i obrazovanje, Svezak 9, Broj 2 (2025)

Politehnika
Polytechnica
https://politehnika.uniri.hr
cte@uniri.hr

DOI: https://doi.org/10.36978/cte.9.2.1

Izvorni znanstveni članak
Original scientific paper

AI u veterinarskoj oftalmologiji: segmentacija slika i veliki jezični modeli u dijagnostici očnih bolesti pasa

Matija Burić

Sveučilište u Rijeci, Fakultet informatike i digitalnih tehnologija

Radmile Matejčić 2, 51000 Rijeka

Matija.buric@uniri.hr

Sažetak

Istraživanje prikazuje primjenu metoda umjetne inteligencije u veterinarskoj oftalmologiji kroz kombinirani pristup računalnog vida i jezične analize. Korišten je prilagođeni U-Net model za detekciju i segmentaciju simptoma očnih bolesti pasa, uključujući zamućenje oka, crvenilo bjeloočnice, pretjerano suzenje i obojeno očno ispupčenje. Dobiveni rezultati poslužili su kao ulazni podaci za velike jezične modele (GPT-4o, Mistral 7B, Gemini2, Llama-3 i Claude 4) s ciljem interpretacije simptoma i davanja preliminarne dijagnoze. Evaluacija je provedena pomoću lingvističkih i semantičkih metrika (MPNet, MiniLM, BERTScore, CLIPScore, BLEU, METEOR, ROUGE i SPICE). Rezultati pokazuju da integracija U-Net segmentacije i analitičkih sposobnosti velikih jezičnih modela (engl. Large Language Model, LLM) omogućuje učinkovitu preliminarnu dijagnozu očnih bolesti pasa. Model s okosnicom ResNet34 pokazao je najveću točnost u prepoznavanju crvenila bjeloočnice, dok je GPT-4o bio najuspješniji u interpretaciji simptoma i postavljanju dijagnoze. Ovaj pristup doprinosi razvoju sustava koji mogu povećati točnost i učinkovitost veterinarske dijagnostike.

Ključne riječi: *veterinarska oftalmologija; računalni vid; U-Net; veliki jezični modeli; dijagnostika pasa.*

1 Uvod

Rana i točna dijagnoza očnih bolesti kod pasa, odnosno kućnih ljubimaca općenito, ključna je za učinkovito liječenje, sprječavanje mogućeg gubitka vida i poboljšanje kvalitete života zahvaćenih životinja. Trenutni napredak u tehnologijama strojnog učenja i obrade slike značajno je unaprijedio dijagnostičke mogućnosti u medicinskim područjima, uključujući i veterinarsku medicinu. Cilj je postići potpunu dijagnostičku procjenu koja može pomoći vlasnicima kućnih ljubimaca da razumiju specifične zdravstvene probleme svojih ljubimaca, a istodobno olakšati veterinarima proces dijagnosticiranja. U ovom radu istraživanje je usmjereno na očne bolesti kod pasa.

Problem automatske dijagnoze očnih bolesti kod pasa je složen. Uz klasične razloge koji povećavaju složenost treniranja modela strojnog učenja poput male površine od interesa koja ponekad zauzima samo nekoliko piksela, različitog položaja kamere i udaljenosti od objekta, zamućenja slike uslijed pokreta, promjena u osvjetljenju, složene pozadine i zaklona, postoje i dodatni izazovi povezani sa značajnim razlikama među pasminama pasa. Nije manje važna činjenica da prikupljanje slika nije jednostavno, jer nije lako ni vlasniku smiriti bolesnog psa i postaviti ga tako da se bolesno oko može snimiti na način da svi simptomi budu jasno vidljivi. Slika 1 prikazuje istu dijagnozu očne bolesti s ružičastim ispupčenjem oka iz unutarnjeg očnog kuta i simptomima pretjeranog suzenja koji upućuju na

Cherry Eye kod različitih pasa, te ilustrira neke od izazova s kojima se moraju nositi modeli za automatsku segmentaciju i interpretaciju očnih bolesti.

Slika 1. Psi različitih pasmina s ispupčenjem ružičaste mase iz medijalnog očnog kuta oba oka, uz prisutne upalne znakove koji uključuju bol i suzenje, oboljeli od „Cherry Eye"(Thamizharasan et al., 2016; Tripathi et al., 2014)

Ovaj rad predstavlja istraživanje s dvostrukim fokusom. Za automatsku oftalmološku dijagnozu bolesti kod pasa i semantičku segmentaciju simptoma korišten je specijalizirani U-Net model (Ronneberger et al., 2015) s prilagođenim okvirom, treniran na posebno izrađenom skupu podataka koji su sastavili stručnjaci iz područja veterinarske oftalmologije. U-Net-ove mogućnosti precizne segmentacije slika kombinirane su sa sposobnostima generiranja i razumijevanja teksta velikih jezičnih modela (LLM-ova) radi poboljšanja točnosti dijagnoze i pomoći u tumačenju medicinskih simptoma. Analiziran je potencijal različitih LLM-ova, kao što su GPT-4o (Savage et al., 2024), Mistral 7B (Jiang et al., 2023), Gemini2 (Gemini Team et al., 2024), Llama-3-3 (Touvron et al., 2023) i Claude 4 (PBC, 2024), za interpretaciju medicinskih simptoma, uzimajući u obzir da svaki LLM ima specifične prednosti u generiranju teksta. GPT-4o se ističe u obradi složenih medicinskih podataka kroz konverzacijsku fluentnost, Mistral 7B je prilagođen rješavanju specifičnih upita (engl. prompt), Gemini2 je učinkovit u dinamičnim interakcijama, Llama-3 se ističe u multimodalnom razumijevanju, dok Claude 4 nudi uravnotežen pristup analizi podataka.

Za usporedbu rezultata dobivenih različitim LLM modelima u pogledu točnosti dijagnoze u medicinskom kontekstu korištene su standardne evaluacijske metrike za lingvističke i semantičke modele. One uključuju unaprijed istrenirane modele MPNet (Song et al., 2020) i MiniLM(Wang et al., 2020), te evaluacijske metrike BERTScore (Zhang et al., 2019), CLIPScore (Hessel et al., 2021), BLEU(Papineni et al., 2001), METEOR(Denkowski & Lavie, 2014), ROUGE (Ganesan, 2018) i SPICE(Anderson et al., 2016), koje su odabrane kako bi se pravilno procijenile razlike u jezičnom razumijevanju potrebne za točnu medicinsku dijagnozu.

Prvi sljedeći odjeljak rada daje pregled postojećih metodologija i tehnologija relevantnih za dijagnozu očnih bolesti kod pasa. Odjeljak nakon, nazvan *„Ekstrakcija simptoma oftalmoloških bolesti pasa korištenjem U-Net modela"*, daje detaljan opis U-Net arhitekture, pripreme skupa podataka i evaluacijskih metrika za segmentaciju slike. Odjeljak *„Interpretacija simptoma pomoću LLM-ova"* objašnjava integraciju LLM-ova za razumijevanje simptoma i metrike korištene za njihovu evaluaciju. Odjeljak *„Rezultati i rasprava"* prikazuje analizu rezultata istraživanja. Na kraju, rad se zaključuje sažetkom doprinosa istraživanja i smjernicama za budući rad.

2 Pregled povezanih istraživanja

Područje dijagnosticiranja očnih bolesti kod pasa značajno se razvilo s integracijom konvolucijskih neuronskih mreža (CNN), koje se ističu u zadacima prepoznavanja slika. Arhitektura U-Net, mreža tipa autoenkoder-dekoder, poznata je po učinkovitoj ekstrakciji značajki i mogućnostima segmentacije koje su ključne za analizu medicinskih slika. Naknadni napreci uključuju integraciju naprednih CNN okosnica poput ResNet (K. He et al., 2015), EfficientNet (Tan & Le, 2020), VGG (Simonyan & Zisserman, 2015) i Inception (Szegedy et al., 2015), koje su poboljšale lokalizaciju kliničkih simptoma u različitim medicinskim kontekstima (Huang et al., 2023; Sreng et al., 2020). Nedavna istraživanja i dalje potvrđuju konkurentne performanse modela temeljenih na U-Netu, naglašavajući njegovu trajnu važnost u medicinskoj analizi slika (Azad et al., 2022; Petit et al., 2021). Nadalje, usporedbe između U-Neta i arhitektura temeljenih na transformerima za medicinsku registraciju slika pokazale su prilagodljivost i konkurentnost U-Neta (Azad et al., 2022).

Istraživanja očnih bolesti kod pasa i dalje su ograničena, s naglaskom na bolesti poput glaukoma (Boevé & Stades, 1985; Johnsen et al., 2006; Strom et al., 2011), što ukazuje na oskudnost i nedostupnost sveobuhvatnih skupova podataka. Iako postoji potencijal za stvaranje skupova podataka pomoću sintetskih slika, kao što je to primijenjeno u drugim područjima (Burić, Paulin, et al., n.d.; Dandekar et al., 2018), ovaj pristup trenutačno je nepraktičan zbog činjenice da se istraživanja još uvijek provode na stvarnim životinjama (Deane et al., 2021).

Istodobno, procjena učinkovitosti velikih jezičnih modela (LLM-ova) u zdravstvenim primjenama još je ograničena zbog njihove rane faze razvoja. Postojeće metrike za evaluaciju izvedbe LLM-ova često ne obuhvaćaju složenost potrebnu za medicinske primjene i dalje zahtijevaju stručno znanje za određivanje stvarne točnosti (Thirunavukarasu et al.,

2024). Postoje i istraživanja koja ocjenjuju rezultate LLM-ova korištenjem medicinske terminologije (Z. He et al., 2024; Katic et al., 2021), no u okviru ovog istraživanja testirat će se njihova učinkovitost u području veterinarske oftalmologije kritičkom evaluacijom sličnom onoj koju su proveli (González-Chávez et al., 2023; Hrga & Ivasic-Kos, 2024).

Metode prethodnog treniranja, koje pokazuju obećavajuće rezultate, poput MPNet-a (Bucur, 2023) i MiniML-a (Sasazawa et al., 2023), također su analizirane.

3 Ekstrakcija simptoma oftalmoloških bolesti pasa korištenjem U-Net modela

Za poboljšanu klasifikaciju i semantičku segmentaciju bolesti, podaci su grupirani na temelju najčešćih simptoma - zamućenje oka, crvenilo bjeloočnice, pretjerano suzenje i obojeno očno ispupčenje. Ova strategija grupiranja usmjerava model na ključne vizualne simptome koji upućuju na više različitih poremećaja. Na primjer, ako su prisutni i pretjerano suzenje i obojeno očno ispupčenje, bolest bi mogla biti *Cherry Eye*, dok prisutnost samo pretjeranog suzenja može ukazivati na *začepljeni suzni kanal*. Ovakav pristup omogućuje modelu bolju generalizaciju među različitim pasminama pasa i očnim patologijama.

3.1 Priprema skupa podataka

Skup podataka DogEyeSeg4 (Burić, Ivašić-Kos, et al., n.d.) prikupljen u koordinaciji s autorom publikacije Atlas bolesti očiju pasa i mačaka (Grozdanić et al., 2020), posebno je izrađen za ovo istraživanje i namijenjen je treniranju i testiranju modela za semantičku segmentaciju slika s četiri medicinska simptoma oka. Sastoji se od 145 slika, svaka dimenzija 320x320 piksela. Slike su prikupljene iz specijaliziranih veterinarskih oftalmoloških klinika, a pregledali su ih stručnjaci iz područja veterinarske medicine (Animal Eye Consultants of Iowa). Skup obuhvaća širok raspon pasmina i manifestacija bolesti. Svaka slika sadrži jedinstvenu masku u boji koja označava četiri ključna simptoma ovisno o njihovoj prisutnosti: zamućenje oka, crvenilo bjeloočnice, pretjerano suzenje i obojeno očno ispupčenje. Zastupljenost različitih klasa održana je proporcionalno kroz cijeli skup.

Kako bi se povećala varijabilnost i poboljšala sposobnost generalizacije modela, primijenjene su tehnike augmentacije podataka, uključujući horizontalna zrcaljenja, rotacije i translacije. Zumiranje nije korišteno kako bi se sačuvao integritet maski. Nakon augmentacije, skup je proširen na 200 slika. Distribucija pojavljivanja simptoma prosječno iznosi 136 ±18, što proizlazi iz činjenice da pojedine slike prikazuju više simptoma istovremeno.

3.2 Poboljšanja i evaluacija U-Net modela

Različite konfiguracije U-Net modela, unaprijeđene pomoću više pozadinskih CNN mreža (backbone), podešene su na prilagođenom skupu podataka za semantičku segmentaciju oftalmoloških simptoma kod pasa. Cilj je bio poboljšati fazu ekstrakcije značajki i postići bolje rezultate segmentacije u odnosu na tradicionalnu U-Net arhitekturu.

Prilikom treniranja korišteni su slijedeći parametri: 100 epoha, veličina grupe (batch size) 16, stopa učenja (learning rate) 0.0001, Adam optimizator te funkcije gubitka Categorical cross-entropy, Dice i Focal kao i njihove kombinacije. Skup podataka podijeljen je na dio za treniranje i za testiranje u odnosu 80:20, dok je evaluacija provedena pomoću metrika IoU (Jaccard), Dice, točnost piksela (Pixel Accuracy), osjetljivost (Sensitivity) i specifičnost (Specificity). Rezultati treniranja i pojedinačnih metrika detaljnije su opisani u radu (Buric et al., 2024).

Integracija ResNet34 pozadine uvodi rezidualno učenje (deep residual learning), koje omogućuje učinkovitije treniranje složenijih mrežnih struktura kroz preskakajuće (skip) veze. Arhitektura Inception V3 uključena je zbog svoje sposobnosti učinkovitog hvatanja među-kanalnih korelacija uz smanjenu računalnu složenost. VGG16, poznat po jednostavnoj ali dubokoj konfiguraciji, povećava sposobnost modela u detekciji značajki zahvaljujući svojim sekvencijalnim konvolucijskim slojevima. Dodatno, EfficientNet B3 je dodan radi optimizacije ravnoteže između dubine mreže, širine i rezolucije.

U procjenama performansi, osobito u segmentaciji različitih očnih stanja prikazanih u Tablica 1, korišten je srednji Jaccardov indeks kao glavna mjera točnosti modela u usklađivanju predviđene segmentacije s stvarnim oznakama. U-Net model unaprijeđen pozadinom ResNet34 značajno nadmašuje standardni

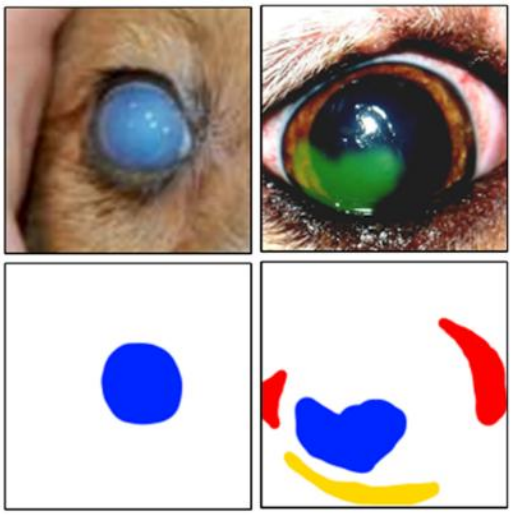
Arhitektura	zamućenje oka	crvenilo bjeloočnice	pretjerano suzenje	obojeno očno ispupčenje
U-Net	38.5	44	0.3	55.8
U-Net + ResNet34	73.9	**80.6**	**38**	**73.9**
U-Net + Inception V3	**78.3**	78.3	37.9	54.2
U-Net + VGG16	75.1	75.7	27.2	54.1
U-Net + EfficientNet	69.7	79.2	36.1	67.5

Tablica 1. Usporedni rezultati U-Net segmentacija s različitim pozadinskim arhitekturama prema Jaccardovim koeficijentima

U-Net u svim testiranim kategorijama, posebno u prepoznavanju crvenila bjeloočnice, s Jaccardovim

indeksom od 80,6%. Slično tome, model s pozadinom Inception V3 također pokazuje visoke performanse, iako zahtijeva dulje vrijeme obrade. VGG16 pruža razuman kompromis između brzine i točnosti, dok se EfficientNet B3 ističe konkurentnim rezultatima, osobito u identifikaciji ispupčenja i crvenila bjeloočnice, ali uz veću računalnu složenost.

Iako su rezultati segmentacije zadovoljavajući, sami vizualni prikazi nisu dovoljni za jasno tumačenje bolesti vlasniku psa, jer prikazuju samo obojena područja označena pripadajućim klasama, kao što je prikazano na slici 2. Naknadna interpretacija pomoću LLM-ova ključna je za pretvaranje informacija iz tih segmentiranih područja u dijagnostičke spoznaje koje se mogu primijeniti u praksi.



Slika 2. Slike glaukoma (lijevo) i ulceracije rožnice (desno) s pripadajućim segmentiranim simptomima (dolje), uključujući zamućenje oka, crvenilo bjeloočnice i pretjerano suzenje.

Segmentacijski rezultati omogućili su jasniju vizualnu identifikaciju simptoma koje veterinar može potvrditi pregledom. Na taj način model ne djeluje izolirano, već služi kao pomoćni alat koji poboljšava brzinu i dosljednost uočavanja očnih promjena.

4 Interpretacija simptoma pomoću LLM-ova

Glavni izazov velikih jezičnih modela (LLM-ova) leži u njihovoj sposobnosti da točno dešifriraju širok raspon opisa simptoma i izdvoje relevantne medicinske činjenice kako bi pružili koherentne odgovore koji su u skladu s utvrđenim medicinskim znanjem.

Kako bi se postigla što veća točnost u interpretaciji simptoma očnih bolesti i izbjeglo širenje pogrešaka, u analizu su uključeni samo simptomi čija je točnost segmentacije pomoću U-Net modela bila veća od 98%. S obzirom na subjektivnost i varijabilnost tekstualnih opisa simptoma, koji često sadrže nejasne ili djelomične informacije, precizno formulirana pitanja ključna su pri korištenju LLM-a kako bi se dobili sažeti i točni odgovori koji uključuju ispravnu dijagnozu.

Kako bi se taj rizik smanjio, upit (prompt) je oblikovan tako da LLM-u pruža multimodalne podatke (sliku bolesnog oka ili masku segmentacije te popis simptoma), ako je to bilo moguće, ili tekstualne podatke (popis simptoma) za daljnju analizu. Od LLM-a se zatim tražilo da procijeni najvjerojatnije dijagnoze (najviše dvije) na temelju zadanih ulaznih podataka.

4.1 Metrike za evaluaciju interpretacije simptoma pomoću LLM-ova

Evaluacija uspješnosti LLM-ova u interpretaciji simptoma provedena je pomoću nekoliko poznatih metrika. Za izračun sličnosti rečenica korišteni su MPNet i MiniLM, pri čemu je izračunata kosinusna sličnost između vektorskih reprezentacija (embeddinga) generiranih ovim modelima. Ovi modeli odabrani su zbog svoje učinkovitosti i točnosti u generiranju smislenih vektorskih prikaza rečenica, što omogućuje precizno mjerenje sličnosti.

MPNet koristi strategiju maskiranog i permutiranog predtreniranja kako bi poboljšao razumijevanje jezika, dok MiniLM primjenjuje duboku distilaciju pažnje (self-attention distillation) za kompresiju predtrenirane arhitekture transformera, čime postiže visoku učinkovitost i skalabilnost.

Uz ove modele korišten je i niz drugih metrika radi sveobuhvatnije procjene.

BERTScore koristi BERT(Devlin et al., 2018), unaprijed istreniran model jezične reprezentacije, za stvaranje kontekstualnih embeddinga svakog tokena u rečenici, te računa kosinusnu sličnost između kandidata i referentnih tokena.

CLIPScore procjenjuje tekst stvaranjem vektorskih reprezentacija pomoću CLIP-a, multimodalnog modela koji kombinira vizualne i jezične informacije, te izračunava kosinusnu sličnost između vektora teksta i referentnih opisa.

BLEU računa geometrijsku sredinu n-gram preciznosti između kandidata i referentnih opisa, uz korekciju duljine rečenice kako bi se izbjeglo generiranje prekratkih rečenica.

METEOR računa harmonijsku sredinu unigramske preciznosti i odziva između kandidata i referenci, koristeći WordNet sinonime i tablice parafraza.

ROUGE određuje najdulji zajednički podniz tokena između kandidata i referentne rečenice, bez zahtjeva za uzastopnim podudaranjem.

SPICE ocjenjuje opise slika usporedbom scenskih grafova kandidata i referenci na temelju njihovog semantičkog sadržaja i relacija među entitetima.

Ove metrike pružaju sveobuhvatnu procjenu učinkovitosti LLM-ova u interpretaciji medicinskih

simptoma, osiguravajući robustnost i pouzdanost dijagnostičke podrške koju ovi modeli mogu pružiti.

4.2 Usporedna evaluacija dijagnoza temeljenih na LLM-ovima nakon detekcije simptoma U-Net modelom kod očnih bolesti pasa

Kako bi se procijenile mogućnosti LLM-ova u interpretaciji i postavljanju dijagnoza, korišteno je 15 različitih medicinskih slučajeva očnih bolesti pasa, odabranih iz atlasa Medical Atlas of Canine and Feline Eye Diseases (Grozdanić et al., 2020). Svaki slučaj sadržava jedinstvenu kombinaciju simptoma: zamućenje oka, crvenilo bjeloočnice, pretjerano suzenje i obojeno očno ispupčenje, u rasponu od jednog do četiri simptoma, bez ponavljanja. Ti specifični simptomi, kao što je ranije opisano, prepoznati su na slikama pomoću U-Net modela i poslužili su kao ulazni podaci za LLM-ove pri određivanju osnovnih dijagnoza. Generirani tekst koji su proizveli LLM-ovi uspoređen je s referentnim, stvarnim dijagnozama.

Izlaz velikih jezičnih modela uvelike ovisi o dobro strukturiranim promptovima. Kako bi se to istražilo, primijenjeno je nekoliko pristupa. Prvi pristup uključivao je generiranje pitanja od strane vlasnika psa bez medicinskog znanja i njihovu usporedbu s pitanjima temeljenima na simptomima koje je prepoznao U-Net, simulirajući upite koje bi mogao koristiti veterinarski specijalist.

Za ovaj zadatak prikazani su rezultati koje je generirao Mistral 7B LLM (Tablica 2).

Upit	Dijagnoza generirana pomoću Mistral 7B LLM-a
Od strane korisnika[1]	Žao mi je što vaš pas ima problema s očima. Zamućene i ružičaste oči kod pasa mogu biti uzrokovane različitim čimbenicima, poput **alergija, ozljeda oka ili infekcija** (bakterijskih, virusnih ili gljivičnih). Važno je što prije dobiti točnu dijagnozu od veterinara jer neka očna stanja mogu brzo napredovati i potencijalno dovesti do gubitka vida.
Korištenjem U-Net simptoma[2]	Simptomi zamućenja oka, crvenila bjeloočnice i pretjeranog suzenja kod psa mogu biti povezani sa stanjima kao što su **prednji uveitis, luksacija leće ili glaukom**, te je potreban stručni veterinarski pregled radi postavljanja točne dijagnoze i odgovarajućeg liječenja.

„[1]Primijetio sam nešto neobično s okom svog psa i jako sam zabrinut. Oko mu je zamućeno i ružičasto. Trebam li se zabrinuti?“
„[2]Pas ima sljedeće simptome: zamućenje oka, crvenilo bjeloočnice i pretjerano suzenje. Navedi moguću dijagnozu.“

Tablica 2. Primjer rezultata koje je generirao Mistral 7B LLM, na temelju upita vlasnika psa bez navođenja simptoma, uspoređen je s rezultatima dobivenima segmentacijom slike pomoću U-Net modela, gdje su simptomi korišteni kao što bi to učinio veterinarski specijalist. Glaukom je u ovome slučaju točna dijagnoza

Drugi pristup uključivao je pružanje LLM-u slike bolesnog oka posebno s maskom koju je generirao U-Net i bez nje s opisom te maske. Primjer koji je generirao GPT-4o, sposoban za rad s multimodalnim ulaznim podacima, prikazan je u

Tablica 3.

Ulazna slika(e)	Upit (prompt)	Dijagnoza
	Postavi dijagnozu na temelju priložene slike.	Slika koju ste priložili prikazuje pseće oko sa zelenkastim obojenjem rožnice, što može upućivati na nekoliko mogućih očnih stanja, poput **edema rožnice, infekcije ili upalnog odgovora**.
	Postavi dijagnozu na temelju slike i simptoma opisanih u priloženoj maski, gdje crvena označava crvenilo bjeloočnice, plava zamućenje oka, a žuta pretjerano suzenje.	Na temelju vidljivih znakova i označenih simptoma na priloženoj slici, najvjerojatnije dijagnoze, uzimajući u obzir sve anotirane pokazatelje, jesu **konjunktivitis, ulceracija ili edem rožnice**.

Tablica 3. Primjer odgovora modela GPT-4o na slikovni ulaz s i bez segmentacijske maske. Točna dijagnoza je ulceracija rožnice, pri čemu se postižu bolji rezultati kada se koristi maska.

U-Net se ističe u lokalizaciji simptoma unutar slike, ali nema sposobnost razumjeti što ti simptomi predstavljaju. Suprotno tome, LLM-ovi mogu interpretirati simptome, ali ih ne mogu lokalizirati na slici, što otežava promatračima, osobito nestručnjacima, potvrđivanje prisutnosti određenih simptoma, osobito ako su oni suptilni ili teško uočljivi.

Zapažanja pokazuju da se rezultati poboljšavaju kada su navedeni simptomi, u usporedbi sa slučajevima kada je dostupna samo slika bolesnog oka. Sukladno tome, provedeni su daljnji eksperimenti koristeći simptome segmentirane pomoću U-Net modela. Osim toga, budući da samo mali broj LLM-ova posjeduje sposobnost analize multimodalnih ulaza koji uključuju slike, ulazni podaci za LLM-ove bili su ograničeni na tekstualne informacije (oznake klasa) dobivene iz U-Net segmentacije.

LLM	Cherry Eye	Ocular Trauma	1. diag.	1. i 2. diag.
GPT-4o	**Cherry Eye**, Orbitalna neoplazija	Nuklearna skleroza, **Trauma**	1	2
Gemini2	Protruzija, **Cherry Eye**	Glaukom, **Očna Trauma**	0	2

Upit: „Sljedeće pitanje odnosi se na istraživanje očnih bolesti kod pasa. Uzimajući u obzir četiri simptoma: S1-zamućenje oka, S2-crvenilo bjeloočnice, S3-pretjerano suzenje i S4-obojeno očno ispupčenje, potrebno je postaviti dijagnozu na temelju svih mogućih kombinacija ovih simptoma.
Vrati dvije najvjerojatnije dijagnoze bez ikakvog objašnjenja, odvojene zarezom.“

Tablica 4. Primjer prompta korištenog za dobivanje dijagnoze iz LLM-ova te primjer bodovanja u slučaju jedne i dvostruke dijagnoze. U prvom retku Cherry Eye je najvjerojatnija dijagnoza koja se podudara s referentnom dijagnozom (ground truth). Trauma je druga najvjerojatnija dijagnoza koja se također podudara s referentnom dijagnozom. Ako se procjenjuje jedna dijagnoza, rezultat (ocjena) iznosi 1, dok ako se procjenjuju dvije najvjerojatnije dijagnoze, rezultat iznosi 2.

Kako bi se dobio željeni izlaz, upiti upućeni LLM-ovima strukturirani su u tri dijela: kontekstualna pozadina, izravni upit i očekivani izlaz, kao što je prikazano u Tablica 4.

U konačnici, segmentacijski model generira maske koje označavaju područja četiri definirana simptoma (S1-S4). Ti se rezultati prevode u tekstualne opise koji predstavljaju semantički sažetak segmentiranih područja, npr. postojanje simptoma S1 – „zamućenje oka“ i S2 – „crvenilo bjeloočnice“. Ti se opisi zatim uvrštavaju u unaprijed definiran predložak upita koji sadrži tri dijela: (1) kontekstualnu pozadinu koja sadrži kombinaciju simptoma, (2) izravni upit i (3) očekivani izlaz (Slika 4. Prosječna distribucija rezultata svih LLM-ova u slučaju jedne dijagnoze, evaluirana pomoću MiniLM metrike za svaku kombinaciju simptoma. Simptomi su: S1-zamućenje oka, S2-crvenilo bjeloočnice, S3-pretjerano suzenje i S4-obojeno očno ispupčenjeTablica 4). Na taj način modelu se prenosi strukturirani tekstualni prikaz segmentiranih simptoma, čime se osigurava konzistentnost i usporedivost između različitih slučajeva.

5 Rezultati i rasprava

Dobiveni rezultati potvrđuju da se predloženi pristup može primijeniti u kliničkom kontekstu, primjerice za preliminarno usmjeravanje dijagnostike prije detaljnog oftalmološkog pregleda. Analiza uspješnosti LLM-ova u dijagnosticiranju očnih bolesti pasa, mjerena pomoću više metrika u scenarijima jedne i dvostruke dijagnoze, kako je prikazano u Tablica 5., otkriva nekoliko važnih zapažanja.

U slučaju jedne dijagnoze, GPT-4o pokazuje najbolju dijagnostičku sposobnost prema svim metrikama. Slijede ga Claude 4, Mistral 7B, Gemini2 i na kraju Llama-3. Metrike MPNet, BERTScore i CLIPScore, za razliku od ostalih koje pokazuju širu distribuciju rezultata, imaju preklapajuće vrijednosti među različitim LLM-ovima. Rezultati metrika ROUGE i SPICE pokazuju slične vrijednosti zaokružene na dvije decimale.

LLM	MPNet	MiniLM	BERTScore	CLIPScore	BLEU	METEOR	ROUGE	SPICE
Jedna dijagnoza								
GPT-4o	**0.66**	**0.64**	**0.80**	**0.91**	**0.52**	**0.42**	**0.56**	**0.56**
Gemini2	0.52	0.46	0.77	0.85	0.20	0.13	0.20	0.20
Mistral 7B	0.55	0.51	0.75	0.86	0.27	0.19	0.27	0.27
Claude 4	0.62	0.55	0.76	0.87	0.40	0.26	0.40	0.40
Llama-3	0.54	0.43	0.69	0.85	0.07	0.06	0.07	0.07
Dvostruka dijagnoza								
GPT-4o	**0.59**	**0.57**	**0.69**	0.86	**0.22**	**0.25**	**0.27**	**0.43**
Gemini2	0.52	0.46	0.68	0.85	0.15	0.20	0.19	0.23
Mistral 7B	0.50	0.48	0.66	0.85	0.10	0.11	0.12	0.20
Claude 4	0.54	0.50	0.66	0.84	0.19	0.21	0.24	0.28
Llama-3-3	0.54	0.48	0.69	**0.87**	0.07	0.07	0.09	0.15

Tablica 5. Evaluacija LLM-ova u dijagnosticiranju bolesti na temelju kliničkih simptoma u slučajevima jedne i dvostruke dijagnoze. U slučaju jedne dijagnoze uzima se najvjerojatnije medicinsko stanje, dok se u slučaju dvostruke dijagnoze uzimaju dvije najvjerojatnije dijagnoze i uspoređuju s referentnim vrijednostima.

Najveća razlika između najvišeg i najnižeg rezultata zabilježena je upravo kod metrika ROUGE i SPICE, dok je najmanja razlika (0,06) uočena kod CLIPScore.

U slučaju dvostruke dijagnoze, GPT-4o i dalje ostvaruje najviše rezultate.

Prema metrikama BERTScore, najviše rezultate ostvaruju GPT-4o i Llama-3, pri čemu je razlika između najvišeg i najnižeg rezultata još manja, svega 0,03.

Promjena između rezultata jedne i dvostruke dijagnoze najveća je kod metrika BLEU i ROUGE, dok je najmanja razlika uočena kod CLIPScore i MPNet.

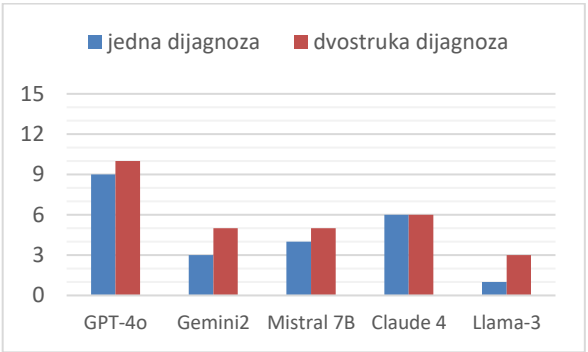
Slika 3. prikazuje rezultate dobivene ljudskom evaluacijom prema Tablica 4. Rezultati ljudske evaluacije u velikoj se mjeri podudaraju s rezultatima dobivenima metrikama sličnosti rečenica (vidi Tablica 5.), uz manja odstupanja.

Najuspješniji LLM je GPT-4o, koji ima najveću točnost kada se koristi dvostruka dijagnoza od svih ostalih modela u scenarijima jedne i dvostruke dijagnoze.

Model Claude 4 nije pokazao poboljšanje pri proširenoj dijagnozi, dok su najveći napredak ostvarili Gemini2 i Llama-3.

U slučaju jedne dijagnoze, metrike MPNet, BERTScore i CLIPScore imaju identične rezultate za pojedine LLM-ove, što je u suprotnosti s ljudskom evaluacijom, gdje svaki model ima različit broj točnih dijagnoza.

Metrike MiniLM, BLEU, METEOR, ROUGE i SPICE najvjernije prate poredak LLM-ova prema rezultatima ljudske procjene, uz nekoliko iznimaka.



Slika 3. Točno predviđene dijagnoze LLM-ova prema ljudskoj evaluaciji u slučajevima jedne i dvostruke dijagnoze

Na temelju evaluacijskih rezultata, koji demonstriraju uspješnost LLM-a (npr. Claude 4 s vrijednošću 6 točnih analiza) u odnosu na 15 promatranih slučajeva, Slika 3 i Slika 4 koriste metriku MiniLM kao referencu za daljnju analizu.

Kao što je prikazano na Slika 4, dijagnoza koja je najtočnije predviđena kod svih LLM-ova je Cherry Eye,

s prisutnim samo jednim simptomom - obojenim očnim ispupčenjem.

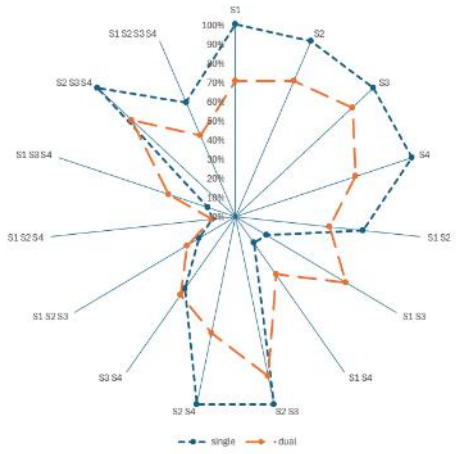
Slika 4. Prosječna distribucija rezultata svih LLM-ova u slučaju jedne dijagnoze, evaluirana pomoću MiniLM metrike za svaku kombinaciju simptoma. Simptomi su: S1-zamućenje oka, S2-crvenilo bjeloočnice, S3-pretjerano suzenje i S4-obojeno očno ispupčenje

Najzahtjevnija dijagnoza za predviđanje je bakterijska infekcija, koja je opisana kombinacijom triju simptoma.

Kombinacija simptoma zamućenje oka i obojeno očno ispupčenje pokazala se najtežom za ispravnu dijagnozu, osim u slučajevima kada su uključeni svi simptomi.

Na Slika 5 prikazana je razlika između rezultata dobivenih korištenjem jedne i dvostruke dijagnoze, evaluiranih pomoću MiniLM metrike na modelu GPT-4o.

Jednostruki slučaj dijagnoze ima manju uspješnost u pronalaženju točne dijagnoze, ali veću sigurnost u predviđanju.



Slika 5. Razlika između rezultata dobivenih korištenjem jedne i dvostruke dijagnoze, evaluiranih pomoću MiniLM metrike na modelu GPT-4o, za različite kombinacije simptoma. Simptomi su: S1-zamućenje oka, S2-crvenilo bjeloočnice, S3-pretjerano suzenje i S4-obojeno očno ispupčenje

To nije neočekivano, jer se dodatne pogreške javljaju kao posljedica proširenja mogućih dijagnoza.

Poboljšanja su posebno vidljiva kod dijagnoza koje uključuju simptom pretjeranog suzenja. U stvarnim slučajevima model je točno predložio dijagnozu u 10 od 15 analiziranih primjera, što pokazuje njegov potencijal u potpori kliničkom odlučivanju, iako se ne može koristiti kao zamjena za stručno mišljenje veterinara.

6 Zaključak

Ovo istraživanje pokazuje potencijal integracije naprednih modela strojnog učenja, posebno prilagođene U-Net arhitekture i različitih velikih jezičnih modela (LLM-ova), u dijagnostici očnih bolesti pasa. Model U-Net, unaprijeđen različitim pozadinskim arhitekturama poput ResNet34 i Inception V3, pokazuje značajna poboljšanja u segmentaciji očnih stanja u odnosu na tradicionalni U-Net model. Posebno se ističe arhitektura ResNet34, koja postiže visoku točnost u prepoznavanju crvenila bjeloočnice s visokim Jaccardovim indeksom, što potvrđuje njezinu učinkovitost u medicinskoj segmentaciji slika.

Evaluacija LLM-ova, uključujući GPT-4o, Mistral 7B, Gemini2, Llama-3 i Claude 4, pokazuje njihovu korisnost u interpretaciji informacija o simptomima dobivenima U-Net segmentacijom te u davanju preliminarnih dijagnoza.

LLM-ovi su procijenjeni pomoću različitih lingvističkih i semantičkih metrika tj. MPNet, MiniLM, BERTScore, CLIPScore, BLEU, METEOR, ROUGE i SPICE, kako bi se osigurala robusnost i pouzdanost njihovih dijagnostičkih rezultata. Model GPT-4o dosljedno postiže najbolje rezultate u scenarijima jedne i dvostruke dijagnoze, što ukazuje na njegovu sposobnost obrade medicinskih podataka i generiranja točnih dijagnostičkih prijedloga.

Kombinacija precizne segmentacije slika U-Net modela i naprednih sposobnosti generiranja i razumijevanja teksta LLM-ova povećava točnost i učinkovitost dijagnostike u veterinarskoj praksi.

Ograničenja istraživanja odnose se na relativno malen i nedovoljno raznolik skup podataka te osjetljivost modela na promjene u uvjetima snimanja i prikazu simptoma. Takvi čimbenici mogu utjecati na sposobnost modela da generalizira na različite kliničke slučajeve i uvjete snimanja.

Buduća istraživanja mogla bi se usmjeriti na povećanje raznolikosti i kvalitete skupa podataka te na primjenu suvremenih pristupa koji omogućuju istodobnu obradu vizualnih i tekstualnih informacija. Poseban potencijal imaju tehnike koje kombiniraju računalni vid i jezične reprezentacije unutar jedinstvenih multimodalnih sustava, što bi omogućilo usporedbu njihove dijagnostičke točnosti s tradicionalnim pristupima temeljenima isključivo na tekstu. Nadalje, uključivanje evaluacije u stvarnim kliničkim uvjetima, uz povratne informacije

stručnjaka, doprinijelo bi procjeni stvarne primjenjivosti razvijenog sustava.

Takav integrirani pristup mogao bi u budućnosti omogućiti razvoj inteligentnog dijagnostičkog asistenta koji povezuje računalni vid, generativne metode i stručno znanje iz veterinarske oftalmologije, čime bi se unaprijedila točnost, pouzdanost i dostupnost dijagnostičkih alata u praksi. Daljnji razvoj trebao bi se usmjeriti i na stvaranje jednostavnijih korisničkih sučelja kako bi sustav bio razumljiv i primjenjiv i u manje specijaliziranim ambulantama.

Literatura

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). *SPICE: Semantic Propositional Image Caption Evaluation.* https://doi.org/10.48550/ARXIV.1607.08822

Anderson, P., Fernando, B., Johnson, M., & Gould, S. Boevé, M., & Stades, F. (1985). Glaucoma in dogs and cats. Review and retrospective evaluation of 421 patients. I. Pathobiological background, classification and breed predisposition. *Tijdschrift Voor Diergeneeskunde, 110*(6), 219—227.

Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., & Merhof, D. (2022). Medical Image Segmentation Review: The success of U-Net. https://doi.org/10.48550/ARXIV.2211.14830

Bucur, A.-M. (2023). Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms from Social Media. https://doi.org/10.48550/ARXIV.2307.02313

Buric, M., Grozdanic, S., & Ivasic-Kos, M. (2024). Diagnosis of ophthalmologic diseases in canines based on images using neural networks for image segmentation. Heliyon, e38287. https://doi.org/10.1016/j.heliyon.2024.e38287

Burić, M., Ivašić-Kos, M., & Grozdanić, S. (n.d.). DogEyeSeg4: Dog Eye Segmentation 4-Class Ophthalmic Disease Dataset (No. urn:nbn:hr:195:405214) [Dataset]. Faculty of Informatics and Digital Technologies, University of Rijeka. Retrieved August 22, 2024, from https://urn.nsk.hr/urn:nbn:hr:195:405214

Burić, M., Paulin, G., & Ivašić-Kos, M. (n.d.). Object Detection Using Synthesized Data. In *Proceedings of the ICT Innovations Conference*, Ohrid, North Macedonia, 15.

Dandekar, A., Zen, R. A. M., & Bressan, S. (2018). A Comparative Study of Synthetic Dataset Generation Techniques. In S. Hartmann, H. Ma, A. Hameurlain, G. Pernul, & R. R. Wagner (Eds.), *Database and Expert Systems Applications* (pp. 387–395). Springer International Publishing. https://doi.org/10.1007/978-3-319-98812-2_35

Deane, J., Kearney, S., Kim, K. I., & Cosker, D. (2021). DynaDog+T: A Parametric Animal Model for Synthetic Canine Image Generation (No. arXiv:2107.07330). arXiv. http://arxiv.org/abs/2107.07330

Denkowski, M., & Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 376–380. https://doi.org/10.3115/v1/W14-3348

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

Ganesan, K. (2018). *ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks.* https://doi.org/10.48550/ARXIV.1803.01937

Gemini Team, Reid, M., Savinov, N., Teplyashin, D., Dmitry, Lepikhin, Lillicrap, T., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A., Millican, K., Dyer, E., Glaese, M., … Vinyals, O. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.* https://doi.org/10.48550/ARXIV.2403.05530

González-Chávez, O., Ruiz, G., Moctezuma, D., & Ramirez-delReal, T. A. (2023). Are metrics measuring what they should? An evaluation of image captioning task metrics (No. arXiv:2207.01733). arXiv. http://arxiv.org/abs/2207.01733

Grozdanić, S., Đukić, S., Luzhetskiy, S., Milčić-Matić, N., & Lazić, T. (2020). *Atlas bolesti oka pasa i mačaka*. Oculus Vet.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition (No. arXiv:1512.03385; Version 1). arXiv. http://arxiv.org/abs/1512.03385

He, Z., Bhasuran, B., Jin, Q., Tian, S., Hanna, K., Shavor, C., Arguello, L. G., Murray, P., & Lu, Z. (2024). Quality of Answers of Generative Large Language Models vs Peer Patients for Interpreting Lab Test Results for Lay Patients: Evaluation Study. *Journal of Medical Internet Research, 26*, e56655. https://doi.org/10.2196/56655

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2021). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, 7514–7528. https://doi.org/10.18653/v1/2021.emnlp-main.595

Hrga, I., & Ivasic-Kos, M. (2024). Measuring the Sensitivity of Image Captioning Metrics to Caption Perturbations. In X.-S. Yang, R. S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of Eighth International Congress on Information and Communication Technology* (Vol. 696, pp. 1053–1063). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3236-8_85

Huang, K.-W., Yang, Y.-R., Huang, Z.-H., Liu, Y.-Y., & Lee, S.-H. (2023). Retinal Vascular Image Segmentation Using Improved UNet Based on Residual Module. *Bioengineering, 10*(6), 722. https://doi.org/10.3390/bioengineering10060722

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B*. https://doi.org/10.48550/ARXIV.2310.06825

Johnsen, D. A. J., Maggs, D. J., & Kass, P. H. (2006). Evaluation of risk factors for development of secondary glaucoma in dogs: 156 cases (1999–2004). *Journal of the American Veterinary Medical Association, 229*(8), 1270–1274. https://doi.org/10.2460/javma.229.8.1270

Katic, T., Pavlovski, M., Sekulic, D., & Vucetic, S. (2021). Learning Semi-Structured Representations of Radiology Reports (No. arXiv:2112.10746). arXiv. http://arxiv.org/abs/2112.10746

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

PBC, A. (2024). *Claude LLM (Version 3) [Computer software]*. Anthropic PBC.

Petit, O., Thome, N., Rambour, C., & Soler, L. (2021). U-Net Transformer: Self and Cross Attention for Medical Image Segmentation. https://doi.org/10.48550/ARXIV.2103.06104

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation (No. arXiv:1505.04597). arXiv. http://arxiv.org/abs/1505.04597

Sasazawa, Y., Yokote, K., Imaichi, O., & Sogawa, Y. (2023). Text Retrieval with Multi-Stage Re-Ranking Models. https://doi.org/10.48550/ARXIV.2311.07994

Savage, C. H., Park, H., Kwak, K., Smith, A. D., Rothenberg, S. A., Parekh, V. S., Doo, F. X., & Yi, P. H. (2024). General-Purpose Large Language Models Versus a Domain-Specific Natural Language Processing Tool for Label Extraction From Chest Radiograph Reports. *American Journal of Roentgenology*, AJR.23.30573. https://doi.org/10.2214/AJR.23.30573

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition (No. arXiv:1409.1556). arXiv. http://arxiv.org/abs/1409.1556

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. https://doi.org/10.48550/ARXIV.2004.09297

Sreng, S., Maneerat, N., Hamamoto, K., & Win, K. Y. (2020). Deep Learning for Optic Disc Segmentation and Glaucoma Diagnosis on Retinal Images. *Applied Sciences, 10*(14), 4916. https://doi.org/10.3390/app10144916

Strom, A. R., Hässig, M., Iburg, T. M., & Spiess, B. M. (2011). Epidemiology of canine glaucoma presented to University of Zurich from 1995 to 2009. Part 1: Congenital and primary glaucoma (4 and 123 cases). Veterinary Ophthalmology, 14(2), 121–126. https://doi.org/10.1111/j.1463-5224.2010.00855.x

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision (No. arXiv:1512.00567; Version 3). arXiv. http://arxiv.org/abs/1512.00567

Tan, M., & Le, Q. V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks (No. arXiv:1905.11946). arXiv. http://arxiv.org/abs/1905.11946

Thamizharasan, A., Murugan, M. S., & Parthiban, S. (2016). Surgical Management of Cherry Eye in a Dog. *Intas Polivet, 17*(II), 420-421.

Thirunavukarasu, A. J., Mahmood, S., Malem, A., Foster, W. P., Sanghera, R., Hassan, R., Zhou, S., Wong, S. W., Wong, Y. L., Chong, Y. J., Shakeel, A., Chang, Y.-H., Tan, B. K. J., Jain, N., Tan, T. F., Rauz, S., Ting, D. S. W., & Ting, D. S. J. (2024). Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digital Health, 3*(4), e0000341. https://doi.org/10.1371/journal.pdig.0000341

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama-3 2: Open Foundation and Fine-Tuned Chat Models (No. arXiv:2307.09288). arXiv. http://arxiv.org/abs/2307.09288

Tripathi, R. M., Kashyap, D. K., & Giri, D. K. (2014). Surgical Management of Cherry Eye in a Dog. *Intas Polivet, 15*(1), 131-132.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*. https://doi.org/10.48550/ARXIV.2002.10957

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). *BERTScore: Evaluating Text Generation with BERT*. https://doi.org/10.48550/ARXIV.1904.09675

AI in Veterinary Ophthalmology: Image Segmentation and Large Language Models in the Diagnosis of Canine Eye Diseases

Abstract

This study presents the application of artificial intelligence methods in veterinary ophthalmology through the combination of image segmentation and language understanding models. A customized U-Net model was used for detecting and segmenting canine ocular symptoms, including ocular opacity, sclera redness, excessive tearing, and colored ocular protrusion. The obtained results were used as input for large language models (GPT-4o, Mistral 7B, Gemini2, Llama-3, and Claude 4) to interpret symptoms and provide preliminary diagnoses.

The evaluation was performed using linguistic and semantic metrics (MPNet, MiniLM, BERTScore, CLIPScore, BLEU, METEOR, ROUGE, and SPICE). The results show that integrating U-Net segmentation with LLM analytical capabilities enables effective preliminary diagnosis of canine eye diseases. The ResNet34-based model achieved the highest accuracy in identifying sclera redness, while GPT-4o performed best in interpreting symptoms and suggesting diagnoses. This approach contributes to the development of intelligent systems that can improve the accuracy and efficiency of veterinary diagnostics.

Keywords: *U-Net; large language models; image segmentation; veterinary diagnostics; canine eye diseases.*