VILNIUS TECH
Elektroninių sistemų katedra

Rūta Jankevičiūtė

AKIŲ LIGŲ KLASIFIKAVIMO METODŲ TAIKOMŲ TINKLAINĖS VAIZDAMS TYRIMAS

INVESTIGATION OF EYE DISEASE CLASSIFICATION METHODS FROM RETINAL IMAGES

Baigiamasis magistro darbas

Informacinių elektroninių sistemų studijų programa, valstybinis kodas 6211BX015

Dirbtinio intelekto sistemų specializacija

Informatikos inžinerijos studijų kryptis

Vilnius, 2024

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
ELEKTRONIKOS FAKULTETAS
ELEKTRONINIŲ SISTEMŲ KATEDRA

TVIRTINU
Katedros vedėjas

_____
(*parašas*)
Prof. dr. A. Serackis
2024 m. _____ mėn. ___ d.

Rūta Jankevičiūtė

AKIŲ LIGŲ KLASIFIKAVIMO METODŲ TAIKOMŲ TINKLAINĖS VAIZDAMS TYRIMAS

INVESTIGATION OF EYE DISEASE CLASSIFICATION METHODS FROM RETINAL IMAGES

Baigiamasis magistro darbas

Informacinių elektroninių sistemų studijų programa, valstybinis kodas 6211BX015
Dirbtinio intelekto sistemų specializacija
Informatikos inžinerijos studijų kryptis

Vadovas dr. V. Abromavičius _____
(*pedag. vardas, moksl. laipsnis, vardas, pavardė*) (*parašas*) (*data*)

Lietuvių kalbos konsultantas dr. Aušra Žemienė _____
(pedag. vardas, moksl. laipsnis, vardas, pavardė) (parašas) (data)

Vilnius, 2024

TURINYS

PAVEIKSLŲ SĄRAŠAS

LENTELŲ SĄRAŠAS

ĮVADAS

Darbo aktualumas ir tikslas

Rega yra vienas svarbiausių žmogaus pojūčių. Akių ligos, ypač tokios, kaip glaukoma arba diabetinė retinopatija, kurios gali sukelti apakimą, gali turėti rimtas pasekmes pacientas ir kardinaliai pakeisti jų gyvenimus. Dėl to ankstyvas akių ligų aptikimas ypatingai svarbus, nes neretais atvejais ankstyvose stadijose tokias ligas galima sustabdyti arba net atstatyti jų padarytą žalą. Šiame darbe siekiama ištirti šiuolaikinius, giliaisiais neuroniniais tinklais paremtus, metodus akių ligų aptikimui naudojant akies tinklainės nuotraukas. Dėl to, kad rasti medicininių vaizdų rinkinius, kurie būtų klasifikuoti medicinos srities specialistų, yra labai sudėtinga ir dauguma akių ligų reikalauja daugiau tyrimų, nei tik tinklainės fotografijos analizės, buvo nuspręsta eksperimentams naudoti tik vienos ligos duomenų rinkinius – diabetinės retinopatijos. Ši liga diagnozuojama pagrinde naudojant tinklainės nuotraukų analizę ir yra laisvai prieinamų rinkinių, skirtų būtent jos lygio nustatymui skirtų modelių apmokymui.

Vaizdų klasifikavimas yra viena fundamentinių vaizdų apdorojimo sričių, kuri plačiai taikoma įvairiose industrijose ir visokioms problemoms spręsti. Medicinoje vaizdų klasifikavimas gali padėti atpažinti įvairių ligų požymius iš visokių skirtingų formų vaizdų (ultragarso, magnetinio rezonanso, kompiuterinės tomografijos ir pan.) [1]. Nors medicinoje vaizdo klasifikavimas turi unikalių iššūkių, iš principo; jam gali būti taikomi tokie patys modernūs metodai, kaip ir kitiems klasifikavimo uždaviniams, t. y. giliųjų neuroninių tinklų modeliai.

Tyrėjai pastaraisiais metais ypač daug dėmesio skiria šiems metodams dėl jų našumo, universalumo, daugiadisciplininės naudos ir gebėjimo apibendrinti informaciją. Atsiranda vis daugiau skirtingų modelių su savais privalumais ir trūkumais. Taikoma vis daugiau mokymo metodų, skirtų skirtingoms užduotims atlikti ir modelio atliktai užduočiai vertinti. .

Kaip matoma iš mokslinės literatūros apžvalgos, konvoliuciniai tinklai yra labai populiarus būdas klasifikuoti vaizdą [1] ir tokie modeliai tiriami jau daugiau nei 10 metų, t. y. nuo 2012 metų, kai jie *IMAGENET* konkurse buvo pradėti taikyti pirmą kartą [2]. Šie modeliai vis dar vyrauja ir medicininių vaizdų klasifikavimo procesuose. 2017 metais buvo pristatyti nauji, vadinamieji transformerio modeliai [3], kurie 2020 metais irgi pritaikyti vaizdams klasifikuoti [4], Transformerio modeliai susilaukia vis didesnio susidomėjimo dėl savo unikalių privalumų:

- gebėjimo išgauti globalų kontekstą nuo pat pirmų sluoksnių;
- dėmesio sau algoritmo;

- galimybės juos plėsti ir efektyviai naudoti duomenis;
- mažesnio šališkumo.

Šiame darbe siekiama atrinkti ir palyginti kelis skirtingų architektūrų klasifikavimo modelius akių ligoms klasifikuoti naudojant tinklainės vaizdus. Analizė tikslas yra nustatyti, kurie iš atrinktų modelių veikia optimaliai tikslumo ir ekonominiu aspektais.

Darbo tikslas – ištirti metodus, skirtus atpažinti akių ligoms iš akių tinklainės nuotraukų.

Darbo uždaviniai

1. Atlikti šiuolaikinių modelių, naudojamų medicininiams vaizdams klasifikuoti, literatūros bei architektūros apžvalgą.
2. Pasirinkti tinkamiausią akių tinklainės vaizdų duomenų rinkinį, apžvelgti jo sandarą ir paruošti naudoti apmokant tinklus.
3. Ištirti pasirinktų metodų efektyvumą ir palyginti gautus rezultatus.

Naudoti tyrimo ir analizės metodai

- Mokslinės literatūros analizė;
- Praktinis pritaikymas;
- Lyginamoji analizė.

Darbo naujumas ir praktinė nauda

Diabetinės retinopatijos tyrimuose, naudojančiuose dirbtinius neuroninius tinklus, regos transformeriai dar nėra išbandyti, apie jų bandymus nėra rašoma mokslinėje literatūroje ir jie nėra naudojami konkursuose, kuriuose dominuoja konvoliuciniai neuroniniai tinklai. Šiuo darbu siekiama išbandyti šiuos modelius ir palyginti juos su vienu plačiai naudojamu konvoliucinių tinklų modeliu naudojant, kiek įmanoma, tokias pačias strategijas ir aplinką ir prieiti prie išvadų apie modelių pritaikomumą šiam uždaviniui spręsti.

Darbo struktūra

Darbą sudaro įvadas, keturi skyriai, išvados, literatūros sąrašas ir priedai.

Pirmajame skyriuje atliekama mokslinės literatūros apie pagrindines giliųjų neuroninių tinklų architektūras, naudojamas pastaruosius penkerius metus medicininiams vaizdams klasifikuoti, apžvalga. Analizuojama, kokiems duomenų rinkiniams jie buvo taikomi ir aptariami kelių tyrimų rezultatai.

Antrajame skyriuje aptariami pagrindiniai klasifikavimo uždaviniams spręsti naudojami mokymo būdai ir jų taikymo galimybės.

Trečiajame skyriuje aprašomas darbe naudojamas duomenų rinkinys, jo paruošimo būdai ir modeliai, kurie bus naudojami bandymams.

Ketvirtajame skyriuje aptariami empirinio tyrimo rezultatai, analizuojamos sąlygos, kuriomis jie buvo gauti, ir atliekama visų skirtingų bandymų konfigūracijų lyginamoji analizė.

1. VAIZDŲ KLASIFIKAVIMO MODELIŲ MEDICINOJE LITERATŪROS APŽVALGA

Vaizdui klasifikuoti, nepriklausomai nuo užduoties, galima naudoti įvairiais sluoksniais paremtas architektūras. Iš principo, net perceptronai kai kuriais atvejais gali būti geras pasirinkimas, bet sudėtingiems uždaviniams, kai mes norime modelio, kuris sugeba rasti ir prisitaikyti prie ypatybių, kurios leis jam klasifikuoti nuotrauką, kurios gali pasirodyti skirtingose formose ir skirtingose vaizdo vietose, mums reikia sudėtingesnių, daugiau resursų reikalaujančių modelių.

Šiame skyriuje aptariamos kelios architektūrų kategorijos, kurios dažniausiai, pasak literatūros [5], naudojamos medicininiams vaizdams klasifikuoti, kodėl jos naudingos ir koks jų populiarumas literatūroje ir medicininių vaizdų apdorojimo srityje.

1.1. Konvoliuciniai neuroniniai tinklai

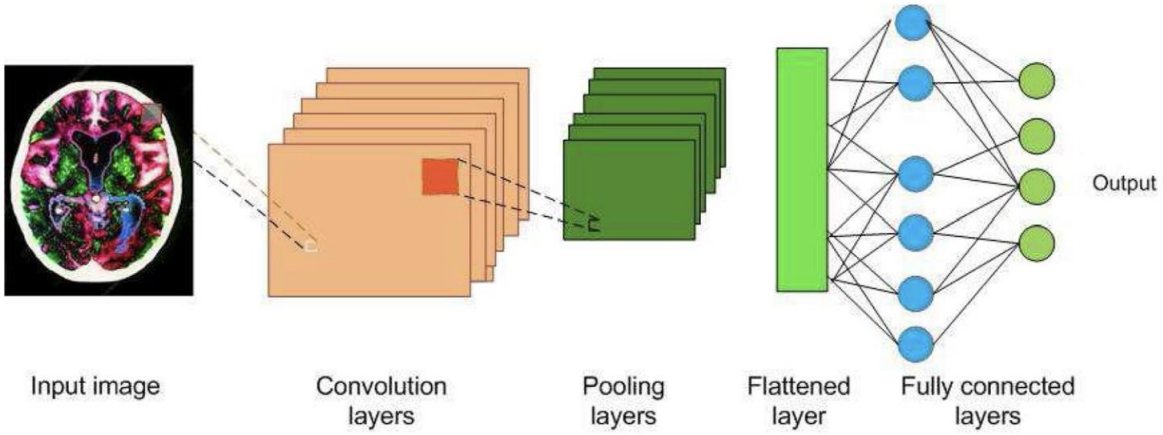
Konvoliuciniai neuroniniai tinklai (angl. *Convolutional neural network, CNN*) naudojami medicininiams vaizdams atpažinti [1], [6], [7]. Tokio tipo tinklai atsirado dar XX a. pabaigoje [8], bet jų taikymo vaizdų apdorojimo srityje tyrimai išpopuliarėjo, kai 2012 metais buvo išleistas ir konkurse ILSVRC-2012 panaudotas ALEXNET modelis [1], [2], [6], [7]. Naudojantis šiuo modeliu buvo pasiektas 15.3% klaidos lygis atpažįstant *IMAGENET* duomenų rinkinyje esančius vaizdus, kai antros vietos nugalėtojai pasiekė 26.2% [2].

Toliau apžvelgiama konvoliucinių neuroninių tinklų bazinė architektūra, ne konkretaus modelio architektūra, bet dažniausiai skiriasi modelių gylis (sluoksnių skaičius), filtrų plotis, išankstinio apdorojimo (ang. Preprocessing) strategijos ir apmokymo strategijos [7], [9]. Po to plačiau apžvelgiamas tokio tipo tinklų naudojimas medicininiams vaizdams klasifikuoti.

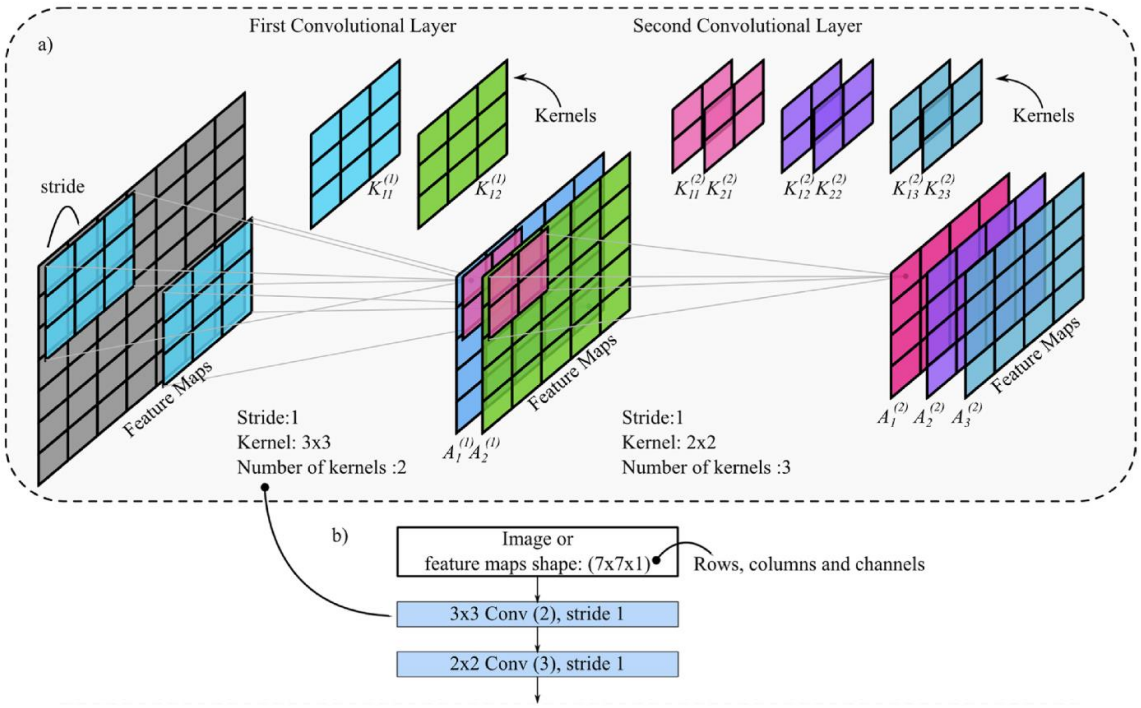
1.1.1. Konvoliucinių tinklų architektūros apžvalga

Konvoliuciniai neuroniniai tinklai specializuojami apdorojant duomenis, kurie turi į tinklelį panašią topologiją, pavyzdžiui, nuotraukas. Konvoliucinius tinklus (**1 pav.**) sudaro konvoliucinis sluoksnis, surinkimo sluoksnis (ang. *Pooling*) ir pilnai sujungtas (angl. *fully-connectd)* sluoksnis.

Konvoliucinio sluoksnio schema matoma ***2 pav.*** Pagrindinis konvoliucinio sluoksnio tikslas yra išgauti vaizdo ypatybes. Tai yra daroma su apmokomais filtrais (angl. *kernel*), kurių skaičius žymi sluoksnio gylį, pavyzdžiui (žr. ***2* pav.**) pirmasis konvoliucinis sluoksnis turi du filtrus, o antras – tris. Šių filtrų dydis yra $x \times y \times d$, kur d – gylis (įvestyje, tai būtų spalvų kanalų skaičius, filtrų atveju, tai prieš tai buvusio sluoksnio gylis).



1 pav. Konvoliucinio neuroninio tinklo struktūra [7]



2 pav. Dviejų sluoksnių konvoliucinis neuroninis tinklas su dvejais ir trejais filtrais. Paveikslėlio viršuje, a) dalyje, matoma detali tinklo schema, b) dalyje - tokio pačio tinklo supaprastinta schema [1]

Konvoliucinio sluoksnio tikslas yra slenkant filtrą, kuris yra mažesnis, nei įvestis (pavyzdžiui, *ALEXNET* atveju filtrai buvo 11×11 , o įvesties nuotraukos - 256×256 RGB vaizdai [2]), sugeneruoti ypatybių žemėlapius (angl. *feature map*), kurie reikalingi teisingai klasifikuoti vaizdą. Kitaip tariant, filtrai apmokomi taip, kad keliaujant per įvestį reikšmes būtų apskaičiuotos taip, kad daugiau „dėmesio“ būtų duota toms įvesties ypatybėms, kurios reikalingos mūsų sprendžiamam uždaviniui. Šiai operacijai naudojama **1 lygtis**.

1 lygtis. Ypatybių žemėlapio skaičiavimas

$$\sum_{i}^{N} \sum_{j}^{M} X[i,j] \cdot Y[i,j]$$

Surinkimo sluoksnio funkcija yra sumažinti kiekvienos įvesties dydį ir taip kontroliuoti persimokymą ir sumažinti tinklo parametrų kiekį, taip pagreitinant tinklo apmokymą ir apmokymo efektyvumą (kitaip tariant, greitį). Iš esmės, šios operacijos metu, mes įvestį skaidome į konkretaus dydžio regionus ir atliekame operaciją kiekvienam iš jų, tą rezultatą saugodami atitinkamoje išvesties matricos vietoje. Dažniausios operacijos yra regiono maksimumo arba vidurkio radimas ir saugojimas.

Galiausiai paskutinio konvoliucinio arba surinkimo sluoksnio išvestis paduodama į plokštinimo (angl. *flatten*) sluoksnį, kuris kelių dimensijų matricą paverčia į vienmatį vektorių, kuris paduodamas į pilnai sujungta sluoksnį. Šiame sluoksnyje apskaičiuojama aktyvavimo funkcija ir nustatomos tikimybės, kuriai klasei priklauso įvestis.

Aptarta bazinė architektūra šiek tiek skirtingomis formomis naudojama įvairiuose modeliuose. Tokios struktūros tikslas (kelių dimencijų matricų analizavimas dalimis) išlieka, bet keičiant sluoksnių kiekius ar kitus parametrus galima pasiekti skirtingų rezultatų atitinkamoms problemoms.

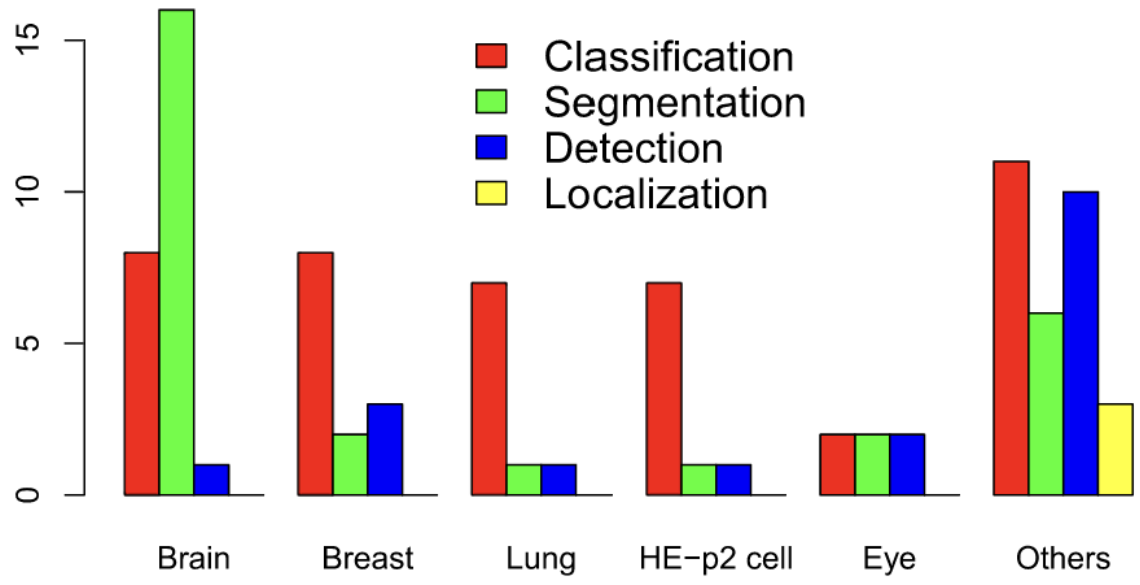
1.1.2. Konvoliucinių tinkle naudojimas medicininių vaizdų klasifikavimui

Medicinoje konvoliuciniai tinklai plačiai taikomi įvairių tipų vaizdams apdoroti (*x-ray*, CT, MRT, EKG, ultragarso ir kitų) [1]. Pagrindiniai tyrimai atliekami įvairių vėžinių ląstelių [6], [7], COVID-19 ir kitų plaučius veikiančių ligų [7], akių ligų atpažinimui [7], bei širdies anomalijų ir smegenų ligų, kaip Alzhaimerio ligos atpažinimui [7]. Minėtoms sritims naudojami įvairūs skirtingi modeliai ir įvairios architektūros, nes medicininių vaizdų duomenų rinkiniai skirtingoms problemoms gali žymiai skirtis, ypač kalbant apie duomenų rinkinių klasių balansą [7], [9]. Skirtingi modeliai sprendžia skirtingas problemas ir vienas modelis gali grąžinti kitokius rezultatus, kai yra pritaikomas šiek tiek kitokiai problemai.

Iš esmės, tai gerai iliustruoja **3 pav.** ir **4 pav**., kur matomos efektyviausios architektūros skirtingoms medicininių vaizdų apdorojimo problemoms naudotos architektūros, bei kiek straipsnių buvo peržiūrėta atsižvelgiant į vaizdų apdorojimo bei medicininę sritį. [7].

Organ (modality)	Diseases	Tasks	Best architecture
Brain MRI	Alzheimer	Segementation	Lenet-5, GoogleNet
		Detection	Multipath CNN
	Epilepsy	Classification	CNN with seven conv and
	Schizophrenic, bipolar disorder	Segementation	Patch based CNN with 4 conv, 2 maxpool and softmax regression
Breast mammogram	Tumour	Segementation	(Conv + maxpool) × 2 + 2 FC, Softmax
		Detection	Scaled VGGnet or GoogleNet
		Classification	Alexnet
		Localization	Semi Supervised Deep CNN
Heart (CT)	Coronary artery calcium scoring	Segmentation	Two path CNN
Heart (ECG)	ECG	Classification	Two path CNN, one along spatial and the other along temporal and fused together finally
Heart (MRI) left ventricle		Localization	(Conv, maxpool) ×2, FC
Lung (CT)	Cancer	Classification	Ensemble of overfeat across axial, saggital and coronal plane
	Nodule	Detection	Lenet, Overfeat
	COVID-19	Classification	multi scale multi encoder ensemble CNN model
	ILD	Classification	Any CNN architecture like Alexnet
Hep-2 cell		Classification	(Conv, maxpool) ×3, 2 FC, softmax regressor
Eye	Haemorrhage	Detection	(Conv, ReLu, maxpool) ×5, FC, softmax classifier
	Glaucoma	Detection	(Conv, ReLu, maxpool) ×2, FC with adaboost
	Retinopathy	Segmentation	(Conv, ReLu, maxpool) ×10, 3 FC, softmax classifier
Colon	Polyp	Classification	Any simple CNN architecture
	Polyp	Detection	Ensemble of CNN
Skin	Melanoma	Detection	(Conv, ReLu, maxpool) ×2, FC
Liver (CT)		Classification	(Conv, maxpool) ×2, 3 conv, max pool, 3 FC, softmax
Abdomen (US scan)	Fetus	Localization	(Conv, ReLU, maxpool) ×2 conv ×3, maxpool, FC ×3

3 pav. Efektyviausios modelių architektūros, paremtos konvoliuciniais neuroniniais tinklais, pasak 2022 metais atliktos literat8ros apžvalgos [7]



4 pav. 2022 metais atliktos literat8ros apžvalgos peržiūrėtų straipsnių kiekis ir pasiskirstymas [7]

Iš kitos pusės, tai rodo ir tokio tipo architektūrų populiarumą ir didelį, bei toliau augantį, tyrimų kiekį, kuris padeda spręsti problemas, kylančias su tokiais modeliais.

Kai kurios problemos ypač opios medicininių vaizdų apdorojimui. Pavyzdžiui, konvoliuciniai tinklui apmokyti reikia didelių kompiuterinių skaičiavimo resursų [7], [10], [11]. Taip pat yra

įrodymų, kad gilesni tinklai gali pasiekti geresnius rezultatus [12], [13], bet kuo daugiau sluoksnių tinkle, tuo sunkiau jį apmokyti.

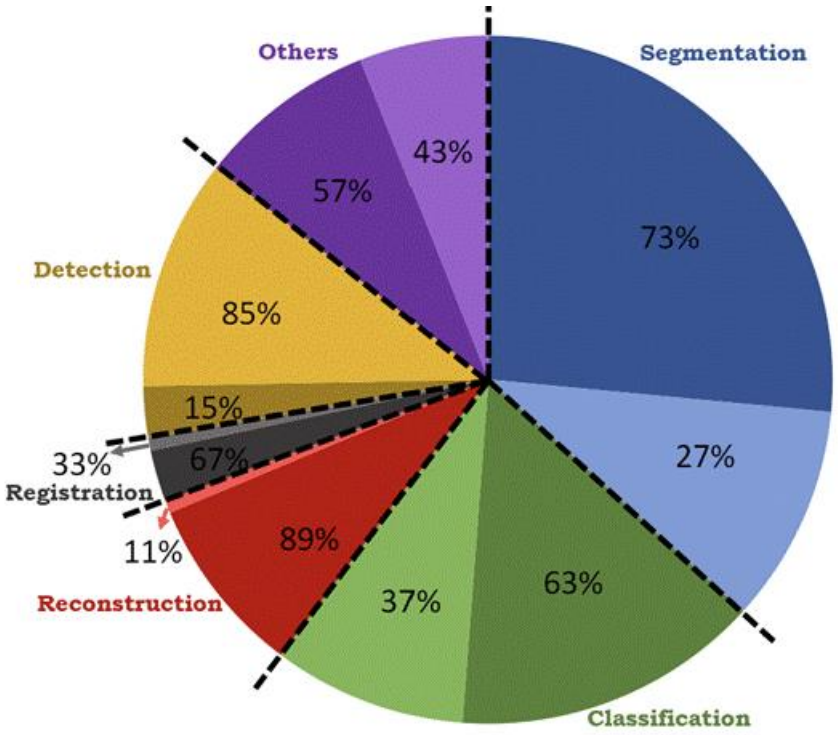
Priedu prie to, tokio tipo tinklai negali efektyviai „atsiminti“ santykių tarp labai nutolusių ypatybių, kas gali žymiai padėti klasifikuojant vaizdus [10]. Iš to, taip pat, ryškėja viena iš būtent medicininiams vaizdams apdorojimui specifinių problemų – duomenų rinkinių klasių disbalanso problema [9], dėl kurios tinklai per daug dėmesio skiria didžiausioms klasėms, per daug prisilygina prie jų (angl. *overfit*), ir jų rezultatai tampa blogesni, nes jie išmoksta klasifikuoti didžiausią klasę ir tų ypatybių negali tiesiogiai pritaikyti teisingam mažesnių klasių atskirimui.

Yra konvoliucinių tinklų, kurie padeda spręsti šias problemas, kaip RESNET [14] ir EfficientNet [15], ir kitokių apmokymo strategijų, kaip perkeliamas mokymasis (angl. *transfer learning*) [16], [17], kuriame tinklas apmokomas naudojant didelį, paruošta duomenų rinkinį, kaip *IMAGENET* ir tuomet koreguojamas su mažesniais, bet specifiniais sprendžiamai problemai duomenų rinkiniais.

1.2. Transformeris

Transformerio modelis pirmą kartą aprašytas 2017 metais [3]. Originaliai šis modelis buvo naudojamas natūralios kalbos apdorojimo uždaviniams, konkrečiai – vertimui. Vaizdo apdorojimui pradžiai buvo naudojami hibridiniai modeliai, kurie taiko modelio dalis, kaip dėmėsio sau (angl. *self-attention*) algoritmą kartu su konvoliucinių tinklų ypatybių žemėlapiais [18] vaizdo klasifikavimui, arba taikė šį algoritmą vaizdų aproksimavimui [19], bet originali architektūra vaizdo apdorojimui buvo pritaikyta tik 2020 metais su Regos transformerio (angl. *vision transformer*, *ViT*) modeliu [4], kuris buvo pritaikytas vaizdams klasifikuoti.

Medicininių vaizdų apdorojimo srityje susidomėjimas transformerių pritaikymui didelis, kaip matoma **5 pav.** priedu prie to, įvairiose mokslinės literatūros apžvalgose, jie dažnai minimi kaip viena iš naujų, ypač perspektyvių technologijų medicininių vaizdų klasifikavimui [1], [11], [20], [21], bet tai nereiškia kad nėra atvirų klausimų ar panašų problemų jų pritaikymui, kaip yra su konvoliuciniais neuroniniais tinklais.

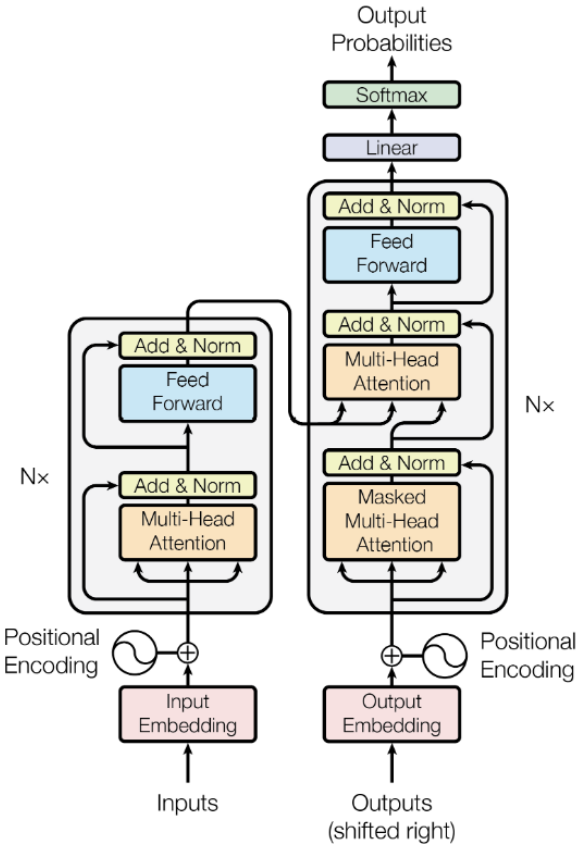


5 pav. Skirtingų tipų medicininių vaizdų apdorojimo straipsnių pasiskirstymas, kur tamsesnė spalva - *ViT* paremti straipsniai išleisti 2021 metais, o šviesesnė - konvoliuciniais tinklais nuo 2012 iki 2015 metų [11], [21]

Kituose skyreliuose apžvelgiama transformerio architektūra, kaip ji keičiasi norint ją pritaikyti vaizdų apdorojimui ir plačiau apžvelgiamas taikymas medicininių vaizdų apdorojimui.

1.2.1. Transformerio architektūros apžvalga

Transformerio modelis yra gylus neuroninis tinklas ir originalioje versijoje buvo sudarytas iš dviejų blokų – koduotojo (angl. *Encoder*) ir dekoduotojo (angl. *Decoder*). Originalios architektūros schema matoma **6 pav**.

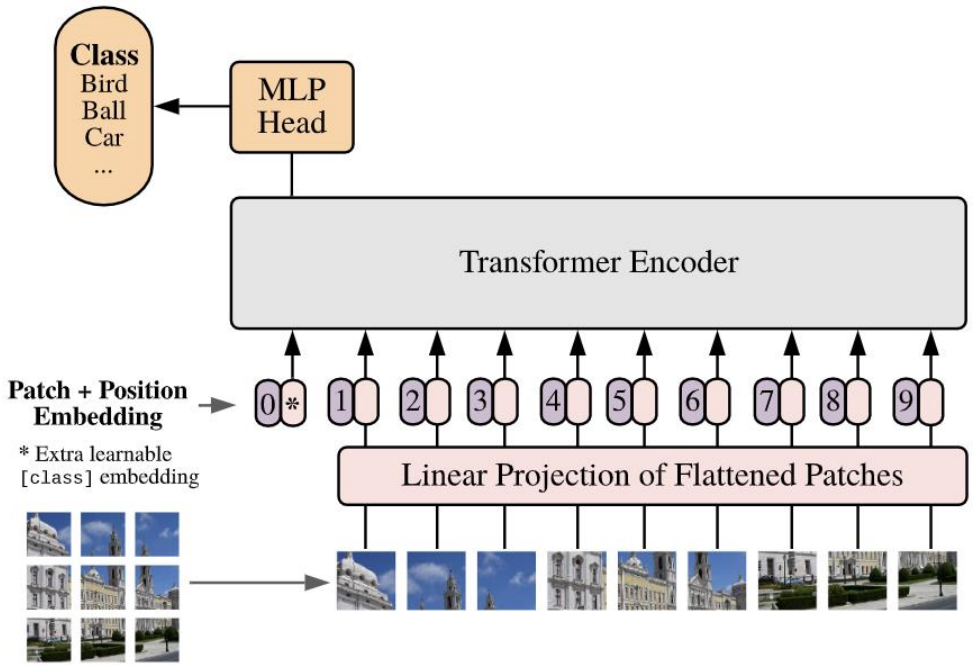


6 pav. Originali transformerio modelio architektūros schema [3]

Iš principo, kodavimo blokas koduoja įvesties seką $\{x_1 \dots x_n\}$ į tokio pačio ilgio išvesties seką $\{z_1 \dots z_n\}$, kuri paduodama dekodavimo blokui, kuris generuoja $\{y_1 \dots y_n\}$ išvestį kiekvienam elementui, naudodamas praėjusią išvestį, kaip dar vieną įvestį. Galima teigti, kad kodavimo blokai skirti analizuoti, kaip įvesties sekos dalys yra susijusios tarpusavyje, o dekodavimo blokai – atspėti koks „žetonas“ (angl. token), turėtų būti sekantis sekoje, imant kodavimo bloko sugeneruota kontekstinę informaciją.

Vaizdų klasifikavimo uždaviniams pagrinde naudojamas tik kodavimo blokas [4], [11], [19], [21], [22], nes užtenka klasifikuoti pagal kodavimo bloko išvestį ir dekodavimo bloko sekos žetonų spėjimas nereikalingas.

Vienas pirmųjų vaizdo klasifikavimo modelių, panaudojusių gryną transformerio modelį buvo ViT, aprašytas 2020 metais [4], jo architektūros schema matoma **7 pav.**

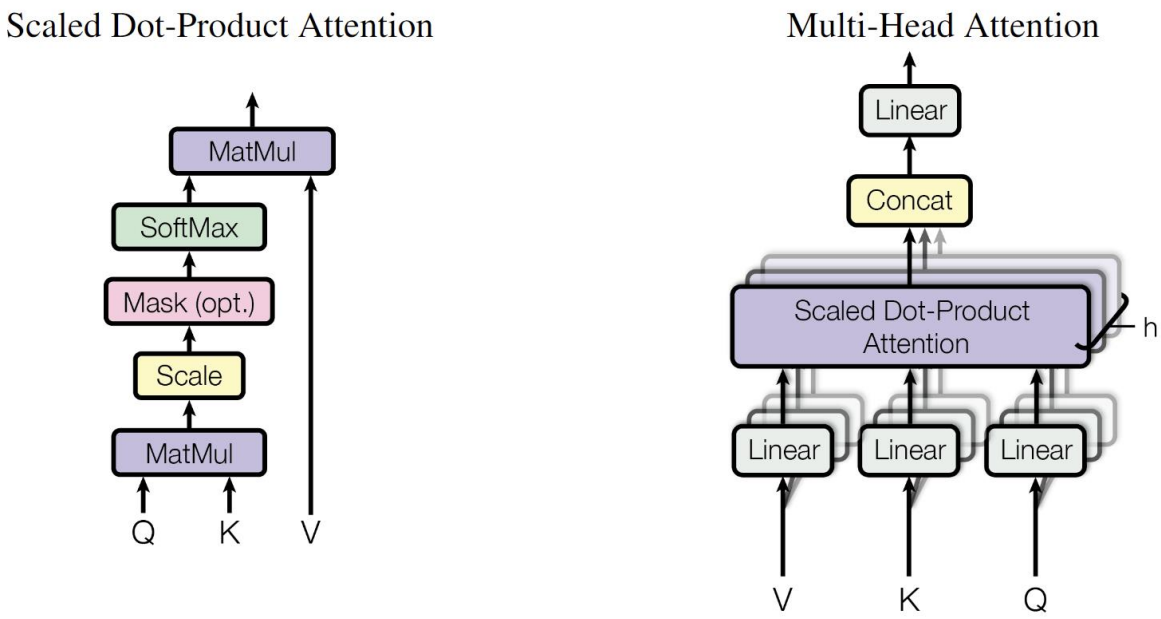


7 pav. ViT modelio schema [4]

ViT modelyje iš esmės keitėsi tik įvesties forma. Originaliai modeliui paduodamas 1D žetonų seka, bet tai buvo naudojama teksto vertimui [3]. Šiuo atveju paduodami suplokštinti 2D nuotraukų $x \in R^{H\times W\times C}$ lopinėliai $x_p \in R^{N\times(P^2\cdot C)}$, kur (H, W) – originalios nuotraukos rezoliucija, C – kanalų skaičius, (P, P) – lopinėlio rezoliucija, o $N = HW/P^2$ – lopinėlių skaičius, kuris taip pat yra transformerio įvesties sekos ilgis.

Vėliau, tokia įvestis paduodama į transformerio kodavimo bloką, kuris dažniausia sudarytas iš $N = 6$ identiškų sluoksnių, kuriuos sudaro daugiagalvis dėmesio sau (angl. *multi-head self-attention*) mechanizmas ir perdavimo (angl. *feed-forward*) tinklas su *ReLU* aktyvavimo funkcija [3].

Dėmesio mechanizmas, kurio schemos matomos **8 pav.**, yra viena pagrindinių šio modelio dalų. Dėmesio funkciją galima apibūdinti kaip užklausos ir rakto-reikšmių porų rinkinio susiejimą su išvestimi, kur užklausa, raktai, reikšmės ir išvestis yra vektoriai. Išvestis apskaičiuojama kaip svertinė reikšmių suma, kur kiekvienai reikšmei priskirtas svoris apskaičiuojamas pagal užklausos suderinamumo funkciją su atitinkamu raktu [3].



8 pav. kairėje - skalerinės sandaugos dėmesio mechanizmo schema; dešinėje - daugiagalvos dėmesio sau mechanizmo schema [3]

Skaliarinės sandaugos dėmesys gali būti aprašomas formule (žr. **2 lygtis**) kur Q – užklausų matrica, K ir V – raktų ir reikšmių matricos, o d_k – Q ir K matricų dydis. Daugiagalvis savęs dėmesys paraleliai h kartų naudoja dėmesio mechanizmą su skirtingomis išmoktomis projekcijomis W^Q, W^K, W^V (žr. **3 lygtis**) konkatonuoja jo išvestis kartu su projekcijos matrica W^O (žr. **4 lygtis**), ir transformuoja į reikiamo dydžio vektorių [3]. Intuityviai, tai leidžia paskirstyti dėmesį skirtingoms sekos dalims.

2 lygtis. Dėmesio apskaičiavimo formulė [3]

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

3 lygtis. "galvos" apskaičiavimo formulė daugiagalvio dėmesio sau mechanizme [3]

$$head = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

4 lygtis. daugiagalvio dėmėsiu sau apskaičiavimo formulė [3]

$$Multihead = concat(head_1 \dots head_h)W^O$$

Pasibaigus kodavimo procesui *ViT* atveju naudojamas daugiasluoksnis perceptronas galutinei paveikslėlio klasei nustatyti [4].

Kaip minėta prieš tai, kaip konvoliucinių tinklų atveju, jau yra nemažai skirtingu transformerio modelio pritaikymo variantų. *ViT* labiausiai atitinka originalią architektūrą ir iš principo turi visas dalis, kurios būtų naudojamos kitų, transformeriais paremtų, modelių.

1.2.2. Transformerių naudojimas medicininiams vaizdams klasifikuoti

Kaip minėta skyriaus pradžioje ir matoma *4* pav**.** susidomėjimas transformerių panaudojimu medicininių vaizdų apdorojimo ir atpažinimo tikslais nėra mažas, ir, kaip matoma *9* pav**.,** prieš tai aptartas *ViT* modelio panaudojimas standartiniams vaizdo duomenų rinkiniams klasifikuoti gali būti ypač tinkamas [4] ir jis gali atlikti klasifikavimą geriau, nei konvoliuciniais tinklais paremti modeliai, net jei skirtumas nėra didelis. Priedu prie to, transformerių privalumai, kaip greitesnis apsimokymas [3], [4], [10], gebėjimas juos plėsti (angl. *Scale*) [3] ir mažesnis induktyvus šališkumas (angl. *Inductive bias*) [4] daro juos patikimu pasirinkimu tolesniems tyrimams ir taikymams.

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

9 pav. *ViT* palyginimas su *RESNET* ir *EfficientNet* [4]

Vis dėl to, kaip ir konvoliucinių tinklų atveju, medicinių vaizdų apdorojimo atveju validžių didelių duomenų rinkinių trūkumas yra esminė problema. Pasak, kai kurių tyrėjų, ši problema dar aktualesnė naudojant transformerius, nei konvoliucinius tinklus, nes jų efektyvumas apmokant naudojant tik mažesnius duomenų rinkinius gali būti gerokai mažesnis, nei konvoliuciniais tinklais paremtų modelių, pvz., *RESNET* [10], [11], [21], [22], tai galima pamatyti ir **10 pav**., kur matomi 2021 metais atlikto tyrimo, lyginančio *RESNET50* ir transformeriais paremto *DEiT-S* modelių efektyvuma su skirtingomis pradinio apmokymo strategijomis ant kelių skirtingų medicininių vaizdų duomenų rinkinių.

Initialization	**Model**	**APTOS2019**, $\kappa \uparrow$	**ISIC2019**, Recall ↑	**DDSM**, ROC-AUC ↑
Random	ResNet50	0.849 ± 0.022	0.662 ± 0.018	0.917 ± 0.005
	DeiT-S	0.687 ± 0.017	0.579 ± 0.028	0.908 ± 0.015
ImageNet (supervised)	ResNet50	0.893 ± 0.004	0.810 ± 0.008	0.953 ± 0.008
	DeiT-S	0.896 ± 0.005	0.844 ± 0.021	0.947 ± 0.011
ImageNet (supervised) + Self-supervised with DINO [4]	ResNet50	0.894 ± 0.008	0.833 ± 0.007	0.955 ± 0.002
	DeiT-S	0.896 ± 0.010	0.853 ± 0.009	0.956 ± 0.002

10 pav. Konvoliucinio tinklo *RESNET50* ir transformerio modelio *DeiT-S* palyginimas naudojant skirtingas pradinio apmokymo strategijas. Rodomas galutinis efektyvumas medicininių vaizdų klasifikavimui [10]

Prieš tai minėtame šaltinyje bei paveikslėlyje taip pat matoma, kad transformerio modelių efektyvumas žymiai gerėja, jeigu jie prieš tai apmokomi ant didelių duomenų rinkinių, kaip *IMAGENET*. Tas pats yra tiesa ir konvoliuciniams tinklams, tačiau priklausomai nuo problemos baziniai transformerio privalumai gali žymiai praversti.

1.3. Skyriaus išvados

Konvoliuciniai neuroniniai tinklai jau daugiau nei 10 metų plačiai naudojami vaizdo apdorojimo uždaviniams, įskaitant ir vaizdų klasifikavimą. Medicininių vaizdų klasifikavimui jie taip pat plačiai paplitę, skirtingiems duomenų rinkiniams ir skirtingoms problemoms spręsti. Bandoma taikyti nemažai skirtingų modelių.

Konvoliucinių tinklų esminė architektūra labai tinka vaizdo klasifikavimo uždaviniams, nes ji skirta apdoroti kelių dimensijų matricas taip, kad atfiltruotumėm mums reikalingas ypatybes į mažesnę formą, nei mūsų įvestis ir prarastumėm kuo mažiau informacijos. Skirtingi modeliai taikantys tokius tinklus turi kitokį kiekį sluoksnių, kitokias strategijas duomenų apdorojimui, apmokymui ar pan., bet jų esmė lieka tokia pati.

Tokio tipo tinklai gali reikalauti ypač daug resursų (tiek duomenų, tiek kompiuterių), juos sunku „didinti“ (angl. *Scale*) ir, nors jie tinka vaizdų apdorojimo uždaviniams, dėl nedidelių filtrų ir surinkimo sluoksnio savybių jiems sunku išlaikyti santykį tarp nutolusių ypatybių, kas gali paveikti jų efektyvumą. Yra įvairių būdų tai spręsti naudojant kitokius konvoliucinų tinklų modelius, tačiau tokios problemas potencialiai sprendžia ir visai kitos architektūros tinklai – transformeriai.

Transformeriai vaizdo klasifikavimui pradėti naudoti visai nesenai, bet turi aiškių privalumų, kaip greitesnis apmokymas, geresnis sąsajų tarp nutolusių ypatybių išsaugojimas. Juos veikia panašios problemos, kaip konvoliucinius tinklus, ypač kalbant apie medicininių vaizdų klasifikavimą, bet yra įrodymų, kad jau naudotos strategijos, kaip perduodamas mokymasis, padeda ir transformeriams [10].

Priedu prie to, nors buvo aptarta tik bazinė transformerio architektūra buvo minėta, kad naudojamos ir hibridinės architektūros [4], [11], [18], [21]. Taikant tokias architektūras galimai įmanoma pasiekti geresnį efektyvumą su mažiau duomenų, bet išlaikyti transformerio daugiagalvio dėmesio sau mechanizmo privalumus.

2.MEDICININIŲ VAIZDŲ KLASIFIKAVIMUI NAUDOJAMŲ MOKYMO BŪDŲ APTARIMAS

Šiame skyriuje trumpai aptariame keli pagrindiniai būdai, kaip mokomi klasifikavimo modeliai ir kaip jei taikomi medicininių vaizdų analizės srityje. Konkrečiai, aptariamas prižiūrimas mokymas ir perkeliamas mokymas.

Perkeliamas mokymas dažnai nelaikomas atskiru mokymo būdų, o labiau papildoma metodika, tačiau šio darbo tikslui tai patariama tame pačiame skyriuje, nes metodai bus lyginami vėliau.

Klasifikavimo uždaviniams metodai, kaip neprižiūrimas mokymas, priešpriešinis mokymas ar kiti dažniausiai nėra taikomi, nes mes siekiame išmokyti modelį priskirti klasę vaizdui ir tai efektyviausia leidžiantis modeliui keisti svorius pagal teisingą atsakymą.

2.1. Prižiūrimas mokymas

Prižiūrimas mokymas yra vienas iš pagrindinių mašininio mokymosi šaknų, kurio tikslas – mokyti modelį prognozuoti arba klasifikuoti duomenis, naudojant priežiūrą arba pavyzdinius duomenis su jau žinomomis reikšmėmis. Klasifikavimo uždaviniams ši mokymo strategija dažniausiai naudojama [23], nes mes norime, kad modelis išmoktų nustatyti konkrečią klasę vaizdui, o ne išgalvotų ją pats, ypač kai tai nėra binarinis klasifikavimas (turintis tik dvi galimas klases).

Prižiūrimas mokymas reikalauja, kad duomenų rinkiniai, kurie naudojami mokymo procese mokymui ir validacijai, turėtų etiketes (klasių identifikatorius) ir kad jie būtų paduodami modeliui įvestyje. Medicininių vaizdų apdorojimo srityje surinkti pačias nuotraukas duomenų rinkiniams yra sudėtinga, o gauti teisingas etiketes dar sudėtingiau, nes tam reikia atitinkamos medicinos srities eksperto/ekspertų. Dėl šių priežasčių paruošti rinkiniai dažniausiai būna sąlyginai maži (keli šimtai ar keli tūkstančiai nuotraukų). Priedu, tokie vaizdai ir ypatybės, kurias modelis turi išgauti, pakankamai sudėtingi ir modeliams išmokti gerai klasifikuoti naudojant tik prižiūrimą mokymą ir tokius duomenų rinkinius gali būti labai sunku. Dėl to, prižiūrimas mokymas gali būti kombinuojamas su kitu metodu – perkeliamu mokymu, kuriam galima naudoti didelius duomenų rinkinius, kaip *IMAGENET,* kuriuos surinkti paprasčiau ir jų klasifikavimui nereikia srities ekspertų.

Perkeliamas mokymas plačiau aprašomas kitame skyrelyje, bet šiame darbe išbandomas ir paprastas prižiūrimas mokymas.

2.2. Perkeliamas mokymas

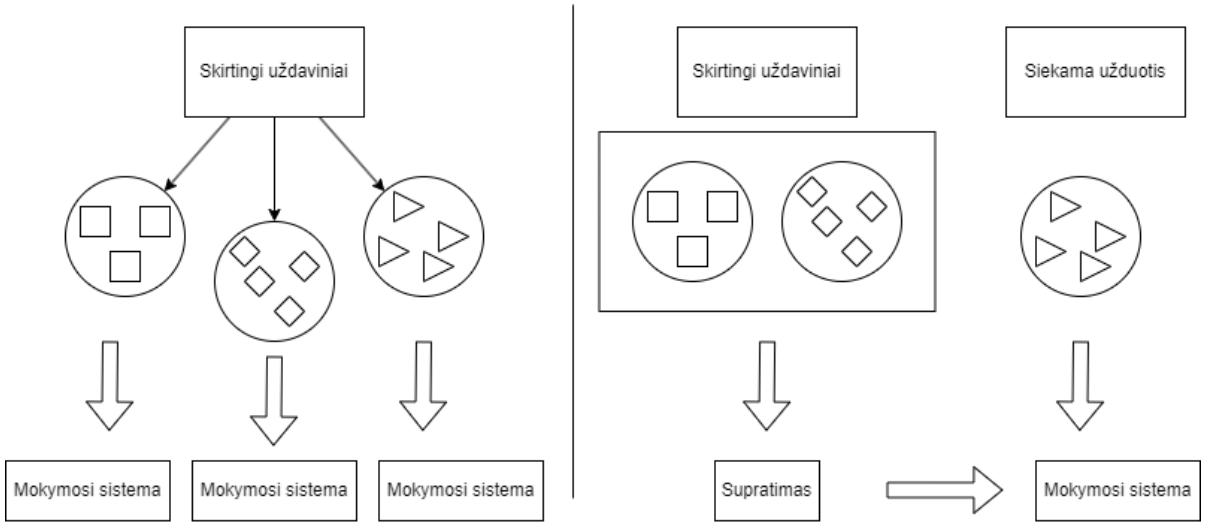
Perkeliamas mokymasis nėra naujiena mašininio mokymo srityje ir ši metodika sulaukė didesnio susidomėjimo jau nuo 1995 metų [24]. Dabar šis mokymo tipas kartais kategorizuojamas į

induktyvų, transduktyvų ir neprižiūrimą perdavimą ir į dar smulkesnes kategorijas [24], [25] priklausomai nuo to, ar pradiniai ir naujos užduoties duomenys, išvestys ir reikalingos ypatybės yra panašios, bet, nors kategorijos egzistuoja, skaitant literatūrą atrodo, kad taip klasifikuoti perkeliamą mokymą nėra labai populiaru.

Šiame skyriuje plačiau apžvelgiami skirtumai tarp tradicinio mašininio mokymo ir perkeliamo mokymo, plačiau aprašomas apibrėžimas ir pagrindiniai metodai.

2.2.1. Skirtumai tarp perkeliamo mokymo ir tradicinio mašininio mokymo

Tradicinėse mokymo technikose modeliai mokomi vienam uždaviniui nuo pradžių. Kaip pavyzdį galima naudoti *MNIST* ranka rašytų skaitmenų klasifikavimo uždavinį, kur mes turime 60000 ranka rašytų skaitmenų nuotraukų ir 10 klasių (nuo 0 iki 9), kurias norime suklasifikuoti nuotraukas. Modelis pradėtų nuo atsitiktinių svorių ir besimokydamas naudojant duomenų rinkinį juos naujintų ir gerintų. Galiausiai, turėtumėme modelį gebanti klasifikuoti tik duomenų rinkinį ir nuotraukas panašias į jį. Tai iliustruoja **11 pav.** Standartinis mašininio mokymo procesas (kairėje) ir procesas, naudojojant perkeliamą mokymą (dešinėje) (Paveikslėlis kairėje esanti diagrama.



11 pav. Standartinis mašininio mokymo procesas (kairėje) ir procesas, naudojojant perkeliamą mokymą (dešinėje) (Paveikslėlis paremtas Sinno Jialin Pan ir Qiang Yang perkeliamo mokymo tyrimo medžiaga [24])

Iš kitos pusės, kaip matoma **11 pav.** Standartinis mašininio mokymo procesas (kairėje) ir procesas, naudojojant perkeliamą mokymą (dešinėje) (Paveikslėlis dešinėje pusėje, perkeliamas mokymas naudoja jau apmokytą modelį, dažniausiai ant didelių duomenų rinkinių, kaip *IMAGENET*.

Pasak straipsnių apie perkeliamo mokymo naudojimą [26], [27], [28], mokymui naudojamos modelio aukštesniuose sluoksniuose išmoktos bendros ypatybės, kurias galima pritaikyti ir kitokiems duomenims, o jo paskutinis sluoksnis dažniausiai pakeičiamas nauju su tiek klasių, kiek reikia naujai užduočiai. Tuomet aukštesni sluoksniai „šaldomi“, t. y. nebus keičiami mokymo metu ir mes

mokysime tik pridėtą naują paskutinį sluoksnį ant naujos užduoties duomenų. Matematiškai, tai galima apibrėžti pagal [24], taip: Imant šaltinio domeną D_S ir mokymo užduotį T_S, siekiamą domeną D_T ir mokymo užduotį T_T, perkeliamas mokymas siekia pagerinti siekiamos prognozavimo funkcijos $f_T(\cdot)$ mokymą ir D_T su žiniomis iš D_S ir T_S, kur $D_S \neq D_T, arba\ T_S \neq T_T$. Domeną galima suprasti, kaip ypatybes X ir tikimybes P(X), o užduotis, kaip etiketes y ir objektyvią prognozavimo funkciją $f_{\square}(\cdot)$.

Medicininių vaizdų klasifikavimo atveju naudojant perkeliamą mokymą pagrinde naudosime bazinius modelius, kurių domenas sutampa su mūsų užduotimi, t. y. bazinis modelis bus apmokytas klasifikuoti vaizdus.

2.2.2. Perkeliamo mokymosi metodai

Perkeliamas mokymas gali būti skirstomas į du bendrinius metodus – ypatybių ištraukimą (ang. *Feature extraction*) ir tikslų derinimą (ang. *Fine-tuning*). Kiekvienas iš jų dar gali būti skaidomas į dvi subkategorijas, kurios matomos ***12*** **pav.** Keturi perdavimo mokymo būdai: žaliai pažymėti mokomi sluoksniai, baltai – užšaldyti. MM –mašininio mokymo, PS – pilnai sujungtas sluoksnis. Paveikslėlis [29].

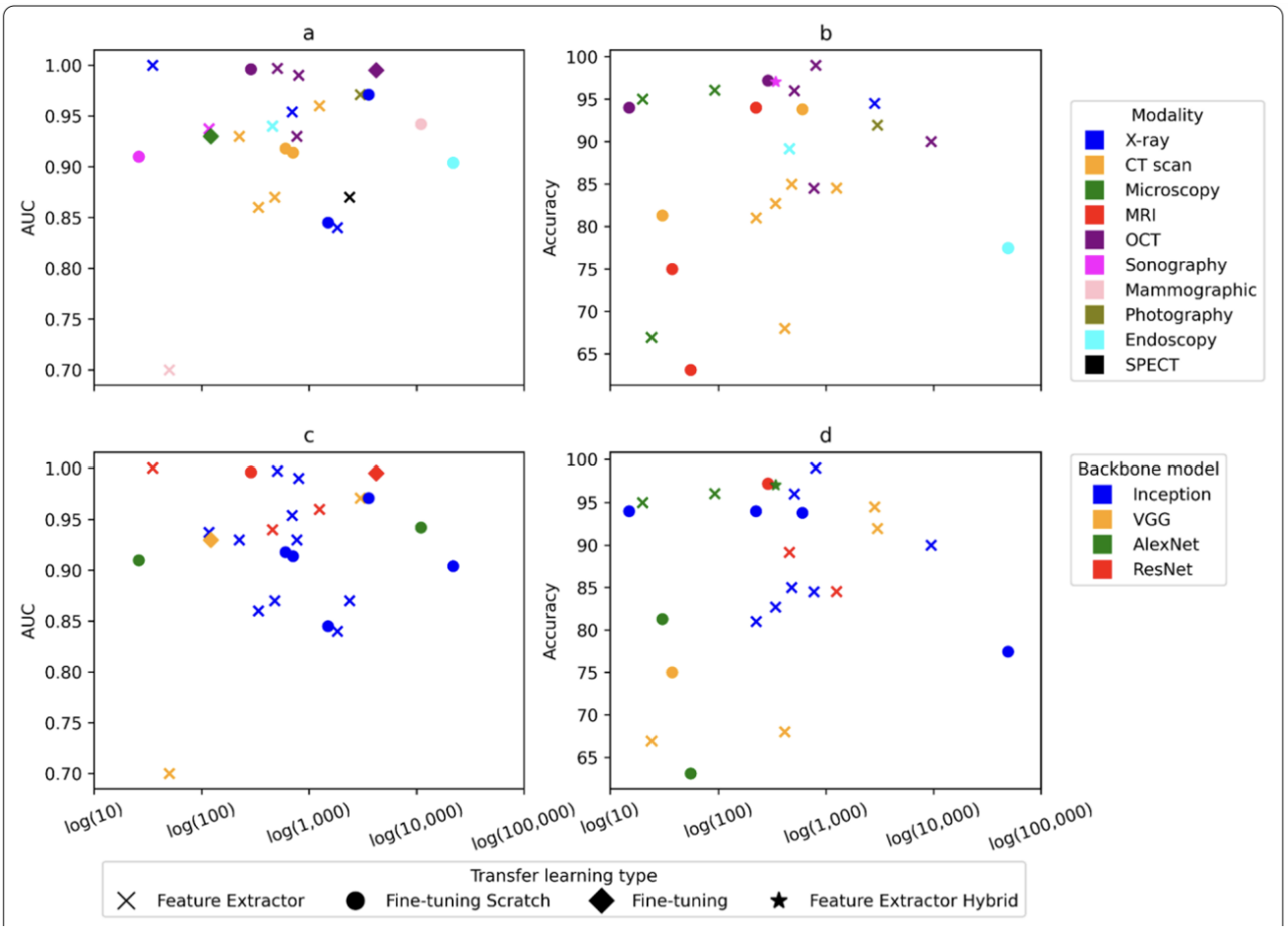
12 pav. Keturi perdavimo mokymo būdai: žaliai pažymėti mokomi sluoksniai, baltai – užšaldyti. MM – mašininio mokymo, PS – pilnai sujungtas sluoksnis. Paveikslėlis paremtas iliustracija iš [29]

Paveikslėlio dešinėje pusėje ypatybių ištraukimo metodai, kurie paremti tuo, kad bazinio, apmokyto modelio svoriai „šaldomi“ ir nebus keičiami, kai modelį mokysime su mažesniais duomenų rinkiniais. Priedu prie to, iš bazinio modelio bus išimti paskutiniai klasifikavimo sluoksniai ir vietoj jų bus pridėtas arba mažas modeliais su konvoliuciniais sluoksniais arba pilnai sujungtas sluoksnis klasifikavimui. Kad ir kuris variantas pasirinktas mokant su naujais duomenimis bus mokomi tik nauji paskutiniai sluoksniai.

Šis variantas geriau veikia, kai pradiniai ir nauji duomenų rinkiniai panašūs, nes tada prieš tai buvusios ypatybės geriau pritaikomos naujiems duomenims, bet ypatybių ištraukimo metodai reikalauja gerokai mažiau resursų, dėl to pasak Hee E. Kim et al perkeliamo mokymo medicininių

vaizdų klasifikavimui literatūros analizės [29] dažniausiai rekomenduojama pradėti nuo šio metodo, ir tik tada bandyti daugiau resursų reikalaujančius metodus.

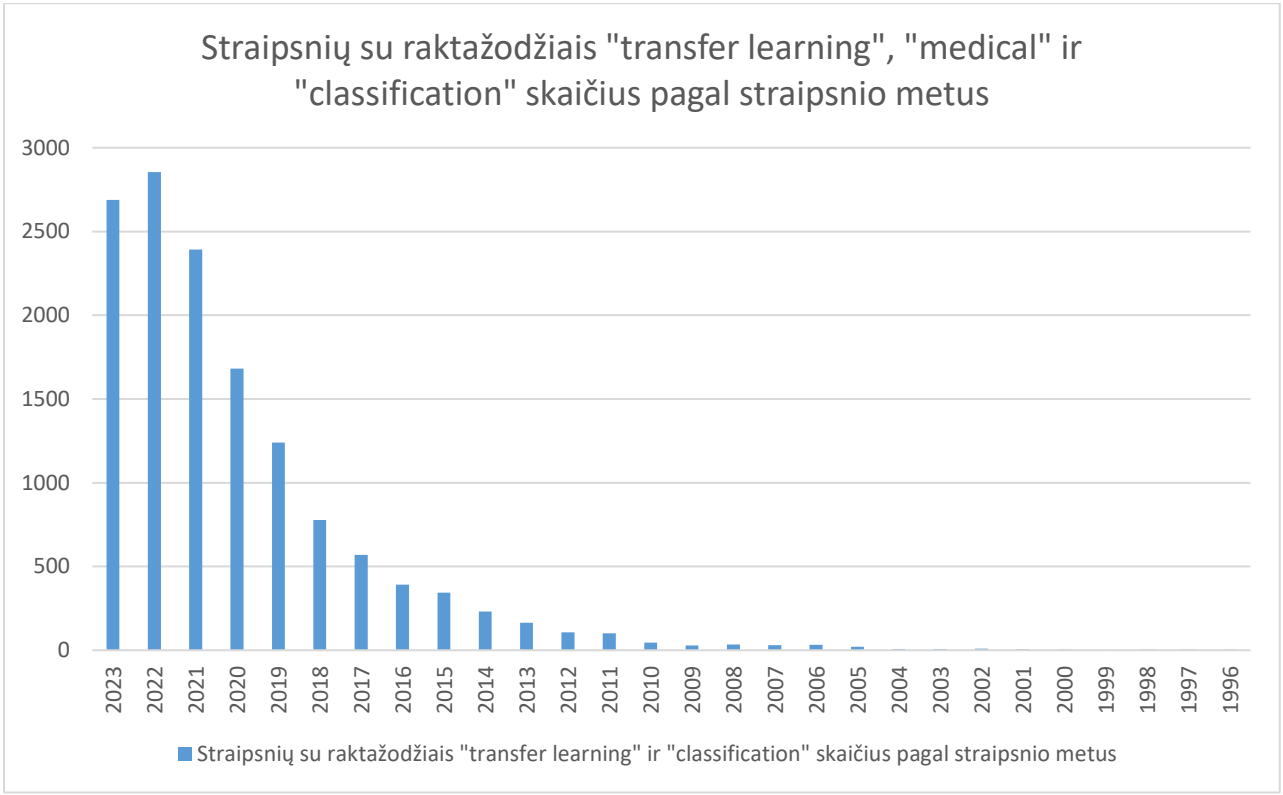
Kaip matoma **13 pav.**, medicininių vaizdų klasifikavimui, bent pagal straipsnių analizę [29], autoriai dažniausiai renkasi ypatybių ištraukimą, o antroje vietoje – tikslų derinimą nuo pradžių, t. y. sunkiausią resursų prasme metodą, bet apmokanti visą modelį, dėl to galinti labiau prisitaikyti.



13 pav. medicininių vaizdų klasifikavimo straipsnių rezultatai iš 2022 metų straipsnių analizės [29] pagal vaizdų tipą (a ir b), naudotus apmokytus modelius (c ir d) ir perkeliamo mokymo tipą (X – ypatybių ištraukimas; ● – tikslus derinimas nuo pradžių; ◆ – tikslus derinimas; ⋆ – hibridinis ypatybių ištraukimas)

2.2.3. Perkeliamas mokymas medicininių vaizdų klasifikavimui

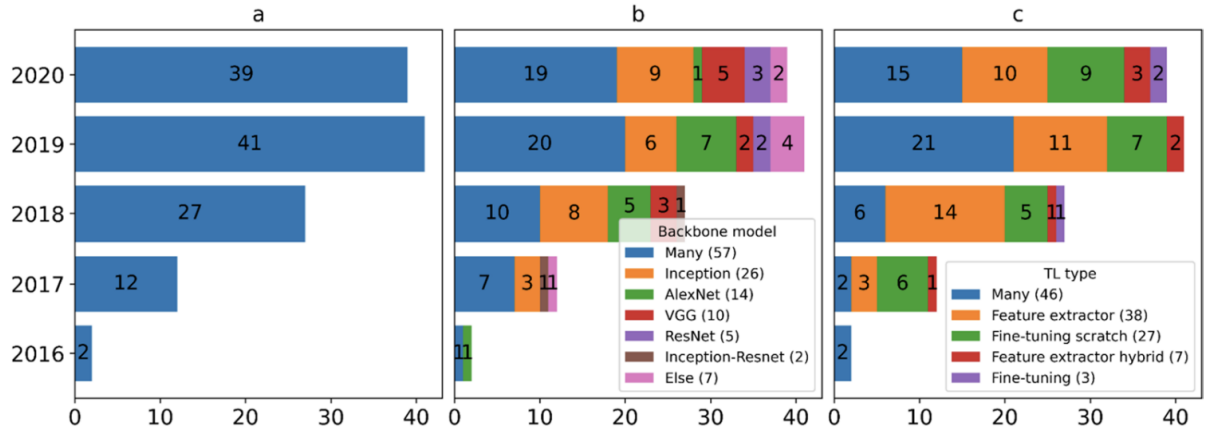
Sunku surasti kada išaugo susidomėjimas perkeliamu mokymusi, bet, pasak Litjens G. Et al [30], jau 2016 metais medicininių vaizdų klasifikavimo srityje susidomėjimas augo, o pasak Wang L. et al [31] dar labiau sustiprėjo dėl COVID-19 pandemijos 2020 metais. Pagal paieškos pagal raktažodžius „*Transfer learning*“, „*Medical*“ ir „*Classification*“ duomenis iš *PubMed* [32] nuo 1996 iki 2023 metų, kurie matomi **14 pav.** panašu, kad teiginiai teisingi ir referato rašymo metu ir straipsnių perkeliamo mokymo tema, skaičius iki šiol tik auga ir bent su pasirinktais paieškos kriterijais 2023 metų gruodį straipsnių skaičius jau siekė daugiau, nei septynis tūkstančius.



14 pav. Straipsnių kiekis pagal paieškos raktažodžius „*transfer learning*", „*medical*“ ir „*classification*" iš *PubMed* [32] svetainės pagal straipsnio leidimo metus, surinkta 2023-12-16

Nors naudojimas ir susidomėjimas pasak straipsnių autorių ir straipsnių kiekio *PubMed* auga, nedaug autorių lygina ypatybių ištraukimo ir tikslaus derinimo rezultatus [30] arba perkeliamo mokymosi metodas buvo taikomas savavališkai [31]. Šiame skyriuje bandoma plačiau aptarti metodo taikymą šiam uždaviniui ir, kiek įmanoma, aptarti palyginimus tarp ypatybių ištraukimo, tikslaus derinimo ir modelio mokymo nuo pradžių.

Medicininių vaizdų analizei, pagal 2022 Kim H. et al atlikta literatūros analizę [29], ir, kaip matoma **15 pav.** vieno bazinio modelio su geriausiais rezultatais visoms užduotims nėra ir autoriai dažnai naudoja skirtingus.



15 pav. Grafikas, rodantis mokslinių tyrimų kiekį per metus (a), naudojant skirtingus bazinius modelius (b) ir skirtingus perkeliamo mokymo tipus (c) [29]

Imant kelis straipsnius palyginimui, net tikrinant tuos pačius modelius skirtingoms užduotims, skirtingi autoriai gavo skirtingus rezultatus. Kaip matoma **16 pav.** ir **17 pav.**, vienu atveju geriausią rezultatą autoriai pasiekė naudodami *InceptionV3* modelį su perkeliamu mokymu [33], kitu – *VGG19* [34]. *VGG* yra mažesnis modelis, nei modeliai, kaip *Inception* ar *RESNET*, dėl to geri rezultatai naudojant jį yra ypač aktualūs, bet nedaug autorių jį rinkosi palyginimui [29]. Antony J. et al [35] taip pat rinkosi *VGG* kelio ostioartritio lygio detektavimui su tiksliniu derinimu ir pasiekė geresnių rezultatų, kai kuriais atvejais siekiančių 99%, nei modelis be šio metodo.

Model	Specificity	Sensitivity	Accuracy	AUC	F1
ResNet50	88.74%	77.39%	84.94%	0.91	0.78
Xception	87.16%	77.44%	84.06%	0.90	0.76
InceptionV3	**89.06%**	**77.44%**	**85.13%**	**0.91**	**0.78**
CNN3	79.22%	63.19%	74.44%	0.78	0.60
Traditional model	74.61%	58.10%	70.55%	0.73	0.49

16 pav. Perkeliamam mokymui naudotų modelių ir tradicinio konvoliucinio tinklo be perkeliamo mokymo rezultatų palyginimas krūties navikų ieškojimui [33]

Classifier	Class	Precision	Recall	F1 score	Testing Accuracy (%)
VGG16	COVID	0.923	0.900	0.911	88.57
	Normal	0.855	0.940	0.895	
	Pneumonia	0.891	0.820	0.854	
VGG19	COVID	0.950	0.950	0.950	**89.3**
	Normal	0.800	0.960	0.873	
	Pneumonia	0.975	0.780	0.866	
ResNet50	COVID	0.654	0.850	0.739	64.3
	Normal	0.673	0.660	0.666	
	Pneumonia	0.590	0.460	0.517	
ResNet50V2	COVID	0.846	0.825	0.835	72.85
	Normal	0.650	0.780	0.709	
	Pneumonia	0.732	0.600	0.659	
ResNet101	COVID	0.682	0.750	0.714	65.0
	Normal	0.702	0.660	0.680	
	Pneumonia	0.571	0.560	0.565	
ResNet101V2	COVID	0.837	0.900	0.867	84.28
	Normal	0.800	0.880	0.838	
	Pneumonia	0.905	0.760	0.826	
ResNet152	COVID	0.733	0.550	0.628	62.85
	Normal	0.654	0.680	0.666	
	Pneumonia	0.552	0.640	0.593	
ResNet152V2	COVID	0.875	0.875	0.875	77.14
	Normal	0.662	0.900	0.762	
	Pneumonia	0.875	0.560	0.683	
DenseNet121	COVID	0.680	0.850	0.756	71.43
	Normal	0.735	0.720	0.727	
	Pneumonia	0.732	0.600	0.659	
DenseNet169	COVID	0.707	0.725	0.716	71.42
	Normal	0.689	0.840	0.757	
	Pneumonia	0.763	0.580	0.659	
DenseNet201	COVID	0.805	0.825	0.814	75.71
	Normal	0.714	0.800	0.755	
	Pneumonia	0.767	0.660	0.709	
MobileNetV1	COVID	0.733	0.825	0.776	61.43
	Normal	0.541	0.800	0.645	
	Pneumonia	0.619	0.260	0.366	
XeptionNet	COVID	0.735	0.625	0.675	71.43
	Normal	0.733	0.880	0.799	
	Pneumonia	0.674	0.620	0.646	
InceptionV3	COVID	0.745	0.875	0.805	82.14
	Normal	0.833	0.800	0.816	
	Pneumonia	0.889	0.800	0.842	
InceptionResNetV2	COVID	0.652	0.375	0.476	62.14
	Normal	0.691	0.760	0.724	
	Pneumonia	0.548	0.680	0.607	

17 pav. Perkeliamam mokymui naudotų modelių palyginimas COVID-19 identifikavimui iš rentgeno nuotraukų [34]

Iš straipsnių, kurie lygino tradicinių mokymą ir perkeliamą mokymą, kaip prieš tai buvusiame **16 pav.** ir Badaqi T. et al straipsnyje apie akių ligų klasifikavimą [36] ir **18 pav.** matoma, kad

perkeliamo mokymo naudojimas su tinkamai pasirinktu modeliu gali žymiai pagerinti klasifikavimo rezultatus.

Model	Class	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
CNN	cataract	87	96	91	84
	Diabetic retinopathy	100	100	100	
	Glaucoma	67	85	75	
	Normal	86	56	68	
Transfer Learning	cataract	97	96	96	94
	Diabetic retinopathy	100	99	99	
	Glaucoma	94	85	89	
	Normal	86	96	91	

18 pav. Konvoliucinio modelio mokyto tradiciškai ir mokyto naudojant perkeliamą mokymą palyginimas akių lygų klasifikavimo uždaviniui, kur matoma, kad perkeliamas mokymas gali pagerinti modelių rezultatus per daugiau nei 10% [36]

Tačiau iš rastų straipsnių, ir, kaip paminėta [29], kartais sunku suprasti, kokią konfigūraciją autoriai naudojo, kodėl tokią pasirinko ir panašiai. Badaqi T. et al, pavyzdžiui, neminėjo, su kokiais duomenimis apmokė bazinį tinklą, kitur nebuvo minima, kodėl buvo pasirinktas vienas bazinis rinkinys ar tinklas.

2.3. Skyriaus išvados

Šiame skyriuje buvo aptartas pagrindinis klasifikavimo užduočių mokymo metodas – prižiūrimas mokymas, bei perkeliamo mokymo metodika.

Prižiūrimas mokymas yra dažniausiai naudojama mokymo būdas klasifikavimo uždaviniams, ypač kai turime daugiau nei dvi klases, nes mes modeliui mokymo metu paduodame ir klasę, kurią jis turėjo nustatyti ir pagal tai, ką modelis nustatė ir ką turėjo nustatyti gali būti koreguojami svoriai. Pačiam akių ligų klasifikavimui iš tinklainės vaizdų tai taip pat tinkanti metodika, bet dėl to, kad medicininių vaizdų duomenų rinkiniai neretais atvejais yra labai maži vien šio mokymo būdo gali nepakakti.

Perkeliamas mokymas yra kita metodika, kur naudojamas jau apmokytas modelis. Turint modelį, kuris apmokytas klasifikuoti didelį duomenų rinkinį, kaip *IMAGENET*, mes galime jį pritaikyti prie kitokio duomenų rinkinio leisdami jam mokytis su labai nedideliu mokymosi greičiu. Tai gali būti ypač naudingo šio darbo užduočiai, nes mes pradėsime ne nuo atsitiktinai parinktų svorių, o jau nuo modelio gebančio surasti konkrečias ypatybes vaizduose.

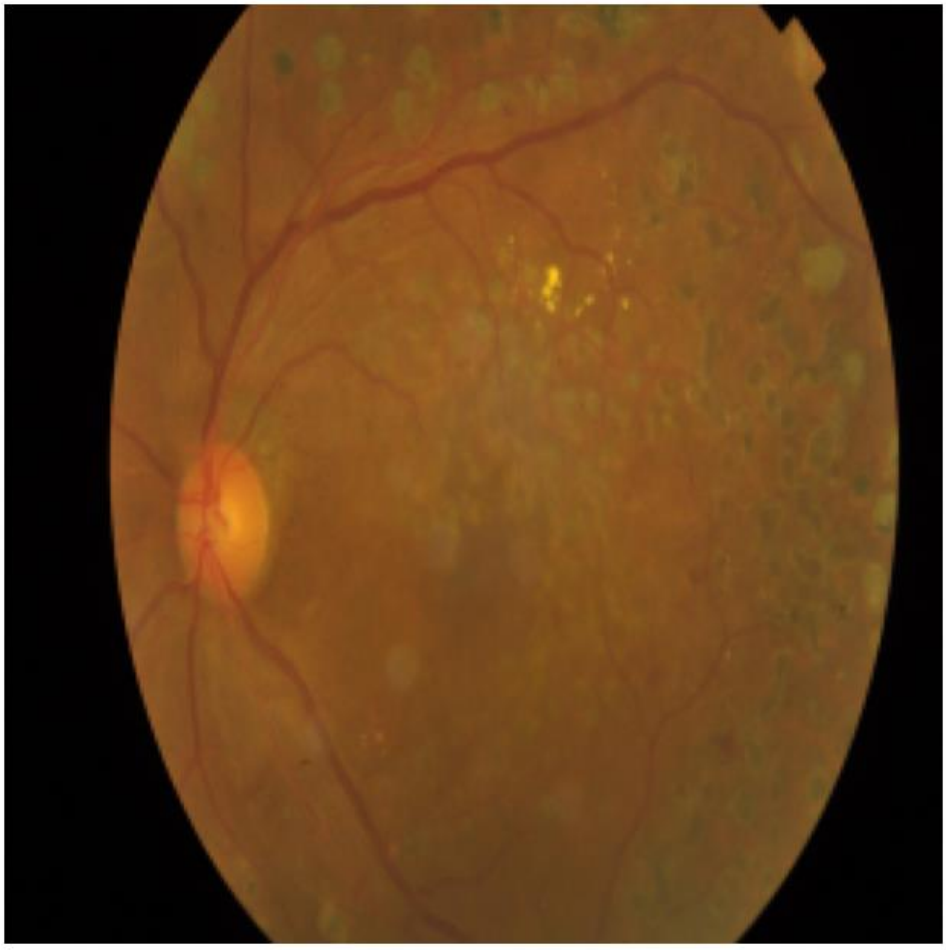
3.DUOMENŲ RINKINIO IR METODŲ ATRANKA

Darbui atlikti buvo naudotas *APTOS 2019* [37] duomenų rinkinys. Toliau skyriuje aptariamas pats duomenų rinkinys, kokie išankstinio apdorojimo metodai buvo išbandyti ir galų gale pritaikyti.

3.1. APTOS 2019

APTOS 2019 (angl. *Asia Pacific Tele-Ophthalmology Society*) duomenų rinkinys yra naudojamas moksliniuose tyrimuose, siekiant nustatyti ir klasifikuoti diabetinę retinopatiją [37]. Šis duomenų rinkinys buvo pagrindinis duomenų šaltinis 2019 metų „Aklumo aptikimo“ konkurse, kuris buvo vykdomas populiariame dirbtinio intelekto konkursų platformoje *Kaggle*.

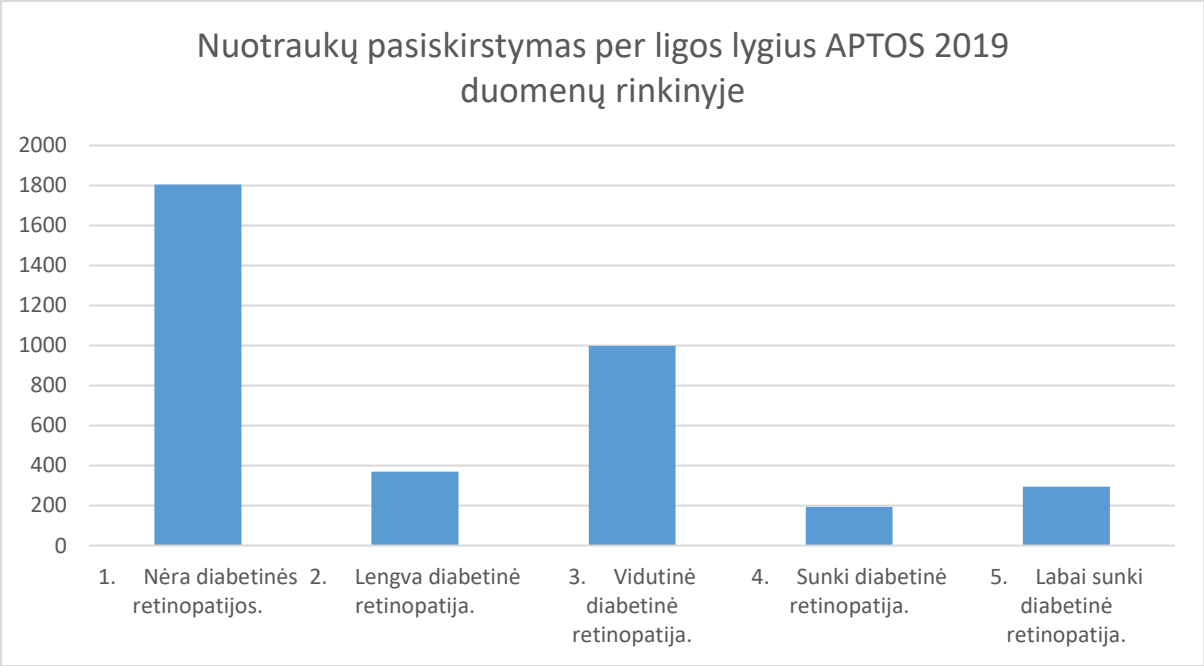
Duomenų rinkinys susideda iš akių dugno nuotraukų, kurios buvo daromos naudojantis funduskameromis. Nuotraukos varijuoja kokybe ir aiškumu, nes jos buvo darytos skirtingose sąlygose ir naudojantis įvairia įranga, kas yra būdinga tikrovės klinikinėms aplinkybėms, bet dauguma nuotraukų rinkinyje yra geros kokybės ir yra ryškios ir aiškios, kad pagrindinės reikalingos detalės, kaip kraujagyslės ir jų anomalijos, gerai matomos. Nuotraukos pavyzdys matomas **19 pav**.



19 pav. *APTOS 2019* duomenų rinkinio nuotraukos pavyzdys

Kiekviena nuotrauka mokymo rinkinyje buvo žmogaus eksperto įvertinta ir priskirta vienai iš penkių kategorijų, kurių pasiskirstymas rinkinyje matomas Error! Reference source not found. **pav.**, p riklausomai nuo diabetinės retinopatijos progresijos laipsnio:

1. Nėra diabetinės retinopatijos;
2. Lengva diabetinė retinopatija;
3. Vidutinė diabetinė retinopatija;
4. Sunki diabetinė retinopatija;
5. Labai sunki diabetinė retinopatija.



20 pav. *APTOS 2019* duomenų rinkinio nuotraukų pasiskirstymas per klases/ ligos lygius

Mokymo rinkinyje, kuris turi klasių etiketes nustatytas ekspertų, yra 3662 nuotraukos. Šiame darbe buvo naudota tik ši rinkinio dalis, nes likusios nuotraukos nėra klasifikuotos (neturi paruoštų etikečių) ir nebūtų buvusios galimybės įvertinti, ar modelis klasifikavo nuotraukas teisingai ar ne.

3.1.1. Išankstinio apdorojimo metodai

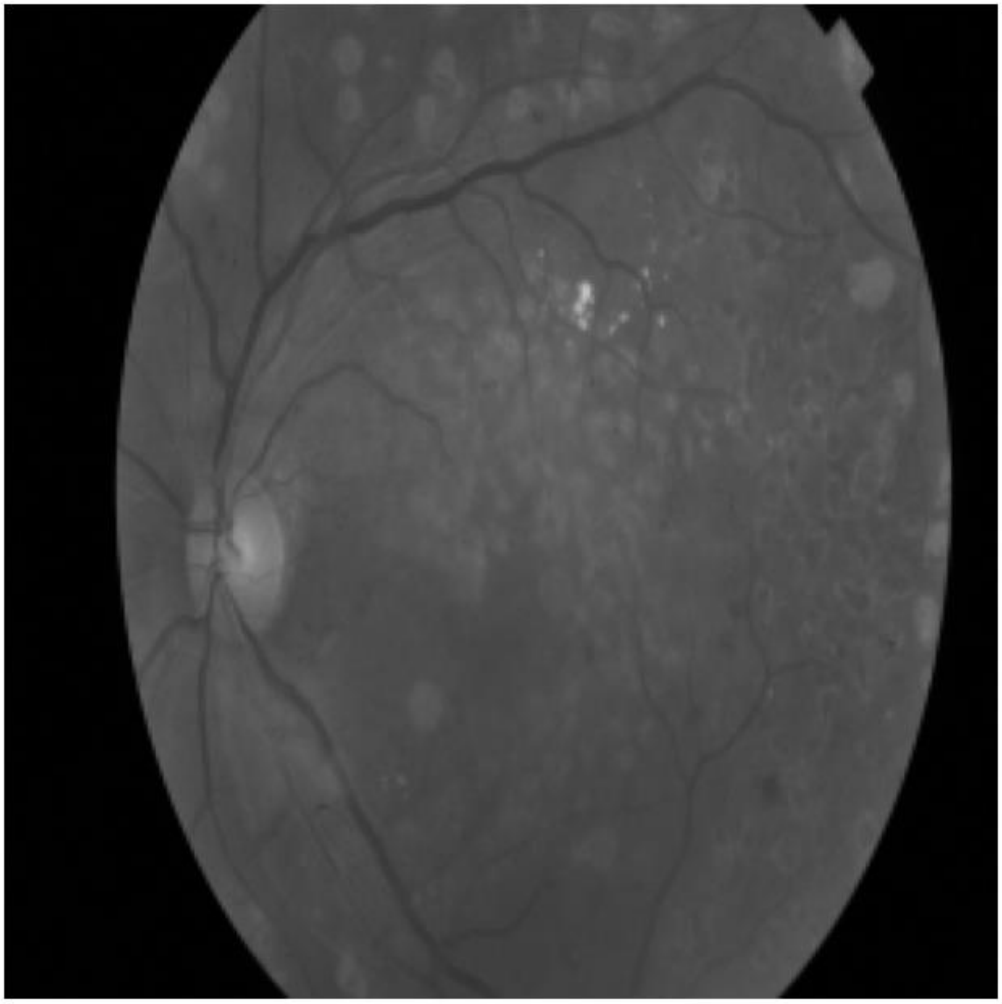
Išankstinis apdorojimas vaizdų klasifikavimo uždaviniuose – ypač svarbus žingsnis. Apdorojimas visų pirma paruošia duomenis modeliams, kurie dažniausiai sukurti priimti tik konkretaus dydžio bei formato vaizdus ir vėliau kitos technikos gali būti naudojamos siekiant išryškinti ypatybes, į kurias reikia atkreipti dėmesį sprendžiant užduotį.

Pirminiai ir paprasčiausi apdorojimo metodai – nuotraukų dydžio (ilgio ir pločio) suvienodinimas ir normalizacija. Vaizdo ilgio ir pločio pasirinkimas gali priklausyti nuo pačio

duomenų rinkinio, turimų resursų kiekio arba nuo naudojamo modelio. Šiame darbe dydžiai buvo pasirinkti pagal tai, kokiam vaizdo dydžiui buvo pritaikyti naudoti modeliai.

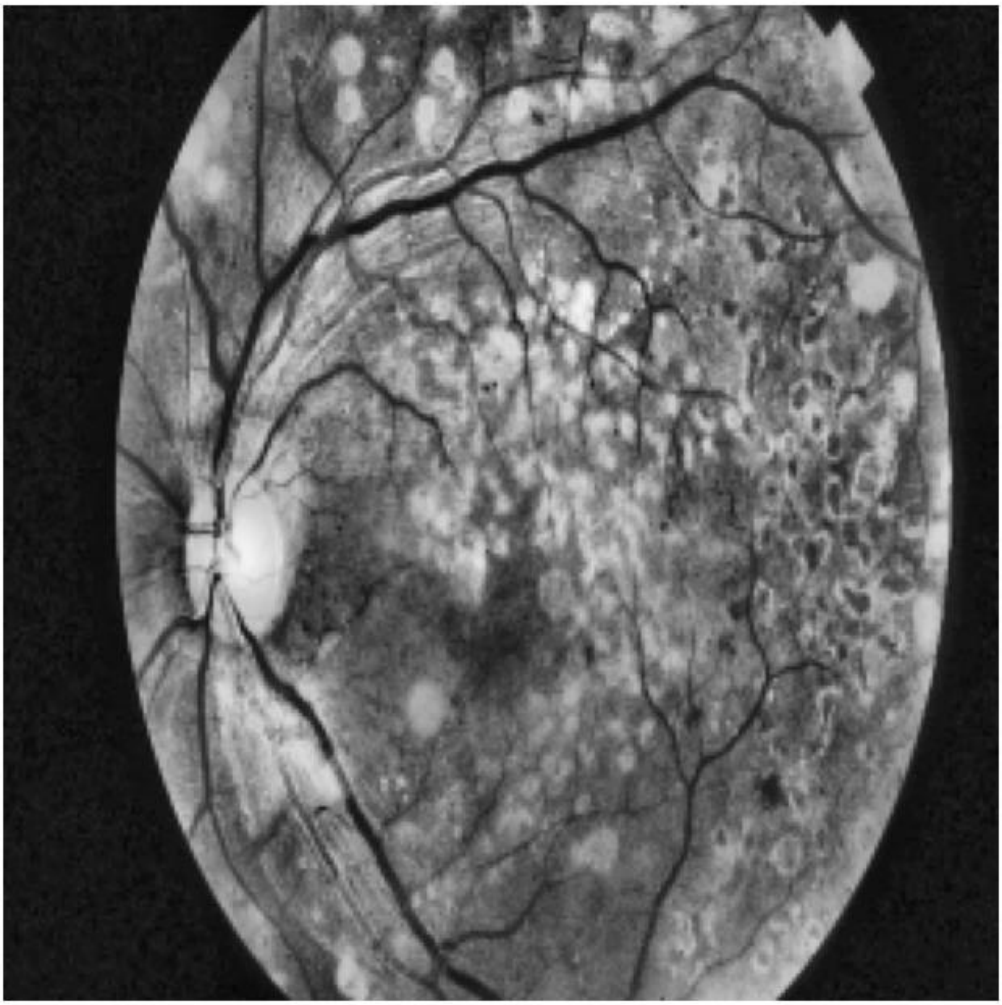
Normalizacija naudojama pakeisti pikselių reikšmes iš sveikųjų skaičių (0 – 255 reikšmių) į realiuosius skaičius (0 – 1), siekiant greitesnio skaičiavimo ir tikslumo, ypač naudojant aktyvavimo funkcijas, kaip sigmoid, kuri pritraukia įvestis į rėžį nuo 0 iki 1.

Visiems šiame darbe atliktiems bandymams buvo taikyti šie išankstinio apdorojimo metodai, bet taip pat buvo atlikti bandymai su metodais, skirtais „pagerinti vaizdus“, tai yra išryškinti ypatybes, į kurias modelis turėtų atkreipti dėmesį bandant klasifikuoti ligos lygį. Konkrečiai diabetinės retinopatijos atveju pagrindiniai ligos požymiai matosi tinklainės nuotraukų kraujagyslėse [38]. Norint išryškinti šias nuotraukų vietas galima ištraukti iš nuotraukų tik žaliąjį spalvų kanalą, taip dažniausiai raudonos kraujagyslės matysis geriausiai [39] ir tokios nuotraukos pavyzdys matomas **21 pav.** Nuotraukoje galima pamatyti, kad kraujagyslės yra tamsesnės, nei aplinkiniai audiniai.



21 pav. *APTOS 2019* rinkinio nuotraukos pavizdys i6traukus tik žaliąjį kanalą

Dar labiau išryškinti kraujagysles, ypač turint juodai baltą vaizdą iš vieno kanalo, galime koreguodami nuotraukos kontrastą. Medicininiams vaizdams pakelti paveikslėlio kontrastą dažniausiai nėra rekomenduojama, nes taip galima prarasti kai kurias detales arba įvesti labai daug „triukšmo“. Tikslesnis būdas taikyti adaptyvius metodus, kaip *CLAHE* [40]. *CLAHE,* (angl. "*Contrast Limited Adaptive Histogram Equalization*") arba Kontrasto ribojama adaptuota histogramos išlyginimo metodika yra vaizdo apdorojimo metodas, skirtas pagerinti kontrastą paveikslėlyje.



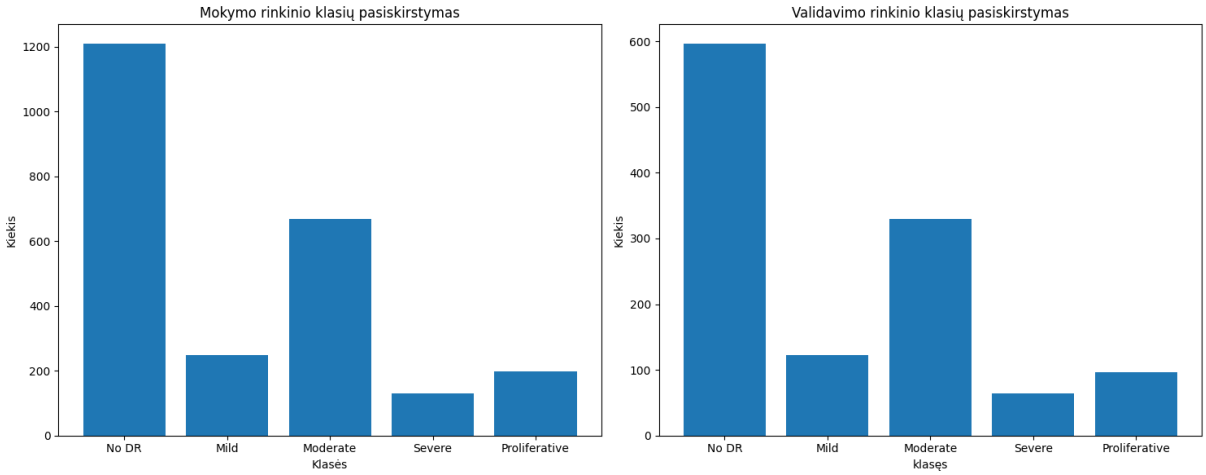
22 pav. *APTOS 2019* duomenų rinkinio nuotraukos pavyzdys su ištrauktu žaliu kanalu ir pritaikyta *CLAHE* metodika

Šis metodas naudoja histogramos išlyginimą, kad būtų išlygintas kontrastas paveikslėlyje, bet prieš tai adaptuodamas keičiamų regionų dydį pagal kontrasto lygį. Pagrindinis *CLAHE* veikimo principas yra suskirstyti paveikslėlį į daugybę mažų regionų, vadinamų skyriais arba langeliais. Kiekvienas skyrius tada yra lyginamas savarankiškai, naudojant histogramos išlyginimo metodą, kad būtų išlaikytas kontrastas tik to skyriaus ribose. Kadangi kiekvienas skyrius yra adaptuojamas prie savo vietinio kontrasto lygio, tai užtikrina, kad kontrastas nebus pernelyg didelis arba per mažas. Kaip matoma **22 pav.,** pritaikius *CLAHE* gerokai pary6kėjo kraujagyslės ir kitos ypatybės paveikslėlyje.

3.1.2. Duomenų rinkinių paruošimas

Modelių mokymui buvo naudojamas *APTOS 2019* duomenų rinkinio mokymo rinkinys, kaip minėta 3.1 skyriuje. Mokymui buvo naudota 80% rinkinio, arba 2453 nuotraukos, o validacijai – likę 20%, arba 1209 nuotraukos.

Pats rinkinys nėra balansuotas, t. y. klasės nėra lygiai pasiskirsčiusios tarp pavyzdžių, bet siekiant tinkamai mokyti modelį ir tinkamai jį vertinti skaidant duomenis buvo siekiama išlaikyti tokį patį klasių pasiskirstymą, kuris matomas **23 pav.**



23 pav. Klasių pasiskirstymas tarp mokymo ir validavimo rinkinių

Rinkiniai buvo saugomi *tfRecords* formatu. *TFRecord* failai yra binariniai failai, kurie optimizuoti greitam įrašymui ir skaitymui. Mokymo rinkinio *tfrecords* failas užima 637 MB, valiavimo – 314 MB, o visos nuotraukos, iš kurių rinkiniai sukurti – 8,01 GB. Dėl to skaityti šiuos failus į atmintį ir su jais dirbti užtrunka mažiau laiko ir resursų.

Prieš perduodant vaizdus modeliui buvo taikomos dvi apdorojimo strategijos:

- Pakeičiamas vaizdų dydis ir atliekama normalizacija, konvertuojant pikselių verte iš skalės nuo 0 iki 255 į skalę nuo 0 iki 1;
- Prie pirmos strategijos dar pridedamas žalio spalvų kanalo ištraukimas ir *CLAE* augmentacija, kaip aprašyta 3.1.1 skyrelyje.

Modeliai buvo apmokyti ir validuoti naudojant tiek vieną, tiek kitą strategiją, siekiant pamatyti ar išankstinis apdorojimas, kuriuo siekiamą išryškinti svarbiausias ypatybes – kraujagysles – padės modeliams geriau ir tiksliau atskirti klases ir prisiderinti prie duomenų, taip pat ar tai padės modeliams ne prisiderinti per daug mažinant nereikalingas ar nesvarbias nuotraukų detales.

3.2. Naudotų modelių atranka

Šiame skyriuje detaliau aprašomi darbe naudoti modeliai, jų parametrai ir architektūros.

3.2.1. Konvoliuciniai modeliai

Konvoliucinių neuroninių modelių yra labai daug ir jų pasirinkimas priklauso nuo turimų resursų kiekio, parametrų kiekio, gebėjimo prisitaikyti, apmokytų modelių prieinamumo ir kitų faktorių. Šiame darbe buvo pasirinkta naudoti *EfficientNet* modelių šeimos modelį *EfficientNetB2* [15] dėl to, kad tai yra gana lengvas ir efektyvus modelis, t. y., jis gali pasiekti didelį tikslumą naudodamas mažai parametrų ir mažiau duomenų, kas reikalauja mažiau skaičiavimo resursų, palyginti su kitais gilių konvoiucinių neuronų tinklų modeliais. Modelių šeimą sudaro 8 modeliai, kurie numeruojami nuo B0 iki B7. B0 yra mažiausias ir paprasčiausias modelis, kuris turi mažiausiai parametrų (apie 5,3 mln.) ir sluoksnių (29), bet jį lengviausia apmokyti, nes jo mažas dydis reikalauja mažiau resursų. Iš kitos pusės B7 – didžiausias modelis, turintis daugiau nei 10 kartų daugiau parametrų (apie 66.7 mln.) ir beveik 10 kartų daugiau sluoksnių (256).

Iš visų 8 modelio versijų *EfficientNetB2* buvo pasirinktas dėl jo dydžio, galimybių ir kad būtų artimesnis sekančiame skyrelyje aprašytame regos transformeriui, atsižvelgiant, kad palyginimas būtų teisingesnis. Ši versija palaiko 260x260 dydžio paveikslėlius, turi apie 9 milijonus parametrų ir reikalauja mažiau resursų, dėl to mokymas truko trumpiau (viena epocha bandymų metu užtruko apie 30 sekundžių).

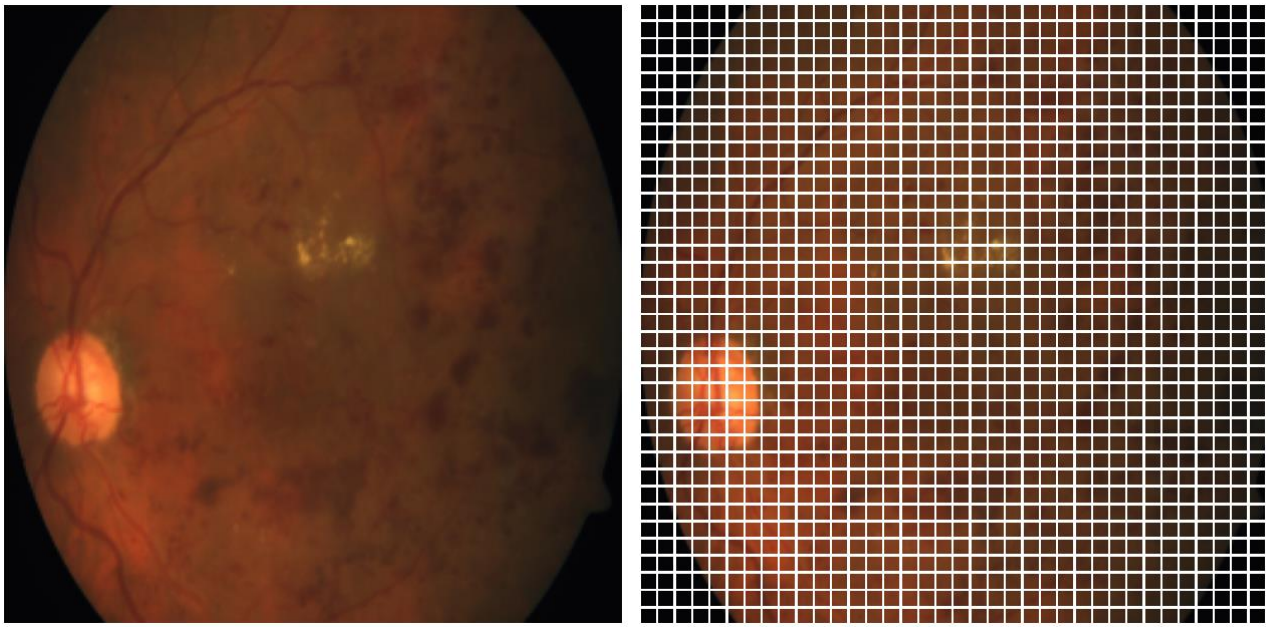
Darbui buvo naudojamas paruoštas modelis *Keras* bibliotekos [41], buvo naudojama tiek neapmokyta modelio versija, tiek versija apmokyta naudojant *IMAGENET* [42] duomenų rinkinį. Vieninteliai keisti modelio parametrai buvo mokymo greitis – 0.0004, ir svorių irimas – 0.00005.

3.2.2. Regos transformeriai

Šiame darbe buvo naudojama, originali regos transformerio architektūra, kuri buvo aprašyta 1.2 skyriuje. Bandymams taip pat buvo naudotas jau paruoštas modelis iš *keras_vit* bibliotekos [43], kurioje yra keturios regos transformerio versijos*: ViT_B16, ViT_B32, ViT_L16* ir *ViT_L32*. Versijos skiriasi gyliu ir lopinėlių dydžių, su kuriuo buvo apmokyti modeliai (aktualu naudojant apmokytą modelį, bet tai yra ir paduodamas hyper-parametras).

Bandymams buvo pasirinkta naudoti *ViT_B16* versiją. Gilesnė versija reikalavo per daug kompiuterinių resursų, o šio modelio lopinėlių dydis (16x16) artimesnis naudojamam dydžiui (7x7), nei *ViT_B32* 32x32 lopinėlių dydis.

Šiam modeliui buvo keičiami keli hyper-parametrai: kaip minėta praeitoje pastraipoje, lopinėlių dydis buvo pakeistas į 7x7 lopinėlius, o mokymo greitis ir svorių yrimas tokie, kaip *EfficientNet* modelio – 0,00001 ir 0,0001.



24 pav. APTOS 2019 rinkinio nuotraukos pavyzdys po perdavimo modeliui su 7x7 lopinėlio dydžiu

Dėl reikalingų resursų kiekio ir transformerio modelio specifikos suvienodinti paveikslėlių dydžio tarp modelių nesigavo ir regos transformeriui paduodami 256x256 dydžio paveikslėliai. **24 pav.** matomas paveikslėlio iš naudojamo duomenų rinkinio pavyzdys po lopinėlių sudarymo operacijos.

3.3. Skyriaus išvados

Šioje dalyje aptariamas *APTOS 2019* duomenų rinkinys, tinkantys išankstinio apdorojimo metodai, duomenų rinkinių paruošimas ir naudotų modelių atranka.

APTOS 2019 yra duomenų rinkinys, kuris naudojamas diabetinės retinopatijos nustatymui ir klasifikavimui moksliniuose tyrimuose. Duomenų rinkinį sudaro akių dugno nuotraukos, gautos naudojant funduskameras. Kiekviena nuotrauka buvo ekspertų įvertinta ir priskirta vienai iš penkių diabetinės retinopatijos laipsnių kategorijų, pradedant nuo retinopatijos nebuvimo iki labai sunkios retinopatijos. Mokymo rinkinyje, turinčiame etiketes, yra 3662 nuotraukos, iš kurių šiame darbe buvo naudota tik ši dalis.

Išankstinis apdorojimas yra svarbus žingsnis vaizdų klasifikavimo uždaviniuose. Tai apima vaizdų dydžio suvienodinimą, normalizaciją ir kitus metodus, skirtus išryškinti svarbias vaizdo ypatybes. Šiame darbe buvo taikyti pirminiai apdorojimo metodai, tokiu būdu paruošiant duomenis modeliams. Be to, buvo išbandyti ir kiti metodai, skirti pagerinti vaizdus, pavyzdžiui, išryškinti

kraujagysles, kurios yra svarbios diabetinės retinopatijos atveju. Tai buvo daroma ištraukiant žaliąjį spalvų kanalą ir taikant *CLAHE* metodą (Kontrasto ribojama adaptuota histogramos išlyginimo metodika), kuris padeda gerinti kontrastą paveikslėlyje. Išsamiai aptariant šiuos metodus, galima geriau suprasti jų veikimo principus ir tai, kaip jie daro įtaką galutiniams klasifikavimo rezultatams. Kadangi diabetinės retinopatijos atveju svarbios yra kraujagyslės, tinkamas išankstinis apdorojimas gali padėti išryškinti šias svarbias ypatybes, taip galimai pagerindamas klasifikavimo tikslumą. Šiame darbe atliekami bandymai tiek naudojant išankstinio apdorojimo metodus, tiek jų nenaudojant siekiant pažiūrėti ar jie daro įtaką rezultatams ir ar jų taikymas padės modeliams geriau apibendrinti duomenis ir pasiekti aukštesnių rezultatų.

Bandymams atlikti, remiantis šiuolaikinės literatūros analize, buvo atrinkti du modeliai: konvoliucinių tinklų modelis *EfficientNetB2* ir regos transformeris. Konkreti *EfficientNet* versija buvo atrinkta, siekiant, kiek įmanoma, suvienodinti abiejų modelių dydį ir klasifikavimo galimybes, bet išlaikant mažesnius kompiuterinių skaičiavimo resursų reikalavimus.

4.REZULTATAI

Šiame skyriuje aptariami diabetinės retinopatijos lygio nustatymo naudojant *EfficientNet* ir regos transformerius rezultatai, aptariami metodai, naudoti juos palyginti, ir pateikiamos įžvalgos apie juos.

4.1. Aplinka

Visi modeliai buvo mokyti naudojant *Google Colaboratory* aplinką su *Nvidea* A100 vaizdo plokšte su 83,5 GB atminties ir 40 GB vaizdo plokštės atminties. Visas kodas buvo rašomas naudojant *Python* programavimo kalbą ir pagrindinis karkasas modelių aprašymui buvo *tensorflow* su *Keras* sąsaja.

4.2. Vertinimo metrikos

Tam, kad būtų galima tiksliau palyginti modelius buvo taikomos trys vertinimo metrikos:

- Tikslumas – santykis tarp teisingai klasifikuotų pavyzdžių ir visų pavyzdžių skaičiaus (žr. **5 lygtis**). Tai matuoja bendrą modelio našumą. Tikslumas yra dažniausiai naudojama metrika vertinant klasifikavimo modelius, dėl to ji buvo taikoma ir čia, tačiau ji tiksliausia, kai duomenų rinkinys yra balansuotas.

5 lygtis. Tikslumo metrikos skaičiavimo formulė

$$tikslumas = \frac{Teisingai\ nustatytų\ pavyzdžių\ skaičius}{Visų\ pavyzdžių\ skaičius}.$$

Bazinis tikslumas, arba tikslumo riba, kurią modelis turi peržengti, kad galėtumėme manyti, kad jis nespėlioja atsitiktinai, skaičiuojama pagal **6 lygtis.** Formulę ir šiame duomenų rinkiniui jis lygus 0.493.

6 lygtis. Bazinio tikslumo skaičiavimo formulė

$$bazinis\ tikslumas = \frac{didžiausios\ klasės\ pavyzdžių\ kiekis}{Visų\ pavyzdžių}$$

- AUC – plotas po sprendimų priimančiojo ypatybių kreive (angl. *area under the reciever operating characteristic curve* arba dažnesniais atvejais *area under the ROC curve*) parodo klasifikatoriaus jautrumą (tikslumą) palyginti su specifiškumu. Jei modelis yra tobulas, t. y. niekas neteisingai neklasifikuoja, *AUC* bus 1. Tai reiškia, kad modelis puikiai atskiria tarp teigiamų ir neigiamų pavyzdžių. Jei modelis tiesiog spėlios, *AUC* bus arti 0,5. Tai reiškia, kad modelis yra bejėgis ir klasifikuoja kaip atsitiktinai. *AUC* nekreipia dėmesio į klasės pasiskirstymą, o vietoj to vertina modelio gebėjimą atskirti

tarp teigiamų ir neigiamų pavyzdžių, dėl to ši metrika tinkamesnė vertinti modelius, mokytus naudojant nebalansuotus duomenų rinkinius.

- F1 - harmoninis vidurkis tarp glaudumo ir atkūrimo (angl. *recall*) (žr. **7 lygtis**.).

7 lygtis. F1 metrikos skaičiavimo formulė

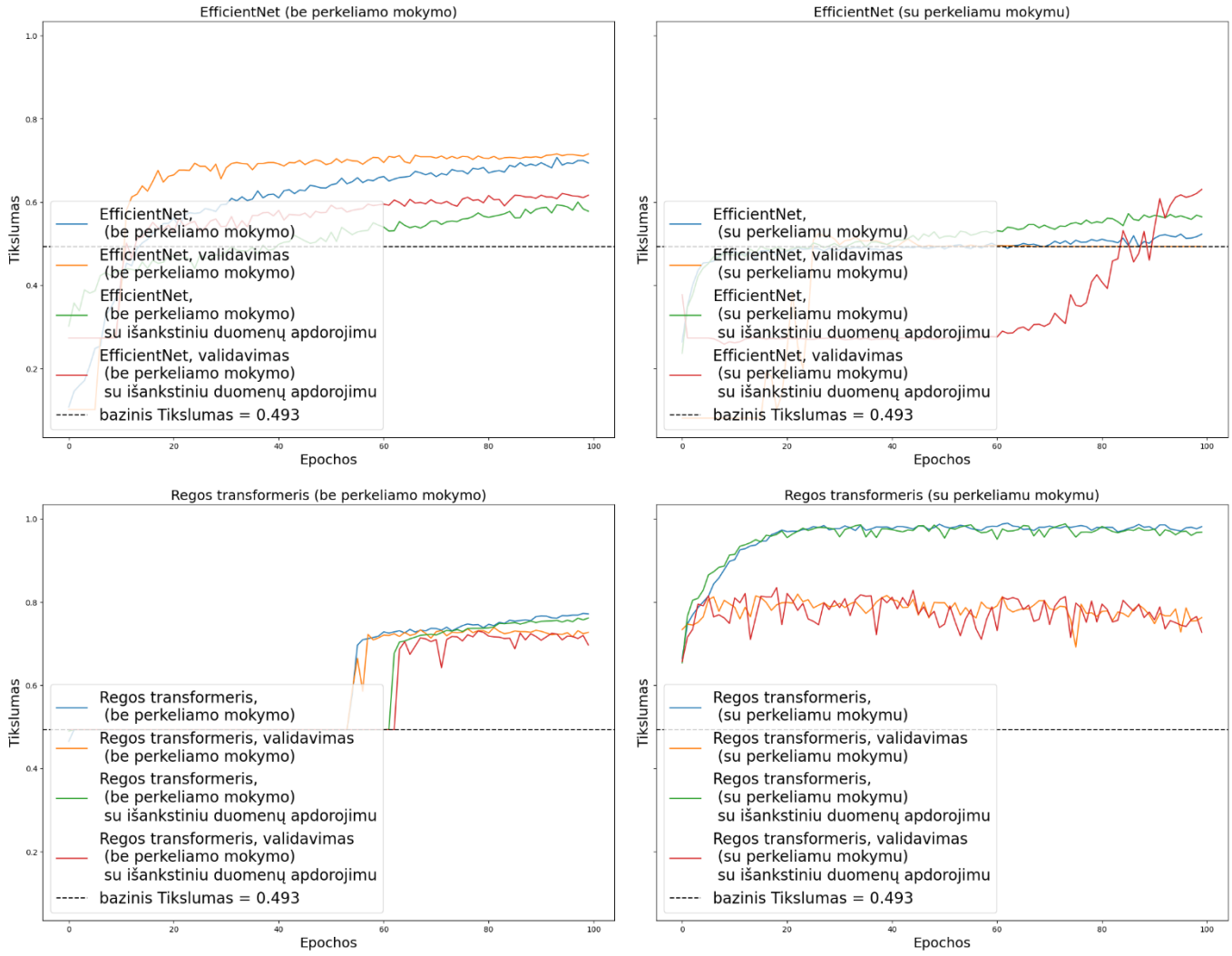
$$F1 = \frac{glaudumas \times atkūrimas}{glaudumas + atkūrimas}$$

F1 metrika nesiveržia į klasės dydžius. Tai reiškia, kad ji yra pakankamai nešališka, kad atliktų objektyvų modelio vertinimą, net jei mūsų duomenų rinkinyje yra disbalansas tarp klasės dydžių. Kitaip tariant, ji vertina modelio veikimą tik pagal jo gebėjimą klasifikuoti teigiamus ir neigiamus atvejus, nepriklausomai nuo klasės skaičiaus. Apskaičiuoti bazinį *F1* sudėtinga, nes tam reikėtų žinoti modelio gaudumą ir atkūrimą, dėl to bazinis *F1* nebuvo skaičiuojamas.

Vertinimo metrikos, įskaitant tikslumą, *AUC* ir *F1*, yra nepakeičiama priemonė modelių efektyvumui įvertinti. Tikslumas suteikia bendrą modelio našumo įžvalgą, tuo tarpu *AUC* gerai vertina klasifikatoriaus gebėjimą atskirti tarp teigiamų ir neigiamų pavyzdžių. *F1* metrika, būdama harmoniniu vidurkiu tarp glaudumo ir atkūrimo, suteikia subalansuotą vertinimą, nepriklausomai nuo klasės dydžių. Šios trys metrikos yra esminės norint tiksliai palyginti modelius ir gauti visapusišką modelio veikimo vertinimą, ypač kai turime nebalansuotus duomenų rinkinius. Taigi, naudojant šias metrikas, galime objektyviai nustatyti modelių stipriąsias ir silpnąsias puses bei priimti pagrįstus sprendimus.

4.3. Rezultatai ir rezultatų palyginimas

Literatūros analizės metu buvo atrinktos dvi modelių architektūros, du mokymo būdai ir du duomenų apdorojimo būdai. Visi bandymai buvo atliekami naudojant abu skirtingus modelius, mokymo būdus ir duomenų apdorojimo būdus, tai vertinamos 8 skirtingos eksperimentų konfigūracijos. Grafikus didesniu formatu galima pamatyti prieduose **1.A**, **1.B**, **1.C**, **1.D**.



25 pav. Modelių tikslumas. Skirtinguose grafikuose pavaizduoti skirtingų modelų su skirtingu mokymo būdų rezultatai, grafikuose pavaizduoti rezultatai su skirtingu duomenų paruošimo būdu

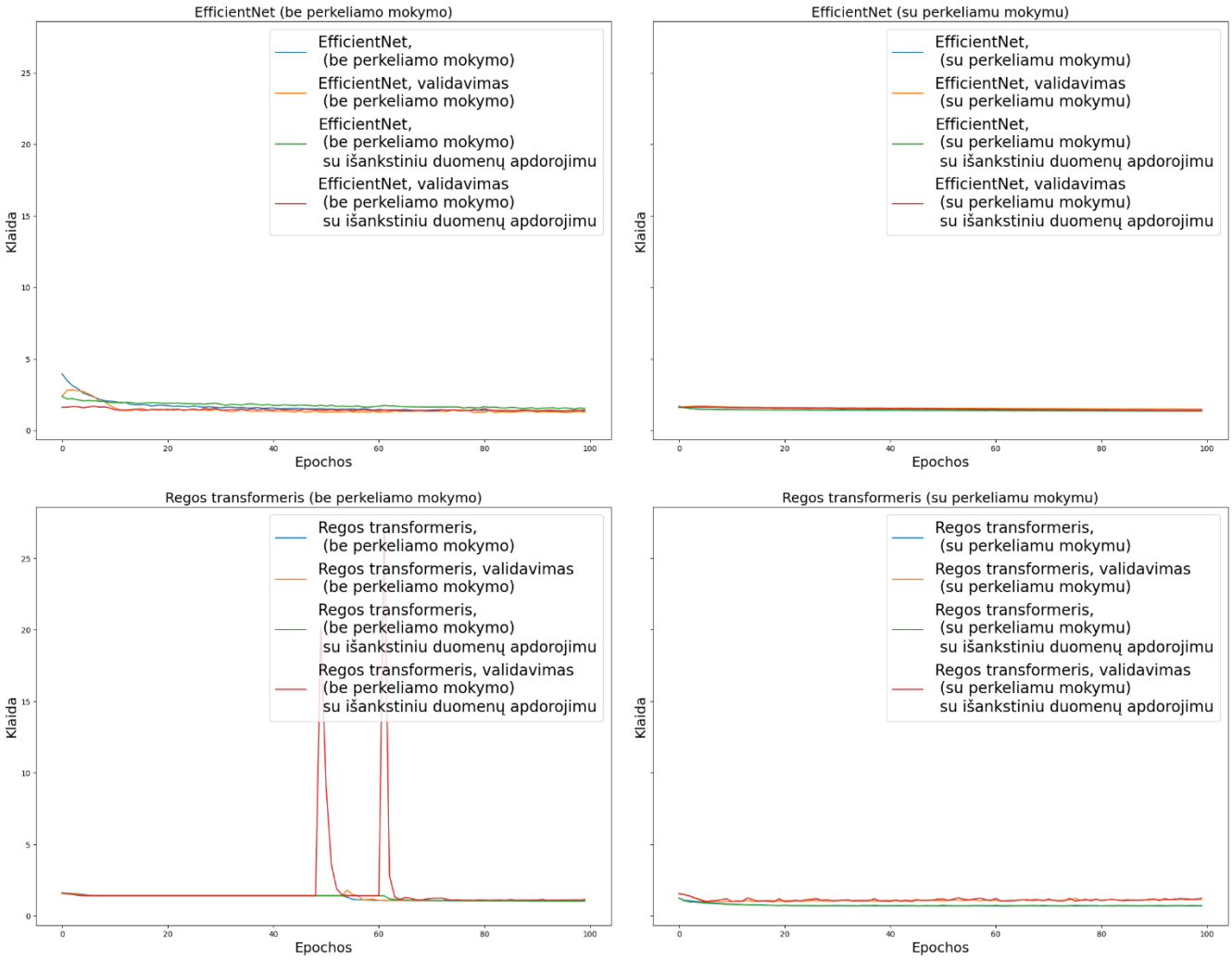
Tikslumas šiems bandymams nėra geriausia metrika dėl stipriai išbalansuoto duomenų rinkinio, bet **25 pav.** galime pamatyti modelių stabilumą ir pradėti matyti, kaip išankstinis duomenų apdorojimas keičia modelių rezultatus.

EfficientNet modelio rezultatuose labiau pasireiškia skirtumai, kai naudojami kitaip apdoroti duomenys. Rezultatai pablogėjo per bent 5% modeliui be perkeliamo mokymo, bet modelis, taip pat stabiliai klasifikavo tiek mokymo, tiek validacijos vaizdus. Iš kitos pusės, *EfficientNet* su perkeliamu mokymu nepriklausomai nuo duomenų paruošimo vos pasiekė bazinį tikslumą, bet variantus su išankstiniu apdorojimu, sunkiau taikėsi ir apibendrino duomenis, nes mokymo ir validacijos tikslumas ženkliai skiriasi.

Regos transformeriams kitokios nuotraukos daug ko nepakeitė ir abiejų variantų tikslumas praktiškai nesiskiria. Modelis be perkeliamo mokymo praktiškai tik spėliojo klases iki 50 epochos, po to jo tikslumas pagerėjo per beveik 25%. Remiantis kitais tyrimais apie transformerio modelius

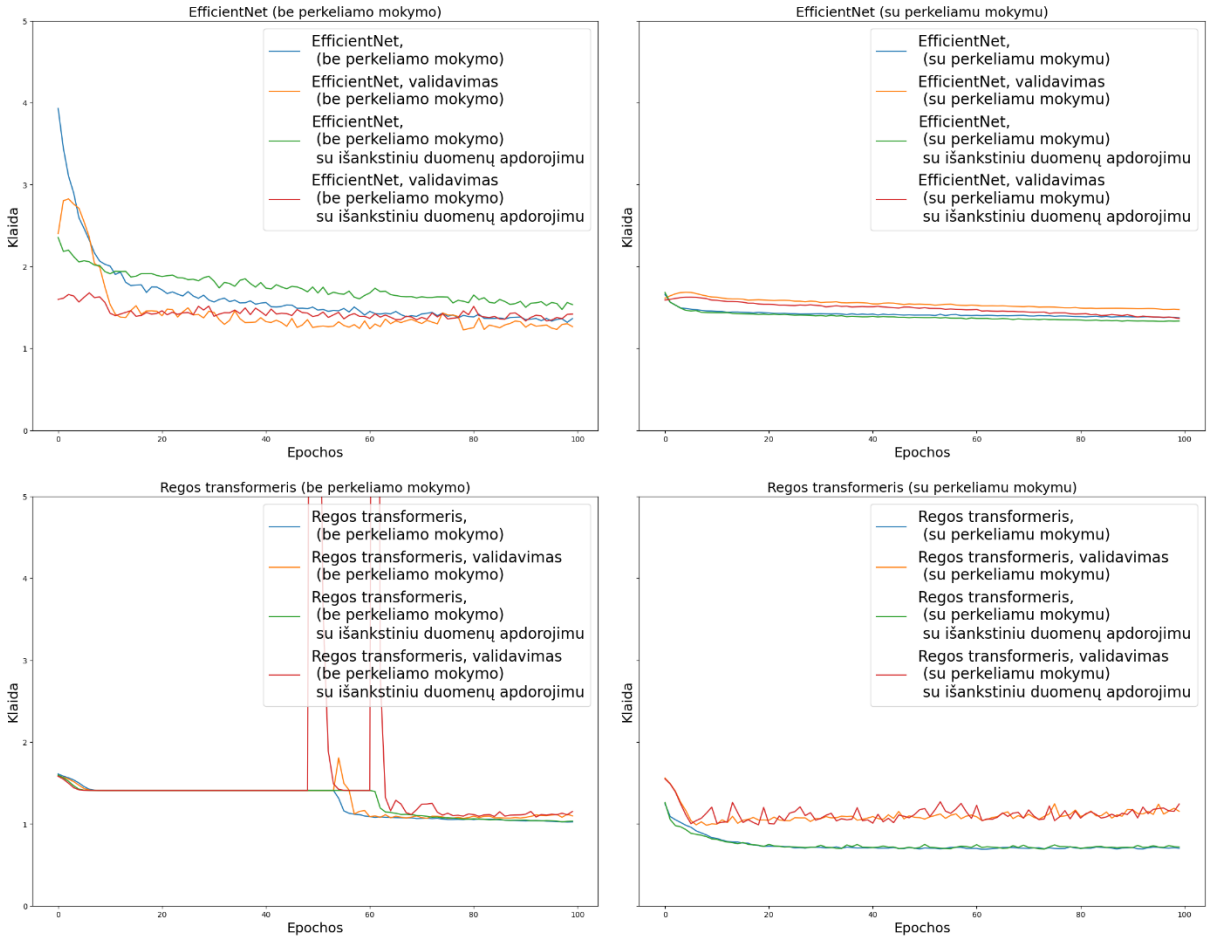
tai gali būti paaiškinta tuo, kad transformeriams prisitaikyti prie duomenų užtrunka daugiau epochų, negu kitokio tipo modeliams.

Transformeriui su perkeliamu mokymu daug epochų pasiekti aukštą tikslumą nereikėjo, tačiau nepriklausomai nuo duomenų, skirtumas tarp validacijos ir mokymo tikslumo buvo apie 20%, remiantis tisklumo aspektu, modeliui buvo sunku prisitaikyti.



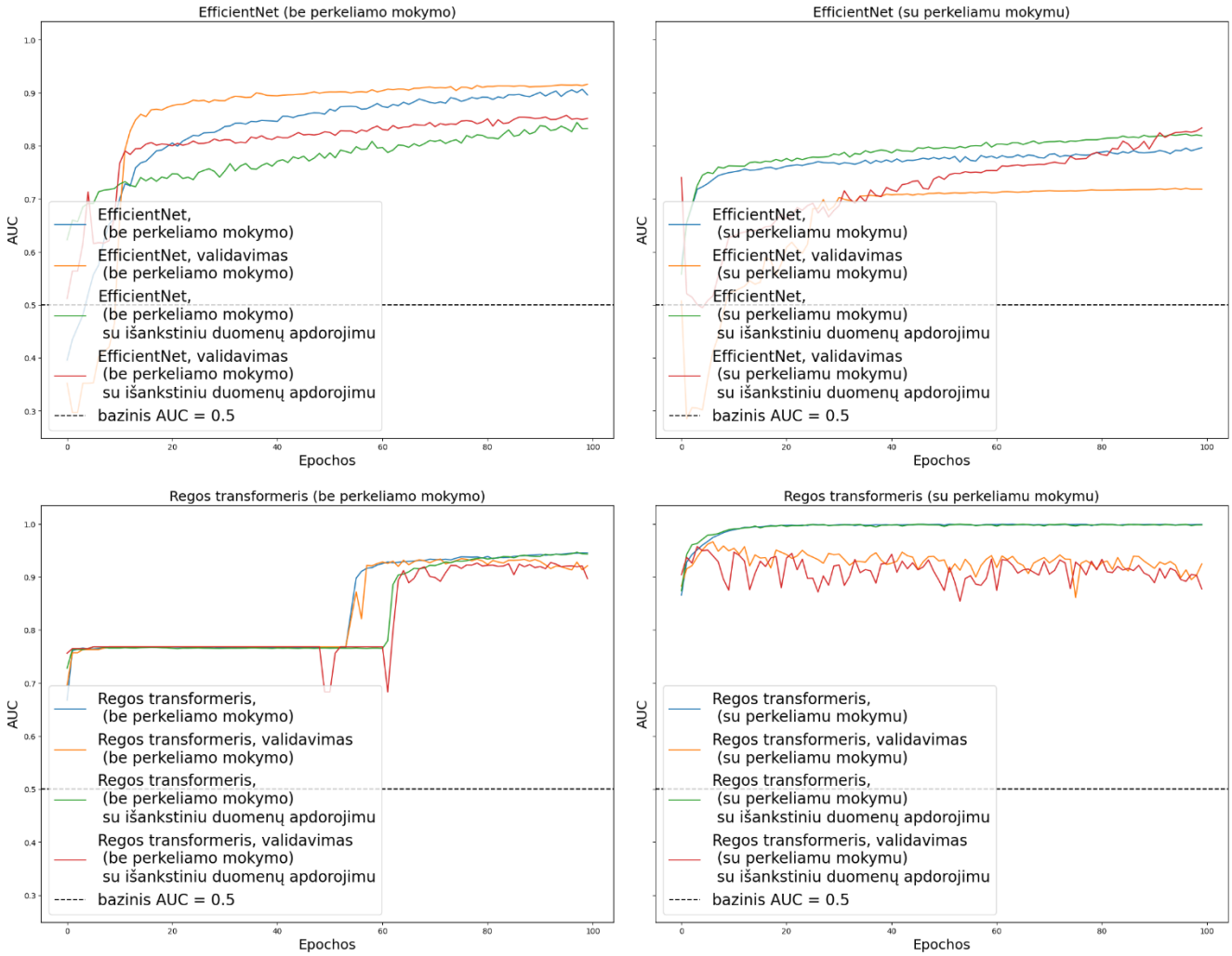
26 pav. Modelių klaida. Skirtinguose grafikuose pavaizduoti skirtingų modelių su skirtingu mokymo būdų rezultatai, grafikuose pavaizduoti rezultatai su skirtingu duomenų paruošimo būdu

Žiūrint į modelių klaidą, kuri matoma **26 pav.,** matome, kad trečio modelio grafike klaida kelis skirtus ženkliai pašoko. Sumažinus y ašies rėžius (žr. **27 pav.**) galime geriau matyti, kaip atrodo kitų modelių klaidos.



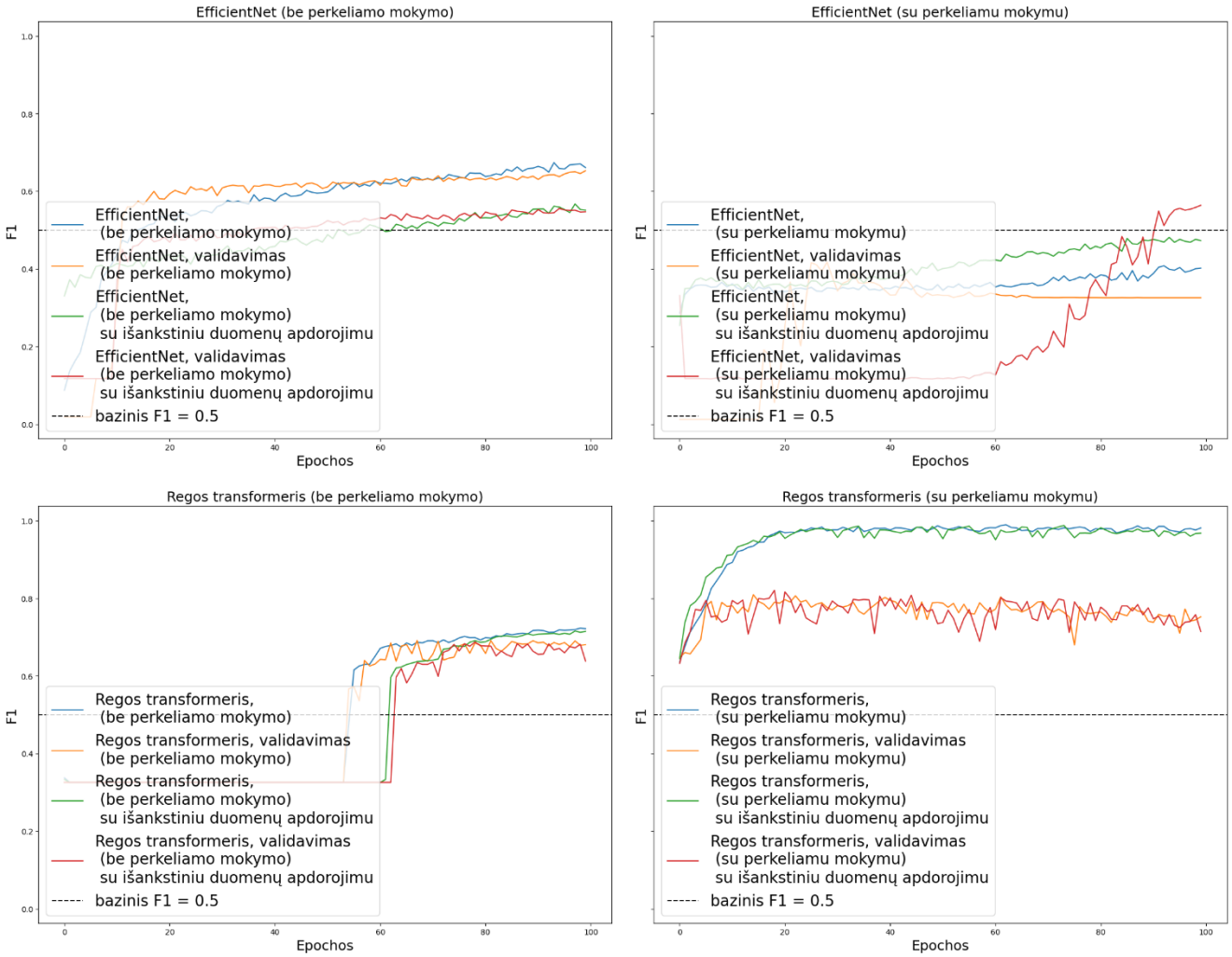
27 pav. Modelių klaida sumažinus y ašies rėžius. Skirtinguose grafikuose pavaizduoti skirtingų modelių su skirtingu mokymo būdų rezultatai, grafikuose pavaizduoti rezultatai su skirtingu duomenų paruošimo būdu

Sumažinus rėžius matome, kad dauguma modelių mokėsi stabiliai, nors klaida krito labai po mažai. Tačiau regos transformeris su perkeliamu mokymusi, prisitaikė prie duomenų, nes jo validacijos klaida didesnė (apie 0,5) negu mokymo klaida per visas epochas.



28 pav. Modelių *AUC*. Skirtinguose grafikuose pavaizduoti skirtingų modelių su skirtingu mokymo būdų rezultatai, grafikuose pavaizduoti rezultatai su skirtingu duomenų paruošimo būdu

Vertinant *AUC*, kuris matomas **28 pav.**, visi modeliai perlipo bazinį *AUC*, tai vidutiniškai visų modelių klasių klasifikavimas nepanašus į atsitiktinį spėliojimą. Apart to dauguma išvadų apie grafikus tokios pat, kaip tikslumo, nes grafikai atrodo panašiai, tačiau regos transformerių rezultatai geresni, nei konvoliucinio modelio. Regos transformeris su perkeliamu mokymu taip pat labiausiai priartėjo prie 1, tai jo klasifikavimo tikslumas labiausiai priartėjo prie puikaus tikslumo.



29 pav. Modelių *F1*. Skirtinguose grafikuose pavaizduoti skirtingų modelių su skirtingu mokymo būdų rezultatai, grafikuose pavaizduoti rezultatai su skirtingu duomenų paruošimo būdu

Paskutinė metrika, *F1* (žr. **29 pav.**), vėl priveda prie panašių išvadų, kaip kiti grafikai, bet ši metrika labiausiai naudinga daugiau nei dviejų klasių nebalansuotiems duomenų rinkiniams. Imant šią ir kitas metrikas galima teigti, kad geriausiai klasifikuoja regos transformeriai.

Pamatyti, kuri konfigūracija buvo geriausia kiekvienai metrikai, vertinant validaciją, galima pamatyti **1 lentelėje**. Trys regos transformerio konfigūracijos pažymėtos, kaip geriausios dėl to, kad *EfficientNet* ir regos transformerio be perkeliamo mokymo rezultatai skiriasi tik per kelis procentus regos transformerio mokymo ir validacijos skirtumas mažesnis. Galima teigti, kad jis geriau prisitaikė prie duomenų. Vertinant rezultatus su perkeliamu mokymu, regos transformerio rezultatai mokyme atrodo beveik puikiai, bet teigti, kad tai geriausias modelis iš keturių yra šiek tiek sudėtinga, nes visose metrikose skirtumas tarp mokymo ir validacijos duomenų didelis, o imant klaidą panašu, kad modelis per daug prisitaikė prie duomenų.

1 lentelė. Kurie modeliai ir su kokia konfigūracija pasiekė geriausius rezultatus kiekvienai metrikai vertinant validaciją

Modeliai ir konfigūracija	**Klaida**	**Tikslumas**	**AUC**	**F1**
EfficientNet (be perkeliamo mokymo, be išankstinio duomenų apdorojimo)	Ne	Ne	Ne	Ne
EfficientNet (su perkeliamo mokymo, be išankstinio duomenų apdorojimo)	Ne	Ne	Ne	Ne
EfficientNet (be perkeliamo mokymo, su išankstinio duomenų apdorojimo)	Ne	Ne	Ne	Ne
EfficientNet (su perkeliamo mokymo, su išankstinio duomenų apdorojimo)	Ne	Ne	Ne	Ne
Regos transformeris (be perkeliamo mokymo, be išankstinio duomenų apdorojimo)	**Taip**	**Taip**	**Taip**	**Taip**
Regos transformeris (su perkeliamo mokymo, be išankstinio duomenų apdorojimo)	Taip	Taip	Taip	Taip
Regos transformeris (be perkeliamo mokymo, su išankstinio duomenų apdorojimo)	Ne	Ne	Ne	Ne
Regos transformeris (su perkeliamo mokymo, su išankstinio duomenų apdorojimo)	Taip	Taip	Taip	Taip

Bandymų rezultatai koreliuoja su kelių autorių darbais, lyginančiais *EfficientNet* ir regos transformerius [44], [45]. Nors duomenų rinkinys kitas ir eksperimentų pobūdis ir modelių konfigūraciją skiriasi iš literatūros galima pamatyti, kad regos transformeriai gali, kad ir ne žymiai, tiksliau klasifikuoti, nei *EfficientNet* tinklai, bent stengiantis juos konfigūruoti kiek įmanoma panašiau. Imant tą pačią užduotį, sprendžiama šiame darbe, t. y. diabetinės retinopatijos detektavimą naudojant *APTOS 2019* rinkinį, nepavyko rasti darbų, kurie naudotų regos transformerius, tačiau vieni iš geriausių rezultatų, vertinant tik tikslumą, apie 20% geresni, nei pavyko išgauti šiame darbe, yra gauti naudojant *EfficientNet* tinklus [46].

IŠVADOS

Šiame darbe buvo apžvelgta literatūra apie medicininių vaizdų apdorojimą, susijusį su klasifikavimo uždaviniais. Taip pat buvo išbandyti ir palyginti du skirtingų architektūrų modeliai: konvoliucinis neuroninis tinklas *EfficientNet* ir regos transfromeris. Taip pat buvo atliekami palyginimai naudojant skirtingus mokymo metodus – paprastą prižiūrimą mokymą ir perkeliamą mokymą, ir skirtingus duomenų paruošimo būdus – grafiškai nekeičiant pačios nuotraukos, ir taikant išankstinio apdorojimo metodus rastus skaitytoje mokslinėje literatūroje. Visi bandymai buvo atlikti su vienu duomenų rinkiniu – diabetinės retinopatijos lygio nustatymui iš tinklainės vaizdų skirtam *APTOS 2019*.

Iš mokslinės literatūros analizės rezultatų matoma, kad medicininių vaizdų klasifikavimas yra populiari, bet sudėtinga vaizdų apdorojimo tyrimų sritis. Konvoliuciniai tinklai vis dar populiariausias tyrėjų pasirinkimas, bet vaizdams apdoroti dar galėtų būti naudojami regos transformeriai, tačiau jie laikomi daug duomenų reikalaujančiais modeliais,. Dėl šios priežasties medicininių vaizdų klasifikavimo srityje, kur gauti didelius duomenų rinkinius yra ypač sunku, regos transformeriai netinka. Atsižvelgiant į architektūrų apžvalginio tyrimo rezultatus, buvo nuspręsta atlikti bandymus, atliekant abiejų modelių (*EfficientNetB2* tinklo ir regos transformerio) konfigūravimo ir našumo lyginamąją analizę.

Ruošiantis bandymams ir ieškant duomenų rinkinio priimtas sprendimas naudoti *APTOS-2019* diabetinės retinopatijos nustatymo rinkinį. Rinkinyje yra aukštos kokybės tinklainės nuotraukos, kurias modeliams lengviau klasifikuoti, palyginti su žemos kokybės nuotraukomis. Analizuojant rinkinį ir literatūrą apie akių ligų klasifikavimą, buvo nuspręsta sudaryti išankstinio apdorojimo eigą, kurios tikslas – išryškinti ligai svarbias ypatybes (pvz., kraujagysles). Išankstiniame vaizdo duomenų apdorojimo etape buvo nuspręsta panaudoti vaizdų žalią kanalą ir pritaikyti kontrasto keitimo algoritmą *CLAHE*.

Bandymai buvo atliekami naudojant du modelius (*EfficientNetB2* ir regos transformerius), du mokymo būdus (prižiūrimą ir perkeliamą mokymą) ir du duomenų apdorojimo būdus (nekeičiant nuotraukų grafiškai ir naudojant išankstinį apdorojimą). Modeliams lyginti buvo naudojamos trys metrikos (tikslumas, AUC ir F1) ir modelių klaida. Lyginant modelių rezultatus, iš atliktų bandymų matoma, kad transformerių architektūros modelių rezultatai buvo geresni ir stabilesni, bet modelis su perkeliamu mokymų per daug prisitaikė prie duomenų t. y. gerai išmoko klasifikuoti mokymo duomenis, bet neišgavo pakankamai abstrakčių ypatybių, kad galėtų jas efektyviai taikyti kitokioms nuotraukoms.

Atsižvelgiant į mokslinės literatūros analizės ir eksperimentinio tyrimų rezultatus galima teigti, kad tiek konvoliuciniai tinklai, tiek regos transformeriai gali būti naudojami akių ligų diagnostikos srityje, analizuojant akių tinklainės vaizdą. Net turint mažą kiekį duomenų, naudojant regos transformerius, pasiekti geri rezultatai (siekiančių 80% ir aukštesnį tikslumą, AUC ir F1), beveik atitinkantys *EfficientNet* modelio rodiklius. Regos transformerių atveju gerų rezultatų pasiekta taikant perkeliamo mokymo būdą. Vis dėlto, reikėtų atlikti daugiau bandymų ir geriau pritaikyti modelį arba rasti daugiau duomenų rinkinių siekiant sumažinti tikimybę, kad modelis per daug prisitaikys.

Pažymėtina, kad šiame darbe atlikti bandymai buvo limituoti: vertinant darbe atliktų tyrimų rezultatus pažymėtina, kad bandymai turėjo apibrėžtas ribas: duomenų rinkinys buvo sudarytas tik iš aukštos kokybės nuotraukų, todėl ne visai atitinka realias tokių modelių pritaikymo sąlygas ir nereikalauja ekstensyvaus išankstinio apdorojimo. Be to, duomenų kiekis buvo per mažas, kad būtų galima turėti tinkamą duomenų rinkinį modeliams patikrinti, todėl buvo apsistota ties mokymo ir validacijos etapais. Taigi ateityje, tęsiant tyrimus ir norint geriau palyginti modelių klasifikavimo galimybes akių ligų nustatymo srityje, ir norint geriau išsiaiškinti, ar modeliai galėtų būti taikomi už eksperimentų ribų, reikėtų pamėginti juos pritaikyti įvairesniems duomenims analizuoti, t. y. ne tik aukštos raiškos nuotraukoms. Be to, svarbu įvertinti modelių panaudojimo galimybes prastesnėmis sąlygomis, priklausomomis nuo skirtingos kokybės įrangos, naudotos joms gauti ir pan. Taip pat reikėtų atlikti daugiau tyrimų taikant išankstinius apdorojimo metodus, kurie galėtų labiau padėti išryškinti kraujagysles tinklainės nuotraukose. Šiame darbe naudoti metodai didelės teigiamos įtakos modeliams neturėjo, bet, remiantis literatūra, geresnis jų pritaikymas turi galimybę padėti modeliams geriau prisitaikyti.

LITERATŪRA

1. A. Anaya-Isaza, L. Mera-Jiménez, and M. Zequera-Diaz, “An overview of deep learning in medical imaging,” Informatics in Medicine Unlocked, vol. 26. Elsevier Ltd, Jan. 01, 2021. doi: 10.1016/j.imu.2021.100723.

2. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” 2012. [Internetinis]. Pasiekiamas: http://code.google.com/p/cuda-convnet/

3. A. Vaswani et al., “Attention Is All You Need,” Jun. 2017, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/1706.03762

4. A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Oct. 2020, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/2010.11929

5. “APTOS2019 Blindness Detection. 1. Description | by Satya Srikar | Medium.” Accessed: Mar. 04, 2024. [Internetinis]. Pasiekiamas: https://medium.com/@nsatyasrikar/aptos2019-blindness-detection-70b749a3d331

6. L. Cai, J. Gao, and D. Zhao, “A review of the application of deep learning in medical image classification and segmentation,” Ann Transl Med, vol. 8, no. 11, pp. 713–713, Jun. 2020, doi: 10.21037/atm.2020.02.44.

7. D. R. Sarvamangala and R. v. Kulkarni, “Convolutional neural networks in medical image understanding: a survey,” Evolutionary Intelligence, vol. 15, no. 1. Springer Science and Business Media Deutschland GmbH, Mar. 01, 2022. doi: 10.1007/s12065-020-00540-3.

8. Y. Lecun, L. Eon Bottou, Y. Bengio, and P. H. Abstract|, “Gradient-Based Learning Applied to Document Recognition.”

9. M. M. Rahman and D. N. Davis, “Addressing the Class Imbalance Problem in Medical Datasets,” Int J Mach Learn Comput, pp. 224–228, 2013, doi: 10.7763/ijmlc.2013.v3.307.

10. C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, “Is it Time to Replace CNNs with Transformers for Medical Images?,” Aug. 2021, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/2108.09038

11. K. He et al., “Transformers in Medical Image Analysis: A Review,” Feb. 2022, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/2202.12165

12. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/1409.1556

13. C. Szegedy et al., “Going Deeper with Convolutions,” Sep. 2014, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/1409.4842

14. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/1512.03385

15. M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” May 2019, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/1905.11946

16. S. Azizi et al., “Big Self-Supervised Models Advance Medical Image Classification,” Jan. 2021, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/2101.05224

17. M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding Transfer Learning for Medical Imaging,” Feb. 2019, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/1902.07208

18. I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, “Attention Augmented Convolutional Networks,” Apr. 2019, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/1904.09925

19. N. Parmar et al., “Image Transformer,” Feb. 2018, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/1802.05751

20. C. Matsoukas, J. F. Haslum, M. Sorkhei, M. Söderberg, and K. Smith, “What Makes Transfer Learning Work for Medical Images: Feature Reuse & Other Factors”.

21. F. Shamshad et al., “Transformers in Medical Imaging: A Survey,” Jan. 2022, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/2201.09873

22. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” Dec. 2020, [Internetinis]. Pasiekiamas: http://arxiv.org/abs/2012.12877

23. A. Aljuaid and M. Anwar, “Survey of Supervised Learning for Medical Image Processing,” SN Comput Sci, vol. 3, no. 4, pp. 1–22, Jul. 2022, doi: 10.1007/S42979-022-01166-1/FIGURES/20.

24. S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10. 2010. doi: 10.1109/TKDE.2009.191.

25. A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, “Transfer learning: a friendly introduction,” J Big Data, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00652-w.

26. “Transfer Learning Guide: A Practical Tutorial With Examples for Images and Text in Keras.” Accessed: Dec. 17, 2023. [Internetinis]. Pasiekiamas: https://neptune.ai/blog/transfer-learning-guide-examples-for-images-and-text-in-keras

27. “A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning | by Dipanjan (DJ) Sarkar | Towards Data Science.” Accessed: Dec. 17, 2023. [Internetinis]. Pasiekiamas: https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a

28. “How Transfer Learning works. Transfer Learning will be the next… | by HAFEEZ JIMOH | Towards Data Science.” Accessed: Dec. 16, 2023. [Internetinis]. Pasiekiamas: https://towardsdatascience.com/how-transfer-learning-works-a90bc4d93b5e

29. H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, “Transfer learning for medical image classification: a literature review,” BMC Medical Imaging, vol. 22, no. 1. 2022. doi: 10.1186/s12880-022-00793-7.

30. G. Litjens et al., “A survey on deep learning in medical image analysis,” Medical Image Analysis, vol. 42. 2017. doi: 10.1016/j.media.2017.07.005.

31. L. Wang et al., “Trends in the application of deep learning networks in medical image analysis: Evolution between 2012 and 2020,” Eur J Radiol, vol. 146, 2022, doi: 10.1016/j.ejrad.2021.110069.

32. PubMed, “About - PubMed,” PubMed. 2021.

33. S. De Yu, L. L. Liu, Z. Y. Wang, G. Z. Dai, and Y. Q. Xie, “Transferring deep neural networks for the differentiation of mammographic breast lesions,” Sci China Technol Sci, vol. 62, no. 3, 2019, doi: 10.1007/s11431-017-9317-3.

34. M. M. Rahaman et al., “Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches,” J Xray Sci Technol, vol. 28, no. 5, 2020, doi: 10.3233/XST-200715.

35. J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," in Proceedings - International Conference on Pattern Recognition, 2016. doi: 10.1109/ICPR.2016.7899799.

36. T. Babaqi, M. Jaradat, A. E. Yildirim, S. H. Al-Nimer, and D. Won, "Eye Disease Classification Using Deep Learning Techniques," 2023.

37. "APTOS 2019 Blindness Detection | Kaggle." Accessed: Apr. 02, 2024. [Internetinis]. Pasiekiama: https://www.kaggle.com/competitions/aptos2019-blindness-detection/discussion/108065

38. C. Hudson, "The clinical features and classification of diabetic Fetinopathy," Physiol. Opt, vol. 16, pp. 43–48, 1996.

39. M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, "Medical image processing, analysis & visualization in clinical research," Proceedings of the IEEE Symposium on Computer-Based Medical Systems, pp. 381–388, 2001, doi: 10.1109/CBMS.2001.941749.

40. P. Vasuki, J. Kanimozhi, and M. B. Devi, "A survey on image preprocessing techniques for diverse fields of medical imagery," Proceedings - 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering, ICEICE 2017, vol. 2017-Gruodis, pp. 1–6, Dec. 2017, doi: 10.1109/ICEICE.2017.8192443.

41. "EfficientNet B0 to B7." Accessed: May 12, 2024. [Internetinis]. Pasiekiamas: https://keras.io/api/applications/efficientnet/

42. "ImageNet." Accessed: May 12, 2024. [Internetinis]. Pasiekiamas: https://www.image-net.org/

43. "keras-vit · PyPI." Accessed: May 12, 2024. [Internetinis]. Pasiekiamas: https://pypi.org/project/keras-vit/

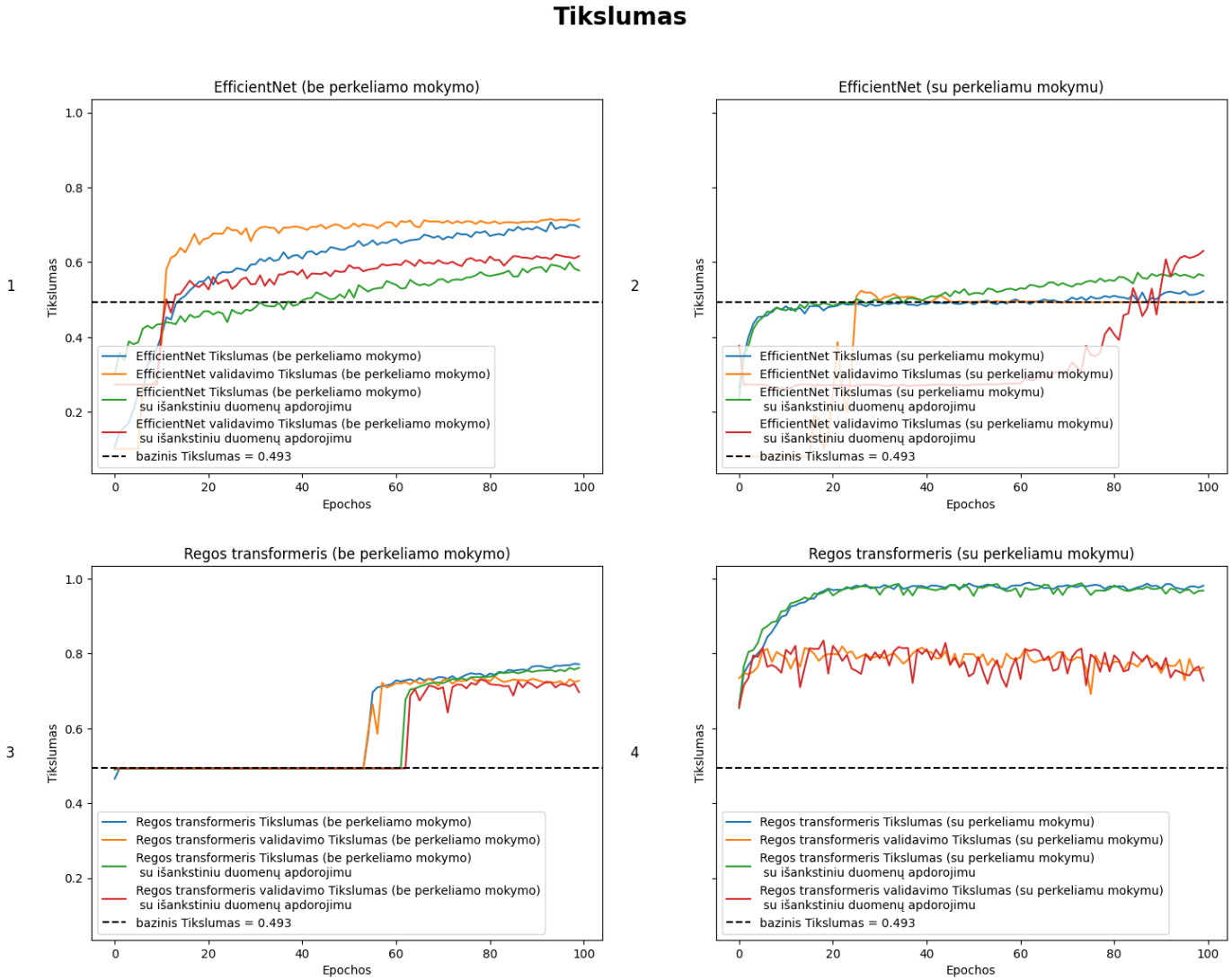
44. P. Byvshev, P. A. Truong, and Y. Xiao, "Image-based Renovation Progress Inspection with Deep Siamese Networks," ACM International Conference Proceeding Series, pp. 96–104, Feb. 2020, doi: 10.1145/3383972.3384036.

45. C. Testagrose, M. Shabbir, B. Weaver, and X. Liu, "Comparative Study Between Vision Transformer and EfficientNet on Marsh Grass Classification," 2023, Accessed: May 15, 2024. [Internetinis]. Pasiekiamas: http://unfail.ccec.unf.edu/marshdata.html

46. “APTOS 2019 Blindness Detection | Kaggle.” Accessed: May 15, 2024. [Internetinis]. Pasiekiamas: https://www.kaggle.com/competitions/aptos2019-blindness-detection/discussion/107926

PRIEDAI

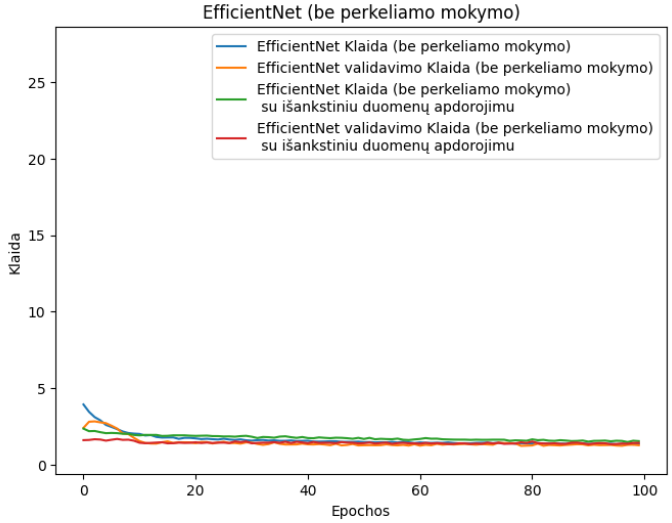
A. Priedas. Tikslumo grafikas



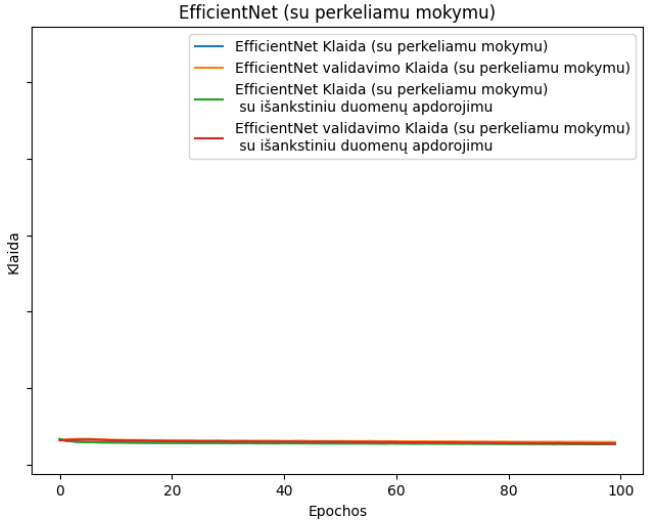
B. Priedas. Klaidos gradikas

Klaida

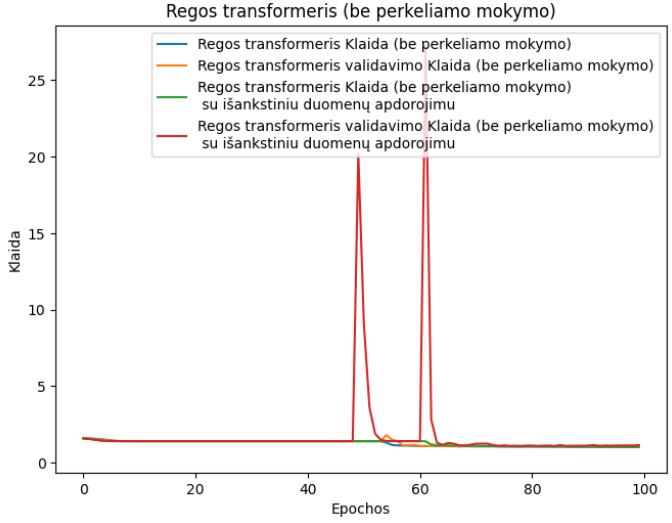
1



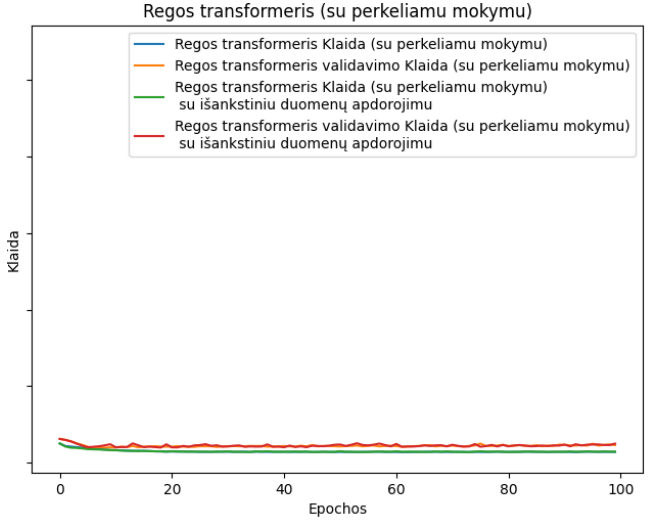
2



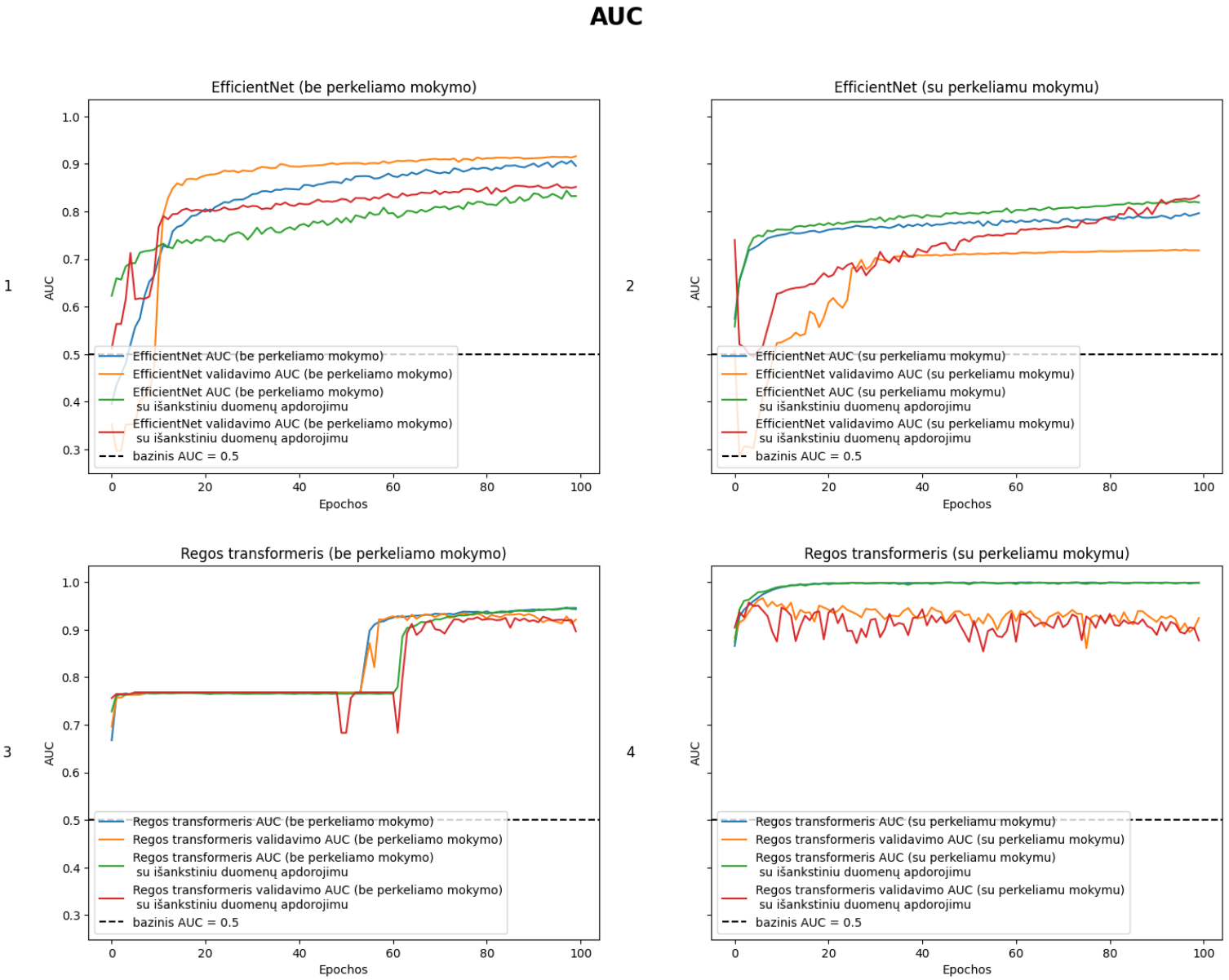
3



4



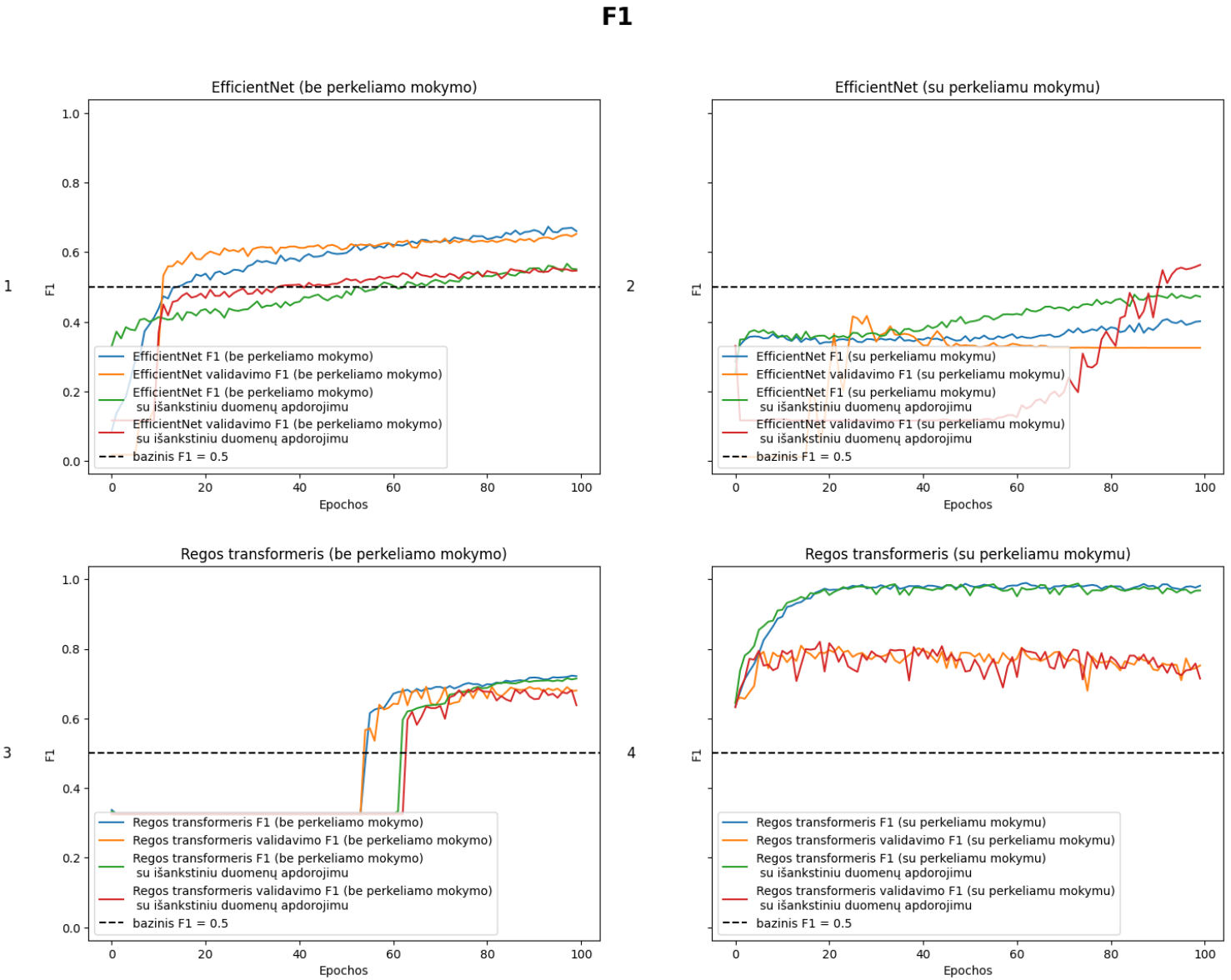
C. Priedas. AUC metrikos grafikas



1 2 3 4

Akių ligų klasifikavimo metodų taikomų tinklainės vaizdams tyrimas

D. Priedas. F1 metrikos grafikas



1 2 3 4

E. Priedas. Pristatymo eStream 2024 konferencijoje sertifikatas

VILNIUS TECH
Elektronikos fakultetas

8th Open International Conference "Electrical, Electronic and Information Sciences"

eStream 2024

CERTIFICATE

Rūta Jankevičiūtė and
Vytautas Abromavičius

Delivered an Oral Presentation

EXPLORING THE EFFECTIVENESS OF VISION TRANSFORMERS IN DIABETIC RETINOPATHY IDENTIFICATION VIA RETINAL IMAGING

Chair of Science Committee Prof. Dr Dalius Navakauskas

Section Chair Assoc. Prof. Dr Dainius Udris

April 25, 2024
Vilnius

F. Priedas. eStream 2024 konferencijos straipsnis

Exploring the Effectiveness of Vision Transformers in Diabetic Retinopathy Identification via Retinal Imaging

Rūta Jankevičiūtė
Department of Electronic Systems
Vilnius Gediminas Technical University
Vilnius, Lithuania
ruta.jankeviciute@stud.vilniustech.lt

Vytautas Abromavičius
Department of Electronic Systems
Vilnius Gediminas Technical University
Vilnius, Lithuania
0000-0003-1588-6572

***Abstract*—Vision transformers (ViTs) have begun to be adopted in medical imaging analysis. However, despite their success in other areas of computer vision, ViTs are still largely unexplored in this specific medical imaging task. In this preliminary study, we aim to assess the performance of the ViT in the detection of diabetic retinal retinopathy. In addition, we seek to identify potential avenues for future research and provide insight into the unique challenges and advantages associated with the use of VITs in the context of the identification of diabetes retinal diseases.**

***Index Terms*—artificial intelligence, diabetic retinopathy, machine learning, medical image classification, vision transformers**

I. Introduction

Medical imaging analysis plays an essential role in early disease detection, prediction, and treatment planning. With the advent of deep learning, the convolutional neural networks (CNN) has become the cornerstone of many state-of-the-art approaches in this area [1]–[4]. However, despite its effectiveness, CNN has inherent limitations in capturing long-range dependence in images, which are often crucial to accurate diagnosis in complex medical imaging tasks.

Interestingly, the potential of visual transformers (ViTs) in medical image analysis remains largely unknown. Unlike CNN, which relies on local convolutional operations, ViT models are capable of capturing global dependencies in the image by treating the image as a sequence of patches. CNNs have been the dominant model architecture for image classification tasks for many years. Although they are still the most popular in the literature, especially when it comes to medical image classification, at least by their popularity in PubMed publications, as seen in Fig. 1, Vision Transformers have slowly caught up.

ViT has yet to be as widely used in the variety of datasets available for medical image analysis, which can be due to many reasons, including how recent they are, the lack of extensive papers on how they should be optimized for each different dataset and their general need of large amounts of data [5].

The transformer architecture itself was first introduced in 2017 [6] and was initially used for natural language processing tasks. The architecture was only applied in hybrid forms, mostly as a mix of convolutional neural networks and parts of the transformer, such as its self-attention algorithm [7], until 2020, when Dosovitskiy et al. introduced the vision transformer (ViT) [8].

The ViT model was introduced as an image classification model and has been adopted quickly, even if, as mentioned before, it's still not nearly as widely researched as CNN models. ViTs are considered a very perspective type of model, especially for medical image classification tasks [9]–[12], even if it has been pointed out that, in many cases, it is hard to reach state of the art results [11]. ViTs with their ability to capture long-term dependencies through self-awareness mechanisms, provide an important way of investigating the identification of diabetic retinopathy. Surprisingly, despite their success in other areas of computer vision, ViTs are still largely unexplored in this specific medical imaging task.

Therefore, our main objective is to conduct a research into the effectiveness of ViTs for the diagnosis of diabetic retinal

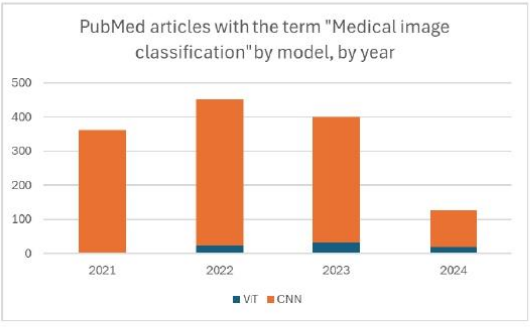


Fig. 1. Count of PubMed articles, by relevant year with the search terms "medical image classification" and the respective model ("ViT" or "CNN").

retinopathy through retinal imaging. Through this preliminary study, we aim to assess the performance of the ViTs in the detection of diabetic retinopathy tasks. In addition, we seek to identify potential avenues for future research and provide insight into the unique challenges and advantages associated with the use of VITs in the context of the identification of diabetic retinal diseases.

A. ViT architecture

Transformers themselves are deep neural networks with a decoder-encoder architecture. Both blocks contain multi-head attention blocks and feed-forward blocks. The encoder, essentially, analyzes how parts of the input sequences, made into tokens, associate with each other, while the decoder tries to guess which token should be next in the output sequence, taking into account the encoders' augmented contextual information.

ViT only uses the encoder, as the decoder output is not necessary for classification tasks. Unlike traditional convolutional neural networks operate directly on raw pixel values, ViT requires structured input embedding. The input image is divided into fixed-size patches, as seen in Figures 2 and 3. If the input image has dimensions $H \times W$ and each patch has dimensions $P \times P$, then the number of patches is

$$N = \frac{HW}{P^2}. \tag{1}$$

Each patch is linearly projected into a lower-dimensional space to obtain the embedding of the patch. These patch embedding are then flattened into sequences where each patch corresponds to a token.

The transformer encoder consists of multiple layers, each containing a multi-head self-attention layer followed by a position-wise feed-forward network. The self-attention mechanism computes the attention scores between all pairs of tokens in the input sequence. Given a sequence of token embedding X with dimensions (N, d_{model}), the self-attention mechanism computes the attention scores $\text{Attention}(Q, K, V)$ as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where $Q = XW_Q$, $K = XW_K$, $V = XW_V$, and W_Q, W_K, W_V are learnable linear projections. Each token embedding passes through a position-wise feed-forward network, which applies two linear transformations with an activation function (commonly ReLu) in between. The output embedding of the last transformer layer includes a special classification token, which is fed into a classification head to predict the class label. The classification head typically consists of a linear layer followed by a softmax activation function to output class probabilities.

II. Methodology

The following section describes the data, pre-processing, model, and training used for the experiment described in this paper.

Fig. 2. Example image from the APTOS 2019 dataset, not patched [13].

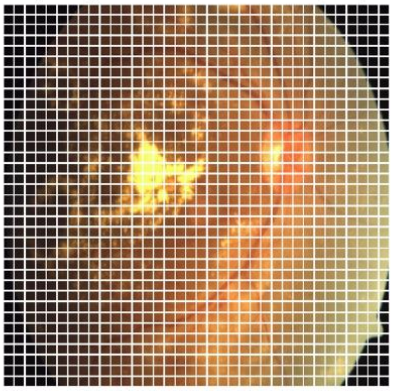


Fig. 3. Example image from the APTOS 2019 dataset [13], separated into 7×7 pixel patches.

A. APTOS 2019 dataset

The APTOS 2019 dataset [13] is a collection of retinal images acquired from patients with diabetic retinopathy (DR) to develop automated algorithms to detect and diagnose diabetic retinopathy. This dataset was provided as part of a Kaggle competition organized by the Asia Pacific Tele-Ophthalmology Society (APTOS) in 2019.

DR is primarily diagnosed and graded through the use of retinal images, also known as fundus images, or images of the interior of the eye, by spotting red spots known as microanuerysms and bright yellow lesions called exudates [14], both of which are features visible in the dataset and that should be able to be identified by image classification models.

B. Dataset description

The dataset comprises high-resolution color fundus photographs captured using various imaging devices. Each image is labeled with a severity grade that indicates the presence

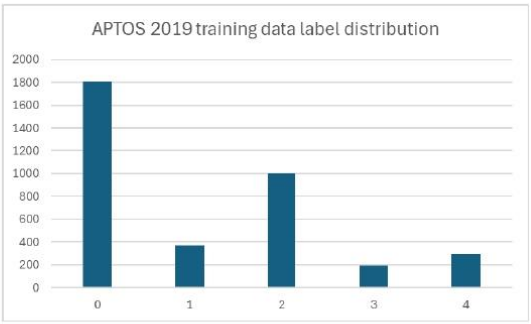


Fig. 4. The image distribution between the five possible labels in the training dataset of the APTOS 2019 dataset [13].

and severity of diabetic retinopathy. The severity grades are defined as follows:

- 0: No DR;
- 1: Mild DR;
- 2: Moderate DR;
- 3: Severe DR;
- 4: Proliferative DR;

The training dataset, used for the model in this paper, consists of 3662 images and their label distribution is shown in Fig. 4.

The dataset exhibits class imbalance, with a greater proportion of images belonging to lower severity grades (e.g., grades 0 and 1) compared to higher severity grades (e.g., grades 3 and 4). In addition, the images in the dataset can vary in terms of illumination and focus, reflecting the real-world variations encountered in clinical practice.

C. Image pre-processing

The images used for training and validation were not extensively augmented, only image normalization, resizing to a common size of 256×256 and geometric transforms, such as horizontal flipping, rotation and random zoom, were applied.

While more extensive pre-processing for these types of medical image datasets is not unheard of, geometric transforms like the ones used and described above have a much lower potential of losing necessary information, such as the contrast and colors that help distinguish between a healthy retina and signs of DR. Furthermore, even though the dataset has some variation in quality, the images are high enough resolution that cleanup techniques, such as de-noising, seemed unnecessary.

D. Model and training

This paper uses the transformer architecture described in Section I-A, from the original paper [8]. The parameters used for the model are described in Table I. The learning rate and weight decay were both set at 0.0001, the model ran for 100 epochs with a batch size of 16 and each image was divided into 7×7 pixels patches, resulting in 1286 patches per image. The model used 4 attention heads and 8 transformer layers with a projection dimension of 64. For training, the Adam optimizer was used.

TABLE I
PARAMETERS USED FOR THE VIT MODEL IN THIS PAPER

Parameter	Value
Learning rate	0.0001
Weight decay	0.0001
Batch size	16
Epochs	100
Patch size	7×7
Patches per image	1296
Projection dimension	64
Number of attention heads	4
Number of transformer layers	8

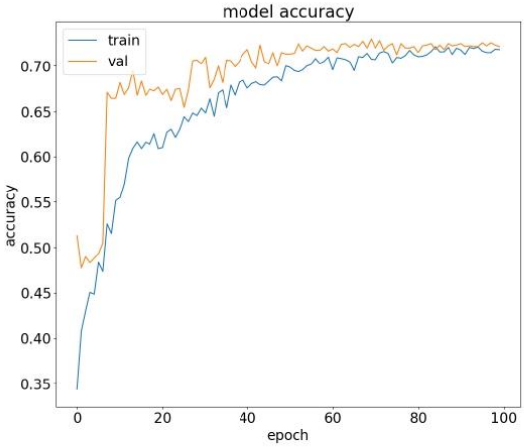


Fig. 5. Training and validation accuracy of the investigated model during the 100 epoch run.

III. EXPERIMENT RESULTS

As described in Table I, the model was run for 100 epochs with a batch size of 16 on the training dataset. The data was split into 2453 images for training and 1209 for validation. The model was trained using an A100 Nvidia GPU in the Google Colaboratory environment, with each epoch taking approximately 23 seconds to run.

Training ended with a final loss of 0.8279 and an accuracy of 0.7295. As seen in figures 6 and 5, the training was steady, as both the training and validation graphs stayed close together throughout the run, meaning the model didn't over or underfit. The accuracy increased significantly during the first half of training, but leveled off afterwards.

There can be a variety of reasons for the plateau, including the considerably small amount of training data an not enough hyperparameter tuning.

IV. DISCUSSION AND CONCLUSIONS

Compared to some other work, such as Bodapati et al. use of deep CNN models [15], or the models that have placed high on the Kaggle competition leader boards, the results of

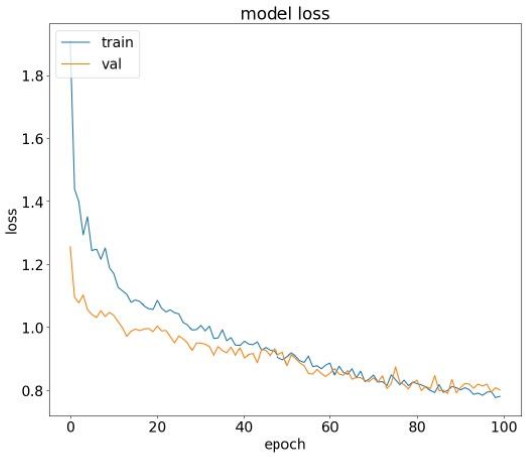


Fig. 6. Training and validation loss of the investigated model during the 100 epoch run.

the model in this paper are worse, with some models reaching accuracies of over 90%.

ViT models generally perform better with more data than CNN models, so combining multiple datasets or using transfer learning might be helpful to achieve more state-of-the-art results. Moreover, better pre-processing, such as focusing on the green channel, as described in [16], could help the model tune to the data better and find more features.

Hyper parameters and the shallowness of the model could also affect its ability to accurately classify the images, however, through testing with more attention heads (up to 8, double the amount used for the results shown), more layers (up to 12 from the 8 used in the paper) there was little to no improvement in the models performance, with the accuracy and loss staying about the same. Deepening the model does make it use more resources, but for no benefit to it's score.

In conclusion, our preliminary study into the effectiveness of ViTs for identifying diabetic retinal retinopathy via retinal imaging has provided valuable insights into the potential applications of ViTs in this crucial medical imaging task. Although our results may not show ViTs superiority over conventional CNN methods, they nevertheless shed light on ViTs' feasibility as an alternative architecture for diabetic retinal detection. By our investigation we aim to inspire further research in this direction.

Regarding the future directions, a promising way forward could be better tuning, studies of the impact of data enhancement techniques, especially for retinal images, applying transfer learning with a larger, pre-trained ViT, or simply exploring the use of more data by combining multiple diabetic retinopathy datasets.

REFERENCES

[1] B. Cicėnas and V. Abromavičius, "Investigation of pneumonia detection using convolutional neural networks," in *2022 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*. IEEE, 2022, pp. 1–4.

[2] R. Maskeliūnas, A. Kulikajevas, R. Damaševičius, J. Griškevičius, and A. Adomavičienė, "Biomac3d: 2d-to-3d human pose analysis model for tele-rehabilitation based on pareto optimized deep-learning architecture," *Applied sciences*, vol. 13, no. 2, p. 1116, 2023.

[3] S. Hussain, I. Mubeen, N. Ullah, S. S. U. D. Shah, B. A. Khan, M. Zahoor, R. Ullah, F. A. Khan, M. A. Sultan *et al.*, "Modern diagnostic imaging technique applications and risk factors in the medical field: a review," *BioMed research international*, vol. 2022, 2022.

[4] M. Elsharkawy, M. Elrazzaz, A. Sharafeldeen, M. Alhalabi, F. Khalifa, A. Soliman, A. Elnakib, A. Mahmoud, M. Ghazal, E. El-Daydamony *et al.*, "The role of different retinal imaging modalities in predicting progression of diabetic retinopathy: A survey," *Sensors*, vol. 22, no. 9, p. 3490, 2022.

[5] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: a comprehensive review," *Medical Image Analysis*, p. 103000, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*. PMLR, 2018, pp. 4055–4064.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] A. Anaya-Isaza, L. Mera-Jiménez, and M. Zequera-Diaz, "An overview of deep learning in medical imaging," *Informatics in Medicine Unlocked*, vol. 26, 1 2021.

[10] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[11] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it time to replace cnns with transformers for medical images?" *arXiv preprint arXiv:2108.09038*, 2021.

[12] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, p. 102802, 2023.

[13] S. D. Karthik, Maggie, "Aptos 2019 blindness detection," 2019. [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection

[14] S. D. Yu, L. L. Liu, Z. Y. Wang, G. Z. Dai, and Y. Q. Xie, "Transferring deep neural networks for the differentiation of mammographic breast lesions," *Science China Technological Sciences*, vol. 62, 2019.

[15] N. Sikder, M. S. Chowdhury, A. S. M. Arif, and A.-A. Nahid, "Early blindness detection based on retinal images using ensemble learning," in *2019 22nd International conference on computer and information technology (ICCIT)*. IEEE, 2019, pp. 1–6.

[16] P. Vasuki, J. Kanimozhi, and M. B. Devi, "A survey on image preprocessing techniques for diverse fields of medical imagery," in *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*. IEEE, 2017, pp. 1–6.