

Vještačka inteligencija

*Tema: Klasifikacija medicinskih slika koristeći
ručnu implementaciju VGG*

Uradili: Memić Alem 18827
Žuga Ernad 18814

Project Backlog

Faza 1: Izbor teme i opis problema	3
Opis problema koji se rješava kroz projektni zadatak	3
Definisanje osnovnih pojmova i koristi koje rješavanje tog problema donosi (primjena)	4
Kratki pregled postojećih dataset-ova povezanih sa problemom koji se rješava	5
Faza 2: Pregled stanja u oblasti	6
Analiza trenutnog stanja u oblasti problema	6
Opseg problema rješavan do sada i korištene metode vještačke inteligencije	7
Postignuti rezultati	8
Faza 3: Izbor, analiza i pretprocesiranje dataset-a	9
Dataset koji će biti korišten za rješavanje problema i njegova detaljna analiza	9
Osnovni pregled izabranog dataset-a	10
Objasniti metode pretprocesiranja podataka i objasniti zašto se koriste i koje su potrebne	11
Identifikovati rizike (npr. uzorci iz dvije klase su slični jedni drugima, imbalans klase – nekih klase ima znatno manje od ostalih)	12
Pretprocesiranje podataka	13
Faza 4: Odabir, formiranje, treniranje i testiranje modela	13
Način (metoda) koja će biti korištena za rješavanje problema	13
Tehnologije koje će biti korištene	14
Opisati kako je model treniran	14
Opis metrika koje su korištene	18
Poređenje metrika i ograničenja accuracy metrike	19
Dobijena rješenja i osvrt na istaknute rizike	21
Izlaganje modela podacima koji su mu nepoznati	23
Faza 5: Cjelokupni osvrt na problem i dobijeno rješenje	24
Osvrt na postignute rezultate	24
Uporediti sa radovima iz prethodne faze	24
Prodiskutovati šta se moglo bolje uraditi	25
Zaključak	26

Faza 1: Izbor teme i opis problema

Za projektni zadatak iz predmeta Vještacka Inteligencija, odabrali smo temu **Klasifikacija medicinskih slika koristeći ručnu implementaciju VGG**. S obzirom da je ovo široka tema i pokriva više medicinskih grana, odabrali smo da se fokusiramo na klasifikaciju plućnih rendgenskih slika. U ovom dokumentu će biti opisan projekat i njegovi dijelovi, a sam kod će biti pisan u Google Colabu, (MedMNISTVGG 2025), na linku: [MedMNIST-VGG](#)

Opis problema koji se rješava kroz projektni zadatak

U ovom projektnom zadatku fokus je na razvoju sistema za automatsku klasifikaciju plućnih rendgenskih snimaka (CXR) korištenjem konvolucijskih neuronskih mreža (CNN), konkretno ručno implementirane VGG arhitekture. Cilj je identifikacija prisustva jedne ili više od 14 mogućih plućnih patologija na osnovu CXR slika. Ovaj problem se zapravo formalizira kao multi-label-klasifikacijski zadatak, gdje jedan snimak može biti povezan sa više dijagnoza istovremeno.

Sami rendgenski snimci grudnog koša su ključni i koriste se kao dijagnostički alat u medicini. Međutim njihova interpretacija može biti izazovna zbog subjektivnosti i varijabilnosti među radiolozima. Stoga automatizacija procesa interpretacije može poboljšati tačnost dijagnoza, omogućiti širu dostupnost dijagnostičkih alata i naravno smanjiti vrijeme obrade. Sve ove navedene prednosti automatizacije mogu biti od izuzetne koristi, pogotovo kada se radi o sredinama sa ograničenim resursima.

Implementacijom ovog sistema, može se omogućiti brža i konzistentnija analiza velikog broja CXR snimaka. Također ovo može koristiti i kao dodatni alat za edukaciju medicinskog osoblja, ali naravno može i unaprijediti i doprinijeti smanjenju troškova zdravstvene zaštite.

Definisanje osnovnih pojmova i koristi koje rješavanje tog problema donosi (primjena)

Prvo ćemo navesti ključne pojmove:

Konvolucijske neuronske mreže (CNN): To je zapravo tip neuronskih mreža, koji je posebno efikasan za obradu slika. Koriste se konvolucijski slojevi za automatsko izdvajanje značajki iz datih ulaznih podataka.

VGG arhitektura: To je specifična CNN arhitektura, ona je poznata po svojoj jednostavnosti i dubini istovremeno. Koriste se mali (3x3) konvolucijski filteri i slojevi se nadovezuju jedan na drugi.

Multi-label klasifikacija: To je klasifikacija gdje jedan ulazni podatak, npr. slika, može biti povezan sa više izlaznih oznaka, jedan od primjera je prisustvo više bolesti.

Primjena i koristi:

Poboljšanje dijagnostike: Automatska klasifikacija može pomoći u ranom otkrivanju plućnih bolesti, to kasnije naravno pomaže u pravovremenom liječenju.

Dostupnost dijagnostike u udaljenim područjima: Ovakvi sistemi mogu omogućiti osnovnu dijagnostiku u nekim regijama gdje nedostaje stručnog medicinskog osoblja.

Smanjenje opterećenja na radiologe: Korištenjem automatizovanih sistema, omogućava radiolozima da se fokusiraju na složenije i teže slučajeve, tako što će sistem preuzeti dio posla ako se radi o nekim osnovnim ili lakšim slučajevima.

Edukacija i istraživanje: Korištenje automatizovanih sistema može poslužiti kao dobar alat za obuku medicinskog osoblja ali i studenata. Moguće je lakše pružiti uvid u analize, koje također mogu biti automatizovane, ali i uvid u potencijalne dijagnoze.

Kratki pregled postojećih dataset-ova povezanih sa problemom koji se rješava

Postoji nekoliko javno dostupnih dataset-ova koji se koriste konkretno za temu klasifikacije plućnih rendgenskih snimaka, nabrojat ćemo neke od njih i kratko opisati.

ChestMNIST

Opis: Sadrži 112120 frontalnih CXR slika od 30805 pacijenata, sa oznaka za 14 plućnih bolesti. Slike koje se koriste su standardizirane na veličinu od 28x28 piksela, što omogućava brzu obuku modela i eksperimentisanje sa različitim arhitekturama.

Primjena: Pogodan za edukativne svrhe i brz prototip modela.

NIH ChestX-ray14

Opis: ovaj dataset sadrži preko 220000 CXR slika sa oznakama za 14 različitih plućnih bolesti, koje uključuju i informacije o nesigurnosti u oznakama.

Primjena: Koristi se za istraživanje i razvoj naprednih modela za detekciju plućnih bolesti.

CheXpert

Opis: sadrži preko 220000 CXR slika sa oznakama za 14 različitih patologija, uključujući i informacije o nesigurnostima u oznakama.

Primjena: Koristi se za razvoj modela koji mogu upravljati nesigurnostima u medicinskim podacima.

MIMIC-CXR

Opis: otvoreni skup podataka sa više od 370000 CXR slika i pripadajućim radiološkim izvještajima. To omogućava istraživanje i razvoj modela koji kombinuju slike i tekstualne podatke.

Primjena: Pogodan za istraživanje u oblasti kombinovanja vizualnih i tekstualnih podataka u medicinskoj dijagnostici.

Faza 2: Pregled stanja u oblasti

Analiza trenutnog stanja u oblasti problema

Analiza trenutnog stanja u oblasti klasifikacije plućnih rendgenskih snimaka (CXR) pomoću konvolucijskih neuronskih mreža (CNN), s fokusom na VGG arhitekturu, pokazuje značajan napredak u primjeni dubokog učenja u medicinskoj dijagnostici. Ova tehnologija omogućava automatsko prepoznavanje i klasifikaciju različitih plućnih patologija na osnovu CXR slika, što može poboljšati tačnost dijagnoza i smanjiti opterećenje na radiologe. Nekoliko ključnih naučnih radova, doprinijelo je napretku u ovoj oblasti:

(Rajpurkar et al. (2017) – CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning 2017): *Prvi veliki rad koji koristi 121-slojni DenseNet za detekciju pneumonije na NIH ChestX-ray14 datasetu.*

(Irvin et al. (2019) – CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison 2019):

Detaljno predstavljen CheXpert dataset i izazovi klasifikacije sa nesigurnim oznakama.

(Ouyang et al. (2020) – Learning to diagnose from scratch by exploiting dependencies among labels 2020): *Fokus na modeliranju veza između dijagnoza u multi-label kontekstu, efikasan za realne medicinske scenarije.*

(arXiv 2019): *Razvijen je algoritam koji može detektovati pneumoniju iz CXR slika na nivou koji nadmašuje prosječnog radiologa. CheXNet je 121-slojna konvolucijska neuronska mreža trenirana na ChestX-ray14 datasetu.*

(arXiv 2017): *Ova studija je pokazala da manji modeli poput VGG-16 mogu postići slične performanse kao i dublje mreže poput DenseNet-201, uz manju računarsku složenost.*

(ScienceDirect 2021): *U ovom istraživanju, ResNet-50 je postigao višu tačnost u klasifikaciji pneumonije iz CXR slika u poređenju s VGG-19 i modelima treniranim od nule.*

VGG arhitektura je starija CNN arhitektura, ali je i dalje korisna kao početna tačka, zbog svoje jednostavnosti za implementaciju, također omogućava lako razumijevanje rada konvolucijskih mreža i može se efikasno koristiti na datasetovima kao što su ChestMNIST ili manji subsetovi NIH/ChestXpert. Ipak,

danas se više koriste ResNet, DenseNet i EfficientNet, koji daju bolje rezultate uz manju količinu parametara.

Datasetove smo već naveli ranije, oni se najčešće koriste i najviše su upotrebljivi, ali možemo dodati još jedan. To je **PadChest (2019)**, sadrži 160000 slika i više sa oznakama i tekstualnim izvještajima na španskom jeziku.

Što se tiče repozitorija, izdvojit ćemo neke:

(covid-chestxray-dataset 2021): *popularan dataset i model za COVID klasifikaciju.*

(Datasets 2020): *opšti resurs za medicinske datasetove.*

Na osnovu prethodno analiziranih stavki, može se zaključiti da primjena konvolucijskih neuronskih mreža (CNN), posebno VGG arhitekture, u klasifikaciji plućnih rendgenskih snimaka predstavlja značajan napredak u oblasti medicinske dijagnostike. Korištenje javno dostupnih datasetova, omogućilo je razvoj i evaluaciju modela koji postižu visoke performanse u detekciji plućnih patologija. U budućnosti, fokus bi trebao biti na razvoju modela koji su robusni na različite izvore podataka i koji mogu pružiti pouzdane rezultate u stvarnim kliničkim okruženjima. Također, integracija dodatnih informacija, poput kliničkih podataka pacijenata, može dodatno unaprijediti tačnost i korisnost ovih modela u praksi.

Opseg problema rješavan do sada i korištene metode vještačke inteligencije

U dosadašnjim istraživanjima klasifikacije plućnih rendgenskih snimaka (CXR) pomoću metoda vještačke inteligencije, fokus je bio na detekciji i klasifikaciji različitih torokalnih bolesti, uključujući pneumoniju, tuberkulozu, COVID-19 i pneumotoraks. Primijenjene metode obuhvataju različite arhitekture dubokih konvolucijskih neuronskih mreža (CNN), pri čemu su VGG, ResNet, DenseNet, MobileNet i EfficientNet među najčešće korištenim. VGG arhitektura je korištena u 26,5% studija, dok su ResNet, MobileNet i DenseNet korišteni u 20,6%, 14,7% i 14,7% respektivno. VGG je ostvario 93,62% tačnosti. Nedavna istraživanja su također uključila napredne tehnike poput mehanizama pažnje i samonadziranog učenja za poboljšanje interpretabilnosti i performansi modela.

Transfer učenje je široko primijenjeno sa 61,8% studija koje su koristile ovu tehniku za poboljšanje performansi modela, posebno u slučajevima ograničenih skupova podataka.

Uprkos značajnom napretku, izazovi ostaju u pogledu generalizacije modela na različite izvore podataka i kliničke scenarije. Buduća istraživanja trebaju se fokusirati na razvoj robusnih modela koji mogu pouzdano djelovati u stvarnim kliničkim okruženjima, uz integraciju dodatnih kliničkih informacija za poboljšanje dijagnostičke tačnosti.

Postignuti rezultati

Kao što smo ranije rekli, u dosadašnjim istraživanjima klasifikacije plućnih rendgenskih snimaka (CXR) pomoću metoda vještačke inteligencije, fokus je bio na detekciji i klasifikaciji različitih torakalnih bolesti, uključujući pneumoniju, tuberkulozu, COVID-19 i pneumotoraks. Primijenjene metode obuhvataju različite arhitekture dubokih konvolucijskih neuronskih mreža (CNN), pri čemu su VGG, ResNet, DenseNet, MobileNet i EfficientNet među najčešće korištenima.

Primjena ovih metoda rezultirala je značajnim uspjesima u klasifikaciji plućnih bolesti:

- **CheXNet**, baziran na DenseNet-121 arhitekturi, postigao je AUC od 0.841 na ChestX-ray14 datasetu, što je usporedivo s performansama radiologa u detekciji pneumonije.
- Na CheXpert datasetu, **ResNet-152** i **DenseNet-161** modeli ostvarili su AUC od 0.882 i 0.881, respektivno, što ukazuje na visoku tačnost u detekciji torakalnih patologija.
- **CheXDet**, model treniran s detaljnim anotacijama lezija, nadmašio je CheXNet u klasifikaciji fraktura na NIH-Google datasetu (AUC: 0.67 naspram 0.51) i PadChest datasetu (AUC: 0.78 naspram 0.55), demonstrirajući važnost preciznih anotacija za poboljšanje generalizacije modela.
- Na MIMIC-CXR datasetu, modeli poput DenseNet-121 i ResNet-50 postigli su AUC vrijednosti iznad 0.84 za detekciju edema i pleuralnog izljeva, što potvrđuje njihovu efikasnost u kliničkim scenarijima.

Dakle, dosadašnja istraživanja demonstriraju da budoke konvolucijske neuronske mreže, posebno arhitekture poput DenseNet i ResNet, mogu postići visoke performanse u klasifikaciji plućnih rendgenskih snimaka. Međutim, izazovi poput generalizacije modela na različite izvore podataka i potreba za detaljnim

anotacijama ostaju pristupačni. Buduća istraživanja, trebaju se fokusirati na razvoj robusnih modela koji mogu pouzdano djelovati u stvarnim kliničkim okruženjima, uz integraciju dodatnih kliničkih informacija za poboljšanje dijagnostičke tačnosti.

Faza 3: Izbor, analiza i pretprocesiranje dataset-a

Dataset koji će biti korišten za rješavanje problema i njegova detaljna analiza

Za rješavanje zadatka klasifikacije plućnih rendgenskih snimaka, izabrali smo ChestMNIST dataset iz MedMinst kolekcije. Ovaj skup podataka sadrži 112 120 frontalnih CXR slika, pri čemu su sve iste veličine (28x28 piksela) i u grayscale formatu, što značajno pojednostavljuje ulazne podatke i smanjuje računarske zahtjeve. ChestMNIST je posebno prilagođen za edukativne i istraživačke svrhe jer dolazi sa već zadanim podjelama na skupove za trening (70%), validaciju (10%) i testiranje (20%).

Svaka slika nosi oznake iz multi-label vektora dužine 14, pri čemu svaka od 14 bolesti, može biti prisutna ili odsutna neovisno o ostalim. Takva konfiguracija zadatka čini ga zahtjevnim od klasične binarne klasifikacije, jer model mora istovremeno naučiti prepoznati i često preklapajuće ili podudarajuće patološke obrasce. Iz perspektive karakteristika dataset-a, ukupna veličina od oko 30 MB, čini ga vrlo dostupnim i prikladnim za lokalni razvoj bez potrebe za moćnim GPU klastreima. Međutim, snižena rezolucija od 28x28 piksela znači da suptilne morfološke promjene na snimcima mogu biti izgubljene. Zbog toga je u predprocesiranju ključno pažljivo odabrati metode normalizacije i augmentacije koje neće dodatno degradirati informacije ključne za dijagnostiku.

Analiza same distribucije podataka otkriva da je najčešća oznaka „No Finding“ u približno 60 361 primjeru, dok su neke patologije zastupljene samo s nekoliko stotina do hiljadu primjera. Takav disproporcionalan omjer može dovesti do pristranosti modela, pa iz tog razloga bi u evaluaciji trebamo koristiti i metrike osjetljivosti (recall) i F1-score, a ne samo ukupnu tačnost.

ChestMNIST smo odabrali jer nudi dobar kompromis između veličine, složenosti i praktične primjene. Malen format omogućava brzi razvoj i iteracije s različitim arhitekturama (kao što je naša ručna implementacija VGG), dok bogatstvo oznaka pruža realističan izazov multi-label klasifikacije. U sljedećim fazama projekta, nakon što pripremimo i pretprocesiramo podatke, fokusirat ćemo se na gradnju, treniranje i evaluaciju VGG modela, s posebnim naglaskom na rješavanje uočenih rizika niske rezolucije i neuravnoteženosti klasa.

Osnovni pregled izabranog dataset-a

Izvor skupa podataka: ChestMNIST je dio MedMNIST kolekcije, dostupne na službenoj web-stranici MedMNIST <https://medmnist.com/>

Format u kojem je dostupan i način za preuzimanje: Grayscale slike rezolucije 28x28 piksela, pohranjene kao PNG fajlovi ili NumPy nizovi. Preuzimanje je moguće putem Python paketa medmnist ili direktno sa GitHub repozitorija.

Broj instanci: 112 120 frontalnih CXR slika.

Broj atributa: Svaka slika predstavlja vektor od 784 atributa (28x28 piksela). Ne postoje dodatni nestrukturani atributi unutar ChestMNIST-a.

Broj klasa: 14 mogućih plućnih stanja (npr. Infiltration, Effusion, Atelectasis, ...), plus implicitna klasa „No Finding“.

Broj instanci po klasama: No Finding – 60361, Infiltration – 19894, Effusion – 13 317, Atelectasis – 11 559, Nodule – 6 331, Mass – 5 782, Pneumothorax – 5 302, Consolidation – 4 667, Pleural Thickening – 3 385, Cardiomegaly – 2 776, Emphysema – 2 516, Edema – 2 303, Fibrosis – 1 686, Pneumonia – 1 431, Hernia – 227.

Količina podataka: otprilike 30 MB.

Podjela na trening, validaciju i test:

-Trening: 70% (približno 78 484 slike)

-Validacija: 10% (približno 11 212 slika)

-Test: 20% (približno 22 424 slike)

Objasniti metode pretprocesiranja podataka i objasniti zašto se koriste i koje su potrebne

Metode pretprocesiranja podataka za ChestMNIST obično obuhvataju sljedeće korake:

Učitavanje i inicijalna provjera: učitavanje slika i labela iz NumPy nizova ili PNG fajlova. Provjera integriteta podataka, to znači osigurati da nema nedostajućih ili oštećenih fajlova, provjeriti dimenzije i tip.

Normalizacija piskela: Min-max skaliranje: vrijednosti x koje su u opsegu od 0 do 255 skalirati u vrijednosti od 0 do 1 uz formulu $x' = x/255$.

Konverzija tipa podataka: nakon skaliranja, konvertovati polje piksela u tip float32 ili float64, kako bi se izbjeglo zaokruživanje tokom računa u neuronskoj mreži.

Kodiranje labela (multi-label vektor): ChestMNIST koristi multi-label vektore dužine 14. Potrebno je implementirati one-hot ili multi-hot enkodiranje tako da se za svaku klasu čuva 0/1 vrijednost. Ovaj vektor se direktno koristi u loss funkciji.

Balansiranje klasa: **Oversampling** – povećati broj primjera rijetkih klasa dupliciranjem ili generiranjem sintetičkih. **Undersampling** – smanjiti broj primjera dominantnih klasa kako bi se izbjegla pristranost.

Augmentacija podataka: primjenjuje se isključivo na trening skup kako bi se povećala varijabilnost i spriječilo preučenje. Tu spadaju rotacije, translacije, skaliranje, horizontalni flip, dodavanje šuma. Svaka augmentacija se obično nasumično bira za svaku epohu, čime se simulira širok spektar uvjeta snimanja.

Podjela skupa i priprema za trening: Shuffle cijelog skupa prije podjele kako bi se izbjegao redoslijed po pacijentima. Podjela na trening, validaciju i testiranje. Batching, formiranje mini-batcheva koji sadrže reprezentativnu mješavinu klasa.

Dodatne tehnike: Denoising, ukoliko je potrebno primijeniti filtere za uklanjanje artefakata. Histogram equalization, poboljšanje kontrasta, iako oprezno, jer može promijeniti ključne medicinske signale. Maskiranje, fokus na region interesa, ako su dostupne segmentacione maske.

Identifikovati rizike (npr. uzorci iz dvije klase su slični jedni drugima, imbalans klasa – nekih klasa ima znatno manje od ostalih)

Postoji nekoliko rizika koje smo prepoznali, a to su:

Neuravnoteženost klasa – neke patološke oznake imaju desetine hiljada primjera, dok za rijetke bolesti postoji samo nekoliko stotina slika. Takav neravnomjeran omjer, može uzrokovati da model zanemari rijetke klase i postane pristran prema dominantim,

Sličnost između klasa – pojedine bolesti imaju slične radiološke karakteristike, pa model može imati poteškoća u razlikovanju finih razlika u teksturi i obrascima, to dalje može dovesti do lažno pozitivnih ili lažno negativnih predikcija.

Niska rezolucija slika – standardizacija na veoma mali format smanjuje detalje koji su nekad ključni za otkrivanje sitnih abnormalnosti. Time se ograničava maksimalna moguća tačnost modela.

Kvaliteta i pouzdanost oznaka – ChestMNIST koristi oznake naslijeđene iz ChestX-ray14, koje su automatski izdvojene iz radioloških izvještaja putem algoritama. Tu naravno postoji rizik od pogrešnih ili nesigurnih oznaka, što može smanjiti performanse modela.

Overfitting – manji broj primjera za rijetke klase i ograničeni broj varijacija povećava rizik da model zapamti specifične uzorke umjesto da nauči generalne značajke. Potrebno je pažljivo koristiti augmentaciju i regularizaciju.

Domain shift / razlike u izvoru slika – ChestMNIST uzorci dolaze iz različitih ustanova i uređaja, pa može postojati varijabilnost u kontrastu, osvjetljenju i orijentaciji. Model treniran na ovom skupu možda neće odmah generalizirati na CXR snimke iz drugih izvora.

Odsustvo dodatnih kliničkih podataka – samo su slike dostupne bez povezanih meta – podataka, što ograničava mogućnost izgradnje multimodalnih modela koji bi mogli bolje interpretirati kontekst pacijenta.

Etička i regulatorna pitanja – automatska dijagnoza nosi rizik pravne i etičke odgovornosti ukoliko model pogriješi. U kliničkoj primjeni, uvijek je potreban nadzor i potvrda stručnog radiologa.

Pretprocesiranje podataka

Što se tiče samog pretprocesiranja podataka, ono će biti obavljeno kroz kod u skripti, a ovdje ćemo dati neki osnovni tekstualni opis. Dakle ono što ćemo uraditi jeste konvertovanje slike iz PIL.Image formata u tenzor pri čemu će biti urađeno i skaliranje piksela od 0 do 1. Uradit ćemo i blage transformacije sa RandomRotation i RandomAffine radi veće generalizacije. Također uradit ćemo i standardizaciju vrijednosti u opseg $[-1, 1]$ zbog stabilnijeg rada konvolucione mreže. Koristit ćemo i DataLoader zbog ubacivanja podataka u model u mini-batchevima.

Faza 4: Odabir, formiranje, treniranje i testiranje modela

Način (metoda) koja će biti korištena za rješavanje problema

Za rješavanje problema klasifikacije rentgenskih (CXR) snimaka pluća iz ChestMNIST dataseta, odabrana je **konvolucijska neuronska mreža (CNN)**, i to konkretno **ručna implementacija VGG arhitekture**. Ova arhitektura se zasniva na istraživanju iz rada „Very Deep Convolutional Networks for Large-Scale Image Recognition“ (Simonyan & Zisserman, 2014), gdje se pokazalo da sekvencijalno nadslojavanje jednostavnih konvolucijskih slojeva malog filtera (3x3) uz povremeno reduciranje dimenzionalnosti kro max-poding može ostvariti izuzetne rezultate na zadacima klasifikacije slika.

U ovom projektu VGG se neće koristiti kao unaprijed trenirani model, već se kreće od nule, gdje ćemo mi zapravo trenirani model kako nam je potrebno. Time se omogućava bolje razumijevanje arhitekture i same dinamike treniranja neuronskih mreža. Ova metoda je prikladna jer ChestMNIST slike imaju rezoluciju (28x28 piksela), što znači da su i potrebe za memorijom i brojem slova umjerene pa je jednostavnija CNN arhitektura poput VGG-a dovoljno snažna za rješavanje ovog problema.

Tehnologije koje će biti korištene

Za implementaciju i testiranje izabran je sljedeći softverski stack:

- **Python:** glavni programski jezik zbog svoje čitljivosti i bogate podrške za rad s dubokim učenjem
- **PyTorch:** otvorena biblioteka za automatsku derivaciju i duboko učenje. PyTorch omogućava jednostavnu konstrukciju i treniranje neuronskih mreža, te je poznat po svojoj fleksibilnosti i jasnoći koda.
- **MedMNIST:** biblioteka koja omogućava jednostavan pristup medicinskim datasetovima. Omogućava standardiziran način učitavanja podataka bez potrebe za ručnim raspakivanjem ili formatiranjem.
- **Google Colab:** Cloud okruženje koje omogućava besplatan pristup GPU-u. Idealno je za rad sa dubokim učenjem jer ne zahtjeva lokalnu instalaciju i dolazi sa već instaliranim bibliotekama (PyTorch, NumPy, Matplotlib itd.)
- **Matplotlib & Seaborn:** biblioteke za vizualizaciju, korisne za prikaz uzoraka iz dataseta, vizualizaciju gubitka i tačnosti tokom treniranja i prikaz matrice konfuzije.

Ovakva kombinacija alata omogućava brzu izradu, treniranje i evaluaciju modela uz jednostavno testiranje u različitim fazama. Također sve tehnologije su besplatne i dobro dokumentovane, što je dodatna prednost u kontekstu obrazovnog projekta.

Opisati kako je model treniran

U ovom projektu, kako smo i ranije naveli, koristili smo ChestMNIST dataset, koji sadrži 112 120 grayscale slika dimenzija 28x28 piksela, podijeljenih na 70% trening (78 484 slike), 10% validaciju (11 212) i 20% test (22 424). Prije treniranja, svaki primjer je prošao kroz sljedeće korake:

1. Skaliranje i normalizacija

ToTensor(): transformiše uint8 vrijednosti piksela (0-255) u tenzore float32 u opsegu [0,1].

Normalize(mean = 0.5, std=0.5): centriranje na [-1, 1] radi stabilizacije učenja i brže konvergenције.

2. Augmentacija (samo na trening skupu)

Nasumična rotacija do $\pm 15^\circ$, kako bi model bio otporniji na blage promjene položaja snimanja.

Nasumična translacija do $\pm 10\%$ dimenzija, simulirajući različite okvire snimanja.

Horizontalni flip sa 50% vjerovatnoće, da model nauči da prepozna obrasce u obje orijentacije.

Sve transformacije su sastavljene u **torchvision.transforms.Compose**, a DataLoaderi su kreirani sa **batch_size = 64** i **shuffle = True** za trening, te **shuffle = False** za validaciju i test.

Arhitektura modela („VGG-like“) ima dva bloka konvolucijskih slojeva s po dva 3x3 filtera i po jednim MaxPool slojem, što iz ulaza 1x28x28 sažima do 64x7x7 značajki, te zatim klasifikacioni dio koji flatten-uje izlaz, projektuje ga linerarno na 256 neurona uz ReLU i dropout 0.5 i konačno na 14 izlaza sa Sigmoid aktivacijom, dajući za svaku klasu vrijednost između 0 i 1.

Za optimizaciju smo odabrali Adam($lr = 1 \times 10^{-3}$) zbog njegove samoadaptivne prilagodbe brzine učenja, a kao loss funkciju koristili smo **BCELoss**, idealnu za multi-label zadatke gdje se svaka od 14 bolesti procjenjuje neovisno.

Treniranje je obavljeno na Google Colab GPU-u (Tesla T4) i trajala je oko 30 sekundi do minute po epohi. Korišteni su sljedeći hiperparametri: **Optimizator** Adam sa početnom brzinom učenja $lr=0.001$, **Funkcija gubitka** BCELoss pogodna za nezavisnu binarnu klasifikaciju po labeli, **Veličina batcha** 64 i **Broj epoha je 20**. Proces za 20 epoha je izgledao ovako:

- **Epoha 1:**

Train Loss: 0.1848

Val Loss: 0.1737

- **Epoha 2:**

Train Loss: 0.1783

Val Loss: 0.1717

- **Epoha 3:**
Train Loss: 0.1764
Val Loss: 0.1695
- **Epoha 4:**
Train Loss: 0.1751
Val Loss: 0.1730
- **Epoha 5:**
Train Loss: 0.1741
Val Loss: 0.1683
- **Epoha 6:**
Train Loss: 0.1731
Val Loss: 0.1664
- **Epoha 7:**
Train Loss: 0.1772
Val Loss: 0.1693
- **Epoha 8:**
Train Loss: 0.1713
Val Loss: 0.1658
- **Epoha 9:**
Train Loss: 0.1708
Val Loss: 0.1662
- **Epoha 10:**
Train Loss: 0.1704
Val Loss: 0.1643
- **Epoha 11:**
Train Loss: 0.1697
Val Loss: 0.1647
- **Epoha 12:**
Train Loss: 0.1693
Val Loss: 0.1637
- **Epoha 13:**
Train Loss: 0.1689
Val Loss: 0.1641
- **Epoha 14:**

Train Loss: 0.1686

Val Loss: 0.1630

- **Epoha 15:**

Train Loss: 0.1681

Val Loss: 0.1637

- **Epoha 16:**

Train Loss: 0.1677

Val Loss: 0.1627

- **Epoha 17:**

Train Loss: 0.1673

Val Loss: 0.1624

- **Epoha 18:**

Train Loss: 0.1674

Val Loss: 0.1628

- **Epoha 19:**

Train Loss: 0.1672

Val Loss: 0.1619

- **Epoha 20:**

Train Loss: 0.1669

Val Loss: 0.1626

U svakoj epohi, **model.train()** je omogućio računanje gradijenta i ažuriranje težina, dok je validacija radila pod **model.eval()** i **with torch.no_grad()**, isključujući gradijente radi brže i determinističke provjere gubitka na neviđenim podacima. Primjetan je pad vrijednosti i Train Loss i Val Loss iz epohe u epohu, što znači da model uči i poboljšava se. Također kako nisu velike razlike u vrijednostima, to znači da model ne pamti podatke nego ih postepeno generalizuje. Također nema overfittinga, jer validacijski gubitak nije veći od trening gubitka, a s druge strane ne pokazuje ni rast što nam daje informaciju da model nije prenaučen tj. da nema overfittinga kako smo i rekli. Na kraju, u posljednjoj epohi, razlika između vrijednosti je mala, što govori da je model dobro izbalansiran.

Ovaj trening proces pokazuje stabilnu konvergenciju bez značajnog razilaženja između **train_loss** i **val_loss**, što ukazuje na dobru regularizaciju i minimalni overfitting. Također, važno je napomenuti da je trening izvršen na **Google Colab** okruženju koristeći **Tesla T4 GPU**, što je značajno ubrzalo proces učenja modela i

omogućilo efikasnu obradu podataka čak i uz složenije transformacije i augmentacije.

Opis metrika koje su korištene

U projektu smo koristili sljedeće metrike:

1. Precision (preciznost)

Preciznost mjeri udio ispravno predviđenih pozitivnih primjera u odnosu na sve primjere koje je model označio kao pozitivne. Za klasu i izračunava se kao:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

Pri čemu je:

TP_n (True Positives) broj ispravno predviđenih primjera klase i ,

FP_n (False Positives) broj primjera koje je model krivo označio kao klasu i .

Visoka preciznost znači da model rijetko daje lažno pozitivne rezultate.

2. Recall (odziv, osjetljivost)

Recall mjeri udio ispravno predviđenih pozitivnih primjera u odnosu na ukupan broj stvarnih pozitivnih primjera. Za klasu i :

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

Pri čemu je:

TP već objašnjen (True Precision) u dijelu za Precision,

FN_n (False Negatives) broj primjera koje je model krivo proglasio negativnima. Visok recall znači da model rijetko propušta pozitivne primjere.

3. F1-score

F1-score je harmonijska sredina preciznosti i recall-a, i daje jedinstvenu mjeru koja balansira ova dva aspekta:

$$F1_i = 2 * \frac{\text{Precision}_i * \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

F1-score je posebno koristan kod neuravnoteženih skupova podataka jer penalizira i lažno pozitivne i negativne rezultate.

4. AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

AUC-ROC mjeri sposobnost modela da rangira pozitivne primjere iznad negativnih bez obzira na konkretni prag. Vrijednost AUC od 1.0 označava savršenu separaciju, dok 0.5 znači nasumično pogađanje.

- **ROC krivulja** prikazuje odnos True Positive Rate (TRP) i False Positive Rate (FPR) za sve moguće pragove.
- **AUC** je površina ispod te krivulje i odražava ukupnu sposobnost razlikovanja klasa.

5. Micro-averaged F1

Pri konstruktiranju **micro-F1**, svi TP, FP i FN se sumiraju preko svih klasa prije izračuna F1-score. Time se daju jednaki težinski doprinosi svim primjercima, pa dominantne klase imaju veći uticaj.

6. Macro-averaged F1

Macro-F1 računa F1-score pojedinačno za svaku klasu, a zatim uzima aritmetičku sredinu tih vrijednosti. Time se svim klasama daje jednak težinski doprinos, bez obzira na broj primjera, što naglašava performanse na rijetkim klasama.

Možemo reći da ove metrike zajedno omogućavaju sveobuhvatnu procjenu modela, pri čemu:

- **Precision/Recall/F1** pokazuju konkretne kvalitetne odnose predikcija na nivou klasa,
- **AUC** ukazuje na sposobnost rangiranja i stabilnost modela kroz različite pragove,
- **Micro-F1** i **Macro-F1** daju globalne pokazatelje performansi s različitim pogledima na neravnotežu klasa.

Poređenje metrika i ograničenja accuracy metrike

U kontekstu zadatka multi-label klasifikacije sa neuravnoteženim klasama, kao što je to slučaj sa ChestMNIST datasetom, metrika *accuracy* nije pouzdana i ne daje realan prikaz stvarne performanse modela. Accuracy mjeri procenat ukupno tačno klasifikovanih instanci, ali ne uzima u obzir distribuciju klasa. Na primjer, ako jedna

klasa čini 90% podataka (npr. "No Finding"), model koji sve predikcije klasifikuje kao tu klasu može imati 90% accuracy, ali bi potpuno ignorisao sve druge bolesti, što je neprihvatljivo u medicinskoj dijagnostici.

Zbog toga u ovom projektu koristimo sveobuhvatniji skup metrika koje bolje odražavaju performanse modela po klasama:

- **Precision (preciznost)** - mjeri koliko su predikcije koje je model označio kao pozitivne zaista tačne. Visoka preciznost je važna kako bi se izbjegle lažne uzbune (false positives), npr. označavanje zdrave slike kao bolesne.
- **Recall (odziv, osjetljivost)** - mjeri koliko stvarnih pozitivnih slučajeva model uspije prepoznati. Ovo je naročito važno u medicinskoj domeni gdje je neprihvatljivo propustiti slučaj bolesti (false negatives).
- **F1-score** - predstavlja harmonijsku sredinu između preciznosti i recall-a. Pogodna je metrika za neuravnotežene podatke jer penalizira modele koji zapostavljaju jednu stranu (tačnost ili odziv).
- **AUC-ROC** - mjeri sposobnost modela da razlikuje između pozitivnih i negativnih instanci kroz sve moguće klasifikacione pragove. $AUC > 0.75$ je dobar pokazatelj da model prepoznaje osnovne obrasce.
- **Micro-F1** - agregira sve true positives, false positives i false negatives preko svih klasa i izračunava F1 globalno. Time dominiraju klase sa većim brojem primjera.
- **Macro-F1** - računa F1-score za svaku klasu pojedinačno i zatim izračunava njihov prosjek, tretirajući sve klase jednako. Ova metrika je ključna za praćenje učinka na rijetkim klasama.

Zbirno, ove metrike pružaju potpuniju i realniju evaluaciju učinka modela, pogotovo u domenu kao što je medicinska dijagnostika gdje greške mogu imati ozbiljne posljedice. Accuracy se može koristiti samo kao okvirni pokazatelj, ali nikako kao glavni kriterij za ocjenu kvaliteta modela.

Dobijena rješenja i osvrt na istaknute rizike

- **Proces treniranja:** Trening je proveden kroz 20 uzastopnih epoha, pri čemu se nadziralo tri ključne metrike: trenirajući gubitak (Train Loss), validacijski gubitak (Val Loss) i validacijski AUC (Val AUC). Na početku, u prvoj epohi, trenirajući gubitak iznosio je 0.1848, dok je gubitak na validacijskom skupu bio 0.1737, uz odgovarajući Val AUC od 0.6506. U narednim epohama vidimo brzi pad vrijednosti gubitka – do desete epohe trenirajući gubitak opada na 0.1704, a validacijski na 0.1643, dok Val AUC raste na 0.7339. Taj ubrzani napredak u prvih deset epoha ukazuje na uspješnu inicijalnu fazu učenja. Kako trening odmiče, od obilježene epohe 10. prema 20., smanjenje gubitka postaje sve sporije i stabilnije, a Val AUC prolazi kroz postupno usporeni rast: već u epohi 19 ostvarena je najviša vrijednost od 0.7516, dok je u posljednjoj epohi 20 Val AUC blago pala na 0.7492. Ova oscilacija može upućivati na to da je model dosego tačku zasićenja, a razlika između trenirajućeg i validacijskog gubitka u kasnijim epohama (manja od 0.005) pokazuje da nema drastičnog overfittinga.
- **Vizualizacija gubitka i AUC tokom epoha:** Grafički prikazi oba ključa treniranja pružaju dodatni uvid. Krivulje gubitka pokazuju blizak slijed trenirajućeg i validacijskog gubitka – obje vrijednosti se konvergiraju već oko epoha 12–15, što znači da model stabilno usvaja obrasce iz skupa podataka bez značajnog odstupanja. S druge strane, krivulja validacijskog AUC jasno razdvaja dvije faze: strmi rast u prvim epohama do približno 0.73, te sporiji rast prema vršnoj vrijednosti od oko 0.75. Takav tok sugerira da bi eventualno povećanje broja epoha donijelo samo minimalne koristi, te se valja razmisliti o ranijem zaustavljanju (eng. early stopping) ili o promjeni arhitekture za daljnja poboljšanja.
- **Ocjena modela na testnom skupu:** Nakon treniranja, model je evaluiran na neovisnom testnom skupu. Generirani klasifikacijski izvještaj razvrstio je 14 klasa prema precision, recall i F1-mjerama. Primjetno je da je najveći broj klasa (12 od 14) potpuno neprepoznat (precision i recall jednaki nuli), dok su za klasu 2 i klasu 3 zabilježene mjerljive vrijednosti: klasa 2 ima precision od 0.64, ali vrlo nizak recall od 0.10 (F1-mjera 0.17), a klasa 3 pokazuje visoku preciznost od 0.78, ali nulti recall. Takvi rezultati ukazuju

na snažnu neravnotežu u učenju jer model, iako je u stanju vrlo precizno odrediti određene uzorke, gotovo u potpunosti izostavlja većinu pozitivnih primjera. Neravnoteža je dodatno potvrđena podrškom (support) svake klase: dok za dominantne kategorije poput klase 0, 2 i 3 postoje tisuće uzoraka, za rijetke klase (npr. klasa 13) ima ih tek stotinjak. Posljedično, model ignorira manje zastupljene klase, što je u medicinskoj primjeni osobito opasno jer može rezultirati propuštanjem rijetkih, ali klinički kritičnih stanja.

- **Identificirani rizici:**

- **Neravnoteža skupa podataka:** Kako bi se poboljšala detekcija rijetkih klasa, predlaže se primjena tehnika balansiranja skupa poput oversamplinga, SMOTE algoritma ili ciljanih augmentacija.
- **Ograničena generalizacija:** Uprkos umjerenom Val AUC-u (~ 0.75), model nije dovoljno robustan za kliničku upotrebu. Nužno je istražiti napredne arhitekture (ResNet, EfficientNet) i uključiti metode regularizacije (dropout, weight decay) te transfer learning.
- **Procjena stabilnosti:** Preporučuje se provođenje k-fold unakrsne validacije i uključivanje vanjskih, nezavisnih skupova podataka radi procjene konzistentnosti performansi.
- **Klinička i regulatorna usklađenost:** Prije implementacije u stvarnim uvjetima, potrebno je proći klinička ispitivanja i dobiti potvrdu relevantnih regulatornih tijela (npr. FDA, CE certifikat). AUC vrijednosti ispod 0.85–0.90 smatraju se nedovoljnima za pouzdanu dijagnostiku.

Dakle, trening VGG-modela na MedMNIST skupu pokazao je da se osnovne CNN arhitekture mogu brzo prilagoditi medicinskim slikovnim podacima i ostvariti početnu ACU razinu od približno 0.75. Ipak, dubinske analize klasifikacijskog izvještaja i vizualizacija treniranja ukazuju na ozbiljna ograničenja, model ne prepoznaje većinu klasa zbog neravnoteže podataka, a postignute performanse

nisu dovoljne za kliničku primjenu. Sljedeći koraci uključuju preradu skupa podataka, eksperimentisanje s modernijim arhitekturama i sveobuhvatnu evaluaciju kako bi se osigurala sigurnost i pouzdanost u medicinskom kontekstu.

Izlaganje modela podacima koji su mu nepoznati

U svrhu evaluacije sposobnosti treniranog modela da generalizuje na nove, do tada neviđene primjere, izvršeno je testiranje na svim klasama. Za svaku klasu je spašeno po 150 slika, osim za jednu gdje je spašeno 42 slike. Sve slike su prethodno transformisane kako bi bile kompatibilne s ulaznim zahtjevima modela.

Rezultati klasifikacije za sve slike prikazuju vrlo slične vrijednosti vjerovatnoće za svih 14 klasa bolesti. Takvi rezultati ukazuju na nekoliko potencijalnih problema. Model je vjerovatno naučio da izbjegava donošenje čvrstih zaključaka, te kako svaki AUC, za svaku klasu osciluje oko 0.5, model je nesiguran.

Može se zaključiti da u trenutnom obliku, model nije pouzdan za donošenje dijagnostičkih odluka na novim slikama i nije spreman za kliničku primjenu. Ali ipak za naše potrebe ovi rezultati su prihvatljivi, jer fokus stavljamo na razvoju modela na njegovom samom početku, kako zapravo to izgleda te koji su koraci potrebni. U trenutnoj fazi, sistem može poslužiti kao ilustracija principa rada dubokih modela u detekciji bolesti iz medicinskih slika. Naravno ako bi se ovdje uradile neke dodatne optimizacije kao što je bolje balansiranje podataka, podešavanje klasifikacionih pragova i sl., model bi tada bio spreman za korištenje u medicinske svrhe.

Faza 5: Cjelokupni osvrt na problem i dobijeno rješenje

Osvrt na postignute rezultate

Za analizu sposobnosti modela da detektuje 14 klasa u ChestMNIST području, izdvojeno je po 150 slika za svaku klasu (osim klase “Hernia” koja je u testnom skupu imala samo 42 primjera), a model je potom evaluiran kroz standardni klasifikacijski izvještaj i AUC metrike. Najbolji **precision** ostvarile su klase poput “Cardiomegaly” (0.50) i “Atelectasis” (0.10), ali uz zanemarivo nizak **recall** (0.04, odnosno 0.21). Izuzetak je klasa “Hernia” (samo 42 primjera), za koju je model postigao perfektan recall (1.00), no izrazito nizak precision (0.02), što implicira veliki broj lažno pozitivnih predikcija. Vrijednosti AUC-a variraju između 0.4703 (“Pneumonia”) i 0.6169 (“Effusion”). Perfektan recall za “Hernia” uz vrlo nizak precision pokazuje da model jednostavno klasificira (previše) primjera kao “Hernia”, vjerojatno zato što je te podatke učio na vrlo malom broju primjera i nije uspio naučiti robusne značajke.

Ovi rezultati, iako nedovoljni za kliničku primjenu, zadovoljno demonstriraju sve faze izrade i evaluacije multi-label modela na medicinskim slikama—uključujući izdvajanje vlastitih testnih uzoraka, metodičku transformaciju podataka, kvantitativnu procjenu performansi i kritički osvrt na prepoznate slabosti. To čini projekat potpunim za akademske potrebe i pokazuje razumijevanje izazova u primjeni dubokog učenja na stvarne medicinske podatke.

Uporediti sa radovima iz prethodne faze

U toku treniranja VGG-baziranog modela na ChestMNIST skupu podataka postignut je stalan napredak u metrikama: trenirajući gubitak je s prve epohe (0,1848) pao na 0,1669 u dvadesetoj, dok se validacijski gubitak spustio s 0,1737 na približno 0,1626, a validacijski AUC porastao s 0,6506 na vršnu vrijednost od 0,7516 u devetnaestoj epohi. To ukazuje na uspješno učenje i generalizaciju modela do otprilike petnaeste epohe, nakon čega je rast AUC-a usporio i došlo je do zasićenja performansi.

Evaluacija na službenom ChestMNIST test skupu donijela je mješovite rezultate. Iako su vrijednosti AUC-a po klasama bile solidne – krećući se od 0,6633 do 0,8562

– klasična binarna klasifikacija s pragom od 0,5 gotovo da nije prepoznavala prisustvo bolesti: precision i recall za većinu klasa bili su 0,00, izuzev Effusion (precision 0,64, recall 0,10) i Infiltration (precision 0,78, recall 0,00), što se odrazilo na mikro-F1 oko 0,03 i makro-F1 oko 0,01. Drugim riječima, model je znao relativno dobro rangirati primjere po vjerovatnosti, ali je pri konkretnom pragiranju bio previše konzervativan da bi označio pozitivan slučaj. Kad je model primijenjen na vlastitu kolekciju od 150 slika po klasi (osim Hernia koje je imalo 42 primjera), njegovo ponašanje dodatno se pogoršalo. Micro-F1 iznosio je tek 0,075, macro-F1 0,065, a weighted-F1 oko 0,07. AUC-ovi su pali u raspon od 0,47 (Pneumonia) do 0,62 (Effusion), a klasifikacija je proizvela vrlo niske recall vrijednosti za gotovo sve klinički značajne klase (npr. Pneumonia, Consolidation, Edema su imale recall 0,00). Jedini izuzetak bio je Pneumothorax s recallom 0,32 i precisionom 0,09 te Hernia s perfektnim recallom od 1,00, ali vrlo slabim precisionom (0,02) zbog brojnih lažno pozitivnih. Ova usporedba jasno pokazuje da iako je model tokom treniranja i na službenom test skupu ostvario zadovoljavajući AUC, njegova stvarna sposobnost donošenja pouzdanih klasifikacija pri konkretnom pragiranju ostaje nedovoljna.

Prodiskutovati šta se moglo bolje uraditi

U postojećoj implementaciji model postiže razumnu vrijednost AUC, ali se pokazao nedovoljno pouzdanim pri konačnom pragiranju izlaza na 0,5. Za značajno poboljšanje performansi mogu se primijeniti sljedeće mjere:

Prvo, važno je riješiti problem neravnoteže klasa. To se može postići oversamplingom rijetkih klasa ili SMOTE sintetizacijom dodatnih uzoraka, kao i uvođenjem ponderiranog gubitka (BCEWithLogitsLoss s parametrima pos_weight) ili primjenom focal loss funkcije, čime bi se model snažnije usmjerio na učenje iz ograničenih primjera bolesti. Drugo, umjesto jedinstvenog praga od 0,5 za sve klase, preporučuje se optimizirati pragove zasebno za svaku klasu na validacijskom skupu, što može značajno podići recall i F1-mjere, a dodatno se mogu kalibrirati izlazi modela (Plattovo skaliranje ili isotonička regresija) kako bi predviđene vjerojatnosti bolje odražavale stvarnu vjerovatnost bolesti.

Treće, eksperimentiranje s naprednijim arhitekturama i transfer learning pristupom može donijeti vidljiva poboljšanja. Korištenje unaprijed istreniranih modela kao što su ResNet, DenseNet ili EfficientNet i fino podešavanje samo gornjih slojeva često

rezultira bržom konvergencijom i boljom generalizacijom nego treniranje početničkih arhitektura “od nule”. Četvrto, proširene augmentacijske tehnike—uključujući varijacije osvjettjenja, kontrasta, uvođenje šuma, Cutout ili MixUp—mogli bi pomoći modelu da nauči robusne lokalne i globalne značajke plućnih obrazaca, neovisno o razlici u kliničkim skenerima.

Peto, primjena regularizacije poput drop-outa u završnim slojevima i weight decay (L2 regularizacije) može ograničiti overfitting, a dodatno finije podešavanje optimizatora (promjenom learning rate-a, momentum-a te upotrebom scheduler-a) kroz grid search ili Bayesovu optimizaciju može otkriti optimalne hiperparametre. Konačno, za pouzdaniju procjenu performansi potrebno je provesti k-fold cross-validaciju i testirati model na nezavisnim stvarnim Chest X-ray skupovima (npr. NIH ChestX-ray14 ili CheXpert). Takav sveobuhvatan pristup ne samo da će podići tačnost modela, već i njegovu sposobnost generalizacije te cjelokupnu naučnu vrijednost.

Zaključak

U ovom radu prikazana je kompletna analiza i implementacija sistema za klasifikaciju plućnih rendgenskih snimaka korištenjem konvolucijske neuronske mreže zasnovane na VGG arhitekturi. Odabrani dataset ChestMNIST omogućio je brzu i efikasnu obuku modela u edukativnom okruženju, ali i istakao ključne izazove poput neravnoteže klasa, niske rezolucije slika i ograničenja u tačnosti predikcija.

Rezultati treniranja pokazali su stabilan napredak u metrikama kao što su gubitak i AUC, ali i ukazali na potrebu za dodatnim unapređenjima kako bi se postigli rezultati prikladni za kliničku primjenu. Model je pokazao ograničenu sposobnost u prepoznavanju rijetkih patologija, što naglašava važnost kvalitetnog balansiranja podataka i finog podešavanja arhitekture.

Iako trenutna implementacija ne dostiže nivo dijagnostičke pouzdanosti, ona uspješno demonstrira sve ključne korake u razvoju jednog dubokog učećeg sistema za medicinsku dijagnostiku. Ovaj rad predstavlja dobru osnovu za budući razvoj, gdje bi se mogli primijeniti napredniji modeli (npr. ResNet, EfficientNet), transfer learning, kao i dodatne tehnike pretprocesiranja i regularizacije, kako bi se osigurala veća tačnost i robusnost modela u stvarnim uslovima.