EKG Kategorizacija

Komparacija Metoda Mašinskog Učenja i Dubokog učenja

Dragomir Božoki BI55-2020 bozokidragomir@gmail.com Fakultet Tehničkih Nauka Novi Sad

I. Abstrakt

Elektrokardiogram ***(EKG)*** predstavlja grafički i elektronski prikaz rezultata elektrokardiografije – procesa snimanja električne aktivnosti srca putem elektroda koje se postavljaju na određene tačke na koži. Ovaj alat je od vitalnog značaja za procenu stanja srca. Automatska klasifikacija EKG podataka omogućava ranu dijagnostiku, pružajući preliminarne rezultate pre nego što lekari detaljno analiziraju podatke. U oblasti automatske klasifikacije EKG-a, sprovedene su brojne studije i istraživanja. Duboko učenje (engl. Deep Learning) se istaklo kao jedna od najefikasnijih metoda, sa preciznošću koja dostiže čak 99% kada je dostupan dovoljno veliki skup podataka za obuku modela. Istraživanja su pokazala da i konvencionalne metode mašinskog učenja mogu postići zadovoljavajuće rezultate u pogledu tačnosti i drugih metrika, što ukazuje na njihov potencijal u ovoj oblasti.

II. Baza Podataka

Baza podataka se sastoji od dve kolekcije srčanih otkucaja prikupljenih iz dva poznata skupa podataka za klasifikaciju EKG-a: *MIT-BIH Arhythmia Dataset* i *PTB Diagnostic ECG Database*. Broj uzoraka u oba skupa je dovoljno velik za treniranje modela neuronskih mreža. Za obuku i evaluaciju modela korišćen je MIT-BIH Arhythmia Dataset te će se dalji rad referencirati na podatke iz ovog seta podataka. Skup podataka je podeljen na dve baze: jednu za obuku i validaciju modela, i drugu za testiranje modela. Baza podataka za obuku sadrži 87.553 uzoraka, odnosno snimljenih EKG signala sa 186 odbiraka, što znači da baza ima 186 kolona. 187. kolona predstavlja klasifikaciju datog uzorka i posmatraće se kao tražena promenjiva. EKG signali mogu biti klasifikovani u pet različitih kategorija:

- **N:** Non-Ecotic Beats (normalni otkucaji)
- **S:** Supraventricular Ectopic Beats
- **V:** Ventricular Ectopic Beats
- **F:** Fusion Beats
- **Q**: Unknown Beats

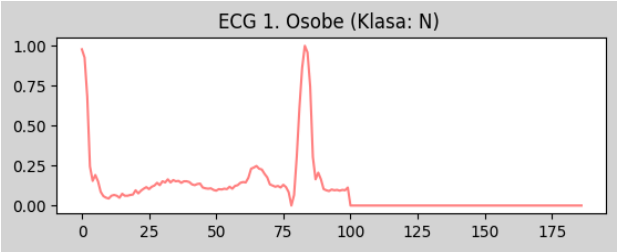


Fig 2.1 EKG Osobe sa zdravim otkucajem srca u trening skupu

Baza podataka za testiranje se sastoji od 21.891 uzorka sa različitom dužinom odbiraka. Međutim, svi uzorci su prošireni (eng. padding) kako bi imali istu dužinu kao i skup za obuku, tj. 186 kolona za podatke i 187. kolonu za klasu uzorka. Proširenje je urađeno kako bi se omogućila obuka neuronske mreže koja zahteva da svi ulazi budu istih dimenzija.

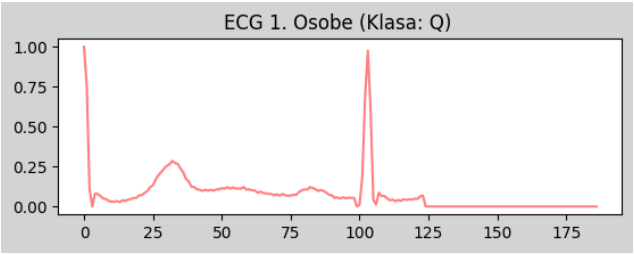


Fig 2.1 EKG Osobe sa anomalijom u otkucaju srca u test skupu

III. PREPROCESIRANJE PODATAKA

Pre obuke modela, potrebno je preprocesirati sirove podatke koji će biti korišćeni za obuku. U analizi baze podataka utvrđeno je da nema nevalidnih i nepostojećih podataka, kao ni ekstremnih vrednosti (outliera). Prvi izazov koji se javlja pri analizi podataka jeste prisustvo disbalansa u izlaznoj promenljivoj, gde različite klase imaju nejednaku zastupljenost uzoraka. Specifično, primetno je da uzorci sa normalnim otkucajem srca imaju značajno veći broj uzoraka u poređenju sa ostalim klasama. Ovaj disbalans može biti problematičan prilikom treniranja modela, jer može dovesti do pristrasnosti modela ka dominantnoj klasi, dok manje zastupljene klase mogu biti nedovoljno reprezentovane. Stoga je od suštinskog značaja primeniti strategije za obradu disbalansa, pri čemu je u ovom istraživanju korišćena metoda resamplinga, konkretno, oversampling. Pristup oversamplinga podrazumeva povećanje broja uzoraka manje zastupljenih klasa tako što se ti uzorci dupliciraju, sve dok zastupljenost svih klasa nije izjednačena.

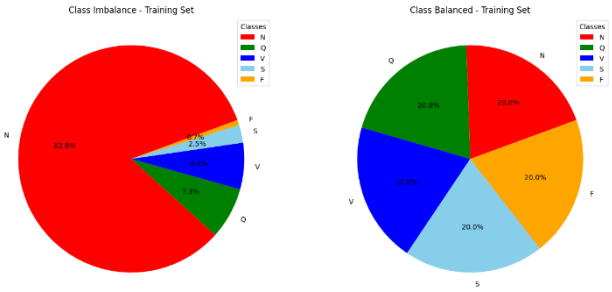


Fig 3.1 Graf raspodele klasa pre i posle oversampling metode

Nakon balansiranja klasa, identifikovan je prisustvo šuma na snimcima elektrokardiograma (EKG-a), što je uobičajena pojava tokom snimanja i može nastati kao posledica različitih artefakata. Ovaj šum može izazvati lažne informacije koje mogu narušiti sposobnost generalizacije modela. Kako bi se suzbio ovaj šum i očuvala relevantna informacija za analizu EKG signala, primenjen je niskopropusni Butterworth filter (NF Butterworth filter) na trening uzorku. U ovom slučaju, sve frekvencije iznad 50 Hz su isečene iz podataka, što je uobičajena praksa za filtriranje EKG signala. Ovim postupkom uklanjaju se visoke frekvencije koje obično predstavljaju šum ili artefakte, dok se zadržava sporopromenjivi deo signala koji sadrži stvarne informacije o srčanim otkucajima.

$$H(s) = \frac{1}{(\omega_c s)^{2N}}$$

3.1 formula niskopropusnog Butterworth filtra

Gde su H(s) – prenosna funkcija filtra, s – kompleksna promenjiva, ω_c – cutoff frekfencija(rad/s) i N – red filtra

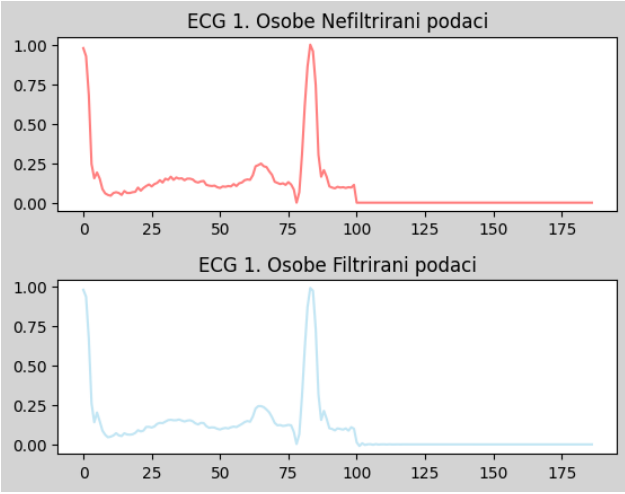


Fig 3.2. Prikaz Filtriranog i Nefiltriranog signala EKG-a

Testiranje modela izvršeno je na neobrađenim uzorcima koji i dalje zadržavaju nebalansiranu distribuciju klasa, kako bi se procenilo kako se modeli ponašaju na podacima koji sadrže artifakte ili druge nepravilnosti. Ova strategija omogućava evaluaciju performansi modela u realnom svetu, uzimajući u obzir potencijalne izazove koji se mogu pojaviti u praktičnim scenarijima primene.

U ovom istraživanju, korišćeni podaci su podeljeni na trening, validacioni i test skup, kako bi se obezbedila pouzdana procena performansi modela mašinskog učenja. Trening skup je dalje podeljen na dve komponente: stvarni trening skup i validacioni skup. Validacioni skup je činio 20% od ukupnog trening skupa, dok je preostalih 80% korišćeno za obuku modela. Ova podela je omogućila finu podešavanje hiperparametara i izbor najboljeg modela bez preklapanja sa test skupom. Test skup je bio rezervisan za konačnu evaluaciju modela, čime se osigurava objektivna procena sposobnosti modela da generalizuje na nove, neviđene podatke.

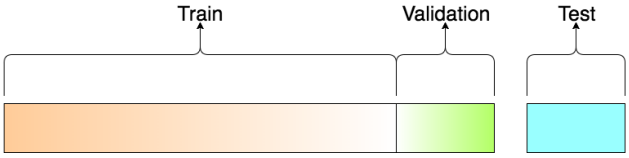


Fig 3.3 Vizualizacija podele podataka

IV. METODE OBRADE PODATAKA

Izabrani algoritmi za klasifikaciju elektrokardiogram (EKG) signala obuhvataju metode iz oblasti dubokog učenja i klasičnog mašinskog učenja. Duboko učenje predstavlja perspektivnu metodologiju koja se ističe po svojoj sposobnosti za kompleksno učenje i ekstrakciju značajki iz podataka. S druge strane, klasični algoritmi mašinskog učenja predstavljaju osnovnu paradigmu za obradu podataka i analizu uzoraka. U ovom radu, cilj je uporediti performanse između algoritama iz ove dve sfere. Duboko učenje se istražuje zbog svoje sposobnosti da automatski nauči reprezentacije podataka visokog nivoa i izvrši klasifikaciju bez potrebe za ručno definisanim značajkama. S druge strane, klasični algoritmi mašinskog učenja, poput kNN-a ili SVM-a, pružaju osnovnu osnovu za poređenje, uzimajući u obzir njihovu računsku efikasnost i interpretabilnost.

A| Metode Mašinskog Učenja

- **k-Najbližih Suseda(kNN)**

Algoritam k-najbližih suseda (k-Nearest Neighbors, k-NN) predstavlja jedan od najjednostavnijih algoritama mašinskog učenja baziranih na nadgledanom učenju. Princip rada ovog algoritma zasniva se na pretpostavci da će novi uzorak biti sličan postojećim uzorcima u skupu podataka. Konkretno, algoritam klasifikuje novi uzorak na osnovu kategorije kojoj pripadaju njegovi najbliži susedi. Za treniranje algoritma k-najbližih suseda, ključni parametri od značaja uključuju broj suseda (k), algoritam pretrage i metriku udaljenosti. Korišćenjem metode GridSearchCV na podskupu od 40 000 uzoraka za trening, optimizovani su ovi parametri. Rezultati optimizacije pokazali su da je najbolji algoritam pretrage 'auto', metrika Menhetn (Manhattan), dok je optimalan

broj suseda (k) 1. Algoritam sa ovim optimalnim parametrima zatim je obučen na celokupnom trening skupu. Rezultati validacije pokazali su zadovoljavajuću preciznost od 96%-99% u zavinosti od klase. Ovaj visok nivo tačnosti ukazuje na efektivnost kNN algoritma u klasifikaciji EKG-a.

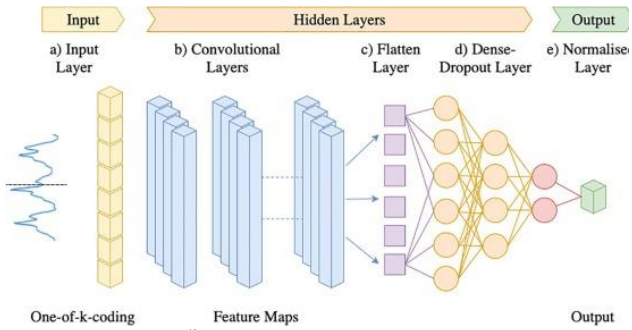
- **Vektori Nosača (SVM)**

Metoda vektora nosača (**Support Vector Machines, SVM**) predstavlja tehniku nadgledanog učenja koja se koristi kako za klasifikaciju, tako i za probleme regresije. Ipak, najčešće se primenjuje u kontekstu klasifikacionih problema. Proces klasifikacije pomoću SVM-a podrazumeva formiranje hiperravni koje na optimalan način razdvajaju klase u prostoru atributa. U ovom istraživanju, korišćen je radijalni bazni funkcioni kernel (rbf) kao hiperparametar. Podaci su fitovani na celokupnom skupu podataka, a validacija je izvršena na posebnom validacionom skupu. Dobijena preciznost na validacionom skupu podataka iznosila je 94%, što je manje u poređenju sa preciznošću od 96% postignutom k-najbližih suseda (k-NN) algoritmom. Ovi rezultati impliciraju da kompleksniji klasifikatori, poput SVM-a, ne moraju nužno davati bolje rezultate u svim slučajevima.

- **Stabla Odluke (Decision Trees - DT)**

Stabla odluke predstavljaju neparametarsku metodu nadgledanog učenja koja se primenjuje u svrhe klasifikacije i regresije. Osnovni cilj ove metode je konstrukcija modela sposobnog za predikciju vrednosti ciljne promenljive kroz učenje sekvenci jednostavnih odluka donetih na osnovu atributa podataka. Ova tehnika funkcioniše tako što se podaci sukcesivno dele na podskupove na osnovu atributa, pri čemu svaka odluka odgovara čvoru u stablu. Odluke se donose tako da minimiziraju unutargrupnu varijabilnost u regresiji ili maksimiziraju separabilnost klasa u klasifikaciji. Proces se ponavlja sve dok se ne dođe do optimalne dekompozicije podataka koja omogućava najprecizniju predikciju ciljne promenljive. Dobijena preciznost na validacionom skupu podataka iznosi 98%, što predstavlja značajno poboljšanje u odnosu na SVM. Ipak, ovaj rezultat je i dalje neznatno lošiji u poređenju sa kNN algoritmom.

B| Metode Dubokog Učenja



4.1 Šablonski prikaz neuralne mreže

U ovom istraživanju razvijen je model kombinovane konvolucione neuralne mreže (**CNN**) i dugoročne kratkoročne memorije (**LSTM**) za klasifikaciju sekvencijalnih podataka. Model, nazvan **CNN-LSTM**, implementiran je koristeći Keras API unutar TensorFlow okvira. Arhitektura modela je dizajnirana kako bi iskoristila prednosti konvolucionih slojeva za ekstrakciju lokalnih karakteristika i LSTM slojeva za modeliranje dugoročnih zavisnosti u podacima. Detalji arhitekture su sledeći:

1. **Ulazni sloj**: Model prima ulazne podatke u obliku tenzora sa tri dimenzije, gde prva dimenzija predstavlja broj uzoraka, druga dimenzija dužinu sekvence, a treća dimenzija broj karakteristika po uzorku.

2. **Prvi konvolucioni sloj**: Konvolucioni sloj sa 64 filtera, veličine kernela 6 i ReLU aktivacijom koristi se za ekstrakciju osnovnih karakteristika iz ulaznih podataka. Nakon konvolucije, koristi se sloj normalizacije (BatchNormalization) kako bi se stabilizovala i ubrzala obuka.

3. **Prvi sloj za maksimizaciju pooling-a (MaxPooling)**: MaxPool1D sloj sa veličinom pool-a 3, strides 2 i padding-om "same" koristi se za redukciju dimenzionalnosti i kontrole prekomernog uklapanja (overfitting).

4. **Drugi konvolucioni sloj**: Konvolucioni sloj sa 64 filtera, veličine kernela 3 i ReLU aktivacijom dalje ekstrakuje karakteristike iz prethodnog sloja. Ponovo se koristi BatchNormalization za stabilizaciju obuke.

5. **Drugi sloj za maksimizaciju pooling-a**: MaxPool1D sloj sa veličinom pool-a 2, strides 2 i padding-om "same" koristi se za dalju redukciju dimenzionalnosti.

6. **Treći konvolucioni sloj**: Još jedan konvolucioni sloj sa 64 filtera, veličine kernela 3 i ReLU aktivacijom dodatno pojačava ekstrakciju karakteristika. BatchNormalization se koristi i u ovom sloju.

7. **Treći sloj za maksimizaciju pooling-a**: MaxPool1D sloj sa veličinom pool-a 2, strides 2 i padding-om "same" koristi se za dalju redukciju dimenzionalnosti.

8. **Prvi LSTM sloj**: LSTM sloj sa 64 jedinice i tanh aktivacijom koristi se za modeliranje dugoročnih zavisnosti u sekvencijalnim podacima, vraćajući sekvencu kao izlaz.

9. **Drugi LSTM sloj**: LSTM sloj sa 32 jedinice i tanh aktivacijom koristi se za dalju obradu sekvencijalnih podataka, vraćajući jedinstveni vektor za svaki ulazni uzorak.

10. **Sloj poravnavanja (Flatten)**: Sloj za poravnavanje koristi se za transformaciju

dvodimenzionalnog izlaza iz LSTM sloja u jednodimenzionalni vektor.

11. **Prvi potpuno povezani sloj (Dense)**: Sloj sa 64 jedinice i ReLU aktivacijom koristi se za učenje složenih obrazaca u podacima.

12. **Drugi potpuno povezani sloj**: Sloj sa 32 jedinice i ReLU aktivacijom dodatno unapređuje mogućnost modela da uči složene obrasce.

13. **Izlazni sloj**: Potpuno povezani sloj sa brojem jedinica jednakim broju klasa u zadatku klasifikacije i softmax aktivacijom koristi se za generisanje verovatnoća pripadnosti svakoj klasi.

Ovaj model kombinuje prostornu i vremensku obradu podataka, omogućavajući efikasnu klasifikaciju sekvencijalnih podataka uzimajući u obzir i lokalne i globalne karakteristike. Uspostavljena arhitektura je testirana i evaluirana na posebno izdvojenom test skupu kako bi se osigurala njena generalizacija i performanse na neviđenim podacima.

- **Konfiguracija Modela**

Za optimizaciju modela korišćen je Adam optimizator, popularan zbog svoje efikasnosti i prilagodljivosti. Kao funkcija gubitka korišćena je unakrsna entropija, što je standardni izbor za višeklasnu klasifikaciju. Tačnost je korišćena kao metrika za evaluaciju performansi modela.

$$H(P^*|P) = -\sum_i P^*_{(i)} \log_{\square} P(i)$$

4.2 formula unakrsne validacije

$$P^*_{(i)} - Prava\ oznaka\ klase$$

$$P(i) - Prediktovana\ oznaka\ kalse$$

- **Obuka modela**

Broj epoha za obuku je postavljen na 10, jer je empirijski utvrđeno da više ili manje epoha od 10 daje lošije rezultate. Veličina serije (batch size) je postavljena na 32, što znači da će model ažurirati svoje težine nakon svakih 32 uzorka. Omogućena je funkcija "eager execution", što znači da se operacije izvršavaju odmah, bez potrebe za građenjem i izvršavanjem grafova. Ovo olakšava debugovanje i praćenje modela tokom obuke. Model je imao tačnost od 98% nakon 10 epoha na validacionom skupu podataka.

V. REZULTATI

- **Algoritam Mašinskog učenja**

Biće prikazani samo rezultati najboljeg algoritma klasičnog mašinskog učenja, k-Nearest Neighbors (kNN) algoritma. Ovaj algoritam je postigao ukupnu tačnost od 94%. Međutim, tačnost varira u zavisnosti od klase. Posebno je problematična Klasa 3, koja je klasifikovana sa najmanjom tačnošću, jer je ima najmanje u testnom skupu. Ova oskudica uzoraka čini da svaka pojedinačna pogrešna klasifikacija značajno utiče na ukupnu tačnost. Stoga je kNN algoritam veoma osetljiv na svaku pojedinačnu pogrešnu klasifikaciju unutar ove klase.

Jedan od mogućih razloga zašto je kNN algoritam postigao najbolje rezultate je njegova sposobnost da prepoznaje lokalne obrasce u podacima. Odluke se donose na osnovu blizine podataka u prostoru karakteristika, što se pokazalo posebno efikasnim kod klasifikacije ECG signala, gde postoje jasni lokalni klasteri podataka. Ova sposobnost omogućava kNN algoritmu da precizno identifikuje i klasifikuje slične obrasce, čime se postiže visoka tačnost u klasifikaciji.

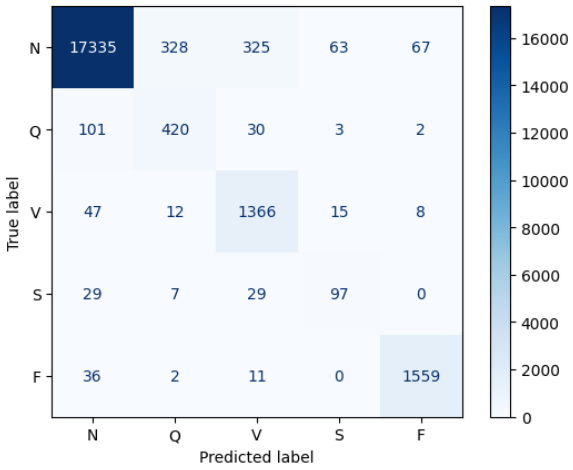


Figura 5.1 Matrica konfuzije kNN algoritma na test skupu

- **Neuralna Mreža**

Neuralna mreža je pokazala značajno bolje rezultate u poređenju sa algoritmima klasičnog učenja, postigavši tačnost od 98% na testnom skupu. Ovo ukazuje na superiornu sposobnost neuralne mreže da prepozna obrasce i pravilno klasifikuje EKG signale u jednu od pet kategorija. Neuralna mreža je takođe pokazala veću efikasnost u klasifikaciji klasa koje su manje zastupljene u testnom skupu, a koje često predstavljaju izazov i imaju značajan uticaj na krajnje rezultate klasifikacije. Kroz iteracije (epohe), primetno je povećanje tačnosti klasifikacije, što je bilo praćeno postepenim smanjenjem funkcije gubitka. Ova pojava ukazuje na to da je model sve bolje usklađen sa podacima tokom treninga, što rezultira manjim gubicima i većom preciznošću u klasifikaciji. Međutim, primećene su

male oscilacije u lošijim klasifikacijama tokom epoha, što može biti posledica kompleksnosti podataka ili specifičnosti algoritma. Ove oscilacije predstavljaju normalan deo procesa treniranja i mogu se očekivati, ali je značajno da su tačnost i funkcija gubitka generalno u poboljšanju tokom vremena.

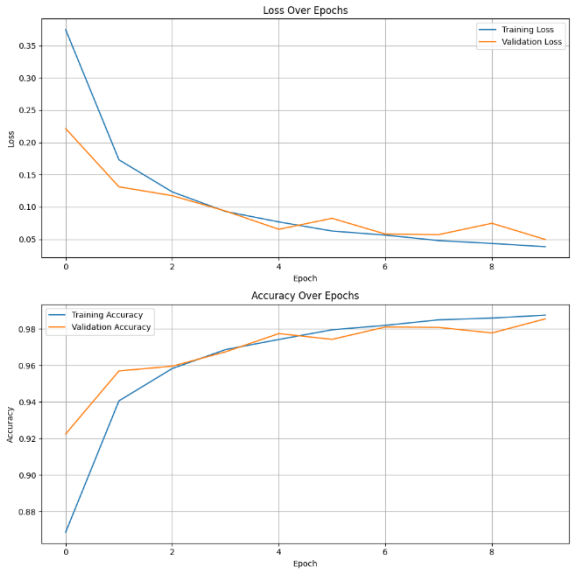


Figura 5.1 Prikaz funkcije gubitka i tačnosti kroz epohe

VI. ZAKLJUČAK

Neuralne mreže su se pokazale kao izvanredan alat za klasifikaciju ECG-a, a budući radovi na ovu temu bi trebalo da se fokusiraju na dodatno obučavanje i fino podešavanje postojećih modela kako bi se postigla što bolja i preciznija klasifikacija. Sa druge strane, klasične metode mašinskog učenja, iako često smatrane jednostavnijim, pokazuju odlične rezultate kada osetljivost nije ključna (tj. kada broj lažnih negativa u celokupnoj populaciji nije od presudnog značaja). Ove metode mogu biti posebno korisne kada je baza podataka mala i nije dovoljno obimna za efikasno obučavanje neuralnih mreža.

Kvalitetna klasifikacija ECG-a je od suštinskog značaja iz nekoliko razloga. Prvo, precizna dijagnostika omogućava ranu detekciju srčanih abnormalnosti, što može značajno poboljšati ishode lečenja i smanjiti mortalitet. Drugo, pouzdana klasifikacija može smanjiti broj nepotrebnih medicinskih intervencija, čime se smanjuju troškovi zdravstvene zaštite i smanjuje opterećenje za zdravstveni sistem. Konačno, tačna analiza ECG-a pruža lekarima dragocene informacije koje mogu koristiti za donošenje informisanih odluka o daljem lečenju pacijenata. U ovom kontekstu, napredak u tehnologiji klasifikacije ECG-a ima potencijal da unapredi celokupno zdravlje i dobrobit populacije.

VII. REFERENCE I KOD

[1] **Praktikum za Mašinsko učenje** – *Tijana Nosek, Branko Brkljač, Danica Despotović, Milan Sečujski, Tatjana Lončar Turukalo*

[2] **Mašinsko ucenje Pytorch I Scikit-Learn** - *Sebastian Raschka, Yuhi (Hayden) Liu, Vahid Mirjalili*

[3] **Comparison of two artificial intelligence-augmented ECG approaches: Machine learning and deep learning** – *Anthonz H Kashou, Adam M May, Peter A Noseworthy*

[4] **ECG Classification: ML & DL Comparative Study** - *Ismael Elfhussein*

[5] **Link za Kod**