Detekcija srčanog udara na EKG slikama

Matko Barbić, Dora Bilić-Pavlinović, Antun Jurelinac, Mia Krstičević, Jakov Samardžija

Analiza slika u biomedicini
FER
Siječanj, 2025.

Sadržaj

1. Uvod

Srčana oboljenja, uključujući srčani udar, vodeći su uzrok invaliditeta i smrtnosti diljem svijeta. Podatci WHO (Svjetske zdravstvene organizacije) nam ukazuju na to da su kardiovaskularne bolesti uzrok 30% smrtnih slučajeva globalno. Srčani udar, naravno, predstavlja značajan udio tog postotka. Iz tog razloga, predviđanje srčanog udara postalo je jedno od najvećih područja istraživanja u medicinskoj i zdravstvenoj industriji. Pravovremeno identificiranje rizika razvoja srčanog udara te razvoj metoda prevencije i liječenja od velike su važnosti u smanjenju smrtnosti modernog društva.

Kako bismo razvili precizne prediktivne modele, potrebno nam je mjeriti utjecaj pojedinih faktora na razvoj srčanog udara. Mnogi značajni faktori su već općepoznati zbog iznimno velike količine provedenih prediktivnih istraživanja. Naš rad je također jedno takvo istraživanje. Razumijevanje faktora rizika poput krvnog tlaka, kolesterola , pušenja, tjelesne težine, povijesti bolesti i sl. omogućava nam razvoj prediktivnih modela korištenjem algoritama strojnog učenja.

Motivacija za ovaj rad leži u vrijednosti koje strojno učenje može unijeti u društvo, osobito u medicinskom području. Naime, razvojem boljih i preciznijih metoda za procjenu rizika od oboljevanja od srčanog udara imamo priliku spasiti mnoge ljudske živote. Napretkom tehnologije i sve većom dostupnošću podataka sve je veća mogućnost razvitka modela strojnog učenja. Takvi modeli omogućuju medicinskim profesionalcima lakše prepoznavanje pacijenata pod velikim rizikom te daljnji tretman istih.

U našem radu, analizirat ćemo kakav utjecaj na oboljevanje od srčanog udara imaju sljedeći faktori:

- spol ("Muško", "Žensko" ili "Ostalo")
- starost
- hipertenzija- povišen tlak (0 ako pacijent nema hipertenziju, 1 ako ima hipertenziju)
- srčana bolest (0 ako pacijent nema srčanu bolest, 1 ako ima)
- bio/la je u braku ("Ne" ili "Da")
- tip posla ("Djeca", "Javni sektor", "Nikada nije radio/la", "Privatni sektor", "Samozaposlen/a")
- tip prebivališta ("Ruralno" ili "Urbano")
- prosečna količina glukoze u krvi

- BMI: indeks telesne mase
- status pušenja ("Bivši pušač", "Nikada nije pušio", "Puši" ili "Nepoznata")

U radu ćemo koristiti model potpornih vektora (SVM) kako bismo se suočili s ovim prediktivnim problemom. Razvojem ovakvih alata, moguće je unaprijediti preventivne mjere, smanjiti troškove zdravstvene zaštite, ali i značajno poboljšati kvalitetu života pacijenata sa srčanim bolestima.

2. Pregled postojećih metoda

Predikcija srčanih udara je tema primamljiva istraživačima diljem svijeta. Prema podatcima PubMed baze podataka, do danas je objavljeno preko 10 000 znanstvenih radova na ovu temu. Naravno, s obzirom da imamo pregršt postojećih radova i metoda ne možemo dati pregled svih njih no možemo na nekoliko primjera prikazati postojanost i raznovrsnost postojećih metoda.

"Predicting heart disease using machine learning techniques" – Kaur, P., & Bhatia, M. (2019)

Ovaj rad istražuje različite algoritme strojnog učenja, uključujući SVM, k-NN i logističku regresiju, kako bi se predvidio rizik od srčanih bolesti. Autori koriste podatke o pacijentima, kao što su starost, krvni tlak, kolesterol i druge relevantne varijable, kako bi izgradili prediktivni model.

"Heart disease prediction using machine learning algorithms" – Kumar, A., & Agrawal, A. (2020)

Rad koristi metode poput nasumičnih šuma, SVM i regresije, za predviđanje srčanog udara. Autori analiziraju podatke koji uključuju faktore rizika kao što su pušenje, fizička aktivnost, i povijest bolesti, kako bi razvili precizne prediktivne modele.

"A machine learning approach to predicting heart disease" – Zhang, H., & Li, J. (2021)

U ovom istraživanju, autori koriste duboko učenje i druge tehnike strojnog učenja za razvoj modela koji predviđa vjerojatnost nastanka srčanog udara. Istražuju koristeći metode kao što su duboke neuronske mreže (DNN) i SVM na osnovu podataka o pacijentima.

"Prediction of cardiovascular disease risk using machine learning algorithms" – Sharma, S., & Singh, V. (2019)

Ovaj rad se fokusira na primenu strojnog učenja za predviđanje rizika od kardiovaskularnih bolesti, uključujući srčani udar. Autori koriste algoritme poput SVM, k-NN, i regresije kako

bi razvili modele koji mogu efikasno identificirati pacijente sa visokim rizikom za srčane bolesti.

"Heart attack prediction using machine learning algorithms" – Khan, A., & Shah, N. (2021)

Ovaj rad istražuje primenu algoritama strojnog učenja za predviđanje srčanog udara na temelju medicinskih podataka. Autori koriste SVM i nasumične šume za razvoj modela koji identificira pacijente sa visokim rizikom za srčani udar, uzimajući u obzir faktore poput starosti, kolesterola, krvnog tlaka i životnih navika.

Vidimo da su korištene metode i algoritmi raznovrsni. Međutim, možemo uočiti da je model potpornih vektora dosta često u upotrebi u analiziranim radovima. Naime, SVM se pokazuje pogodnim za uporabu u području biomedicine. U našem radu, mi ćemo koristiti upravo model potpornih vektora za analizu promatranog problema.

3. Opis rješenja problema

Nakon što je dataset prethodno očišćen i pripremljen za analizu, istraženi su različiti algoritmi strojnog učenja kako bi se naposljetku odredilo koji model najbolje odgovara ovom problemu. Sveukupno je istrenirano 7 različitih modela, od kojih su dva duboki modeli – konvolucijska neuronska mreža (CNN) i unaprijed istrenirani model VGG16. Ostali modeli temelje se na tradicionalnim pristupima strojnog učenja i statističkim metodama.

Prvi korak uključivao je primjenu tradicionalnih modela strojnog učenja kako bi se procijenilo kako skup podataka (dataset) reagira na klasične pristupe. Među njima pokazalo se da logistička regresija daje iznimno dobre rezultate, što će biti detaljno objašnjeno u nastavku kroz opis eksperimentalnih rezultata. Logistička regresija se inače koristi za klasifikaciju na temelju linearnog odvajanja klasa i pokazao se prikladnim za problem detekcije srčanog udara zbog svoje jednostavnosti i učinkovitosti nad podacima. Logistička regresija implementirana je s parametrima ‘multinomial’ za višeklasnu klasifikaciju i algoritam ‘lbfgs’ za optimizaciju.

Uz logističku regresiju, isproban je i stroj potpornih vektora (SVM), koji se temelji na maksimiziranju margine između klasa. U ovom slučaju, korištena je linearna jezgra s regularizacijskim parametrom C=0.001. Ovaj model je također postigao visoku točnost.

Nadalje, isprobane su i statističke metode koje se razlikuju od tradicionalnih po tome što pretpostavljaju statističke distribucije podataka. Korišten je Naivni Bayesov klasifikator, točnije GaussianNB (pretpostavlja normalnu distribuciju podataka). Parametri modela ostavljeni su na zadanim vrijednostima. Ovaj model je koristan za prepoznavanje obrazaca u EKG slikama.

Za dodatnu usporedbu korišten je i model slučajne šume (Random Forest), s 100 stabala (n_estimators=100) i maksimalnom dubinom od 10 (max_depth=10). Model se sastoji od više stabala odlučivanja, gdje svako stablo pojedinačno donosi odluku, a konačna klasifikacija temelji se na većinskom glasanju svih stabala. Ovaj pristup omogućava modelu da se teže prenauči.

Korišten je model gradijentnog pojačavanja (XGBoost), koji je implementiran s parametrom multi:softmax za višeklasnu klasifikaciju. Ovaj model koristi sekvencijalno učenje stabla odlučivanja, pri čemu svaki novi model nastoji ispraviti pogreške prethodnog. Njegova prednost u detekciji EKG slika leži u njegovoj sposobnosti da uspješno uči složene relacije između značajki.

Uz ovih 5 modela, također su isprobana dva modela dubokog učenja. Prvi model temelji se na konvolucijskoj neuronskoj mreži (CNN), koji je konstruiran kako bi obradio slike EKG podataka. Veličina slika kao inputa je 224x224. Ovaj model sadrži 6 konvolucijskih slojeva (s aktivacijskom funkcijom ReLu), svaki praćen slojem za maksimalnu agregaciju (MaxPooling), a na kraju je korišten potpuno povezani sloj s softmax aktivacijskom funkcijom. Model je treniran s 15 epoha i optimizatorom Adam.

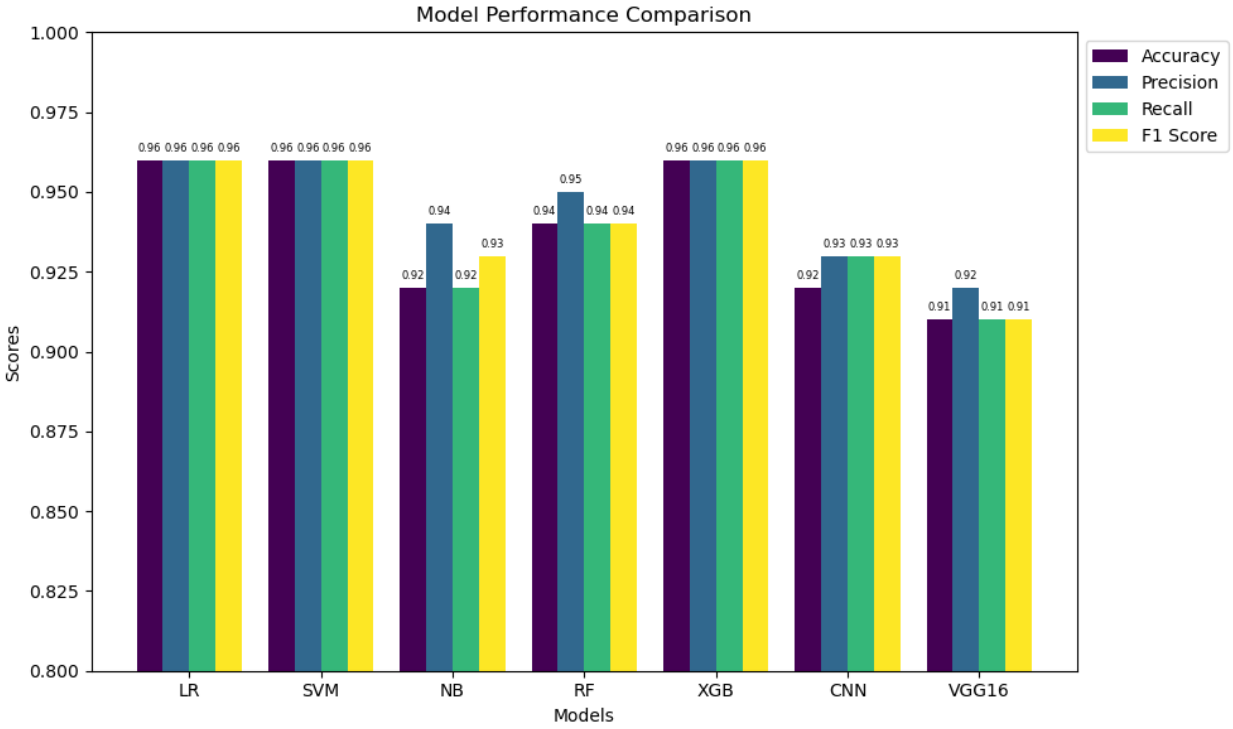
Drugi model duokog učenja je VGG16 model, koji je prilagođen za ovu specifičnu klasifikaciju. VGG16 je duboka konvolucijska mreža koja je prethodno trenirana na velikom skupu podataka, u ovom slučaju ImageNet-a. Model je prilagođen dodavanjem slojeva GlobalAveragePooling2D() i potpuno povezanih slojeva. Optimzator je također Adam i veličine slike je isto 224x224. Korišten je niži stupanj učenja kako bi se spriječilo prekomjerno prilagođavanje postojećim težinama modela. Trenirano je u 20 epoha.

4. Opis eksperimentalnih rezultata

Za svaki od 7 modela koja su istrenirana na skupu za učenje, ispitana je točnost na testnom skupu podataka EKG slika, kako bi se odredilo koji model daje najbolje rezultate u detekciji

srčanih problema na temelju EKG signala. Također, rezultati se mogu proučiti i usporediti na slici dolje.

1. **Logistička regresija:** Model logističke regresije postigao je 96,13% točnosti. Iako jednostavan model, iznimno je efikasan za osnovnu klasifikaciju. Ovaj rezultat ukazuje na visoku preciznost modela. Iz matrice konfuzije vidljivo je da je model pravilno klasificirao većinu uzoraka u sve četiri klase, s minimalnim greškama u klasifikaciji klase 2 (abnormalni otkucaji srca).
2. **Stroj potpornih vektora:** SVM s linearnom jezgrom postigao je 96,13% točnosti. Također, vrlo jednostavan model, ali s izvrnom točnošću. Također, rezultati su vrlo slični rezultatima logističke regresije.
3. **Naivni Bayesov klasifikator** (GaussianNB): Naivni Bayesov klasifikator postigao je 92,49% točnosti. Iako niži od ostalih modela, i dalje je bio uspješan u prepoznavanju određenih obrazaca u EKG slikama. Prema matrici konfuzije, model je imao problema s klasifikacijom klase 1 i 3, gdje se može primijetiti više pogrešnih predviđanja.
4. **Random Forest**: postignuta je točnost od 93,95%. Ovaj model pokazuje vrlo dobre rezultate, s visokom preciznošću u svim klasama. U matrici konfuzije, greške su ponajviše u klasama 1 i 2, ali gotovo neprimjetne.
5. **XGBoost**: 95,64% točnosti, što je jedan od najboljih rezultata među modelima. Matrica konfuzije pokazuje izuzetno visoku preciznost u svim klasama, s minimalnim greškama u svim klasama.
6. **Konvolucijska neuronska mreža** (CNN): CNN model postigao je 92,49% točnosti. Također, iz matrice konfuzije se vidi da je najviše pogreški za klase 1 i 2.
7. **VGG16:** 91,28% točnosti. Iz matrice konfuzije se vidi da je model imao najviše problema s klasom 3, no greške su minimalne.


Model Performance Comparison
Accuracy
Precision
Recall
F1 Score
Scores
Models
LR
SVM
NB
RF
XGB
CNN
VGG16
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.96 0.96 0.96 0.96
0.96 0.96 0.96 0.96
0.92 0.94 0.92 0.93
0.94 0.95 0.94 0.94
0.96 0.96 0.96 0.96
0.92 0.93 0.93 0.93
0.91 0.92 0.91 0.91