SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Alan Jović

DUBINSKA ANALIZA BIOMEDICINSKIH VREMENSKIH NIZOVA ZASNOVANA NA RAČUNALNOM RADNOM OKVIRU ZA IZLUČIVANJE ZNAČAJKI

DOKTORSKI RAD

Zagreb, 2012.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Alan Jović

DUBINSKA ANALIZA BIOMEDICINSKIH VREMENSKIH NIZOVA ZASNOVANA NA RAČUNALNOM RADNOM OKVIRU ZA IZLUČIVANJE ZNAČAJKI

DOKTORSKI RAD

Zagreb, 2012.

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Alan Jović

DATA MINING OF BIOMEDICAL TIME-SERIES BASED ON COMPUTER FRAMEWORK FOR FEATURE EXTRACTION

DOCTORAL THESIS

Zagreb, 2012.

Doktorski rad je izrađen na Sveučilištu u Zagrebu, Fakultetu elektrotehnike i računarstva, Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave.

Mentor: prof. dr. sc. Nikola Bogunović

Doktorski rad ima: 195 stranica

Doktorski rad br.: _____

Povjerenstvo za ocjenu i obranu doktorske disertacije:

1. Dr. sc. Sven Lončarić, redoviti profesor
   Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva

2. Akademik dr. sc. Leo Budin, emeritus, redoviti profesor (u mirovini)
   Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva

3. Dr. sc. Dragan Gamberger, znanstveni savjetnik
   Institut Ruđer Bošković Zagreb

Datum obrane disertacije: 20. travnja 2012.

Sažetak

Analiza biomedicinskih vremenskih nizova (BVN) obuhvaća široko interdisciplinarno područje. Istraživači u tom području nastoje kvantizirati složenost biološkog sustava kako bi se ostvarilo lakše razlikovanje između zdravog i bolesnog organizma. Uloga računarstva u tim istraživanjima je izlučivanje i dubinska analiza značajki.

Cilj ove disertacije je sistematizirati pristupe izlučivanju značajki i predložiti sveobuhvatan postupak analize podataka na putu od baze podataka do konačnog modela koji će omogućiti što bolje rezultate pri razvrstavanju poremećaja BVN. U tu svrhu predlaže se implementacija računalnog radnog okvira koji omogućuje izlučivanje velikog broja značajki za pojedinu vrstu BVN. U okviru disertacije implementacija okvira provedena je za područje analize srčanog ritma. Pritom su detaljno opisane vremenske, frekvencijske, vremensko-frekvencijske i nelinearne značajke korištene u tom području.

Razvijena su dva nova nelinearna postupka za izlučivanje značajki iz BVN: abecedna entropija i napredna analiza slijednog trenda, za koje se pokazuje da poboljšavaju rezultate razvrstavanja poremećaja. Sustavni postupak vrednovan je na dva zasebna problema razvrstavanja, a postignuta je viša točnost od sličnih pristupa navedenih u literaturi. Rezultati dobiveni ovom disertacijom unapređuju područje računalne analize BVN jer daju radni okvir temeljen na najboljoj praksi za postizanje visoke točnosti razvrstavanja poremećaja.

Ključne riječi: dubinska analiza podataka, biomedicinski vremenski nizovi, izlučivanje značajki, računalni radni okvir, nelinearne značajke, odabir atributa, razvrstavanje u više ciljnih razreda, srčani ritam, entropija

Abstract

Analysis of biomedical time-series encompasses a broad interdisciplinary area of scientific research. Numerous methods have been devised and applied in order to quantify the complexity of biological systems. One of the major problems is efficient and accurate classification between the normal state and different disorders. The problems in the area include: lack of reliable data, choice of features, time-series nonstationarities, noise, and others. The aim of this thesis is to provide a systematic approach starting from databases and continuing through feature extraction, data mining, and results evaluation that would ensure more accurate models of disorders in biomedical time-series.

As a part of the systematic approach, an implementation of a computer framework is proposed that consists of a large number of domain-relevant features. The features include: time domain, frequency domain, time-frequency, and nonlinear measures. Additional novel methods for feature extraction are also proposed: alphabet entropy and advanced sequential trend analysis. From the perspective of data mining, the analysis process includes: exclusion of disorders with unsatisfying initial models, filter-based feature selection, feature set optimization with a covering algorithm, classification using several machine learning methods, and construction of models with clear interpretation. The proposed method is evaluated on two classification problems in the cardiac rhythm domain using six open-access internet databases.

The framework was successfully applied for extraction of more than 230 features from cardiac rhythm records. Filter-based feature selection reduced the number of relevant features to 30% of the initial set for the first classification problem and only 20% for the second problem without significant loss of accuracy. Moreover, for the second problem, covering algorithm managed to remove several additional features for most of the inspected segment lengths. The best classification model for the first problem of 9 cardiac arrhythmias' classification was achieved by AdaBoost+C4.5 for 10 s segments (ACC=85.7%). For the second problem of discrimination between healthy persons, patients with any arrhythmia, and patients with congestive heart failure, the highest accuracy was achieved by AdaBoost+C4.5 for 300 s segments (ACC=92.7%). Models with clear interpretation were obtained by C4.5 and RIPPER algorithms. In terms of speed and accuracy, this work recommends random forest and AdaBoost+C4.5 classification algorithms.

The results obtained with the proposed method contribute to the field of computer analysis of biomedical time-series because they provide a framework based on best practice for achieving high discrimination of different types of disorders. Obtaining more data might improve the accuracy of the classification models.

Keywords: data mining, biomedical time-series, feature extraction, computer framework, nonlinear features, feature selection, multi-class classification, cardiac rhythm, entropy

Sadržaj

1 Uvod

1.1 Složenost bioloških sustava

Biološki sustavi pripadaju skupini složenih (engl. *complex*) sustava. Takvi sustavi razlikuju se od zamršenih (engl. *complicated*) sustava. Primjeri zamršenih sustava su listić porezne prijava ili baza podataka. Zamršeni sustav je sa strane promatrača „kompliciran", iako promatrač shvaća da bi, ako bi se dovoljno potrudio, mogao njime potpuno ovladati. Glavna razlika između zamršenog i složenog sustava je ta da je zamršeni sustav moguće razložiti na dijelove, rješavati ili analizirati po dijelovima i sastaviti natrag u logičnu cjelinu u potpunosti. Kod takvog sustava vrijedi princip superpozicije. Složeni sustav je s druge strane nešto više od zbroja pojedinih dijelova. On je cjelovit i hijerarhija njegovih dijelova na višoj razini iskazuje svojstva koja uopće nisu uočljiva s niže razine (tzv. izranjajuća svojstva). Redukcionistički analitički postupci pritom nisu dovoljni da bi se razumijeli složeni mehanizmi prisutni u biološkim sustavima, budući da su biološki sustavi po svojoj prirodi nelinearni. To znači da biološki sustavi iskazuju neočekivane, nagle i često nepredvidljive obrasce ponašanja u nekim situacijama, koje nije moguće opisati linearnim modelima. Najviše što se može učiniti sa složenim sustavom je opisati ga s dovoljno dobrim zamršenim modelom koji će uključivati neke nelinearne komponente. Takav model može se koristiti samo u određenu svrhu i tada će točno opisivati ponašanje složenog sustava u većini situacija.

U istraživanjima složene dinamike biološkog sustava istraživače najčešće zanima kako razlikovati zdravu dinamiku od poremećaja u funkcioniranju. Da bi identificirali ponašanje na nivou sustava koje je kritično za razumijevanje zdrave dinamike i patoloških poremećaja, autori [Peng C. 2009] postavljaju okvir s tri komplementarne hipoteze:

1. Složenost biološkog sustava reflektira njegovu sposobnost da se prilagodi i da funkcionira u stalno promjenjivom okolišu.
2. Biološki sustavi trebaju djelovati preko više vremenskih i prostornih skala. Time je i njihova složenost ujedno višeskalarna i hijerarhijska.
3. Širok razred bolesnih stanja, kao i starenje, čini se da degradira tu biološku složenost i smanjuje adaptivni kapacitet organizma.

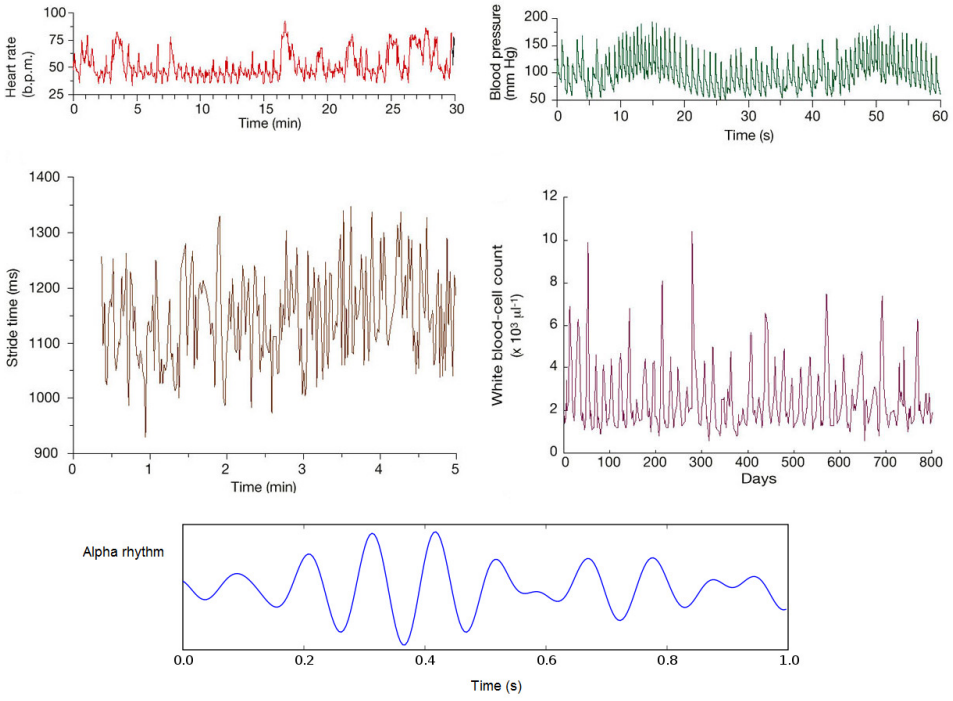
Fokus istraživanja u ovom području je upravo kvantificirati složenost biološkog sustava i na temelju toga odrediti koliko je sustav zdrav.

1.2 Biomedicinski vremenski nizovi

Da bi se odredila razina složenosti biološkog sustava nužno ga je mjeriti. Pritom se razlikuju invazivni i neinvazivni postupci mjerenja. U neinvazivne postupke ubrajaju se svi oni postupci kod kojih mjerenja ne remete cjelovitost biološkog sustava. Primjerice, elektroencefalogram (EEG) mjeren površinskim elektrodama spada u neinvazivni postupak mjerenja stanja mozga, dok se intrakranijalni EEG mjeri postavljanjem iglaste elektrode u mozak i pripada invazivnim postupcima. U pravilu, invazivni postupci daju točniji opis od neinvazivnih. Mjerenjem karakteristika biološkog sustava tijekom vremena s ciljem dijagnostike i liječenja, uz određenu stopu uzorkovanja, dobivaju se vremenski nizovi određenih amplituda koji se nazivaju biomedicinski vremenski nizovi.

Biomedicinski vremenski nizovi (engl. *biomedical time-series*, dalje: BVN) mogu se promatrati kao zasebna skupina vremenskih nizova s naročitim svojstvima koja proizlaze iz prirode bioloških sustava. Ta svojstva često uključuju: periodičnost, kvaziperiodičnost, kaotičnost i slučajnost. Poznati BVN uključuju: srčani ritam, elektrokardiogram (EKG), moždane valne obrasce (alfa ritam, beta ritam, ...), ritam hoda, ritam promjena krvnog tlaka, itd. [Glass 2001], slika 1.1.

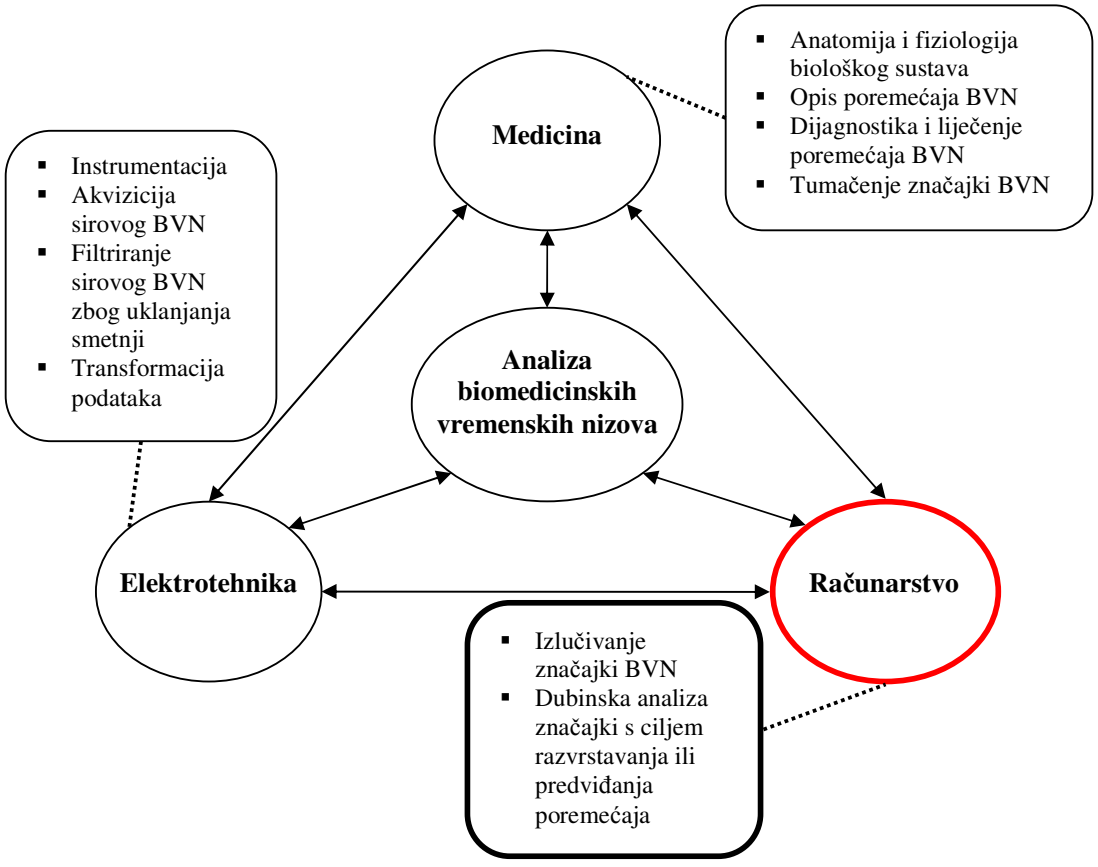
Da bi se neki BVN uspješno analizirao s bilo kojom namjerom, potrebno je uzeti u obzir raznovrsno znanje koje dolazi iz više područja znanosti. Tako se mogu izdvojiti tri glavne perspektive unutar područja analize BVN, a koje se međusobno upotpunjuju. Ta područja su medicina, elektrotehnika i računarstvo.



Slika 1.1. Biomedicinski vremenski nizovi, prilagođeni iz [Glass 2001].

Samu analizu BVN može se smjestiti unutar šireg područja biomedicinskog inženjerstva, ali i unutar računarske znanosti u kontekstu primjene njezinih postupaka. Na slici 1.2 dan je prikaz područja analize BVN zajedno s postupcima i temama kojima se ono bavi. Za potpuno razumijevanje BVN stručnjaci u ovom području koriste se i temeljnim znanostima matematike, fizike, biologije i kemije. Pritom medicina promatra anatomiju i (pato)fiziologiju biološkog sustava, opisuje poremećaj prisutan u BVN i njegov utjecaj na zdravlje. Također, u medicini su definirani postupci za dijagnostiku i liječenje poremećaja BVN. Svaka značajka BVN bi trebala imati neko fiziološko tumačenje, jer je u protivnom teško bilo što racionalno zaključivati o poremećaju.

Elektrotehnika je najčešće zadužena za izradu instrumentacije za dobavljanje i obradu BVN. U novije vrijeme, instrumentacija se povezuje s računalom što omogućuje daljnju napredniju analizu i razmatranje snimljenih zapisa. Provodi se filtriranje šuma, uklanjanje trendova i drugi postupci predobrade podataka. Ovdje se elektrotehnika isprepliće s računarstvom, budući da se na računalu koriste algoritmi posebno prilagođeni za analizu BVN. Fokus ovog rada nije na instrumentaciji, dobavljanju BVN ili njihovoj predobradi, već na računalnoj analizi samih obrazaca. Utoliko je na slici 1.2. istaknuto da se ova disertacija bavi izlučivanjem značajki iz BVN i dubinskom analizom značajki.



Slika 1.2. Postupci područja analize biomedicinskih vremenskih nizova.

1.3 Računalna analiza biomedicinskih vremenskih nizova

Tradicionalno, liječnici se bave proučavanjem značajki BVN, pri čemu opažaju očite i manje očite promjene u nizovima, koje su signal da je došlo do poremećaja. U novije vrijeme, s porasti mogućnosti računala, liječnici počinju koristiti računalne resurse i promatrati složenije značajke za analizu BVN i stoga više surađuju s inženjerima. Istraživanja svojstava determinističkog kaosa od strane fizičara dovela su do rasta zanimanja za proučavanje složenosti BVN. U svojem seminalnom radu u časopisu Lancet, autor [Goldberger 1996] skrenuo je pozornost na činjenicu da BVN imaju svojstvo nestacionarnosti, što znači da sadrže česte neočekivane promjene i odstupanja od periodičnosti. Otprilike do sredine 1990-tih ovakva nelinearnost obrazaca ponašanja nije uzimana u obzir, budući da su linearni modeli u vremenskoj i frekvencijskoj domeni bili dosta razvijeni i približno točni u kliničkoj primjeni [Task Force 1 1996]. Međutim, istraživačima je postalo jasno da tradicionalne značajke BVN kao što su različite statističke značajke, geometrijske značajke razdiobe i spektralne značajke ne daju dovoljno točan opis poremećaja. Razlog tome je taj što te značajke pretpostavljaju da u nizu nema naglih promjena.

Otada pa do danas istraživači sve više proučavaju naprednije značajke koristeći složenije računalne proračune. Tako se primjenjuju postupci kao što su vremensko-frekvencijska analiza, ponajprije analiza valića, s ciljem lokalizacije iznenadnih promjena [Kadbi 2006]. Također, uvidjelo se da su postupci razvijeni za analizu determinističkog kaosa i stohastičkih sustava počeli davati značajnije rezultate pri razvrstavanju i predviđanju poremećaja BVN [Small 2000, Stam 2005]. Otkriveno je primjerice da zdrav srčani ritam posjeduje fraktalna svojstva na više vremenskih skala koja su karakteristična za kaotične sustave [Ivanov 1999], dok je hod zdravog čovjeka ocijenjen kao monofraktalan [Muñoz-Diosdado 2005]. Mjerenjem entropije na različitim vremenskim skalama dobivene su značajke koje su pomogle u uspješnom razlikovanju između mladih i starih zdravih osoba i osoba s oboljenjima kao što su atrijalna fibrilacija i zatajenje srca [Costa 2002].

Velik broj raznih nelinearnih značajki osmišljen je i izlučivan s ciljem boljeg opisa BVN [Richmann 2000, Acharya 2006]. Članci brojnih istraživača u području analize BVN ukazuju na to da je za uspješan opis nekog niza i za predviđanje poremećaja istoga potrebno promatrati kombinaciju više vrsta značajki. Takva kombinacija treba uključivati tradicionalne, linearne značajke, ali i neke od novijih i naprednijih značajki [Asl 2008, Jović 2010 (1)]. Takve kombinacije značajki se dubinski analiziraju da bi se uspješno predvidjela pojava raznih poremećaja.

Velika količina podataka koja je dostupna liječnicima uvelike nadmašuje sposobnost liječnika da sve podatke na vrijeme prouči i donese točnu dijagnozu. Upravo stoga je računalom pogonjena dubinska analiza podataka postala nužna pomoć liječniku za najtočniju procjenu stanja pacijenta [Syed 2011]. Jedan od najčešćih zadataka dubinske analize je izgradnja automatskog sustava za razvrstavanje (klasifikaciju) poremećaja na temelju značajki nekog BVN. Takvi sustavi mogu pomoći liječnicima bilo u fazi

prevencije bolesti, bilo u kliničkoj praksi. Postupci razvrstavanja koji se pritom najčešće koriste su razne vrste umjetnih neuronskih mreža (ANN) [Acharya 2004 (1)], strojevi s potpornim vektorima (SVM) [Asl 2008] i stabla odluke [Exarchos 2007]. Često takvi postupci kombiniraju značajke BVN sa znanjem stručnjaka o poremećaju, čime se dobivaju ekspertni sustavi veće ili manje složenosti [Tsipouras 2005].

1.4 Motivacija i doprinosi disertacije

Fundamentalni problem koji se pojavljuje pri analizi BVN je beskonačna dimenzionalnost prostora značajki. Pri izboru koje značajke izabrati da bi se dobila idealna kombinacija za opis nekog niza i njegovih mogućih poremećaja, istraživači se uglavnom oslanjaju na prijašnja istraživanja i fiziološku interpretaciju značajki s obzirom na promatrane poremećaje. Često ne postoje standardi ni smjernice koje značajke odabrati za koju situaciju pa je stoga izbor značajki stvar informirane, ali subjektivne procjene istraživača. Istraživači odabiru jednu po jednu značajku i slažu ih u skup za koji očekuju, prema dosadašnjim spoznajama, da će dati najbolji opis određenog ritma ili poremećaja. Nedostatak ovog pristupa je u tome što istraživači razmatraju samo malen, ograničen broj postojećih značajki i tako nemaju garanciju da će dotični skup značajki najbolje opisati određeni poremećaj. Takvom pristupu nedostaje sustavnosti razmatranja više značajki istovremeno što dovodi do problema kao što su lošiji rezultati pri modeliranju poremećaja i nemogućnost kvalitetne usporedbe sa srodnim istraživanjima.

Ograničenja s kojima se istraživači često susreću uključuju nedostatak kvalitetnih, referentnih zapisa BVN kao i oslanjanje samo na značajke jednog BVN pri procjeni poremećaja u organizmu. Oprema za snimanje BVN je često nespretna za primjenu u svakodnevnom životu i zahtijeva stručno medicinsko znanje za uporabu. Stoga se pokazuje nužnim razvoj novih paradigmi i računalnih alata pomoću kojih će se na temelju informacija dobivenih samo iz jednostavnih, neinvazivnih BVN omogućiti visoka točnost određivanja poremećaja. Pritom se želi izvući najveća moguća količina informacije iz BVN koja će omogućiti što jednostavniji i razumljiviji model poremećaja.

U ovom radu problemu opisa i klasifikacije poremećaja BVN pristupa se na sustavan način, postupkom odozgo prema dolje. Osnovna ideja je implementirati većinu postojećih značajki korištenih u analizi određene vrste BVN i izlučiti ih uz pomoć računalnog radnog okvira. Tada se na temelju dostupnih internetskih podataka pacijenata s poremećajima BVN može napraviti računalni odabir najboljih značajki korištenjem postupaka dubinske analize podataka te dobiti modele poremećaja korištenjem više algoritama za razvrstavanje. Pritom se želi ustanoviti koji od korištenih postupaka daje najtočniji model, a koji daje model s jasnim tumačenjem za pojedinačne poremećaje. Korišteni postupak u okviru ove disertacije ne garantira sustavnost u

matematičkom smislu, već samo u smislu najbolje prakse za postizanje što točnijih rezultata.

Metodologija korištena u ovom radu kao svoju polaznu točku uzima izradu sveobuhvatnog računalnog radnog okvira za izlučivanje značajki za pojedinu vrstu BVN. U okviru ove disertacije pokazana je primjena metodologije na primjeru srčanog ritma. Pritom su neke značajke specifične za dotičnu vrstu BVN, dok su druge općenitije i primjenjive na više vrsta BVN. Radni okvir pritom uključuje što je moguće veći broj značajki poznat iz literature, i to iz smjernica za analizu te vrste BVN (ako takve smjernice postoje) i iz znanstvenih članaka objavljenih u časopisima ili na poznatim međunarodnim konferencijama. Računalni radni okvir uključuje najmanje sljedeće tri komponente:

1. Učitavanje zapisa za analiziranu vrstu BVN.
2. Postupke za izlučivanje značajki iz analizirane vrste BVN, uz mogućnost odabira parametara za pojedine značajke.
3. Pohranu izlaznih vektora značajki u određenom obliku koji je prilagođen daljnjoj dubinskoj analizi podataka.

Osim izrade radnog okvira, metodologija uključuje i postupke dubinske analize podataka s izradom što točnijih modela za određene poremećaje. Dubinsku analizu može se provesti u bilo kojem poznatom računalnom alatu za otkrivanja znanja, a u sklopu ove disertacije koristi se široko prihvaćeni sustav Weka [Hall 2009].

U disertaciji se razmatraju i dodatna dva nova postupka za dobivanje značajki BVN i to: 1) Abecedna entropija, i 2) Napredna analiza slijednog trenda. Postupci se analiziraju iz teorijskog stajališta i iz stajališta primjene na analizu srčanog ritma.

Cilj istraživanja u okviru ove disertacije je ponuditi potpuniji pristup problemu razvrstavanja poremećaja BVN i time uvesti veći red u postojeće pokušaje istraživača da opišu pravu prirodu određenog biološkog sustava i poremećaja u njegovom radu.

Hipoteze istraživanja unutar ove disertacije su sljedeće:

1. Sustavnim pristupom dobit će se u većini slučajeva točniji rezultati u automatskom razvrstavanju poremećaja BVN nego što je to bilo moguće na temelju vlastitog informiranog odabira značajki.
2. Izradom računalnog radnog okvira omogućit će se lakša nadogradnja postojećeg znanja i ubrzati nova istraživanja u ovom području tako što će se olakšati ponovljivost pokusa i usporedba s prijašnjim istraživanjima.
3. Korištenjem predložene metodologije moći će se uz veću sigurnost tvrditi da određena kombinacija značajki najbolje opisuje pojedini poremećaj ili neki BVN općenito, budući da će većina ostalih značajki biti također uzeta u obzir prilikom izgradnje modela.

1.5 Struktura rada

Čitatelj ove disertacije moći će se kroz nekoliko poglavlja upoznati sa širokim područjem koji obuhvaća analiza BVN. Naglasak disertacije je na perspektivi računarske znanosti, što uključuje izlučivanje i dubinsku analizu značajki. Također, čitatelju će se dati uvid u metodologiju razvijenu u svrhu što točnije analize BVN, s naglaskom na sustavnom postupku obrade i računalnom radnom okviru. Ovdje je dan sažeti opis sadržaja pojedinih poglavlja.

U poglavlju 2 predočena su srodna istraživana u ovome području. Sistematizirani su pristupi izlučivanju značajki i analizi poremećaja te su prikazani problemi pojedinih pristupa. Opisan je uobičajeni postupak za razvrstavanje BVN, navedeni su problemi područja i česte pogreške istraživača.

Poglavlje 3 posvećeno je opisu različitih vrsta značajki BVN. Značajke se opisuju u kontekstu primjene na konkretnu domenu srčanog ritma, a same značajke primjenjive su većinom i na ostale vrste BVN. U ovom poglavlju opisuju se i poremećaji ritma koji se modeliraju dalje u radu. Poseban naglasak u ovom poglavlju dan je raznovrsnim nelinearnim značajkama i njihovoj primjeni.

Izvorno predloženi postupci razvijeni u okviru ovog rada koji izlučuju više značajki iz BVN opisani su u poglavlju 4. To su postupci abecedne entropije i napredne analize slijednog trenda. Opisuje se pozadina koja je značajna za razvoj postupka, teorijski aspekt postupka i primjena u području analize srčanog ritma.

Poglavlje 5 namijenjeno je opisu računalnog radnog okvira koji služi za izlučivanje velikog broja značajki iz BVN. Dan je opis ulaznih i izlaznih podataka iz radnog okvira kao i pregled svih implementiranih postupaka za izlučivanje značajki. U poglavlju se navode parametri radnog okvira i dodatne mogućnosti koje okvir nudi u smislu nadziranog učenja. Razvijeni radni okvir uspoređuje se s drugim pristupima poznatima u literaturi.

U poglavlju 6 dan je pregled područja dubinske analize podataka koja se koristi u okviru metodologije za analizu BVN. Razrađuju se područja: pripreme vektora značajki za analizu, problematike skupova za učenje i testiranje, postupaka za odabir značajnih atributa, algoritama razvrstavanja i mjera vrednovanja izgrađenih klasifikatora.

Poglavlje 7 obrađuje metodologiju izrade modela poremećaja BVN i daje pregled sustavnog postupka. Postupak uključuje: dobavljanje referentnih podataka i njihovu pripremu za analizu, korištenje računalnog radnog okvira za izlučivanje značajki, uklanjanje poremećaja s nezadovoljavajućim rezultatima iz analize, odabir značajnih atributa, razvrstavanje poremećaja, vrednovanje modela i izgradnju modela s jasnim tumačenjem.

U poglavlju 8 vrednuje se na konkretnim podacima na dva klasifikacijska problema predloženi sustavni postupak, kao i zasebno dva izvorno predložena postupka za izlučivanje značajki. Daje se usporedba rezultata s dosadašnjim istraživanjima.

Poglavlje 9 zaključuje ovu disertaciju, daje kritički osvrt i prijedloge za daljnja istraživanja i primjenu metodologije.

2 Srodna istraživanja

2.1 Izbor značajki pri analizi biomedicinskih vremenskih nizova

Biomedicinski vremenski nizovi, kao i većina drugih vremenskih nizova, posjeduju beskonačan skup mogućih značajki koje se mogu promatrati i analizirati. Konačan podskup tog skupa koji bi bio koristan za dijagnostiku nije moguće unaprijed utvrditi, već je on uvijek ovisan o konkretnom problemu. Pri izboru koju značajku uzeti u obzir za opis nekog BVN istraživači se najčešće povode za nekoliko odrednica.

Glavna odrednica je mogućnost fiziološkog tumačenja dotične značajke, što znači da se ustanovljava odnos između iznosa izračunate značajke i karakteristike biološkog sustava na kojem je ta značajka mjerena. Primjerice, analizom srčanog ritma u frekvencijskom području moguće je odrediti dinamiku odnosa između dvije grane autonomnog živčanog sustava (engl. *autonomic nervous system*, kraće: ANS) promatrajući odnos između niskofrekventne i visokofrekventne komponente spektra [Task Force 1 1996]. Druga odrednica je jednostavnost i razumljivost matematičkog proračuna. Radi bržeg izračunavanja značajke, po mogućnosti u stvarnom vremenu, istraživači su skloniji koristiti značajke koje se jednostavnije računaju. Matematički jednostavnije značajke se ujedno i lakše i učinkovitije implementiraju, a i značajka se lakše tumači od strane stručnjaka, čak i kad nema na prvi pogled jasnu fiziološku osnovu. Pritom nema nikakve garancije da će jednostavnija značajka bolje opisati BVN od neke složenije, budući da većina BVN iskazuju vrlo složene obrasce ponašanja. Treća odrednica je robusnost značajke s obzirom na šum u BVN. Iako je pretpostavka da danas postoje uspješni postupci za smanjenje šuma u izvornom nizu, on gotovo nikad nije u potpunosti uklonjen [Clifford 2006]. Neke mjere faznog prostora pretpostavljaju određeni stupanj nelinearnog determinizma za točan izračun (npr. korelacijska dimenzija). Kod slučajeva izraženog šuma u nekim BVN, takve značajke nisu pouzdane [Ding 1993].

Neke teme koje se proučavaju u području BVN uključuju:

1. Izrada modela BVN na temelju jedne ili više značajki [Voss 2007].
2. Automatsko razvrstavanje BVN s obzirom na prisutnost/odsutnost poremećaja [Acharya 2003].
3. Generiranje umjetnog BVN na temelju svojstava stvarnog BVN [Chen Z. 2002].
4. Predviđanje pojave poremećaja BVN u budućnosti [Manis 2007].

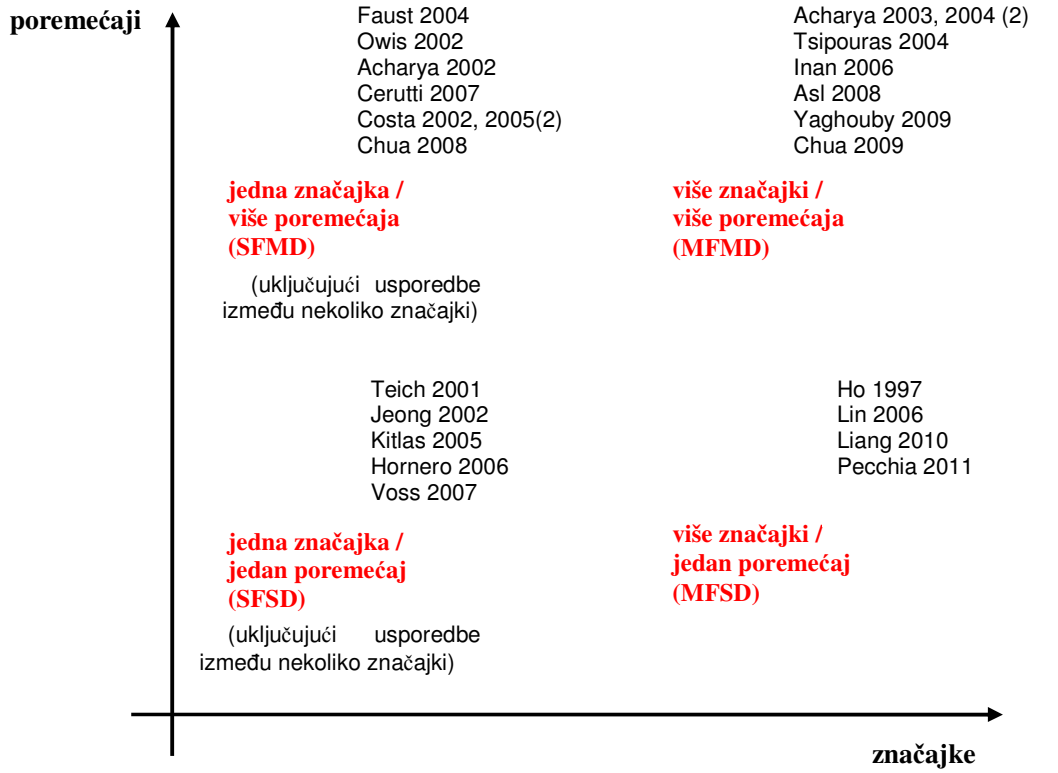
Fokus ove disertacije je na prva dva problema budući da su posebno zanimljivi iz perspektive računarstva. Razrješavanje prvog problema odnosno izrada modela je ujedno i nužan korak za rješavanje svih ostalih problema. Jedino se s uspješnim modelom BVN može biti u određenoj mjeri siguran da će se dalje moći uspješno razvrstavati pojedine nizove s obzirom na prisutnost poremećaja, predviđati pojave poremećaja u budućnosti kao i generirati kvalitetan umjetni BVN. Za predviđanje

poremećaja nužno je ustanoviti koji se obrazac pojavljuje u nekom vremenu prije pojave poremećaja, što se najčešće provodi na temelju nelinearnih značajki [Lehnertz 2008]. Umjetni nizovi su zanimljivi radi određivanja matematičkih svojstava stvarnih nizova, što se provodi usporedbom stvarnog niza s umjetnim nizovima koji imaju matematički definirane karakteristike.

Prilikom razvrstavanja BVN na temelju prisutnosti poremećaja postoji nekoliko pristupa u ovisnosti o tome promatra li se jedna ili više značajki te promatra li se jedan ili više poremećaja istovremeno. Na slici 2.1 prikazan je pregled istraživanja poremećaja BVN na temelju značajki. U pregled su uključeni najznačajniji radovi za pojedino područje. Razlikuju se četiri vrste pristupa:

1. Jedna značajka / jedan poremećaj (engl. *single feature, single disorder*, SFSD)
2. Jedna značajka / više poremećaja (engl. *single feature, multiple disorders*, SFMD)
3. Više značajki / jedan poremećaj (engl. *multiple features, single disorder*, MFSD)
4. Više značajki / više poremećaja (engl. *multiple features, multiple disorders*, MFMD)

1. Kod pristupa SFSD razmatra se jedna značajka i utvrđuje se najčešće: 1) u kakvoj je vezi ta značajka s fiziološkim faktorima koji utječu na BVN, 2) može li se značajka koristiti za opis svojstava BVN ((ne)lineranost, (ne)determinizam)), i 3) koliko je ta značajka uspješna pri određivanju toga postoji li poremećaj u BVN ili je niz normalan. Uspješnost se u slučaju jedne značajke ocjenjuje najčešće opisnom statistikom. Ponekad



Slika 2.1. Pregled istraživanja poremećaja BVN na temelju značajki.

se u ovoj vrsti istraživanja daje i usporedba uspješnosti razvrstavanja između nekoliko pojedinačnih značajki pri čemu se utvrđuje postoji li neka značajka koja najbolje određuje poremećaj. Primjerice, autori [Hornero 2006] istražuju kako značajke približne entropije, složenosti Lempel-Ziv i mjere središnje težnje EEG-a omogućuju razlikovanje zdravih osoba od osoba koje boluju od shizofrenije, pri čemu dobivaju da značajka približne entropije ima najveću osjetljivost (90%), dok složenost Lempel-Ziv ima najveću specifičnost (90%). Pristup SFSD koristan je za relativnu usporedbu kvalitete značajki kao i za razmatranje iz perspektive fiziologije, no problem je u tome što tek u rijetkim slučajevima jedna značajka dovoljno točno razlikuje neki poremećaj. Tako su primjerice autori [Kitlas 2005] pokazali statistički značajnu razliku između zdrave djece i djece s dijabetesom na temelju približne entropije izračunate na nizu srčanih otkucaja. Ipak, jedna značajka, približna entropija, nije mogla razlikovati sve pacijente, a najvjerojatnije zbog malog uzorka nije niti procijenjen postotak uspješnog razlikovanja. Autori [Jeong 2002] su definirali mjeru središnje težnje za graf drugog reda razlike kao mjeru determinizma u BVN i primijenili su je na zapis EEG-a. Pritom su ustanovili da površinski EEG ne sadrži determinizam niskog reda, no time nisu pokazali da je EEG slučajan niz, budući da je šum mogao prekriti visokodimenzionalan red.

2. U slučaju kad autori razmatraju jednu značajku za razvrstavanje više vrsti poremećaja (SFMD) situacija je još nepovoljnija. Osim što dotična značajka treba razlikovati radi li se o normalnom BVN, potrebno je razlučiti i o kojoj se vrsti poremećaja radi. Primjerice, autori [Acharya 2002] koristili su, između ostalih, značajku korelacijske dimenzije srčanog ritma kao postupak kojim su uspješno razlučili između normalnog ritma i dvije vrste srčanih poremećaja. Postigavši gotovo potpuno uspješno razlikovanje, zaključili su da dotična značajka ima potencijala u kliničkoj praksi. Ipak, ovaj rad ima ograničenje u smislu da su se istraživači ograničili na poremećaje koje je relativno jednostavno razlikovati, tj. normalni ritam, atrijalnu fibrilaciju (engl. *atrial fibrillation*, kraće: AF) i potpuni srčani blok (engl. *complete heart block*, kraće: CHB). Autori [Owis 2002] primijenili su pak korelacijsku dimenziju na čitav EKG signal pri razlikovanju pet vrsti obrazaca. Statistička analiza je pokazala da su rezultati značajno različiti između normalnog ritma i različitih aritmija, međutim aritmije se međusobno nisu mogle zadovoljavajuće razlikovati.

Pažljivim razmatranjem uočava se da su prva dva pristupa (SFSD i SFMD) pogodna u slučajevima kad se žele ispitati pojedinačne značajke kao kandidate koji imaju dobar dijagnostički potencijal. Nakon što se za takve značajke pouzdano utvrdi ima li ih smisla izračunavati, istraživači obično dalje upotrebljavaju druga dva pristupa – MFSD i MFMD, s ciljem kombiniranja značajki za što točniji opis poremećaja.

3. Istraživači koji koriste pristup MFSD žele pronaći najtočniji opis jedne vrste poremećaja, odnosno žele pronaći takav složeni model dotičnog poremećaja koji će ga

najtočnije razlikovati od normalnog BVN. Istraživači u ovom pristupu u pravilu ne razmatraju jednostavnost tumačenja dobivenih modela iz perspektive značajki ili složeno fiziološko tumačenje. U odnosu na ostale vrste pristupa, ova istraživanja su nešto rjeđa. Autori [Lin 2006] proučavali su pragove analize valića i energetske spektralne komponente pri razlikovanju zdravih osoba i osoba koje boluju od kongestivnog zatajenja srca (engl. *congestive heart failure*, kraće: CHF), pri čemu se pokazalo da zdrave osobe imaju više vrijednosti pragova i spektralnih komponenti od bolesnih. Istim problemom, ali iz perspektive jednostavnih vremenskih i spektralnih komponenti, bavili su se i istraživači [Pecchia 2011]. Oni su pokazali da korištenjem kratkodjelujućih značajki varijabilnosti srčanog ritma (engl. *heart rate variability*, kraće: HRV) mogu postići zadovoljavajuću točnost (osjetljivost 79.3% specifičnost 100%), pri razlikovanju zdravih osoba i osoba oboljelih od CHF. Liang i koautori [Liang 2010] proučavali su automatsko otkrivanje epileptičnih napada na EEG-u koji su modelirali korištenjem približne entropije i spektralnih komponenti dobivenih autoregresijskim modelom 20. reda. Postigli su najveću ukupnu klasifikacijsku točnost od 98% koristeći stroj s potpornim vektorima s radijalnom funkcijom jezgre (engl. *radial basis function*, kraće: RBF).

4. Pristup MFMD je ujedno i najčešći u novijim istraživanjima. U ovom slučaju koristi se kombinacija značajki za razvrstavanje više vrsti poremećaja odjednom, bez jednostavnog tumačenja. Naglasak pristupa MFMD je na automatskom razvrstavanju raznovrsnih poremećaja sa što većom točnosti. Primjerice, autori [Chua 2009] koriste sedam značajki spektra višeg reda pri razvrstavanju između pet vrsti poremećaja srčanog ritma i dobivaju dobre rezultate (prosječna osjetljivost 90%, prosječna specifičnost 88%). Autori [Inan 2006] koriste vremenski niz EKG-a za otkrivanje tipa srčanog otkucaja. Značajke koje pritom koriste su morfološke značajke analize valića i vremenske značajke. Pritom razlikuju normalni otkucaj, preuranjenu kontrakciju ventrikula (engl. *premature ventricular contraction*, kraće: PVC) i ostale otkucaje s ukupnom klasifikacijskom točnošću od preko 95%.

Pristupi MFSD i MFMD imaju nekoliko nedostataka. Prvo, postoji problem izbora značajki koje se trebaju izlučiti iz BVN-a. Budući da postoji više pojedinačnih značajki koje su u istraživanjima iskazala potencijal, istraživači se najčešće odlučuju za izbor nekolicine od tih značajki, i to na temelju subjektivne procjene i karakteristika samih značajki (fiziološko tumačenje, jednostavnost izračuna, robusnost na šum). Ovakav informirani odabir značajki prije same analize je problematičan u smislu da je moguće da se neka bitna značajka ne uzme u obzir, a istraživači se ponekad ograđuju korištenjem raspoloživih računalnih alata [Melillo 2011]. Drugi problem je taj na koji se poremećaj ili poremećaje treba fokusirati. Najčešće istraživači već unaprijed imaju hipotezu koju žele ispitati, no samim time zanemaruju mogućnost da je sustav primjenjiv i na druge poremećaje osim onih koje oni proučavaju. Nadalje, istraživači iskazuju zabrinutost po pitanju usporedbe njihovih rezultata s rezultatima drugih

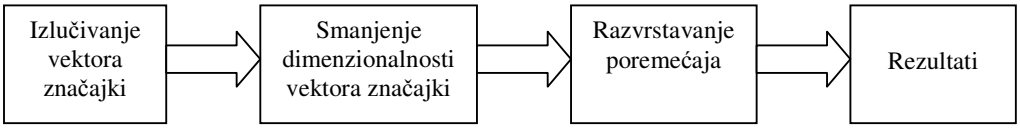
istraživača, pogotovo zato što svatko koristi svoju kombinaciju značajki i razmatra različite poremećaje [Asl 2008]. Posebno je problematično to što najčešće nema službenih smjernica po pitanju koje su značajke najbolje za opis i razlikovanje određenog poremećaja, iako se tome nastoji doskočiti raznim preglednim radovima u kojima se navode prednosti značajke i probleme na koje je dotična značajka dosad primijenjena [Acharya 2006, Xia 2009]. Konačno, problem koji je često zanemaren pri istraživanjima u pristupima MFSD i MFMD je medicinska strana, odnosno ocjena koliko su ti rezultati zaista značajni u smislu kliničke dijagnostike i opisa klinički značajnih poremećaja [Yaghouby 2009].

2.2 Izgradnja modela niza

U slučaju kad se ispituje više od jedne značajke za određivanje modela BVN istraživači koriste postupke dubinske analize podataka. Cilj istraživača je pritom najčešće vezan uz što točnije automatsko razvrstavanje vremenskih nizova [Tsipouras 2004, Asl 2008, Yaghouby 2009, Chua 2009, Ceylan 2009], no koji put cilj je izgraditi model poremećaja sa što jasnijim tumačenjem. Jasno tumačenje moguće je kod sustava zasnovanih na ekspertnom znanju i induktivnom postupku izgradnje klasifikacijskih pravila. Primjerice, autori [Tsipouras 2005] grade ekspertni sustav za otkrivanje srčanih otkucaja i srčanih ritmova. Pritom za otkrivanje vrste otkucaja koriste trointervalni prozor s poznatim liječničkim pravilima za otkrivanje poremećaja, dok za ritmove grade deterministički konačni automat zasnovan također na ekspertnom znanju. Postigli su visoku točnost od 98% točnosti za otkucaje i 94% za srčane ritmove uz jasno tumačenje. Autori [Exarchos 2007] razmatraju metodologiju za automatsku izgradnju neizrazitih ekspertnih sustava. Prema toj metodologiji, prvo se izlučuju izrazita pravila iz stabla odluke na skupu za učenje. Zatim se izrazita pravila transformiraju u neizraziti model. Na kraju se parametri neizrazitog modela optimiraju koristeći globalnu optimizaciju stohastičkom metodom. Autori su primijenili metodologiju na morfološkim značajkama EKG-a i postigli visoku osjetljivost razvrstavanja ishemijskih epizoda (91%) i aritmija (96%). Prethodno već spomenuti rad autora [Pecchia 2011] koristi stabla odluke, i to metodu CART da bi izgradio model poremećaja CHF s jasnim tumačenjem za liječnike, pri tome zadržavajući relativno visoku točnost razvrstavanja.

U ovoj disertaciji razmatra se i automatsko razvrstavanje poremećaja s ciljem što veće točnosti kao i modeli s jasnim tumačenjem uz zadržavanje visoke točnosti. Stoga je najprije potrebno razmotriti uobičajeni postupak kojim se dolazi do rezultata pri automatskom razvrstavanju poremećaja BVN-a. Taj postupak istraživači najčešće provode kroz nekoliko faza, koje su predočene na slici 2.2.

Na početku postupka izlučuju se vektori značajki iz BVN-a. Značajke se najčešće dobivaju analizom po segmentima s određenim trajanjem, najčešće fiksnim [Tsipouras 2004]. Koji puta značajke se izlučuju na čitavom zapisu pacijenta [Teich 2001]. Tako izlučene značajke mogu se izravno koristiti u postupku razvrstavanja. Ipak, mnogi



Slika 2.2. Uobičajeni postupak izgradnje modela BVN pri automatskom razvrstavanju s više značajki

istraživači prije samog razvrstavanja žele smanjiti broj značajki, odnosno smanjiti dimenzionalnost vektora značajki. To smanjenje se provodi iz dva glavna razloga. Prvi razlog je povećanje točnosti konačnog modela, koje se postiže zbog smanjene osjetljivosti na šum kad se uklone suvišne značajke, a drugi razlog je jasnije tumačenje konačnog modela. Nažalost, visoka točnost i mogućnost tumačenja modela su konfliktni zahtjevi u složenim sustavima. Primjenom nekih postupaka za smanjenje dimenzionalnosti istraživači gube na mogućnosti tumačenja (npr. analizom glavnih komponenata (engl. *principal component analysis*, kraće: PCA)) [Minhas 2008, Liang 2010]. Istraživači u području analize BVN rijetko su kad dovoljno upoznati s postupcima odabira značajki koji uklanjaju suvišne značajke bez transformacije podataka i gubljenja mogućnosti tumačenja. Takvi postupci koriste se u ovoj disertaciji i opisani su u poglavlju 6.3.

Prilikom analize BVN najčešće postoji velik broj segmenata ili zapisa koji sadrže normalan obrazac ponašanja. Segmenti kod kojih je odsutan bilo koji poremećaj dominiraju u odnosu na segmente kod kojih je analizirani poremećaj prisutan, često u odnosu 10:1 ili više [Tsipouras 2005]. Iz perspektive algoritama strojnog učenja ova činjenica dovodi do problema neravnoteže razreda prilikom učenja. Razni algoritmi strojnog učenja različito reagiraju na problem neravnoteže razreda, no u pravilu neravnoteža utječe nepovoljno na kvalitetu rezultata [Chawla 2002]. Istraživači u medicinskom području nastoje doskočiti tom problemu uvođenjem različite cijene razvrstavanja za normalne nizove i nizove s poremećajem te kombiniranjem cijene razvrstavanja i vjerojatnosti pojave niza [Darrington 2008]. Također, uvijek se, uz ukupnu točnost razvrstavanja, navode i druge mjere vrednovanja kao što su: osjetljivost, specifičnost, pozitivna prediktivna vrijednost i površina pod krivuljom ROC [Tsipouras 2004, Alvarez 2007]. Ipak, sam problem neravnoteže pri učenju algoritma razvrstavanja ovime nije riješen budući da se korištenim mjerama utječe samo na rezultate, ne i na proces učenja.

Razvrstavanje poremećaja postiže se računalnim algoritmima strojnog učenja. U području analize BVN od algoritama za razvrstavanje daleko najčešće koriste se umjetne neuronske mreže (engl. *artificial neural networks*, kraće: ANN) [Haykin 1998]. Dvije najčešće upotrebljavane arhitekture neuronske mreže su: višeslojni perceptron (engl. *multilayer perceptron*, kraće: MLP) učen algoritmom *backpropagation* [Acharya 2004 (1), Kadbi 2006, Yu S. 2004] i mreža s radijalnom baznom funkcijom [Bezerianos 1999, Manis 2007]. Od ostalih arhitektura ANN u upotrebi treba se spomenuti

neuronsku mrežu s povratnom vezom [Übeyli 2008] i samo-organizirajuću mapu [Lagerholm 2000]. Primjerice, autori [Kadbi 2006] koriste značajke koeficijenata diskretne transformacije valićima, RR-intervale i statističku značajku faktora oblika s kaskadom od dvije ANN za razlikovanje 10 vrsti aritmija kod EKG-a. Autori [Tsipouras 2004] koriste više MLP-ova učenih algoritmom *backpropagation* i glasanje za klasifikaciju srčanih aritmija na temelju vremenskih i vremensko-frekvencijskih značajki varijabilnosti srčanog ritma i postižu osjetljivost i specifičnost od 90% i 93%. Neki istraživači isprobavaju neuronske mreže s neizrazitom logikom, najčešće u kombinaciji s klasičnim MLP-om [Osowski 2001, Acharya 2003]. Neuronske mreže imaju tri značajna nedostatka u odnosu na neke druge postupke strojnog učenja: 1) Modeli im nemaju jasno tumačenje, 2) Postoji mogućnost zaustavljanja u lokalnom optimumu, i 3) Sklone su prenaučenosti ako se pažljivo ne kontrolira izgradnja modela [Lisboa 2000]. Osim toga, neuronske mreže najčešće sporije uče u odnosu na ostale postupke.

U novije vrijeme istraživači eksperimentiraju sa strojevima s potpornim vektorima (engl. *support vector machine*, kraće: SVM) [Platt 1998]. Za razliku od neuronskih mreža, SVM nema problem sa zapinjanjem u lokalnom optimumu, izbjegava prenaučenost kontrolirajući marginu i uči brže od većine arhitektura ANN-a [Lisboa 2000]. Rezultati koji se dobivaju korištenjem SVM-a često su nešto točniji od ANN-a [Osowski 2004, Asl 2008]. Primjerice, autori [Asl 2008] koriste SVM za vrlo točno (99.2%) razvrstavanje šest vrsta srčanih aritmija na temelju kombinacije značajki varijabilnosti rada srca. Pritom transformiraju vektore značajki korištenjem dviju metoda: analize glavnih komponenata (engl. *principal component analysis*, kraće: PCA) i opće diskriminantne analize (engl. *general discriminant analysis*, kraće: GDA), čime gube na sposobnosti tumačenja.

Od ostalih postupaka za izgradnju modela i razvrstavanje BVN koji se rjeđe spominju u literaturi treba spomenuti skrivene Markovljeve modele [de Lannoy 2008], Bayesove mreže [Soman 2005], stabla odluke [Bogunović 2010, Pecchia 2011] i slučajne šume [Jović 2010 (2)] .

2.3 Problemi područja

Područje analize BVN sadrži niz problema koji se mogu uočiti i zasebno razmatrati. Gruba podjela ovih problema je na one svojstvene samom području i na one koje čine istraživači, bilo svjesno bilo nesvjesno, u svojim istraživanjima. Uočeni problemi i ograničenja koja ne ovise o istraživačima su sljedeći:

1. Beskonačna dimenzionalnost prostora značajki.
2. Manjak dostupnih referentnih i visokokvalitetnih zapisa BVN.
3. Brojnost poremećaja koji se mogu analizirati [Hu 1997, Ceylan 2009].
4. Nepredvidljivost naglih promjena u BVN što slijedi iz nelinearnosti bioloških sustava [Goldberger 1996].

5. Niska prediktivna vrijednost pri predviđanju poremećaja u organizmu na temelju analize jedne vrste BVN-a [Chattipakorn 2007].
6. Visoka razina šuma uzrokovana raznim izvorima kod većine vrsta BVN [Clifford 2006].
7. Biološka raznolikost pojedinaca uzrokuje veliku različitost u opaženim obrascima BVN [Hu 1997].
8. Neslaganje stručnjaka oko označavanja nekih vrsta poremećaja, što otežava nadzirano učenje.

Osim ovih problema, postoje i problemi na koje istraživači mogu utjecati, ali se svejedno u radovima na njih ne osvrću, ili ih namjerno zanemaruju. Neki od ovih problema navedeni su u jednom od prijašnjih radova [Jović 2011 (1)]. Problemi uključuju:

1. Nejasan postupak segmentacije zapisa ili označavanja segmenata, kao nužan dio pripreme za postupak izlučivanja značajki [Asl 2008].
2. Izgradnja klasifikacijskih modela na samo jednoj bazi podataka (u preko 90% radova, npr. [Kim 2009, Talbi 2009]).
3. Korištenje javnosti nedostupne baze podataka, bez izravne usporedbe s referentnom, javnosti dostupnom bazom podataka [Wang 2001, Angelini 2007, Chua 2009].
4. Neregularnosti pri izboru skupova za učenje i testiranje [Ceylan 2009].
5. Zanemarivanje standarda bez naročitog opravdanja i to u smislu: 1) izbora parametara pri izračunu značajki [Gamero 2002], 2) ne navođenja svih potrebnih mjera vrednovanja [Acharya 2003].
6. Izračunavanje značajke na segmentu takve duljine za koju dotičnu značajku nema smisla izlučivati [Yaghouby 2009].
7. Nedostatak usporedbe rezultata različitih postupaka razvrstavanja na istom problemu, čime se gotovo sigurno gubi na optimalnosti postignute točnosti razvrstavanja [Asl 2008, Chua 2009].
8. Sva ostala problematika pri korištenju više značajki navedena u poglavlju 2.1.

U predloženom postupku u okviru ove disertacije dan je poseban naglasak na tome da se izbjegnu problemi na koji istraživači mogu utjecati, a opisano je i kako provesti analizu da se što više umanji nepovoljan utjecaj problema na koje istraživač ne može utjecati, budući da su takvi problemi svojstveni upravo području analize BVN.

3 Značajke biomedicinskih vremenskih nizova

Značajke biomedicinskih vremenskih nizova donekle se razlikuju od domene do domene. Iz tog razloga, izgradnja univerzalnog radnog okvira koji bi pokrio sve domene je složena i zasad teško ostvariva. Ipak, primjetno je da mnogi BVN dijele zajedničke karakteristike u pogledu periodičnosti i nelinearnosti koje se mogu iskoristiti za analizu u više različitih domena. U nastavku ovog poglavlja daje se pregled domene varijabilnosti srčanog ritma na koju je dano težište u ovoj disertaciji. Razlog za odabir srčanog ritma pri demonstraciji sustavnog postupka analize BVN je dvojak: 1) Svojstva vremenskog niza srčanih otkucaja kao i primjena raznovrsnih značajki za analizu tog niza najbolje su istražena u znanstvenoj literaturi od svih ostalih vrsta BVN-a; 2) Klinički podaci o srčanom ritmu su, uz ukupni EKG, najdostupniji. Domena varijabilnosti srčanog ritma opisuje se u ovom radu koristeći većinu danas dostupnih pristupa izračunavanja značajki. Mnogi od navedenih pristupa primjenjivi su i na druge BVN, što će biti razvidno tijekom opisa njihovog izračunavanja. Ipak, u ovom poglavlju značajke će se opisivati uglavnom u kontekstu primjene na analizu srčanog ritma.

3.1 Primjer domene primjene: varijabilnost srčanog ritma

3.1.1 Osnovna građa srca

Srce je složen i životno važan mišić koji omogućuje opskrbu svih stanica tijela kisikom i hranjivim tvarima koji se nalaze u krvi. Srce je građeno od četiri unutrašnje komore i to dvije pretklijetke (atrij) i dvije klijetke (ventrikul). U poprečnom presjeku, srce je građeno od četiri sloja: perikard (osrčje), epikard (vanjski sloj), miokard (mišićno tkivo) i endokard (unutarnji sloj). Miokard je najdeblji sloj koji sadrži poprečnoprugaste mišićne stanice i provedbene stanice. Osim mišićnih stanica, srce ima i vlastiti složeni provodni sustav kojemu je zadatak provođenje električnih impulsa i pobuđivanje stanica miokarda.

Provodni sustav počinje sa sinoatrijalnim čvorom (engl. *sinoatrial node*, kraće: SA-čvor), nakupinom provodnih stanica unutar miokarda desnog atrija. Ovo područje zaduženo je za stvaranje električnog impulsa i djeluje kao glavni srčani signal predvodnik (ili pejsmejker) (engl. *pacemaker*) koji diktira brzinu rada srčane pumpe. Brzina otkucaja SA-čvora regulirana je u ovisnosti od informacije koja se dobiva od autonomnog živčanog sustava (ANS) krvožilnog sustava i endokrinog sustava. Signal predvodnik SA-čvora u normalnom režimu radi na 60-100 otkucaja u minuti, s prosjekom od oko 70 [Garcia 2001]. Električni impuls se kroz atrije provodi međučvornim provodnim putevima pri čemu se usput nabijaju stanice miokarda atrija. Provodni putevi se spajaju pri vrhu srčane pregrade u donjem dijelu desnog atrija kod atrioventrikularnog čvora (kraće: AV-čvor). Dalje se impuls provodi kroz brzi Hissov

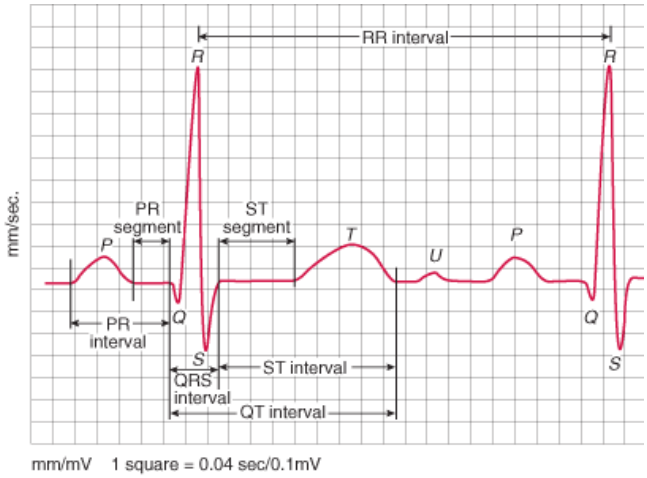
snop uz vrlo malo usporenje sve dok se u centru pregrade snop ne podijeli na lijevu i desnu granu. Snopovi se dalje granaju u sve manje svežnjeve provodnih vlakana sve do samih stanica miokarda ventrikula kako bi osigurali polarizaciju svih dijelova srčanog mišića.

3.1.2 Dobivanje i definicija niza srčanih otkucaja

Srčani otkucaji se u novije vrijeme najčešće dobivaju iz elektrokardiograma (kraće: EKG; engl. *electrocardiogram*, kraće: ECG). EKG je naziv za krivulju koja pokazuje oscilacije u proizvodnji električne struje što je proizvodi srce prilikom procesa depolarizacije i repolarizacije stanica miokarda. Uređaj koji mjeri ove struje naziva se elektrokardiograf. To je osjetljivi galvanometar koji mjeri razlike potencijala između točno definiranih točaka na koži pacijenta. Za određivanje razlike potencijala najčešće se koristi normiranih 12 elektroda (dalje: odvod) koje se prilijepe na točno određena mjesta na koži. Time se dobiju snimke 12 signala EKG-a. Liječnici često ne promatraju sve odvode, već samo njih nekoliko da bi uspješno ustanovili mogući poremećaj.

Električki potencijali jednog srčanog otkucaja prikazani su na slici 3.1. P-val prikazuje depolarizaciju atrija, nakon čega slijedi R-zubac koji predstavlja vrhunac depolarizacije ventrikula. Repolarizacija atrija maskirana je intervalom QRS (ili kompleksom QRS) zbog male amplitude. Nakon intervala QRS slijedi T-val koji predstavlja repolarizaciju ventrikula. Mali U-val koji često nije vidljiv nema općenito prihvaćenog fiziološkog objašnjenja, ali mogao bi predstavljati naknadnu depolarizaciju ventrikula ili repolarizaciju papilarnih mišića srca [Garcia 2001].

Niz srčanih otkucaja najčešće se dobiva iz snimljenog EKG-a. Za to je potreban samo jedan od tragova pod uvjetom da nije bilo značajnih smetnji za vrijeme snimanja. Ono što se otkriva, to je zubac R, koji je najviša (ili najniža, ovisno o odvodu) točka unutar jednog ciklusa srčanog otkucaja. Za pronalaženje R-zupca u signalu koriste se razni



Slika 3.1. Intervali i segmenti srčanog otkucaja.

algoritmi, no najpoznatiji standardni algoritam je onaj Pana i Tompkinsa iz 1985 [Pan 1985]. To je bio prvi algoritam koji je osigurao visoku točnost (99.2%) pri detekciji R-zubaca. Algoritam se zasniva na otkrivanju uspona, amplitude i širine intervala QRS. Noviji algoritmi dostižu i veću točnost otkrivanja R-zubaca, čak i oko 99.8% [Adnane 2009].

Analiza varijabilnosti srčanog ritma (HRV) promatra kolebanja u nizu intervala koji se mjere između dvije depolarizacije ventrikula (RR-intervali). Time se nastoje kvantizirati varijacije u trenutačnom broju otkucaja srca. Takva analiza koristi se za otkrivanje i prvu procjenu ozbiljnosti srčanih poremećaja [Kitney 1982].

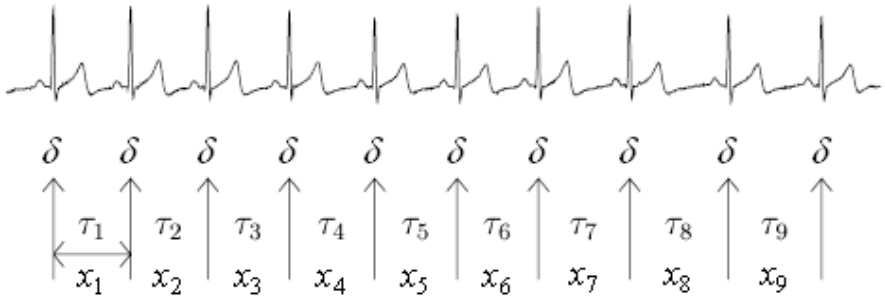
Ponašanje niza srčanih otkucaja može se proučavati tako da se složeni valni oblik QRS-kompleksa pojedinačnih otkucaja zamijeni samo s vremenom kad je nastupila kontrakcija, dakle s vremenom R-zupca. Time se čitav valni oblik zamjenjuje jednim brojem. Na isti način, niz otkucaja srca zabilježen je kao niz brojeva koji označavaju vremenske trenutke pojave R-zubaca. U matematičkom smislu, niz otkucaja je tada neoznačeni točkasti proces. Takav proces može se prikazati izrazom:

$$h(t) = \sum_i \delta(t - t_i) \tag{3.1},$$

gdje je δ Diracova delta funkcija, a $\{t_i\}$ je vremenski niz otkucaja srca (R-zubaca).

Ako se promatra interval između dva otkucaja u procesu u nizu (RR-interval), tada je niz RR-intervala određen s $\{\tau_i\} = \{t_{i+1} - t_i\}$. Očito je da je broj RR-intervala uvijek za jedan manji od broja pojedinačnih otkucaja srca.

U daljnjem radu vremenski niz RR-intervala označavat će se sa $\{x_i\}$, uz to da se podrazumijeva $\{x_i\} = \{\tau_i\} = \{t_{i+1} - t_i\}$ i to da je niz otkucaja neoznačeni točkasti proces, slika 3.2. Prikaz vremenskog niza RR-intervala na kojem je na osi *x* vrijeme (ili redni broj otkucaja), a na osi *y* trenutačna frekvencija otkucaja (bilo u sekundi, bilo u minuti) naziva se tahogram.



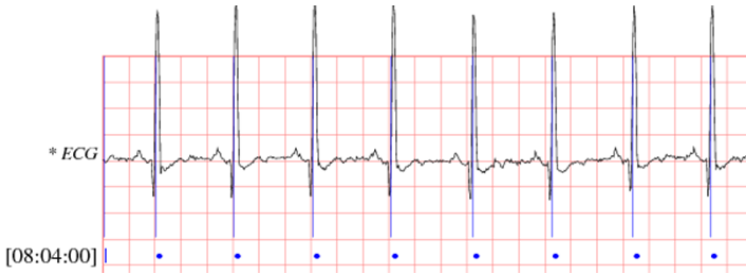
Slika 3.2. Srčani ritam, definiran trenutcima otkucaja, odnosno vremenskim nizom $\{x_i\}$.

3.1.3 Srčani ritmovi i poremećaji

U ovom poglavlju razmatraju se različiti obrasci srčanog ritma koji će se razvrstavati korištenjem sustavnog postupka u okviru ove disertacije. Postoji velik broj raznih vrsta srčanih poremećaja za koje nije dostupno dovoljno podataka ili koji se ne mogu otkriti analizom ritma jer se ne razlikuju bitno od ostalih poremećaja. Stoga će se ovdje razmotriti samo oni poremećaji ritma za koje su dostupni relevantni podaci i koji imaju potencijala da ih se može razvrstati. Svi primjeri srčanih ritmova uzeti su iz zapisa stvarnih pacijenata u internetskim bazama [PhysioNet] i prikazani na slikama u obliku: naziv ritma, baza podataka, broj zapisa, vremenski segment. Od srčanih bolesti razmatrat će se samo klinički značajna bolest kongestivnog zatajenja srca. Prilikom opisa poremećaja koristit će se uvriježeni medicinski nazivi u ovome području. Za medicinske detalje oko poremećaja srčanog ritma preporučuje se literatura [Garcia 2001].

3.1.3.1. Normalan sinusni ritam, tahikardija, bradikardija i respiratorna sinusna aritmija

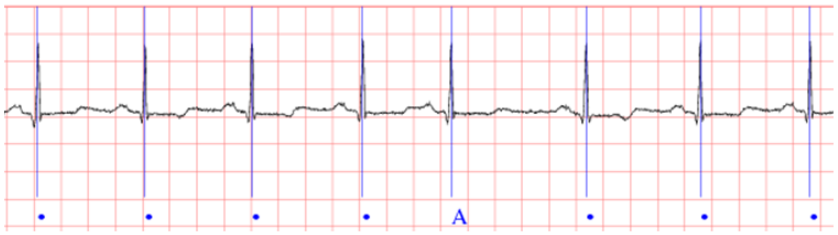
Primjer normalnog sinusnog ritma (engl. *normal sinus rhythm*, kraće: NSR) prikazan je na odsječku EKG-a na slici 3.3. NSR ima pravilnu frekvenciju od 60-100 normalnih otkucaja (N) u minuti. U slučaju da je frekvencija manja od 60 otkucaja u minuti, a ritam i dalje pravilan, tada se govori o sinusnoj bradikardiji (engl. *sinus bradycardia*, kraće: SBR). Ako je frekvencija veća od 100 otkucaja u minuti, uz pravilan ritam, tada je to sinusna tahikardija (engl. *sinus tachycardia*, kraće: ST). Respiratorna sinusna aritmija (engl. *respiratory sinus arrhythmia*, kraće: RSA) je specifičan oblik sinusne aritmije koji nastaje zbog frekvencije disanja i izražava se najčešće kod mlađih ljudi. Karakterizira ju ubrzanja i usporenja ritma u skladu s frekvencijom disanja. Poremećaj nije klinički značajan i u ovoj disertaciji neće se razlikovati od normalnog ritma.



Slika 3.3. Normalan sinusni ritam, *MIT-BIH Normal Sinus Rhythm Database*, 16265, 08:04.

3.1.3.2. Preuranjena kontrakcija atrija

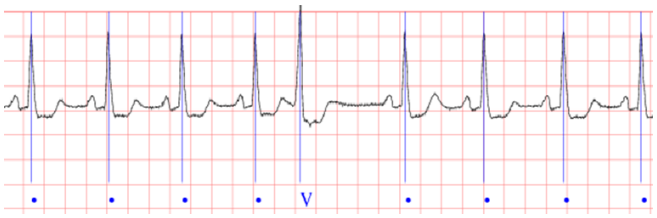
Preuranjena kontrakcija atrija (engl. *premature atrial contraction*, kraće: PAC) je poremećaj pri kojem dolazi do preuranjenog paljenja stanica miokarda u SA-čvoru, zbog raznih razloga. Na EKG-u je prisutan val P, kompleks QRS počinje u istom smjeru kao i kod normalnog otkucaja, duljina QRS-kompleksa je normalna (<0.12 s). Otkucaj je najčešće nekompenziran: udaljenost od PAC do sljedećeg normalnog otkucaja je manja od dvostruke normalne udaljenosti od P do P-vala, slika 3.4.



Slika 3.4. Preuranjena kontrakcija atrija, *MIT-BIH Arrhythmia Database*, 100, 00:03.

3.1.3.3. Preuranjena kontrakcija ventrikula

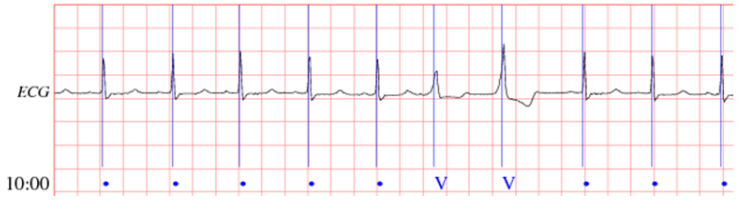
Preuranjena kontrakcija ventrikula (PVC) je poremećaj ritma pri kojem atrij ne potakne kontrakciju srca, već to učini ventrikul. Na EKG-u stoga nema P-vala, QRS-kompleks koji put počinje u suprotnom smjeru od normalnog otkucaja, QRS-kompleks je dugotrajan (>0.12 s) i čudnog izgleda. Otkucaj je najčešće kompenziran: udaljenost od otkucaja prije PVC do onog poslije PVC je približno jednaka dvostrukoj duljini normalnog otkucaja, slika 3.5.



Slika 3.5. Preuranjena kontrakcija ventrikula, *MIT-BIH Arrhythmia Database*, 105, 00:12.

3.1.3.4. Kuplet

Kupleti su parovi nepravilnih otkucaja, koji slijede jedan za drugim. Bilo koja kombinacija je moguća: PAC+PAC, PAC+PVC, PVC+PAC, PVC+PVC. Svaka od kombinacija ima svoja vlastita svojstva po pitanju ritma, ali razlike često nisu jasne ni značajne. Ako je drugi otkucaj PVC trebalo bi doći do kompenzacije, iako ni to nije pravilo. Primjer kupleta prikazan je na slici 3.6.

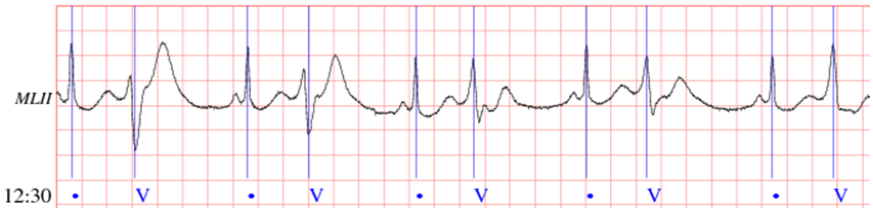


Slika 3.6. Kuplet, *MIT-BIH Supraventricular Arrhythmia Database*, 806, 10:00.

3.1.3.5. Bigeminija

Bigeminija je naziv za izmjenjivanje između normalnog i abnormalnog otkucaja, pri čemu abnormalni otkucaj može biti ili PAC ili PVC. Da bi se takvo izmjenjivanje smatralo bigeminijom, nužno je da postoje barem tri abnormalna otkucaja, npr: N+PAC+N+PAC+N+PAC+N, gdje je N normalan otkucaj. Razlike između atrijalne (engl. *atrial bigeminy*, kraće: ABI) i ventrikularne bigeminije (engl. *ventricular*

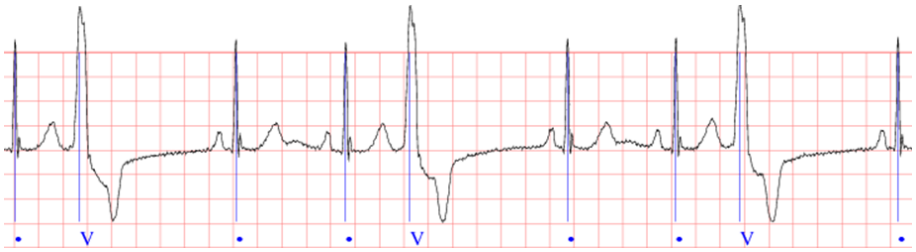
bigeminy, kraće: VBI) postoje po pitanju duljine kompenzacije nakon abnormalnog otkucaja. Ventrikularna bigeminija je i klinički značajnija. Na slici 3.7 dan je primjer ventrikularne bigeminije.



Slika 3.7. Ventrikularna bigeminija, *MIT-BIH Arrhythmia Database*, 106, 12:30.

3.1.3.6. Trigeminija

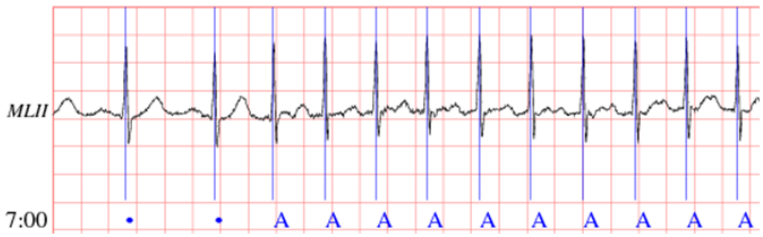
Trigeminija je naziv za izmjenjivanje između dva normalna i jednog abnormalnog otkucaja, pri čemu abnormalni otkucaj može biti ili PAC ili PVC. Da bi se takvo izmjenjivanje smatralo trigeminijom, nužno je da postoje barem tri abnormalna otkucaja, npr: N+PAC+N+N+PAC+N+N+PAC+N. Razlike između atrijalne (ATR) i ventrikularne trigeminije (VTR) postoje po pitanju duljine kompenzacije nakon abnormalnog otkucaja. Ventrikularna trigeminija je i klinički značajnija. Primjer je dan na slici 3.8.



Slika 3.8. Ventrikularna trigeminija, *MIT-BIH Arrhythmia Database*, 119, 05:01.

3.1.3.7. Supraventrikularna tahikardija

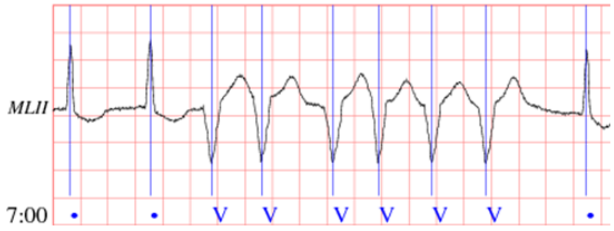
Supraventrikularna tahikardija (engl. *supraventricular tachycardia*, kraće: SVT) je ubrzanje srčanog ritma s izvorom ritma koji se nalazi iznad ventrikula. Poremećaj ima veće značenje od normalne sinusne tahikardije koja se javlja postepeno, a manje značenje od ventrikularne tahikardije (engl. *ventricular tachycardia*, kraće: VT). Poremećaj se može promatrati kao niz od tri ili više PAC-ova koji počinju naglo i naglo prekidaju, slika 3.9.



Slika 3.9. Supraventrikularna tahikardija, *MIT-BIH Arrhythmia Database*, 209, 07:00.

3.1.3.8. Ventrikularna tahikardija

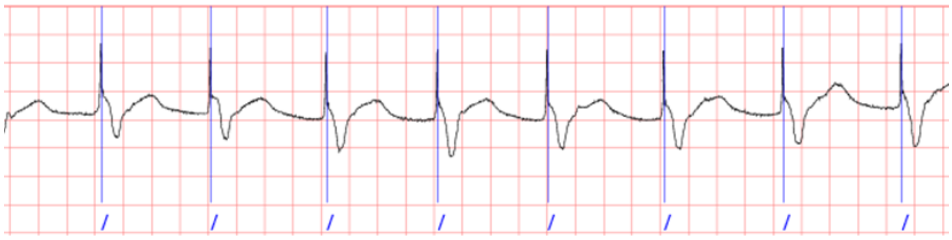
Ventrikularna tahikardija (VT) je ozbiljan poremećaj kod kojeg dolazi do ubrzanog ritma s izvorom u ventrikulima, s mogućnosti prelaska u smrtonosni oblik ventrikularne aritmije - ventrikularnu fibrilaciju (engl. *ventricular fibrillation*, kraće: VF). Poremećaj karakterizira tri ili više uzastopnih PVC-ova koji nenadano počinju i naglo završavaju, slika 3.10.



Slika 3.10. Ventrikularna tahikardija, *MIT-BIH Arrhythmia Database*, 210, 07:00.

3.1.3.9. Umjetno vođeni ritam - pejsmejker

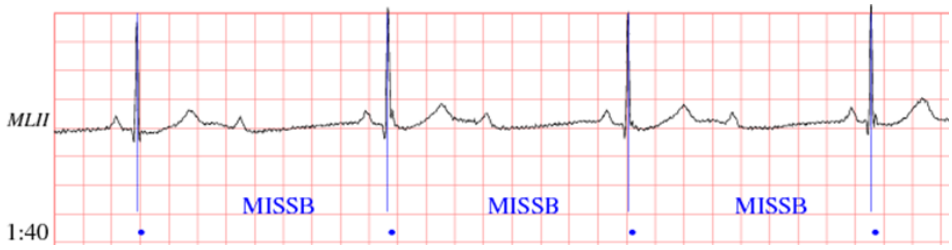
Neki pacijenti imaju u sebi ugrađen ili privremeno prisutan uređaj za kontrolu i vođenje srčanog ritma – pejsmejker. Umjetni pejsmejker ima karakterističan obrazac ritma koji često daje još pravilnija vremena otkucaja od prirodnog sinusnog pejsmejkera iz SA-čvora. Primjer umjetno vođenog ritma dan je na slici 3.11.



Slika 3.11. Vođeni ritam - pejsmejker, *MIT-BIH Arrhythmia Database*, 104, 07:13.

3.1.3.10. Atrioventrikularni (AV) blok drugog stupnja

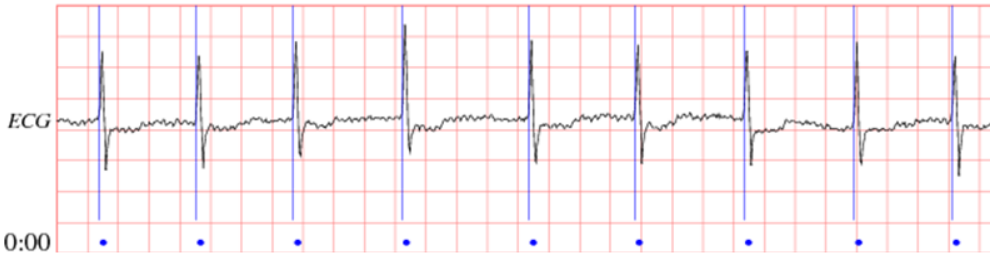
AV-blok drugog stupnja (BII) je poremećaj provodnog sustava u srcu. Postoje dva tipa BII: Mobitz I i Mobitz II, koji se razlikuju po tome je li PR-interval (vidi sliku 3.1) varijabilne duljine (Mobitz I) ili konstantan (Mobitz II). Drugi tip je opasniji za pacijenta. Kod oba tipa BII dolazi do ispuštanja ventrikularnog odgovora (R-zupca i ukupnog QT-intervala) u omjeru *X*:*X*-1, npr. 2:1 (od dva otkucaja, jedan ispušten), slika 3.12.



Slika 3.12. Blok drugog stupnja Mobitz II, 2:1, *MIT-BIH Arrhythmia Database*, 231, 1:40.

3.1.3.11. Atrijalna fibrilacija

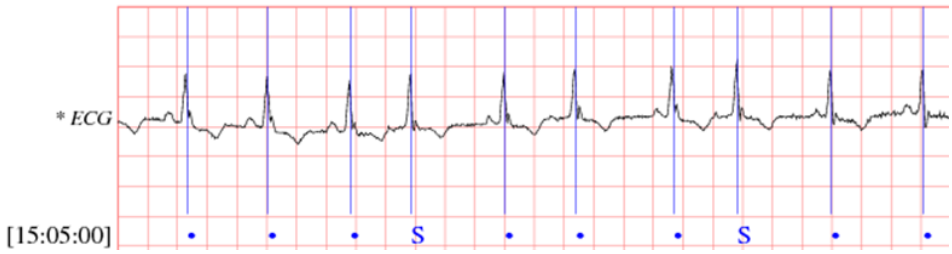
Atrijalna fibrilacija (AF) je obrazac koji nastaje aktivacijom brojnih stanica u atrijima na slučajan i nepredvidljiv način. Ovaj obrazac nema značajnih P-valova, a R-zupci pojavljuju se u nepravilnim vremenskim razmacima, slika 3.13.



Slika 3.13. Atrijalna fibrilacija, *MIT-BIH Supraventricular Arrhythmia Database*, 804, 00:00.

3.1.3.12. Kongestivno zatajenje srca

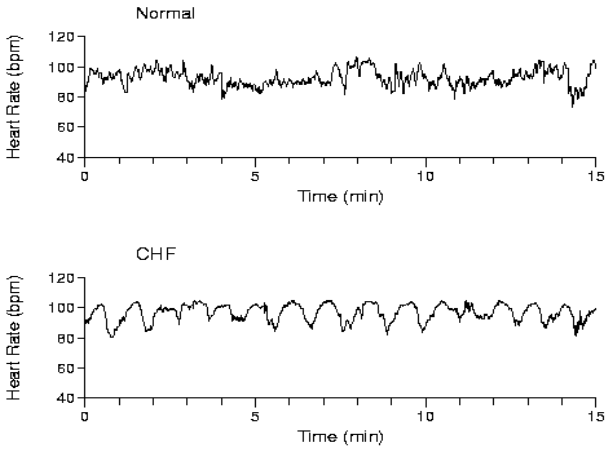
Kongestivno zatajenje srca (CHF) je ozbiljna srčana bolest sa slabom prognozom za pacijente. CHF karakterizira slabljenje srčane pumpe čime dolazi do nakupljanja tekućine u području oko srca, plućima, bubrezima i drugdje u tijelu. Prema simptomatsko-funkcionalnoj klasifikaciji Njujorške udruge za srce (engl. *New York Heart Association*, kraće: NYHA), postoje četiri stupnja zatajenja srca (NYHA I do NYHA IV), od najslabijeg do najtežeg [Task Force 2 2008]. Srčani ritam i EKG karakteriziraju povremene aritmije, smanjenje varijabilnosti i druge nepravilnosti koje su najčešće povezane sa slabljenjem srčane funkcije, slika 3.14.



Slika 3.14. Primjer EKG-a osobe s kongestivnim zatajenjem srca uz dodatne PAC-ove, klasa NYHA III-IV, *BIDMC Congestive Heart Failure Database*, 804, 15:05:00.

3.2 Linearne vremenske značajke

Linearne vremenske značajke su jednostavne značajke koje služe za statistički opis vremenskog niza srčanih otkucaja kao i za opis geometrijskih svojstava vremenske razdiobe otkucaja srca. Tipično se promatraju značajke prosječnog trajanja otkucaja, standardnog odstupanja otkucaja u nekom periodu i sl. Statističke vremenske značajke same po sebi teško mogu razlikovati različite poremećaje rada srca, budući da služe za stacionaran opis signala. Većina poremećaja su nestacionarni, budući da se događaju odjednom uz veću ili manju ponovljivost. Na tahogramu na slici 3.15 prikazana su dva



Slika 3.15. Tahogram NSR i srčanog ritma osobe s CHF, uzeto iz [Goldberger 1996].

Tablica 3.1. Linearne vremenske značajke varijabilnosti srčanog ritma.

Značajka	Kategorija	Literatura gdje je značajka upotrebljena
Srednja vrijednost $\bar{x}$	statistička	[Task Force 1 1996, Ho 1997, Krstačić 2003, Asl 2008, Yaghouby 2009]
SDNN (SDRR)	statistička	[Task Force 1 1996, Tsipouras 2004, Krstačić 2003, Yaghouby 2009, Toichi 1997, Asl 2008, Jović 2010 (1,2)]
RMSSD	statistička	[Task Force 1 1996, Tsipouras 2004, Jović 2010 (1,2)]
SDSD	statistička	[Task Force 1 1996, Asl 2008]
pNN5	statistička	[Mietus 2002, Tsipouras 2004, Asl 2008]
pNN10	statistička	[Mietus 2002, Tsipouras 2004, Asl 2008]
pNN20	statistička	[Mietus 2002, Hutchinson 2003, Jović 2010 (1,2)]
pNN50	statistička	[Task Force 1 1996, Tsipouras 2004, Mietus 2002, Hutchinson 2003, Asl 2008, Yaghouby 2009, Jović 2010 (2)]
SDANN (SENN)	statistička	[Task Force 1 1996, Tsipouras 2004]
Fanov faktor	statistička	[Turcott 1996, Teich 2001, Jović 2010 (2)]
HTI	geometrijska	[Task Force 1 1996, Yaghouby 2009, Jović 2010 (1,2)]
TINN	geometrijska	[Task Force 1 1996]

srčana ritma koje statističke mjere ne mogu razlikovati, a očito je da se ritmovi značajno razlikuju. Naime, statističke mjere definiraju se na nekom vremenskom segmentu i ne analiziraju promjenljivosti u trajanju pojedinih otkucaja unutar tog segmenta. Statističke mjere imaju pokazano veću varijabilnost za poremećaje ritma s većom varijacijom duljine RR-intervala, kao što su PVC, sindrom bolesnog sinusnog čvora (engl. *sick sinus syndrome*, kraće: SSS), i AF, dok su manje korisne za sporo-varirajuće ritmove kao što su ishemijska/dilatacijska kardiomiopatija (engl. *ischemic/dilated cardiomyopathy*, kraće: I/D kardiomiopatija), potpuni srčani blok (CHB) i blok lijeve grane (engl. *left bundle branch block*, kraće: LBBB) [Acharya 2006].

Geometrijske vremenske značajke služe za opis geometrijskog obrasca funkcije gustoće razdiobe srčanog ritma. Tu se najčešće promatra širina ili visina histograma, a moguće su i aproksimacije razdiobe nekim geometrijskim oblikom kao što je primjerice trokut. Geometrijske značajke su relativno robusne s obzirom na analitičke vrijednosti samog signala.

Popis najčešće izlučenih linearnih vremenskih značajki dan je u tablici 3.1. Naveden je naziv značajke, tip značajke (statistička ili geometrijska) i popis značajnije literature gdje je značajka korištena. U nastavku je svaka značajka detaljnije objašnjena.

3.2.1 Statističke značajke

3.2.1.1. Srednja vrijednost $\bar{x}$

Srednja vrijednost niza od *N* RR-intervala određena je izrazom.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.2.}$$

Srednja vrijednost $\bar{x}$ povezana je s pulsom (engl. *heart rate*, HR), koji svaka osoba može odrediti jednostavnim mjerenjem na žili kucavici. Puls se definira kao [Faust 2004, Yaghouby 2009]:

$$HR = \frac{60}{\bar{x}}\left[\frac{otk}{\min}\right] \tag{3.3.}$$

Srednja vrijednost, odnosno puls govori o brzini ritma rada srca, što najčešće nije dovoljno za dijagnozu nekog poremećaja rada srca. Međutim, u kombinaciji s drugim značajkama može biti bitan čimbenik pri razvrstavanju određenih vrsti poremećaja ritma [Asl 2008, Yaghouby 2009].

3.2.1.2. Standardno odstupanje SDNN

Standardno odstupanje RR intervala, u oznaci SDNN (ili SDRR) je značajka koja uzima u obzir sve čimbenike odgovorne za ukupnu varijabilnost ritma unutar određenog perioda. Izračunava se pomoću izraza:

$$SDNN = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} \tag{3.4.}$$

SDNN se može izračunavati za 24-satne snimke kao i za kraće segmente. Prema smjernicama mjeri se za pet-minutne segmente [Task Force 1 1996, Jović 2010 (1)], ali ima je smisla koristiti i za manji broj intervala [Asl 2008].

3.2.1.3. RMSSD

RMSSD (engl. *root-mean-square of successive differences*) je kvadratni korijen sume kvadriranih razlika između *N* slijednih RR-intervala podijeljen s brojem intervala *N* [Tsipouras 2004]. Određuje ga izraz:

$$RMSSD = \frac{\sqrt{\sum_{i=1}^{N-1}(x_{i+1} - x_i)^2}}{N} \tag{3.5.}$$

U odnosu na srednju vrijednost i SDNN, RMSSD se izračunava na kraćem segmentu i govori o brzo djelujućim promjenama ritma unutar promatranog segmenta [Task Force 1 1996]. U smjernicama se navodi da postoji nelinearni odnos između značajki RMSSD

i pNN50 (vidi 3.2.1.5). RMSSD je bolje koristiti nego pNN50 zbog boljih statističkih svojstava [Task Force 1 1996].

3.2.1.4. SDSD

SDSD je standardno odstupanje razlika između slijednih RR-intervala. [Task Force 1 1996]. Ova značajka također govori o brzo djelujućim promjenama ritma unutar promatranog segmenta. Može se izračunati izrazom:

$$SDSD = \sqrt{\frac{\sum_{i=1}^{N-1}\left(x_{diff}(i) - \bar{x}_{diff}\right)^2}{N-2}}, \quad x_{diff}(i) = x_{i+1} - x_i, \quad \bar{x}_{diff} = \frac{\sum_{i=1}^{N-1} x_{diff}(i)}{N-1} \tag{3.6}.$$

3.2.1.5. pNNX

pNNX je skup značajki koje iskazuju omjer broja razlika u trajanju slijednih RR-intervala koje je dulje od *X* ms i ukupnog broja RR-intervala. Može se odrediti izrazom:

$$pNNX = \frac{\sum_{i=1}^{N-1}\delta_i}{N}, \begin{cases} \delta_i = 1, (x_{i+1} - x_i) \geq X \text{ ms} \\ \delta_i = 0, \text{inace} \end{cases} \tag{3.7}.$$

U njihovom radu iz 2002., autori [Mietus 2002] su zaključili da ako se za *X* uzme $X \leq 20\,\text{ms}$, da se u tom slučaju dobiva vrlo dobro razlikovanje između zdravih osoba i pacijenata oboljelih od CHF, kao i između mladih i starih osoba. Standardna vrijednost *X*-a koju preporučuju smjernice pak iznosi 50 ms (pNN50) [Task Force 1 1996]. U radu iz 2003., autor [Hutchinson 2003] je proveo nelinearnu transformaciju mjera pNN20 i pNN50 te tvrdio da su rezultati autora [Mietus 2002] neuvjerljivi. Autor [Hutchinson 2003] ne daje prednost nijednoj od tih dvaju mjera. Prema današnjim saznanjima, ova rasprava nije okončana, tako da neki istraživači koriste pNN20, neki pNN50, a neki i druge vrijednosti [Asl 2008, Tsipouras 2004].

3.2.1.6. SDANN

SDANN (često i SENN [Tsipouras 2004]) je standardno odstupanje prosječnih vrijednosti otkucaja srca svih mjerenih segmenata. Uz pretpostavku da se značajka izlučuje iz *M* segmenata, izraz za SDANN je:

$$SDANN = \sqrt{\frac{\sum_{j=1}^{M}(\bar{x}_j - \bar{x}_{UK})^2}{M-1}} \tag{3.8}.$$

Iako nije nužno da je duljina trajanja svakog od *M* segmenata jednaka, smjernice upućuju na to da duljina segmenata bude jednaka i da iznosi 5 minuta. SDANN vrednuje dugo djelujuće komponente u nizu otkucaja srca [Task Force 1 1996].

3.2.1.7. Fanov faktor

Fanov faktor (engl. *Fano factor*) [Turcott 1996] je omjer varijance broja događaja unutar određenog vremena promatranja i srednje vrijednosti broja događaja u tom

vremenu. To je u stvari mjera korelacije na različitim vremenskim skalama. Mjera je poznata i kao indeks disperzije broja događaja. Ako se promatra segment zapisa srčanog ritma duljine *T*, tada se on podijeli na *k* dijelova, svaki iste duljine τ. Svaki od *k* dijelova sadrži N_i događaja (otkucaja srca). Fanov faktor je definiran kao [Teich 2001]:

$$Fano = \frac{\frac{1}{N-1}\sum_{i=1}^{k}(N_i - \overline{N})^2}{\overline{N}}, N = \sum_{i=1}^{k} N_i \quad (3.9).$$

3.2.2 Geometrijske značajke

3.2.2.1. HTI

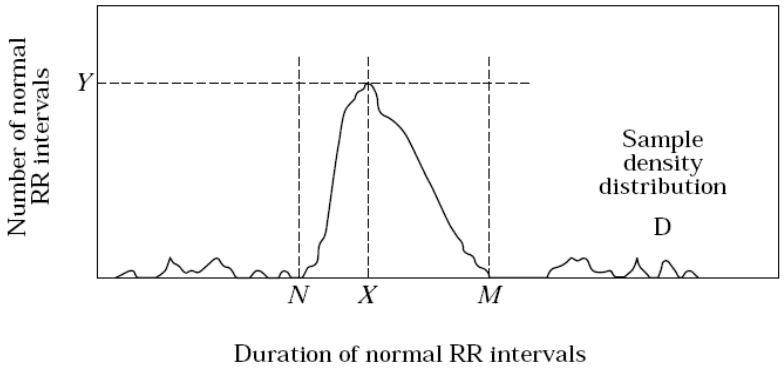
Trokutasti indeks HRV (engl. HRV *triangular index*, kraće: HTI) je geometrijska mjera varijabilnosti srčanog ritma. HTI je omjer ukupnog broja RR-intervala i maksimuma histograma distribucije trajanja RR-intervala. Uz ograničenje na diskretnu skalu, HTI se aproksimira s vrijednosti:

$$HTI = \frac{N}{\max_i n_i(l)} \quad (3.10),$$

pri čemu je *l* širina košarice u histogramu, a $n_i(l)$ frekvencija pojavljivanja duljine RR-intervala u *i*-toj košarici. Širina košarice najčešće je određena frekvencijom uzorkovanja. Ako je frekvencija uzorkovanja različita od 128 Hz, to treba biti navedeno [Task Force 1 1996]. HTI se određuje na relativno dugoj vremenskoj skali, obično većoj od 20 min, budući da je nužno da histogram bude dobro popunjen [Task Force 1 1996].

3.2.2.2. TINN

Trokutna interpolacija RR-intervala (kraće: TINN) je geometrijska značajka varijabilnosti srčanog ritma. TINN je širina bazne linija u histogramu na mjestu koje okružuje uspon prema maksimumu histograma. Budući da u histogramu na tom mjestu postoji trokutasti oblik, TINN je zapravo širina doljnje stranice trokuta i iznosi $M - N$, slika 3.16 [Task Force 1 1996].



Slika 3.16. Histogram RR-intervala koji se koristi za određivanje geometrijskih mjera.

Praktično, TINN se dobiva tako da se otkrije najveća vrijednost u histogramu te se nađe prvu košaricu za koju vrijedi $n_i(l) = 0$ slijeva, što daje vrijednost *M*, kao i prvu košaricu za koju vrijedi $n_i(l) = 0$ zdesna, što daje vrijednost *N*. Slično kao i HTI, TINN iskazuje ukupnu varijabilnost ritma te joj stoga odgovara ukupna snaga spektra. Visoko je koreliran sa značajkom SDNN, no mnogo je manje osjetljiv na ektopične otkucaje i šum zbog toga što uzima u obzir samo središnji dio razdiobe [Acharya 2006].

3.3 Linearne frekvencijske značajke

3.3.1 Frekvencijska analiza

Frekvencijska analiza koristi frekvencijsku domenu vremenskog niza RR-intervala da bi odredila bitna svojstva signala. Izračunava se spektralna gustoća snage u frekvencijskoj domeni (engl. *power spectral density*, kraće: PSD) i iz nje se izlučuju karakteristične komponente koje su pojavljuju u vremenskom nizu srčanog ritma. Za razliku od postupaka u vremenskoj domeni, koji su računarno nezahtjevni, frekvencijski postupci su nešto intenzivniji. Značajke u vremenskoj domeni imaju ozbiljan nedostatak da ne mogu razlikovati utjecaj između simpatičkih i parasimpatičkih komponenti autonomnog živčanog sustava (ANS). U frekvencijskoj domeni liječnici su uočili da pri djelovanju simpatičke komponente ANS-a raste snaga u području niskih frekvencija, dok pri djelovanju parasimpatičke komponente dolazi do porasta snage u području visokih frekvencija [Acharya 2006]. Razlike su zamjetne između odmora i aktivnosti pacijenata, pa čak i između ležanja u podignutom položaju i ležanja u spuštenom položaju [Task Force 1 1996].

Frekvencijske značajke imaju važnu ulogu i u proučavanju posljedica akutnog infarkta miokarda (engl. *acute myocardial infarction*, kraće: AMI) [Chattipakorn 2007]. Naime, pojačana simpatička aktivnost i smanjena parasimpatička aktivnost su pronađene kod pacijenata koji su pretrpili AMI. Simpatička aktivnost smanjuje fibrilacijski prag čime povećava predispoziciju za iznenadnu srčanu smrt (engl. *sudden cardiac death*, kraće: SCD) zbog ventrikularne fibrilacije (VF). Vagalna (parasimpatička) aktivnost pak povećava fibrilacijski prag i čini se da štiti od tahiaritmija ventrikula [Acharya 2006].

Procjena spektralne gustoće snage može se provesti neparametarskim ili parametarskim postupcima, s time da svaki od pristupa ima svoje prednosti i nedostatke. Neparametarski postupci su jednostavniji za implementaciju i daju brže rješenje. Parametarski postupci daju glađi spektar i omogućuju lakše izdvajanje i izračunavanje spektralnih komponenti. Također, parametarski postupci su precizniji, stoga se mogu koristiti na nizu s manjim brojem točaka. Nedostatak parametarskih postupaka je učinkovitost procjene odgovarajućeg reda modela za određeni vremenski niz. Pretpostavka i jednih i drugih postupaka je stacionarnost signala i ne postojanje značajnih nelinearnosti tijekom analize promatranog segmenta [Task Force 1 1996].

3.3.2 Neparametarski postupci frekvencijske analize

Najčešći neparametarski postupak koji se koristi je diskretna brza Fourierova transformacije (engl. *fast Fourier transform*, kraće: FFT). Pretpostavka FFT-a je da se vremenski niz sastoji od 2^n točaka, $n \in N$. U slučaju da ta pretpostavka ne vrijedi, tada je potrebno u postupku predprocesiranja „podrezati" niz na takvu duljinu.

Pseudokod algoritma za određivanja spektra gustoće snage putem FFT-a je ovakav:

1. Podreži niz na dužinu 2^n točaka, $n \in N$.
2. Generiraj niz kompleksnih brojeva čije su realne komponente vrijednosti podrezanog niza. Dobiva se niz x_n, $n = 0, \ldots, N-1$
3. Izračunaj FFT kompleksnog niza pomoću izraza:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk}, k = 0, \ldots, N-1 \tag{3.11}$$

   koristeći brz algoritam, npr. Cooley-Tukey [Cooley 1965]. Dužina niza u frekvencijskoj domeni jednaka je onoj u vremenskoj domeni.
4. Vrijednost spektralne gustoće snage (periodogram) na određenoj frekvenciji odredi izrazom:

$$PSD(f_k) = \frac{|X_k|^2 + |X_{N-k}|^2}{N}, k = 1, 2, \ldots, \left(\frac{N}{2} - 1\right),$$

$$PSD(f_0) = \frac{|X_0|^2}{N}, \quad PSD(f_c) = PSD(f_{\frac{N}{2}}) = \frac{\left|X_{\frac{N}{2}}\right|^2}{N} \tag{3.12},$$

   pri čemu gustoću snage ima smisla izračunavati samo do frekvencije $f_c = f_{\frac{N}{2}}$.

Frekvencije se određuju pomoću izraza:

$$f(k) = \frac{k}{N \cdot \Delta} \text{ [Hz]}, k = 0, 1, \ldots, \frac{N}{2}, \tag{3.13},$$

pri čemu je korak Δ parametar kojim se određuje prozor promatranja u frekvencijskoj domeni [Press 1992].

Iako jednostavna i brza za izračunati, ovakva procjena PSD pati od problema spektralnog curenja (engl. *spectral leakage*) nastalog zbog toga što je niza konačan, pa se dijelovi snage prenose (cure) u najbliže susjedne komponente snage. Time mogu nastati nepravilnosti i nepreciznosti prilikom preciznije procjene. Za problem curenja rješenje predstavljaju funkcije prozora.

Ideja je da se prije FFT vrijednosti točaka pomnože s određenim težinskim faktorima koji bi uzrokovali što je moguće manji gubitak gustoće snage nastalog zbog curenja. U analizi signala varijabilnosti rada srca obično se koristi Hamming (često: Hanning) [Teich 2001] i Hannov [Task Force 1 1996] prozor.

Hammingov prozor zadan je funkcijom:

$$w(n) = 0.56 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), n = 0, \ldots, N-1 \quad (3.14),$$

dok je Hannov prozor određen funkcijom:

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), n = 0, \ldots, N-1 \quad (3.15).$$

U slučaju korištenja funkcije prozora niza RR-intervala $\{x_n\}$ iz koraka 2 pomnoži se za svaku točku s odgovarajućom težinom zadanom funkcijom prozora $w(n)$ i tako izračunata ulazi u korak 3. Za najbržu procjenu PSD-a funkcija prozora se izostavlja.

3.3.3 Parametarski postupci frekvencijske analize

Parametarski postupci primjenjuju se na kraćim nizovima srčanih otkucaja budući da daju bolje izglađeni i precizniji spektar u odnosu na neparametarske postupke. Najčešći parametarski postupak kojim se nastoji procijeniti spektar gustoće snage je autoregresijski model (engl. *autoregressive model*, kraće: AR-model). Osnovna pretpostavka AR-modela je ta da se trenutačna vrijednost duljine RR-intervala može dobiti kao težinska suma *p* prethodnih vrijednosti duljine RR-intervala. Pri tome je *p* parametar koji se naziva red AR-modela. Za AR-model vrijedi izraz:

$$x_n = -\sum_{k=1}^{p} a_k x_{n-k} + \varepsilon_n \quad (3.16),$$

pri čemu su $\{a_k\}$ koeficijenti autoregresijskog modela, a ε_n je bijeli šum. Red AR-modela govori dosta o složenosti samog niza. Autori [Rezek 1998] smatraju da je red AR-modela zasebna mjera složenosti, kojom se može opisati je li vremenski niz slučajan ili određen, te je u tom smislu uspoređuju sa značajkama spektralne entropije i približne entropije. Pokazuje se na temelju usporedbi različitih informacijskih kriterija kao što su Akaikeov informacijski kriterij (engl. *Akaike information criterion*, kraće: AIC) i testova kao što su test najboljeg reda (engl. *optimal order test*, kraće: OOT), da je za red AR-modela kod kraćih nizova HRV signala potrebno biti najmanje $p = 16$ [Boardman 2002]. Tako visok red ujedno znači da je niz varijabilnosti ritma prilično slučajan, neponavljajući i nepredvidiv proces (atmosferski procesi imaju $p > 20$).

Pri procjeni koeficijenata AR-modela najčešće se koristi Burgov algoritam [Task Force 1 1996, Ge 2002]. Burgov algoritam je iterativan. Koeficijenti a_k modela postaju koeficijenti refleksije k_k i procjenjuje ih se izravno koristeći rekurzivni postupak. U svakom koraku rekurzije, procjenjuje se po jedan koeficijent refleksije. Da bi se procijenio sljedeći koeficijent, svi koeficijenti prethodno izračunati drže se fiksnima. Novi koeficijent se određuje tako da se minimizira suma prednjih i stražnjih ostataka (engl. *forward and backward residuals*) [de Waele 2000]. Suma ostataka je minimizirana za $k_{\min}$ koji iznosi:

$$k_{\min} = \hat{k}_k = -\frac{2\sum_{i=1}^{k} v_i^{[k]} w_i^{[k]}}{N(\left\|v^{[k]}\right\|^2 + \left\|w^{[k]}\right\|^2)} \tag{3.17},$$

pri čemu su *v* i *w* vektori prednjih, odnosno stražnjih ostataka dani izrazima:

$$\begin{aligned} v^{[k]} &= \left(f^{[k-1]}(k+1) \cdots f^{[k-1]}(N)\right), \\ w^{[k]} &= \left(b^{[k-1]}(1) \cdots b^{[k-1]}(N-k)\right) \end{aligned} \tag{3.18},$$

a prednji i stražnji ostatci su određeni s:

$$\begin{aligned} f^{[k]}(n) &= x(n) + \sum_{i=1}^{k} k_i x(n-i)\ ,\ n = k+1, \ldots, N \\ b^{[k]}(n) &= x(n) + \sum_{i=1}^{k} k_i x(n+i)\ ,\ n = 1, \ldots, N-k \end{aligned} \tag{3.19}.$$

Nakon što su dobiveni refleksijski koeficijenti, procjena spektralne gustoće snage može se izračunati korištenjem izraza [Bos 2002]:

$$PSD(f) = \frac{\sigma_\varepsilon^2}{\left|1 + k_1 e^{-i2\pi f} + \cdots + k_p e^{-ip2\pi f}\right|^2} \tag{3.20},$$

pri čemu je σ_ε^2 varijanca šuma koju se određuje na temelju izraza 3.16.

3.3.4 Spektralne komponente

Bez obzira na korišteni postupak procjene spektralne gustoće snage, u medicinskoj primjeni za analizu HRV promatraju se iste komponente spektra. Najčešće promatrane komponente navedene su u tablici 3.2, zajedno s frekvencijskim pojasevima u kojima su izlučene kod pojedinih istraživanja. U zadnjem stupcu navedene su linearne vremenske značajke koje približno odgovaraju dotičnoj frekvencijskoj značajki.

Od navedenih značajki, jedino LF i HF imaju jasno tumačenje budući da odgovaraju djelovanju simpatičkog i parasimpatičkog dijela ANS-a. Izražena vrijednost HF je također povezana s mehanizmom disanja i ocrtava respiratornu sinusnu aritmiju (RSA). Na dužim periodima (24 sata), varijacije u odnosu LF i HF-a ukazuju na cirkadijalni ritam, pri čemu je danju jača komponenta LF, a noću HF. Ne postoje slaganja vezana uz fiziološku interpretaciju VLF i ULF spektralnih komponenata. I VLF i ULF komponente ima smisla izračunavati samo za duge segmente, VLF za više od 5 minuta, a ULF samo za 24-satne snimke. Omjer LF/HF je visok za CHB i I/D kardiomiopatiju budući da su varijacije u trajanju RR-intervala male. Omjer je veći kod LBBB, AF, VF i SSS u usporedbi s normalnim ritmom [Acharya 2006].

Primjer spektra procijenjenog neparametarskim i parametarskim postupkom dan je na slici 3.17, prilagođeno iz [Task Force 1 1996]. Smjernice navode da je vrlo važno naglasiti koji se postupak koristio te koje su značajke određene iz spektra. Također, ako se koristio neparametarski postupak potrebno je navesti da li se koristila neka funkcija prozora, a ako se koristio parametarski postupak potrebno je navesti barem red modela, a poželjno i zašto smo se odlučili za taj red. Spektralna gustoća snage niza srčanih

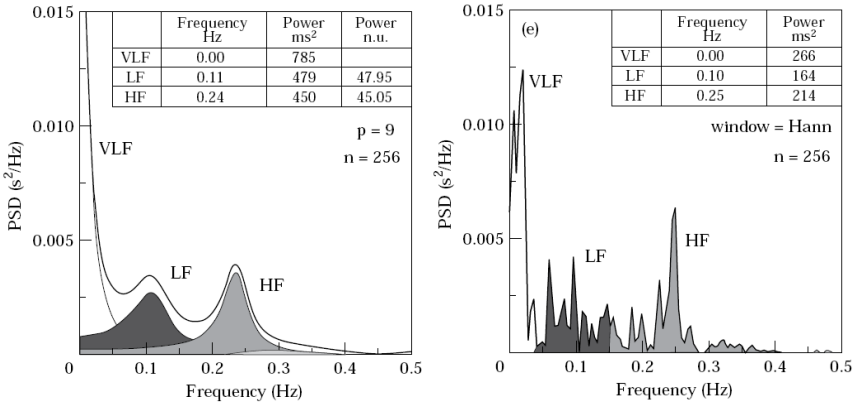
otkucaja izražava se u jedinicama ms^2 (milisekunde na kvadrat). Posebno, LF i HF komponente mogu se izraziti i u bezdimenzijskim normiranim jedinicama (n.u.). To je relativna vrijednost spektralne komponente u odnosu na ukupnu snagu spektra minus komponenta VLF, npr.:

$\text{HF(norm)} = \frac{\text{HF}}{\text{Ukupan PSD} - \text{VLF}}$ [n.u.]. Značenje normiranja je razumljivo s fiziološke strane, budući da jedino LF i HF komponente imaju jasno fiziološko tumačenje kroz ANS.

U okviru nekih novijih istraživanja spektra niza HRV, autori koriste i analiza putem spektralnih značajki višeg reda (engl. *higher order spectra*, kraće: HOS) [Acharya

Tablica 3.2. Linearne frekvencijske značajke varijabilnosti srčanog ritma.

Značajka	Frekvencijski pojas, Hz	Odgovarajuća značajka u vremenskoj domeni
Ukupan PSD	≤0.4 [Task Force 1 1996] 0.001-0.5 [Ho 1997]	SDNN, HTI, TINN
ULF	0.0001-0.003 [Task Force 1 1996]	SDANN
VLF	0.003-0.04 [Task Force 1 1996] 0.001-0.01 [Ho 1997] 0-0.0625 [Gamero 2002] 0-0.07 [Vallverdú 1999]	-
LF	0.04-0.15 [Task Force 1 1996] 0.01-0.15 [Ho 1997] 0-0.07 [Vallverdú 1999] 0.07-0.15 [Gamero 2002]	-
HF	0.15-0.4 [Task Force 1 1996] 0.15-0.5 [Ho 1997] 0.15-0.4 [Vallverdú 1999] 0.125-0.5 [Gamero 2002]	SDSD, RMSSD, pNN50
Omjer LF/HF	LF: 0.04-0.15, HF: 0.15-0.4 [Task Force 1 1996, Asl 2008, Yaghouby 2009]	-



(3a) (3b)

Slika 3.17. Frekvencijska spektralna analiza niza srčanog ritma: (3a) spektar dobiven Burgovim postupkom reda 9, (3b) periodogram dobiven postupkom FFT.

2006]. Od značajnijih mjera potrebno je spomenuti bispektralnu analizu. Bispektar je spektar u dvije frekvencijske dimenzije koji nastaje brzom Fourierovom transformacijom korelacije trećeg reda nekog signala i može ga se prikazati formulom:

$$B(f_1, f_2) = E\left(X_n(f_1)X_n(f_2)X_n*(f_1 + f_2)\right) \quad (3.21),$$

pri čemu je X_n FFT vremenskog niza, $E(\cdot)$ je očekivanje (u praksi srednja vrijednost). Kod bispektra promatraju se značajke srednje magnitude, fazne entropije, spektralne invarijante i druge. Ovi postupci zasad nisu standardni i nemaju neku poznatu fiziološku interpretaciju, no daju vrlo dobre rezultate [Chua 2009].

3.4 Vremensko-frekvencijske značajke

Vremensko-frekvencijske značajke dobivaju se u novije vrijeme uglavnom putem transformacije valićima (engl. *wavelet transform*, kraće: WT) [Vetterli 1992]. Od ostalih postupaka za vremensko-frekvencijsku analizu treba spomenuti Wigner-Villeovu razdiobu [Clariá 2008] i Hilbert-Huangovu transformaciju koja koristi empirijsko razlaganje signala (engl. *empirical mode decomposition*) [Huang 1998]. Prije pojave WT za procjenu vremensko-frekvencijskih značajki koristila se Fourierova transformacija za kratke segmente (STFT) [Portnoff 1980].

STFT ima za cilj lokalizaciju određenog događaja u vremenskom nizu. Za lokalizaciju se koristi prozor konačne širine koji se pomiče u vremenu. Unutar tog prozora provodi se analiza spektralnih komponenti signala. Ovakva analiza dobro funkcionira ako je signal stacionaran i sporo promjenjiv. Međutim, u nizu HRV postoje relativno česte nestacionarnosti i u tom slučaju analiza putem STFT ne uspijeva pronaći zadovoljavajuće rješenje. Ako bi se koristio prozor beskonačne širine (ili širine kao čitav signal), dobila bi se Fourierova transformacija. Da bi dobili određenu stacionarnost u signalu, prozor treba biti što uži. Ako se uzme uzak prozor, dobije se bolja vremenska rezolucija i bolja pretpostavka za postojanje stacionarnosti, međutim spektralna rezolucija je tada preslaba. WT koristi male prozore i visoku frekvenciju kao i dulje prozore i nisku frekvenciju. WT se može podijeliti u dva tipa: diskretnu (engl. *discrete wavelet transform*, kraće: DWT) i kontinuiranu (engl. *continuous wavelet transform*, kraće: CWT). Iako se obje koriste u analizi srčanog ritma, značajnija je DWT pa će se ona dalje opisati.

3.4.1 Diskretna analiza valićima

Valić je određen trajanjem i konačnom energijom i izgrađen je tako da korelira sa signalom kako bi ga što bolje opisao. Prilikom korelacije određuju se koeficijenti valića koji doprinose najboljem opisu u odnosu na signal. Uobičajena definicija DWT-a dana je izrazom [Thurner 1998]:

$$W_{m,n}(x) = 2^{-m/2} \sum_{i=1}^{N} x_i \psi(2^{-m} i - n) \tag{3.22},$$

pri čemu je *m* faktor skaliranja, a *n* faktor posmaka. Valić $\psi(t)$ koji se koristi za koreliranje s vremenskom nizom naziva se osnovni valić (engl. *mother wavelet*). $W_{m,n}$ je dvodimenzionalni obojeni spektar (ili se može promatrati kao trodimenzionalni spektar u kojem je boja-energija treća koordinata), kojemu je na osi *x* redni broj RR-intervala, a na osi *y* faktor skaliranja valića. Najčešće se koriste dijadička DWT, pri kojoj je faktor skaliranja *m* potencija broja dva [Teich 2001]. Kao osnovni valić u diskretnom slučaju upotrebljava se jednostavni Haarov valić dan izrazom [Teich 1998]:

$$\psi(x) = \begin{cases} 1, x \in [0,0.5] \\ -1, x \in (0.5,1] \\ 0, \text{drugdje} \end{cases} \tag{3.23}.$$

Značajka koja se izlučuje iz DWT je standardno odstupanje WT, koja se izračunava posebno za svaku skalu pomoću izraza [Teich 2001]:

$$\sigma_{Wav}(m) = \sqrt{\frac{1}{N-1} \sum_{n=0}^{N-1} \left(W_{m,n}(x) - \overline{W}_{m,n}(x) \right)^2} \tag{3.24},$$

pri čemu se *n* kreće po čitavom rasponu RR-intervala. Može se uočiti, da ako se uzme skala $m = 1$, tada se dobiva već poznata vremenska statistička značajka RMSSD (usporediti s izrazom (3.5)). Standardno odstupanje WT pokazuje se vrlo dobrom u raspoznavanju srčanih bolesti na dužim zapisima pacijenata. Tako [Thurner 1998] koristi četvrtu i petu potenciju ($m = 16$, odnosno $m = 32$ srčana otkucaja) da bi sa 100% točnosti razvrstao zdrave pacijente od pacijenata oboljelih od CHF, dok [Teich 1998] uz visoku točnost razlikuje zdrave pacijente od pacijenata oboljelih od CHF i one kojima je transplantirano srce. Problem kod obje studije je taj što se za točno razlikovanje koristio vrlo velik broj RR-intervala, oko 20 000. Autor [Teich 1998] tvrdi da se za manje od 3500 točaka bolji rezultati dobivaju s mjerom SDNN nego s $\sigma_{Wav}(m)$.

3.5 Nelinearne značajke

Nelinearnost srčanog ritma posljedica je stalnog međudjelovanja simpatičkog i parasimpatičkog sustava, koji reguliraju njegov rad. Srce radi pod izravnom upravom sustava u mozgu, sustava čiji regulacijski mehanizam ovisi o nizu faktora kao što su: disanje, toplina, starost, ishrana, utreniranost, itd. Zbog fizioloških međudjelovanja raznih faktora, dinamika srčanog ritma je izuzetno složena. Problem istraživačima srčanog ritma predstavljaju nestacionarnosti u signalu nastale zbog vanjskih faktora koji djeluju na regulacijski mehanizam ANS-a, a koje se opiru smislenoj analizi, dijagnostičkim procjenama i prognostici. Autor [Xia 2009] navodi da je široko prihvaćena činjenica da je srčani ritam pun nestacionarnosti i nelinearnosti raznih vrsta. Jedna od pomno razrađivanih tema pri analizi srčanog ritma je problem slučajnosti u srčanom ritmu. I dok neki istraživači tvrde da je srčani ritam zdravog srca i većine

srčanih oboljenja gotovo isključivo nelinearan slučajan (stohastičan) proces [Teich 2001], drugi tvrde da u signalu postoji skrivena nelinearna određenost koja daje određenu predvidljivost (determinizam) [Goldeberger 1996].

Otkriće multifraktalnosti u srčanom ritmu objavljeno u časopisu Nature [Ivanov 1999] navodi na djelovanje složenih, nelinearnih procesa u pozadini ritma. Autori [Gutiérrez 2005] proveli su iscrpnu analizu po pitanju određenosti ili slučajnosti u EKG-u te su došli do zaključka da i zdrava i patološka informacija mogu biti slučajne ili određene, može ih se identificirati različitim mjerama i može ih se naći u raznim dijelovima EKG-a. Činjenica je da složenost dinamike srčanog ritma, ali i drugih bioloških ritmova opada s bolesti i sa starosti [Goldberger 1996, Goldberger 2002, Acharya 2004 (2), Hornero 2006, Peng C. 2009], no to ne govori nužno da je na djelu nelinearni determinizam. Istraživanja provedena nedavno na temu postoji li kaos u srčanom ritmu dovela su do miješanih rezultata [Glass 2009]. Glavni zaključak više studija je taj da u zdravom srčanom ritmu postoji multifraktalnost koja degradira s bolesti. Također, pitanje kaosa nije riješeno budući da se kaos uglavnom nije mogao otkriti, no ne u svim slučajevima. Ipak, istraživači zaključuju da samo pitanje postojanja kaosa nije niti bitno, već je mnogo bitnije postoji li dijagnostički potencijal nelinearnih postupaka, na što odgovaraju potvrdno. U ovom dijelu disertacije navest će se ukratko teorija nelinearnih sustava i determinističkog kaosa potrebna za razumijevanje primjene na analizu BVN i opis nelinearnih značajki.

3.5.1 Teorijska pozadina nelinearnosti

Vremenski niz srčanog ritma može se promatrati iz perspektive teorije sustava. Sustavi se u pravilu opisuju putem varijabli stanja. Tako je sustav niza intervala srčanih otkucaja takav sustav u kojem je varijabla stanja intervalno vrijeme koje protječe između dva slijedna otkucaja. Sustav u kojem za varijablu stanja x vrijedi da je njezinu vrijednost x_n moguće izračunati na bilo koji način putem jedne ili više prošlih vrijednosti, naziva se određeni (deterministički) sustav. Sustav u kojem to nije moguće naziva se slučajan (stohastičan) sustav. Prema linearnosti, sustavi se dijele na linearne i nelinearne. Linearan sustav je takav sustav u kojem za svaku varijablu stanja x vrijedi $x_n = a_1 x_{n-1} + \ldots + a_d x_{n-d} + a_{d+1}$, $a_1 \ldots a_{d+1}$ su konstante, d je red ili dimenzionalnost varijable sustava. Nelinearan sustav je takav sustav u kojem veza između trenutne vrijednosti varijable stanja i prethodnih vrijednosti nije linearna, već je dana s: $x_n = f(x_{n-1})$.

Prema stupnju nelinearnosti, sustavi (i njihove varijable stanja) mogu se podijeliti u četiri skupine: linearne, periodične, kvaziperiodične i kaotične. Linearni sustavi imaju poželjno ponašanje. Kod takvih sustava stupanj odziva je proporcionalan snazi stimulansa na ulazu u sustav. Linearne sustave moguće je raščlaniti na djeliće i predvidjeti njegovo ponašanje na temelju djelića budući da u njima nema anomalija. Slični linearnim sustavima su periodični sustavi, u kojem je ponašanje, iako zadano

nelinearnom funkcijom (npr. sinusom) uglavnom predvidljivo. Međutim, u kvaziperiodičnim i kaotičnim nelinearnim sustavima proporcionalnost odziva više ne vrijedi, a ponašanje može biti nepredvidljivo i neočekivano. Takve sustave nije moguće rastaviti na dijelove, proučiti ponašanje dijelova, i očekivati da će se sustav ponašati kao skup dijelova. Biološki sustavi, pa tako i ritam rada srca, mogu pokazivati sva četiri stupnja nelinearnosti, od linearnosti, periodičnosti, kvaziperiodičnosti i kaotičnosti u nekim slučajevima, budući da se interakcije u srčanim stanicama kao i u stanicama mozga koji kontrolira živčani sustav događaju u sprezi, s mogućim naglim promjenama. Stvarni niz srčanih otkucaja je često negdje između kategorija nelinearnosti, a dinamika je maskirana nizom procesa koji se modeliraju kao slučajni iz perspektive srčanog ritma.

Za proučavanje vrste i stupnja nelinearnosti koristi se postupak rekonstrukcije faznog prostora, poznat iz teorije kaosa kao Poincaréov crtež (engl. *Poincaré plot*). Iako je uvriježeni naziv za fazni prostor ujedno i prostor stanja, ovdje se taj naziv neće koristiti budući da se proučava samo jedna varijabla stanja – srčani ritam. Fazni prostor prikazuje ponašanje varijable stanja unutar prostora određenog reda ili dimenzije. Sastoji se od statičkog prikaza samih vrijednosti varijable stanja te dinamičkog prikaza putanje prelaska iz jedne vrijednosti u drugu.

U slučaju srčanog ritma, svaka od *d* osi faznog prostora prati vrijednosti nekog RR-intervala. Formalno, točka u faznom prostoru ima koordinate:

$$\bar{x}_n = \left(x_n \; x_{n+\tau} \cdots x_{n+(d-1)\tau}\right) \tag{3.25},$$

pri čemu je *d* rekonstrukcijska ili ugradbena dimenzija faznog prostora, x_n je duljina *n*-tog RR-intervala, a τ je odgoda ili pomak u faznom prostoru (engl. *delay*, *lag*). Odgoda je u slučaju diskretnog niza srčanih otkucaja određena kao onaj broj RR-intervala koji se preskače za određivanje vrijednosti prve sljedeće koordinate u faznom prostoru. Pri analizi čitavog EKG-a odgoda se zadaje kao vremenski pomak (u sekundama) između promatranih vrijednosti. Optimalnu vrijednost odgode za sustav određene ugradbene dimenzije moguće je odrediti promatranjem smanjenja autokorelacije vremenskog niza na određeni iznos pri variranju duljine odgode [Pritchard 1995] ili korištenjem mjere minimuma zajedničke informacije, više u [Fraser 1989]. Autori [Faust 2004] navode da eksperimentalno utvrđena najbolja vrijednost odgode za niz srčanih otkucaja iznosi $\tau = 5$.

Za prikaz dinamike sustava definira se putanja ili trajektorija (engl. *trajectory*) kao staza između niza točaka u faznom prostoru:

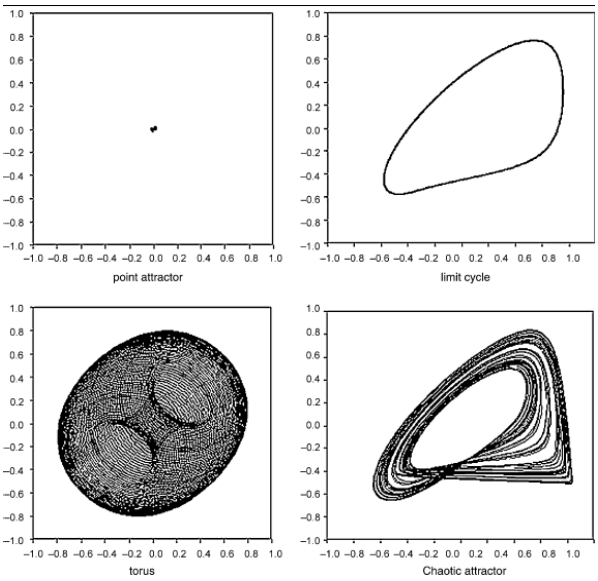
$$\left\{\bar{x}_n\right\} = \left(\bar{x}_n, \bar{x}_{n+1}, \cdots, \bar{x}_{n-L-1}\right), \tag{3.26},$$

pri čemu je *L* duljina putanje. Ako postoji stabilna putanja koja je koncentrirana na nekom dijelu faznog prostora, onda se na tom mjestu govori o atraktoru. Postojanje atraktora može se procijeniti vizualno. Formalno, atraktor je ograničena regija u faznom prostoru za koju se sve dovoljno bliske trajektorije iz slijeva atrakcije privlače asimptotski dovoljno dugo.

Neki od mogućih atraktora u faznom prostoru ugradbene dimenzije $d = 2$ prikazani su na slici 3.18. Točkasti atraktor (engl. *point attractor*) ukazuje na nepromjenjivost u trajanju RR-intervala. Interval je stalno iste duljine i nema primjetne dinamike. Točkasti atraktor je primjer linearnog ponašanja srčanog ritma prisutnog kod naprednog razvoja nekih srčanih bolesti kao što su CHF i I/D kardiomiopatija. U stvarnosti, atraktor nije točkast, već kružnog oblika s vrlo malim radijusom. Periodična dinamika prikazana je putem ograničenog ciklusa (engl. *limit cycle*). Tu se duljina RR-intervala periodički mijenja tijekom ciklusa određene duljine. Neko vrijeme ritam je brži, neko vrijeme je sporiji. Promjene nisu nagle, a ciklus je uvijek isti. Zdravi ritam može prikazivati u nekim situacijama periodičnu dinamiku, iako je takva izražena dinamika češća kod sindroma bolesnog sinusnog čvora, sa stalnim izmjena bržeg i polakšeg ritma. Kvaziperiodična dinamika može se također uočiti na slici 3.18. Karakterističan oblik točaka je zavojnica (torus) ili prsten. U ovom slučaju postoji više prisutnih oscilacija koje se dinamički izmijenjuju. Kvaziperiodična dinamika ima potencijal za prijeći i u periodičnu dinamiku i u točkastu dinamiku, ali i u kaos, kao što je to dokazano za slučaj prijelaza iz ventrikularne tahikardije u ventrikularnu fibrilaciju [Weiss 1999].

3.5.2 Kaos i slučajnost

Kolokvijalna upotreba riječi kaos (engl. *chaos*, od grčkog: χάος) sugerira nepredvidljivo i slučajno ponašanje, često s katastrofalnim posljedicama. Kaos se naveliko proučava još od 40-tih godina 20. stoljeća i dugo se vremena smatrao dubokoumnim problemom (engl. *deep problem*), što znači da je naslovljen kao problem najvećeg stupnja složenosti s kojim se čovječanstvo može susresti [Gleick 1987]. Kaos je samo na prvi pogled sinonim za slučajan proces. Nepravilan red koji se krije iza maske slučajnosti



Slika 3.18. Dinamika varijable stanja u faznom prostoru.

karakteriziraju specifična svojstva. Kaotični atraktor se razlikuje od svih ostalih oblika nelinearne dinamike po sljedećim obilježjima:

1. Aperiodičnost
2. Određenost
3. Ograničenost
4. Osjetljivost na početne uvjete

1. Aperiodičnost označava odsustvo perioda ponavljanja vrijednosti varijable stanja sustava. Tako u faznom prostoru ne postoji jedan ili nekoliko perioda, kao što je slučaj kod periodičkog ili kvaziperiodičkog ponašanja, perioda uopće nema.

2. Svojstvo određenosti je već objašnjeno. Konkretno, vrijednost varijable stanja moguće je odrediti na temelju prethodnih vrijednosti. Ovdje je potreban oprez, budući da iako je ponašanje varijable odredivo, ono je često teško predvidljivo. Vrijednost u budućem stanju je tek u teoriji moguće točno odrediti. U praksi je dovoljno da se vrijednost odredi barem približno točno, što nije moguće kod slučajnog ponašanja.

3. Ograničenost znači da vrijednosti neke varijable stanja uvijek ostaju unutar određene gornje i doljnje granice, tj. da vrijedi: $K < x_n < L, \forall n$, gdje su K i L neke konstante.

4. Osjetljivost na početne uvjete označava da trenutna vrijednost varijable stanja sustava izuzetno ovisi o uvjetima okoline koji su postojali u trenutku kad je sustav započeo s radom ili u trenutku početka mjerenja na sustavu.

Samo mala razlika u četvrtoj, petoj ili nekoj drugoj decimali početnog uvjeta za neku varijablu može dovesti do drastične promjene u putanji varijable i do pojave bifurkacija [Goldberger 1988]. Bifurkacije su značajne kod analize EKG-a budući da one uzrokuju alterniranje određenih obrazaca u signalu, npr. spojnice ST-T, a takvi alternansi se pojavljuju prije VF [Rosenbaum 1994].

Za razliku od nelinearnog određenog sustava u kojem postoji kaos, nelinearan slučajan sustav nema nikakav obrazac ponašanja kojim bi se otkrila sljedeća vrijednost. U nekim slučajevima i na nekim segmentima srčanog ritma teško je odrediti je li on slučajan ili određen. Također, teško je odrediti je li ritam ima linearan atraktor ili nelinearan. Razlog tome je šum u podacima, koji iako je umanjen kod niza HRV u odnosu na čitav EKG [Asl 2008], i dalje postoji, tako da nije lako razlikovati vrstu atraktora.

Da bi se razlikovalo linearno ponašanje od nelinearnog koristi se test za nelinearnost [Theiler 1992]. U tu svrhu generira se novi niz podataka koji je surogat izvornom nizu i to tako da mu se u frekvencijskoj domeni izmiješaju faze na slučajan način. Takav surogatni niz ima istu srednju vrijednost, varijancu, autokorelacijsku funkciju, i histogram kao izvorni niz, međutim fazni odnosi su mu uništeni. Time se surogatni niz ponaša kao Gaussov šum, što znači da je linearan i slučajan. Svojstva izvornog niza se nakon generiranja slučajnog spektra uspoređuju sa surogatnim nizom. Potrebno je ustanoviti postoji li razlika između ta dva niza. Ako nema značajne razlike, tada je izvorni niz linearan slučajan proces, a ako razlika postoji, tada je niz nelinearan, no ne nužno određen.

Da bi se odredilo je li nelinearan sustav slučajan ili određen, potrebno je odrediti postoji li takva ugradbena dimenzija d faznog prostora u kojoj bi kapacitivna (Haussdorfova) dimenzija atraktora D_0 dosegnula svoj maksimum i dalje ne bi rasla. Ako postoji plato dimenzije atraktora, tada se sustav ne ponaša slučajno.

Neka je izvršena podjela d – dimenzionalnog prostora u ćelije l_d. Vjerojatnost p_i nalaženja točke atraktora u ćeliji indeksa i ($i = 1, 2, \ldots, M(l)$) iznosi $p_i = \lim_{N\to\infty} \frac{N_i}{N}$, gdje je N_i broj točaka u i-toj ćeliji, a N ukupan broj točaka na atraktoru. Da bi se opisala nehomogena, statična struktura atraktora, uvodi se beskonačan skup dimenzija D_f koje se odnose prema f – toj potenciji vjerojatnosti p_i^f kao:

$$D_f = \lim_{l\to 0} \frac{1}{f-1} \frac{\log\left(\sum_{i=1}^{M(l)} p_i^f\right)}{\log l}; \quad f = 0, 1, 2, \ldots \tag{3.27},$$

pri čemu je l dimenzija ćelije, a M broj ćelija u prostoru. Za $f = 0$ dobiva se definicija kapacitivne, Hausdorffove dimenzije atraktora:

$$D = D_0 = -\lim_{l\to 0} \frac{\log M(l)}{\log l} \tag{3.28}.$$

Specijalan slučaj kada vrijedi $f = 2$ daje dobru procjenu za dimenziju atraktora. D_2 je nazvana korelacijskom dimenzijom. Dobiva se izraz:

$$D_2 = \lim_{l\to 0} \frac{\log\left(\sum_{i=1}^{M(l)} p_i^2\right)}{\log l} \tag{3.29}.$$

Može se pokazati da uvijek vrijedi $D_2 \le D$. Autori [Ding 1993] pokazali su da je dovoljan (ali ne i nužan uvjet) da se ugradbena dimenzija prostora d treba procijeniti pomoću izraza:

$$d = Ceil(D_2) \tag{3.30},$$

pri čemu je s Ceil(D_2) označena funkcija najmanjeg cijelog broja većeg od D_2. Nakon tako procijenjene dimenzije d korelacijska dimenzija D_2 doseći će plato i neće dalje značajnije rasti ako je sustav određen.

Praktični razlozi kao što su premali broj točaka u nizu i šum u podacima mogu zamaskirati ovaj teoretski prisutan plato.

S druge strane, autokorelacija vremenskog niza može uzrokovati to da se pronađe plato atraktora, a da je sam niz nelinearan slučajan sustav [Schiff 1992]. Autokorelacijski koeficijent niza RR-intervala može se procijeniti pomoću izraza [Teich 2001]:

$$R(x) = \frac{1}{(n-1)\sigma_x^2} \sum_{i=1}^{N-1} \left[x_i - \bar{x}\right]\left[x_{i+1} - \bar{x}\right] \tag{3.31}.$$

Pokazuje se da u prosjeku vrijedi da je $R > 0.8$ za većinu segmenata RR-intervala, što označava visok stupanj autokorelacije. Za uklanjanje utjecaja autokorelacije prilikom vrednovanja kaotičnog ponašanja predlaže se da se gleda niz slijednih razlika RR-

intervala umjesto samog niza RR-intervala, dakle niz: $\{x_2 - x_1, x_3 - x_2, \ldots, x_N - x_{N-1}\}$. Time se autokorelacija gotovo poništava, pa je moguća stvarna procjena kaotičnosti sustava.

Jednom kad je uspostavljeno da je srčani ritam nelinearan i određen, tada je opravdano opisati ritam izlučivanjem značajki koje upravo i služi za opis determinističkog ponašanja. U protivnom, značajke determinističkog kaosa nemaju opravdanu upotrebu, pa bi se za karakterizaciju ritma trebale koristiti samo značajke koje mogu opisivati nered u sustavu, kao što su razne vrste entropija.

3.5.3 Nelinearne značajke

U ovom dijelu bit će prikazana većina postupaka koji se danas koriste za nelinearan opis srčanog ritma. Neke od značajki koriste se za opis nelinearnog determinizma, odnosno kaosa, dok se druge koriste za opis slučajnosti i nereda u podacima. Postoje značajke koje su namijenjene opisu kraćih segmenata u nizu, dok druge zahtijevaju duži segment da bi davale pouzdane rezultate. Većina od ovdje navedenih značajki upotrebljive su i u analizi drugih vrsta BVN, pa i vremenskih nizova općenito.

Popis nelinearnih značajki koje će se u ovom dijelu rada obraditi dan je u tablici 3.3. Naveden je naziv značajke, tip značajke, naziv na engleskom jeziku, te da li se koristi na

Tablica 3.3. Pregled nelinearnih značajki u analizi varijabilnosti srčanog ritma.

Naziv značajke	Tip značajke	Engleski naziv	Duljina vremenskog niza
Ljapunovljevi eksponenti	Fazni prostor, kaos	*Lyapunov exponents*	srednja/duga
Korelacijska dimenzija	Fazni prostor, kaos	*Correlation dimension*	srednja/duga
Standardne devijacije u faznom prostoru	Fazni prostor	*Phase space standard deviations*	sve duljine
Fraktalna dimenzija	Fraktalna	*Fractal dimension*	srednja/duga
Hurstov eksponent	Fraktalna	*Hurst exponent*	srednja/duga
Analiza kolebanja RR-intervala s uklonjenim trendom	Fraktalna	*Detrended fluctuation analysis*	srednja/duga
Približna entropija	Entropijska	*Approximate entropy*	kratka/srednja
Entropija uzorka	Entropijska	*Sample entropy*	kratka/srednja
Shannonova entropija i ispravljena uvjetna entropija	Entropijska	*Shannon entropy and corrected conditional entropy*	kratka/srednja
Entropija na više skala	Entropijska	*Multiscale entropy*	ovisno o skali, srednja ili duga
Spektralna entropija	Entropijska	*Spectral entropy*	srednja/duga
Rényijeva entropija	Entropijska	*Rényi entropy*	srednja/duga
Kolmogorovljeva entropija	Entropijska	*Kolmogorov entropy*	duga
Indeks prostorne popunjenosti	Fazni prostor	*Spatial filling index*	sve duljine
Allanov faktor	Nel. statistička	*Allan factor*	srednja/duga
Indeks asimetrije na više skala	Ostalo	*Multiscale asymmetry index*	srednja/duga
Mjera središnje težnje	Fazni prostor	*Central tendency measure*	sve duljine
Rekurentni crtež	Ostalo	*Recurrence plot*	srednja/duga
Mjera složenosti Lempel-Ziv	Ostalo	*Lempel-Ziv complexity*	kratka/srednja

kraćoj ili dužoj vremenskoj skali. Okvirno, kratka vremenska skala je ona do 100 otkucaja, srednja od 100-1000, a duga je ona duža od 1000 otkucaja.

3.5.3.1. Ljapunovljevi eksponenti

Ljapunovljevi koeficijenti ili eksponenti (engl. *Lyapunov exponent*, kraće: LE) mjere koliko brzo dvije početno bliske točke na putanji u faznom prostoru divergiraju jedna od druge tijekom evolucije sustava. Na taj način, LE daju korisnu informaciju o ovisnosti sustava o početnim uvjetima. Treba se podsjetiti da putanje kaotičnog sustava značajno divergiraju kako vrijeme prolazi, čak i za vrlo bliske točke. Ta nestabilnost jedna je od najboljih mjera koja govori o kaotičnosti procesa koji se promatra.

Ako postoji d – dimenzionalan fazni prostor, tada se mjeri udaljenost između točaka u svakom od d smjerova. Neka je početna udaljenost između dvije najbliže točke u smjeru i jednaka $\left\|\delta(x_{i1}(0), x_{i2}(0))\right\|$, a udaljenost tih istih točaka nakon protjeka vremena t neka iznosi $\left\|\delta(x_{i1}(t), x_{i2}(t))\right\|$. LE u smjeru i definira se kao prosječna stopa rasta λ_i početne udaljenosti [Ubeyli 2008]:

$$\lambda_i = \lim_{t\to\infty} \frac{1}{t} \log_2 \frac{\left\|\delta x_i(t)\right\|}{\left\|\delta x_i(0)\right\|} \tag{3.32}.$$

Postojanje pozitivnog LE ukazuje na kaos u vremenskom nizu [Abarbanel 1991]. Preciznije, bilo koji sustav koji sadrži barem jedan pozitivan LE je po definiciji kaotičan, pri čemu iznos eksponenta određuje vremensku skalu za predvidljivost [Zeng 1991]. Negativan LE označava da se putanja približava određenoj fiksnoj točki, dok LE koji iznosi 0 ili blizu 0 označava stabilni atraktor, u kojoj putanje zadržavaju svoj položaj [Acharya 2004 (1)]. Pozitivan najveći LE (engl. *largest Lyapunov exponent*, kraće: LLE) najsnažniji je pokazatelj da je sustav kaotičan.

Općenito promatrano, LE mogu se izlučiti iz analiziranog vremenskog niza na dva načina. Prvi je taj da se slijedi vremenska evolucija bliskih točaka u faznom prostoru. Ovaj način daje procjenu samo LLE, Drugi postupak je zasnovan na procjeni lokalnih Jakobijevih matrica i sposoban je procijeniti sve LE. Vektor svih LE za neki sustav naziva se i Ljapunovljev spektar [Übeyli 2008].

Većina autora u analizi srčanog ritma koristi izračun samo LLE, budući da je on dovoljan da odredi stupanj kaotičnosti sustava [Seker 2000, Owis 2002, Acharya 2004 (1), Asl 2008, Yaghouby 2009]. Ipak, [Seker 2000] primjećuje da je za pouzdan izračun potreban relativno dug vremenski niz, dok ga autori [Asl 2008] nesmisleno izračunavaju na segmentima od po samo 32 RR-intervala. Autor [Übeyli 2008] koristio se pak čitavim spektrom LE da bi razlikovao između zdravih osoba i osoba oboljelih od epilepsije, a na temelju čitavog EKG signala.

Najjednostavniji i najpoznatiji algoritam izračuna LLE ponudio je Wolf [Wolf 1985]. Algoritam se zasniva jednostavno na tome da se prati evolucija u udaljenosti dvaju točaka i to tako da se slučajno odabere jedna točka, odredi najbliži susjed te se prati putanja od početka pa kroz vrlo dug vremenski period. Ako udaljenost prijeđe neku

unaprijed zadanu granicu, traži se nova najbliža točka u tom trenutku i nastavlja se dalje s izračunom.

Postoji cijeli niz izmjena i nadogradnji osnovnog Wolfovog algoritma [Zeng 1991], od kojih su neki i novijeg datuma i koriste se za izračun čitavog spektra LE [Grond 2003]. Jedan od najkorisnijih algoritama za procjenu LLE predložili su autori [Rosenstien 1993], a u svojem radu ga koriste [Acharya 2004 (1)]. Prednosti ovog algoritma su sljedeće: omogućuje izračunavanje LLE i na srednje dugim vremenskim nizovima od 100-1000 točaka, brz je i jednostavan za implementaciju, otporan na šum, iznos odgode, te ugradbenu dimenziju. Dva postupka karakteriziraju ovaj algoritam:

1. Za svih M točaka u faznom prostoru odrede se najbliže susjedne točke. Prate se putanje od početka pa kroz određeni diskretan broj vremenskih koraka.

2. Kao najbliža susjedna točka uzima se u obzir samo ona točka koja je udaljena za više od prosječnog perioda uzorkovanja vremenskog niza. Time se izbjegava praćenje susjedne točke koja je unutar moguće pogreške uzorkovanja izvornog signala.

Konačan izraz za procjenu LLE dan je kao nagib pravca koji nastaje aproksimiranjem niza točaka funkcije:

$$y(n) = \frac{1}{\Delta t}\left\langle \ln d_i(n) \right\rangle \tag{3.33}$$

postupkom najmanjih kvadrata, pri čemu je Δt diskretan vremenski pomak opažanja na putanji, n je broj vremenskih pomaka, a s $\left\langle \ln d_i(n) \right\rangle$ je označena prosječna vrijednost (po i) konačnih udaljenosti svih točaka od svojih početno susjednih nakon n koraka [Acharya 2004 (1)]. Rosenstienova metoda je ujedno implementirana i u radnom okviru danom u ovoj disertaciji.

3.5.3.2. Korelacijska dimenzija

Korelacijska dimenzija D_2, kao dobra procjena Haussdorfove dimenzije atraktora jedna je od najčešće korištenih mjera za opis nelinearnosti sustava. Kao što je već prije navedeno, korelacijska dimenzija služi da se odredi je li nelinearan sustav slučajan ili određen. Grassberger i Procaccia praktično određuju D_2 iz korelacijskog integrala [Grassberger 1983]:

$$C(N,l) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Theta\left[l - \left|\bar{x}_i - \bar{x}_j\right|\right] \tag{3.34},$$

pri čemu je $\Theta[.]$ Heavisideova funkcija skoka: $\Theta(x) = \begin{cases} 1, & x > 0 \\ 0, & x \le 0 \end{cases}$, l je radijalna udaljenost oko referentne točke x_i u faznom prostoru. Nakon što se odredi $C(N,l)$, D_2 se određuje iz nagiba pravca na grafu $\log_2 C(N,l) / \log_2\left(\frac{l}{l_0}\right)$, pri čemu je l_0 proizvoljan (obično se uzima $l_0 = 1$). Nagib pravca procjenjuje se za male vrijednosti radijalne udaljenosti ($l \to 0$). Prilikom određivanja korelacijske dimenzije postoji problem toga što je reproducibilnost loša. Naime, za male radijalne udaljenosti l nagib pravca teško je

dovoljno točno procijeniti. Da bi se tome doskočilo, autori [Owis 2002] predložili su postupak za automatsko određivanje nagiba. Potrebno je izračunati drugu derivaciju funkcije $\log_2 C(N,l) / \log_2\left(\frac{l}{l_0}\right)$ te potražiti najdulji plato s vrijednostima ispod određenog praga (autori su uzeli prag od 0.1). Ako ima više takvih slično dugih platoa, uzima se onaj koji daje najvišu korelacijsku dimenziju.

O duljini niza koja bi se trebala koristiti za izračun korelacijske dimenzije raspravlja [Seker 2000]. U radu se navodi citat od autora [Kantz 1995] da najmanji broj točaka potreban za ocjenu korelacijske dimenzije iznosi $N_{\min} = 10^{(D2+2)/2}$. To bi značilo da se s 500 točaka može ocijeniti najviše korelacijska dimenzija od $D_2 \approx 3.4$. Za ocjenu sustava visoke složenosti dinamike potrebno je puno točaka. Tako autori uglavnom koriste između 300 i 12 000 točaka za zadovoljavajuću ocjenu složenosti niza srčanog ritma. Manji broj točaka od toga ne bi se trebalo koristiti zato što je iznos korelacijske dimenzije između 0.5 i 2 kod uobičajenih srčanih ritmova [Seker 2000].

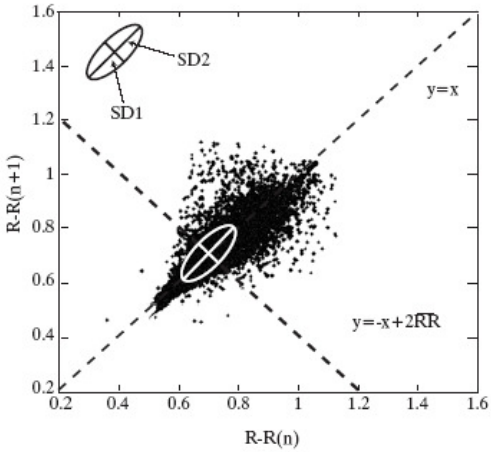
Autori [Small 2000] pružili su dokaz da se prilikom VF događa prijelaz u stanje determinističkog kaosa s korelacijskom dimenzijom koja je iznosila $D_2 = 6$ u prvih 30 sekundi fibrilacije. Autori [Owis 2002] pokazali su da je moguće korištenjem LLE i D_2 uspješno razdvojiti zdrave pacijente od pacijenata s nekom aritmijom (PVC, VBI, VT, VF), međutim ove značajke nisu bile dostatne za razlikovanje aritmija međusobno. Korelacijska dimenzija može se uspješno kombinirati s ostalim značajkama za dobivanje dobrih klasifikacijskih rezultata pri razlikovanju poremećaja rada srca [Acharya 2003, Yaghouby 2009, Jović 2009].

3.5.3.3. Standardne devijacije u faznom prostoru

Prikaz duljine RR-intervala u faznom prostoru daje dobar uvid u nelinearan karakter vremenskog niza. Glavni zadatak značajki LLE i D_2 je procjena stupnja nelinearnosti i kaotičnosti sustava. U dvodimenzionalnom faznom prostoru, procjenjuje se samo dinamika promjene srčanih otkucaja od jednog RR-intervala do njemu susjednog. U takvom prostoru ograničenom na $d = 2$ i s periodom odgode $\tau = 1$ definirane su značajke standardnih odstupanja koje imaju jasnu fiziološku interpretaciju. Kvalitativno promatrano, promatra se elipsa ubačena na mjesto koncentracije točaka u faznom dijagramu, slika 3.19.

Centar elipse pritom je određen prosječnom duljinom RR-intervala: $Center = (\bar{x}, \bar{x})$. Prvo standardno odstupanje SD1 označava raspršenost točaka od osi $y = x$, dok drugo standardno odstupanje SD2 kvantificira raspršenost točaka od osi $y = -x + 2\bar{x}$. [Tulppo 1996, Acharya 2004 (1), Asl 2008, Yaghouby 2009].

Odstupanje SD1 naziva se još i kratkotrajna varijabilnost otkucaja srca i proporcionalna je širini elipse. Na nju utječu i simpatički i parasimpatički dio ANS-a. SD2 je dugotrajna varijabilnost otkucaja srca i određuje duljinu elipse. Na SD2 utječe samo parasimpatički dio ANS-a. SD1 i SD2 se praktično izračunavaju koristeći izraze:



Slika 3.19. Elipsa koja je određena standardnim odstupanjima točaka od osi označenih isprekidanim crtama, preuzeto iz [Kitlas 2005].

$$SD1 = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N-1}(d_i - \bar{d})^2}\,,\ d_i = \frac{1}{\sqrt{2}}\left|x_{i+1} - x_i\right|$$

$$SD2 = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N-1}(d_i - \bar{d})^2}\,,\ d_i = \frac{1}{\sqrt{2}}\left|x_{i+1} + x_i - 2\bar{x}\right| \qquad (3.35).$$

Definira se omjer SD1/SD2 kao nelinearna značajka srčane aktivnosti. Pokazuje se da omjer SD1/SD2 nema značajne korelacije s linearnim vremenskim i frekvencijskim značajkama, dok mala korelacija postoji samo s mjerom približne entropije ApEn [Tulppo 1996]. Autori [Toichi 1997] definirali su nešto drugačije značajke koristeći se standardnim odstupanjima SD1 i SD2. U proučavanju utjecaja lijekova (atropina i propranolola) došli su do zaključka da se značajke SD1 i SD2 mogu iskoristiti za opis utjecaja simpatičkog i parasimpatičkog (vagalnog) dijela ANS na varijabilnost ritma. Tako su odredili značajke simpatičkog srčanog indeksa (engl. *cardiac sympathetic index*, kraće: CSI) i vagalnog srčanog indeksa (engl. *cardiac vagal index*, kraće: CVI) određene izrazima:

$$CSI = \frac{SD2}{SD1}\,,\quad CVI = \log_{10} 16 \cdot SD1 \cdot SD2 \qquad (3.36).$$

Značajke CSI i CVI koristili su i autori [Lin 2010] prilikom izrade sustava za klasifikaciju bolesnika s Parkinsonovom bolesti pomoću značajki HRV. Standardne devijacije smatraju se nelinearnim značajkama budući da se definiraju unutar dvodimenzionalnog faznog prostora. Budući da nisu linearne, u pravilu ih se koristi u kombinaciji s linearnim značajkama prilikom klasifikacije različitih bolesti na temelju srčanog ritma [Acharya 2004 (1), Asl 2008, Jović 2010 (2)].

3.5.3.4. Fraktalna dimenzija

Termin fraktal prvi je uveo Mandelbrot pri analizi ponašanja kaotičnih sustava [Mandelbrot 1983]. Fraktal je skup točaka koje kad se promatraju pri sve manjim i

manjim skalama izgledaju slično kao i kad ih se promatra na višim skalama, što znači da posjeduju svojstvo samosličnosti. Primjeri fraktala u ljudskoj anatomiji uključuju arterijska i venska stabla, grananja srčanih mišića i provodne strukture snopa His i vlakana Purkinje u srcu. Samoslične strukture služe za što učinkovitiji prijenos tvari kroz složeni raspodijeljeni biološki sustav [Goldberger 1996]. Višefraktalni karakter niza srčanih otkucaja označava to da se nepravilna kolebanja pojavljuju na više različitih vremenskih skala i samim time u različitim dijelovima atraktora u faznom prostoru, slično kao što i fraktalni objekt u matematici ima naboranu strukturu na raznim prostornim skalama [Ivanov 1999]. Pojam fraktalne dimenzije (engl. *fractal dimension*, kraće: FD) koji se odnosi na necjelobrojnu ili razlomljenu dimenziju potječe iz fraktalne geometrije. FD je mjera koja opisuje koliko je prostora neki objekt zauzeo između euklidskih dimenzija. U analizi vremenskog niza, FD predstavlja moćnu mjeru za otkrivanje prijelaznih pojava i nestacionarnosti [Acharya 2006]. Postoji nekoliko algoritama za određivanje FD na temelju dostupnog vremenskog niza. U ovom radu obradit će se Higuchijev algoritam koji je primjenjiv na analizu EKG-a i HRV-a [Higuchi 1988], iako postoje i drugi načini za izračun FD, kao što je npr. Katzov algoritam [Katz 1988].

Neka je $\{x_i\}, i = 1, 2, \ldots, N$ niz RR-intervala. Konstruira se *k* novih vremenskih nizova: $x_m^k = \{x(m), x(m+k), x(m+2k), \ldots, x(m+[N-mk])\}$, za $m = 1, 2, \ldots, k$, pri čemu *m* označava početnu vremensku vrijednost, a *k* je diskretan vremenski interval između točaka. Za svaki od *k* vremenskih nizova ili krivulja x_m^k, izračunava se duljina $L_m(k)$ pomoću izraza:

$$L_m(k) = \left(\frac{N-1}{\lfloor a \rfloor k} \right) \sum_{i=1}^{\lfloor a \rfloor} \left| x(m+ik) - x(m+(i-1)k) \right| \tag{3.37},$$

pri čemu $\lfloor a \rfloor$ označava cjelobrojni dio od *a*, $\frac{N-1}{\lfloor a \rfloor k}$ je faktor normiranja, te vrijedi $a = \frac{N-m}{k}$. Prosječna duljina krivulja se izračuna kao prosjek između *k* duljina $L_m(k)$ i to za $m = 1, 2, \ldots, k$. Ovaj postupak se ponavlja za svaki *k* između 1 i nekog $k_{\max}$ tako da se dobije prosjek duljina krivulja za svaki *k*, u oznaci $\langle L(k) \rangle$. Higuchi navodi da se za $k_{\max}$ može uzeti vrijednost $k_{\max} = 8$. Fraktalna dimenzija se definira kao nagib pravca dobivenog postupkom najmanjih kvadrata prosjeka duljina krivulja:

$$FD = \lim_{k \to \infty} \frac{\ln(\langle L(k) \rangle)}{\ln(1/k)} \tag{3.38}.$$

Autori [Acharya 2006] dobili su iznos Higuchijeve fraktalne dimenzije jednak 1.3604 za varijabilnost normalnog srčanog ritma što ukazuje na značajnu fraktalnost niza. Standardne vrijednosti Higuchijeve dimenzije najčešće variraju između 1 i 2. FD za Gaussov šum iznosi 1.5 [Higuchi 1988]. FD se najčešće koristi u relativnom smislu da bi se razlikovale skupine poremećaja, a ne u apsolutnom smislu da se određuje koliko je neki niz deterministčan, a koliko je slučajan. Kao mjera fraktalnosti, Higuchijeva FD

koristi se u pravilu u kombinaciji s ostalim značajkama za procjenu kaotičnosti vremenskog niza za što je potreban velik broj točaka [Higuchi 1988, Cerutti 2007].

3.5.3.5. Hurstov eksponent

Kao i fraktalna dimenzija FD, Hurstov eksponent (engl. *Hurst exponent*, kraće: HE), je mjera za opis samosličnosti i korelacijskih svojstava fraktalnih procesa. HE se najviše koristio za procjenu svojstava razlomljenosti Brownovog šuma, vremenskog niza koji proizvodi fraktalni Gaussov proces. Kod fizioloških vremenskih nizova, HE koristi se za ocjenu prisutnosti dugo djelujućih međuovisnosti. Lokalni trendovi mogu predstavljati problem prilikom procjene stupnja samosličnosti niza, tako da je nužno preskalirati vremenski niz da bi se omogućila procjena fraktalnih svojstava [Acharya 2006].

Za segment od k RR-intervala, razlika između svakog intervala i srednje vrijednosti svih k intervala se izračuna i dodaje kumulativnoj sumi:

$$Sum = \sum_{i=1}^{k} (x_i - \bar{x}_k) \tag{3.39}.$$

Normalizirani domet $\frac{R(k)}{S(k)}$ je razlika između između najveće i najmanje vrijednosti koju postiže kumulativna suma, podijeljena sa standardnom devijacijom segmenta RR-intervala:

$$\frac{R(k)}{S(k)} = \frac{\max(Sum_1, Sum_2, \ldots, Sum_k) - \min(Sum_1, Sum_2, \ldots, Sum_k)}{\sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{x})^2}} \tag{3.40}.$$

$\frac{R(k)}{S(k)}$ se crta u ovisnosti od k. Izraz za fraktalni Hurstov eksponent kao nagib pravca dan je kao [Hurst 1951]:

$$HE = \frac{\log \frac{R(k)}{S(k)}}{\log k} \tag{3.41}.$$

Ovaj izraz izvodi se iz Hurstove opće jednadžbe vremenskog niza koja isto tako vrijedi i za Brownovo gibanje. Ako je $H = 0.5$, tada je ponašanje vremenskog niza slično slučajnom. Za $H > 0.5$ vremenski niz pokriva veći domet od slučajnog hoda, budući da postoji pozitivna korelacija među podacima. Potpuna samosličnost postiže se za $H = 1$. Za $H < 0.5$ vremenski niz pokriva manji doseg od slučajnog hoda pa je niz negativno autokoreliran [Teich 2001]. Odnos Hurstovog eksponenta, korelacijske i ugradbene dimenzije dan je izrazom :

$$HE = d + 1 - D_2 \tag{3.42}.$$

Za zdrave pacijente, HE je relativno visok (autori navode oko 0.61) što indicira visokosloženi karakter zdravog ritma. Za CHB i I/D kardiomiopatiju HE je nizak, budući da je varijacija RR-intervala niska. Za AF, eksponent je još niži budući da je ponašanje gotovo slučajno. Za SSS, Hurstov eksponent je također jako nizak budući da

je ponašanje ritma periodično [Acharya 2006]. Autori [Krstačić 2003] pokazali su da je prije administracija lijeka na bazi blokiranja kanala kalcija HE bio niži (0.62), da bi se nakon 6 mjeseci popeo gotovo na razinu zdravih osoba (0.76).

3.5.3.6. Analiza kolebanja RR-intervala s uklonjenim trendom

Analiza kolebanja RR-intervala s uklonjenim trendom (engl. *detrended fluctuation analysis*, kraće: DFA) pripada u istu porodicu algoritama kao i fraktalna dimenzija i Hurstov eksponent, budući da kvantificira fraktalno skaliranje RR-intervala. Postupak DFA prva je predložila skupina autora [Peng C. 1995]. Oni navode da se zdravi otkucaj srca tradicionalno smatrao reguliranim prema klasičnom principu homeostaze gdje fiziološki sustavi djeluju tako da smanje varijabilnost i postignu stanje slično ravnoteži. Međutim, novija istraživanja otkrila su da pod normalnim uvjetima, kolebanja u otkucajima zdravog srčanog ritma prikazuju dugoročne korelacije koje tipično pokazuju dinamički sustavi daleko od ravnoteže. Vremenski niz srčanih otkucaja od pacijenata koji pate od teškog oblika CHF pokazuju upravo prekidanje tih dugoročnih korelacija.

Autori su predložili algoritam koji se odvija tako da se najprije integrira vremenski niz RR-intervala, ukupne duljine *N*, koristeći formulu:

$$y(k) = \sum_{i=1}^{k} (x_i - \bar{x}),\ k = 1, \ldots, N \tag{3.43},$$

gdje je $y(k)$ *k*-ta vrijednost integriranog niza, Nadalje se tako dobiveni integrirani niz podijeli u manje prozore jednake duljine *n*. Unutar svakog prozora duljine *n*, pravac aproksimira točke postupkom najmanjih kvadrata. Time se ostvaruje prikaz trenda unutar tog prozora. *y*-koordinate ovih linijskih segmenata određene su s $y_n(k)$. Zatim, integriranom vremenskom nizu u svakom prozoru se oduzme trend. Kolebanje ovakvog niza izračunava se izrazom:

$$F(n) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (y(k) - y_n(k))^2} \tag{3.44}.$$

Ovaj proračun se ponavlja za sve veličine prozora da bi se dobio odnos između $F(n)$ i veličine prozora *n*. Obično se prikazuje ovisnost $F(n)$ o *n* na log-log skali. Tipično, $F(n)$ se povećava s veličinom prozora. Kolebanje u prozorima može se okarakterizirati s eksponentom skaliranja (faktorom samosličnosti) α, koji predstavlja nagib pravca $\frac{\log F(n)}{\log n}$.

Za $\alpha = 0.5$ signal ima slučajan hod, sličan bijelom šumu. Za $\alpha < 0.5$ niz je negativno koreliran, tj. povećanje vrijednosti najvjerojatnije slijedi smanjenje vrijednosti, a za $\alpha > 0.5$ postoji pozitivna korelacija. Vrijednost $\alpha = 1.0$ odgovara šumu $1/f$, a 1.5 Brownovom gibanju. Ako je $\alpha > 1.5$, tada postoje dugoročne korelacije koje nisu nužno povezane sa slučajnim procesom već mogu ukazivati na nelinearnu određenost, odnosno kaos [Rodriguez 2008].

Hurstov eksponent povezan je s faktorom α od DFA izrazom [Goldberger 2002]:

$$\alpha = 1 + HE \tag{3.45}$$

Prilikom analize istraživači najčešće koriste zasebno faktore α_1 i α_2. Faktor α_1 koristi se za opis kratkoročnih kolebanja, a faktor α_2 za opis dugoročnih kolebanja. Autori [Huikuri 2000] analiziraju segmente duljine N = 8000 RR-intervala te izračunavaju α_1 za prozor $n < 11$ te α_2 za prozore $n \geq 11$. U čitavom segmentu izračunali su 32 takva faktora, što znači da je pomak između dvaju prozora iznosio 8000/32 = 250 RR-intervala. Pokazali su da je smanjena vrijednost kratkoročnog faktora kolebanja α_1 nezavisni prediktor smrti nakon akutnog infarkta miokarda (AMI), i to i zbog aritmije i zbog razloga koji nisu povezani s aritmijom. Kod autora [Ho 1997] vidljivo je da su oba faktora α_1 i α_2 smanjena kod pacijenata oboljelih od CHF. Postupak DFA sposoban je i za razlikovanje između zdravih pacijenata, pacijenata oboljelih od CHF i pacijenata s transplantiranim srcem [Cerutti 2007]. Značajka DFA uspješno je primijenjena i kod klasifikacije aritmija [Acharya 2002, Asl 2008], klasifikacije zdravih pacijenata i dijabetičara [Kitlas 2005], te pri ispitivanju djelotvornosti lijekova [Krstačić 2003].

3.5.3.7. Približna entropija

Približna entropija (engl. *approximate entropy*, kraće: ApEn), je statistička mjera koja se koristi za kvantifikaciju nepravilnosti u nekom nizu podataka bez *a priori* znanja o procesu. Ovu metodu razvili su [Pincus 1994] i dosta je često korištena u karakterizaciji vremenskog niza, pogotovo na kraćim segmentima [Seker 2000]. Autori su postupak namijenili za bilo koji sustav u kojem je teško ili nemoguće precizno izračunati Kolmogorovljevu entropiju odnosno točno odrediti dimenziju atraktora. Osnovna ideja postupka je opis nereda u smislu da se određuju uvjetne vjerojatnosti pri kojima se slični obrasci u nizu ne ponavljaju. Ako vremenski niz pokazuje složeno, nepravilno ponašanje, onda će imati visoku mjeru približne entropije [Ho 1997]. ApEn ima četiri karakteristike koje ga čine pouzdanom i rado korištenom značajkom pri analizi HRV: na nju ne utječe šum na razini manjoj od radijusa *r*, robusna je na povremene veće ili manje artefakte u vremenskom nizu, daje smislenu informacijsku mjeru za relativno malen broj točaka, i konačna je i za slučajne i za određene procese [Pincus 1994].

Algoritam određivanja približne entropije je sljedeći:

1. Promatra se $N - m + 1$ točaka u faznom prostoru dimenzije *m* (može i *d*, ali za ovaj je algoritam uvriježena oznaka *m*), gdje je *N* duljina segmenta, vidi izraz (3.25).
2. Definira se udaljenost između točaka $\bar{x}_i$ i $\bar{x}_j$, $\delta[\bar{x}_i, \bar{x}_j]$ kao najveća apsolutna razlika između njihovih komponenti po osima, tj. najveća norma :
$$\delta[\bar{x}_i, \bar{x}_j] = \max_{k=1,2,\ldots,m} \left| x_i(i+k-1) - x_j(i+k-1) \right| \tag{3.46}$$
3. Za određenu točku $\bar{x}_i$, prebroje se takve točke $\bar{x}_j$, $j = 1, \ldots, N-m+1$, za koje vrijedi $\delta[\bar{x}_i, \bar{x}_j] \leq r$. Taj broj točaka označi se s $N^m(i)$. Postupak se ponovi $\forall i, i = 1, \ldots, N-m+1$ te se odrede koeficijenti, korelacijski integrali, $C_r^m(i)$ prema

formuli:

$$C_r^m(i) = \frac{N^m(i)}{N-m+1} \tag{3.47}.$$

$C_r^m(i)$ je mjera koja daje, uz toleranciju *r*, frekvenciju točaka koje su bliske nekoj točki $\bar{x}_i$.

4. Izračunaju se prirodni logaritmi za svaki $C_r^m(i)$ i usrednji ih se. Time se dobiva faktor:

$$\phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_r^m(i) \tag{3.48}.$$

5. Dimenzija se poveća na $m+1$. Ponove se koraci 1 – 4 i nađu se $C_r^m(i)$ i $\phi^{m+1}(r)$.
6. Definira se približna entropija kao:

$$ApEn(m,r,N) = \phi^m(r) - \phi^{m+1}(r). \tag{3.49}.$$

Parametri *m* i *r* određuju se u ovisnosti o specifičnom problemu. Ako su već poznati neki dosadašnji rezultati, tada se uzimaju *m* i *r* koji su se dosad pokazali najboljima [Pincus 1994]. U slučaju sustavnog provjeravanja parametara, obično se za početni *m* uzima da iznosi $m = 2$, što znači da se promatra dvodimenzionalni fazni prostor, a za radijus *r* se mogu uzeti neke vrijednosti između $r = 0.1\sigma$ i $r = 0.25\sigma$, pri čemu je σ standardno odstupanje promatranog niza RR-intervala. Izvorni autori mjere [Pincus 1994] preporučuju da se za *r* uzme vrijednost $r = 0.2\sigma$. Moguće je koristiti i više značajki ApEn, svaku za svoju vrijednost parametra *r*. Tako se uzimaju četiri mjere približne entropije, i to za $r = 0.1\sigma$, $r = 0.15\sigma$, $r = 0.2\sigma$, i $r = 0.25\sigma$ [Hornero 2006, Jović 2009]. Time se pokriva veći raspon točaka i postiže bolja preciznost opisa složenosti vremenskog niza.

Autori [Lu 2008] uočili su da često preporučene vrijednosti parametra *r* za ApEn ne daju najtočniju mjeru za procjenu nepravilnosti vremenskog niza te su stoga predložili da se umjesto toga za mjeru nepravilnosti uzima najveća vrijednost od ApEn. Međutim, s obzirom na to da je računarno zahtjevno izračunavati vrijednosti za svaki *r* između 0.0 i 1.0, predlažu empirijski izveden skup jednadžbi koji bi trebao odrediti *r* za kojeg je entropija najveća. Kliničku značajnost najveće vrijednosti ApEn ostaje za istražiti.

Približna entropija može se koristiti i za analizu binarno kodiranih nizova [Cysarz 2000]. Vrijednosti ApEn-a bile su snižene i kod djece oboljele od dijabetesa [Kitlas 2005]. Vrlo važnu primjenu ApEn ima i kao jedna od značajki pri razvrstavanju srčanih ritmova [Acharya 2004 (2), Asl 2008, Chua 2008]. Autori [Acharya 2006] navode da abnormalni srčani ritmovi imaju u pravilu nižu vrijednost ApEn od zdravog ritma, s iznimkom poremećaja bolesnog sinusnog čvora (SSS).

3.5.3.8. Entropija uzorka

Entropija uzorka (engl. *sample entropy*, kraće: SampEn) vrlo je slična mjeri ApEn. Bitna razlika sastoji se u tome što ApEn prilikom izračuna udaljenosti uzima u obzir i to da je referentna točka bliska samoj sebi. Time vrijednost korelacijskog integrala nikad

nije 0, pa je moguće u svakom slučaju izračunati vrijednost entropije uzorka. Budući da se ubrojavanjem i referentne točke na umjetan način smanjuje vrijednost entropije, ApEn je u pravilu nešto malo manjeg iznosa u odnosu na SampEn te pokazuje nešto manju vrijednost entropije nego što ona zaista je. Tu pristranost mjere ApEn proučavali su autori [Richman 2000] te predložili rješenje u vidu mjere SampEn. Osim ne uzimanja u obzir referentne točke pri izračunu, SampEn zahtjeva samo to da je moguće izračunati korelacijske integrale za uvjetnu vjerojatnost B^m, dakle samo u prvih *m* dimenzija.

Algoritam se može prikazati ovako [Richman 2000]:

$$SampEn(m,r,N) = -\ln\left[\frac{A^{\mathrm{m}}(r)}{B^{\mathrm{m}}(r)}\right] \tag{3.50},$$

$$A^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} A_i^m(r), \quad B^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} B_i^m(r) \tag{3.51},$$

$$A_i^m(r) = \frac{\text{broj } j,\ j \neq i,\ j \leq N-m, \text{ takav da } d\left[x_{m+1}(i), x_{m+1}(j)\right] \leq r}{N-m-1},$$

$$B_i^m(r) = \frac{\text{broj } j,\ j \neq i,\ j \leq N-m, \text{ takav da } d\left[x_m(i), x_m(j)\right] \leq r}{N-m-1} \tag{3.52}.$$

pri čemu oznaka x_m označava da se gleda točka *x* kao vektor u *m*-dimenzionalnom faznom prostoru.

Potrebno je zamijetiti da se u (3.52) gledaju samo takvi $x(j)$ koji su različiti od referentne točke $x(i)$. U teoriji, SampEn bi trebao dati bolju procjenu nepravilnosti od ApEn na kratkim nizovima. Za veliki *N*, pristranost ApEn-a nema značajnog utjecaja tako da su vrijednosti SampEn i ApEn približno iste. Kod SampEn mogu se koristiti iste vrijednosti *r* i *m* kao i kod ApEn.

Za razliku od ApEn, SampEn se nije puno koristio u analizi varijabilnosti srčanog ritma. Autori [Chua 2008] koristili su SampEn, ApEn, rekurentni dijagram, i standardne devijacije u faznom prostoru za analizu osam različitih tipova srčanih ritmova. Začudo, ApEn se pokazao nešto točnijim od SampEn pri klasifikaciji različitih abnormalnih tipova ritmova u toj studiji. Autori [Porta 2007] proveli su istraživanje u kojem su pokazali da promjenom nagiba tijela, odnosno glave pacijenta, dolazi do promjena u složenosti niza HRV. Iako se složenost srčanih otkucaja prije povezivala samo s bolesti i starenjem, autori su pokazali da ona ima veze i s djelovanjem simpatičko-parasimpatičkog sustava ravnoteže ANS-a. Što je bio viši položaj glave, to je složenost signala bila manja, što su pokazali korištenjem mjera SampEn, ApEn, i ispravljene uvjetne entropije (engl. *corrected conditional entropy*, kraće: CCE).

3.5.3.9. Shannonova entropija i ispravljena uvjetna entropija

Shannonova entropija (engl. *Shannon entropy*, kraće: ShEn), poznata i kao informacijska entropija, definirana je kao specijalan slučaj Kolmogorovljeve entropije u kojem se promatra prva potencija vjerojatnosti nalaska određenih točaka unutar nekog područja u faznom prostoru.

Neka je u *d*-dimenzionalnom faznom prostoru zadana točka putem izraza (3.25), pri čemu je period odgode $\tau = 1$. Shannonova entropija mjeri prosječan iznos informacije sadržan u uzorku duljine *d* kao:

$$ShEn(d) = -\sum_{i=1}^{N-d} p(\bar{x}_i) \log p(\bar{x}_i) \quad (3.53),$$

pri čemu sumacija ide po vjerojatnostima pojave svih točaka $\bar{x}_i$ unutar neke ćelije u faznom dijagramu [Porta 2007].

Uvjetna entropija ili stopa entropije (engl. *conditional entropy*, *entropy rate*, kraće: CE) definira se kao:

$$CE = ShEn(d) - ShEn(d-1) \quad (3.54)$$

i mjeri informaciju koju donosi *d*-ti uzorak, odnosno *d*-ta komponenta točke u faznom prostoru u odnosu na proteklih *d*-1 komponenti. Procjena uvjetne entropije određuje se u pravilu putem ShEn na kraćem broju točaka (*N* = par stotina).

Da bi se moglo izračunati ShEn, nužno je odrediti širinu ćelije i posljedično, kvantizaciju faznog prostora. U tu svrhu, provodi se uniformna kvantizacija na ξ razina, pri čemu je s ε određena širina ćelije te vrijedi odnos $\varepsilon = (x_{\max} - x_{\min})/\xi$, gdje $x_{\max}$ označava najveću, a $x_{\min}$ najmanju vrijednost komponenti u faznom prostoru. Tako je fazni prostor podijeljen na ξ^d hiperkocaka širine ε. Sve točke koje se nađu unutar jedne hiperkocke neće se razlikovati. Prilikom izračuna ShEn uzimat će se u obzir samo broj točaka unutar hiperkocki. Sad se ponovno definira izraz za ShEn s tolerancijom ε:

$$ShEn(d,\varepsilon) = -\sum_{i=1}^{d\xi} p\left[\bar{x}_i^{\xi}\right] \log p\left[\bar{x}_i^{\xi}\right], \quad p\left[\bar{x}_i^{\xi}\right] = \frac{N_i(d,\varepsilon)}{N-d+1}, \quad (3.55),$$

gdje je s $N_i(d,\varepsilon)$ označen broj točaka unutar *i*-te hiperkocke, $i = 1,\ldots,d\xi$. U skladu s ovom redefinicijom ShEn, može se odrediti i uvjetna entropija kao:

$$CE(d,\varepsilon) = ShEn(d,\varepsilon) - ShEn(d-1,\varepsilon). \quad (3.56).$$

Autori [Porta 1998] primijetili su da je mjera $CE(d,\varepsilon)$ pristrana na sličan način kao i ApEn. Naime, ako se unutar neke hiperkocke nađe samo jedna točka, tada je to slučaj analogan brojanju referentnih točki kao bliskih njima samima kod ApEn-a. Uistinu, same točke unutar hiperkocke u *d*-1 dimenzija ostat će same i u *d* dimenzija. Time je njihov doprinos uvjetnoj entropiji jednak nuli pa stvaraju pristranost u smjeru smanjenja entropije i povećanju pravilnosti u podacima. Da bi se tome doskočilo, autori su predložili mjeru ispravljene uvjetne entropije (CCE) određenu s:

$$CCE(d,\varepsilon) = CE(d,\varepsilon) + CE(1,\varepsilon) \cdot perc(d,\varepsilon) \quad (3.57),$$

gdje je $CE(1,\varepsilon) = ShEn(1,\varepsilon)$, a $perc(d,\varepsilon)$ je postotak samih točaka (jedna točka u hiperkocki) u *d*-dimenzionalnom faznom prostoru u odnosu na ukupan broj točaka, $0 \le perc(d,\varepsilon) \le 1$. Isti autori su pokazali da ako se dimenzija *d* varira, onda vrijede sljedeće opažanja:

1. CCE ostaje stalan u prisutnosti bijelog šuma.
2. CCE se smanjuje prema nuli u prisutnosti potpuno predvidljivog sustava.

3. CCE ima najmanju vrijednost za dimenziju $d_{\min}$, ako se ponavljajući obrasci nalaze ugrađeni u šum.

Autori [Porta 2007] određuju konstantu kvantizacije $\xi = 6$ pa se ε određuje na temelju te konstante i vremenskog niza RR-intervala. Pokazuju da što je bio viši položaj glave pacijenta, to je složenost vremenskog niza bila manja. Autori [Clariá 2008] pokazali su da Shannonova i Rényijeva entropija daju najbolje rezultate u odnosu na tradicionalne značajke i značajke vremensko-frekvencijske analize pri predviđanju koji će pacijent s hipertrofiranom kardiomiopatijom doživjeti iznenadnu srčanu smrt (SCD).

3.5.3.10. Entropija na više skala

Entropija na više skala (engl. *multiscale entropy*, kraće: MSE) noviji je pristup analizi složenosti bioloških sustava koji je razvijen s namjerom opisa biološkog signala na raznim vremenskim skalama. Pokazuje se da je superioran pojednostavljenim mjerama približne entropije i entropije uzorka koje promatraju samo jednu skalu vremenskog niza [Costa 2002].

Neka je dan vremenski niz RR-intervala $\{x_i\}$, $i = 1, \ldots, N$. Iz njega se sagradi niz grubo-usitnjenih vremenskih nizova $\left\{y^{(\tau)}\right\}$, određenih faktorom skaliranja τ kao:

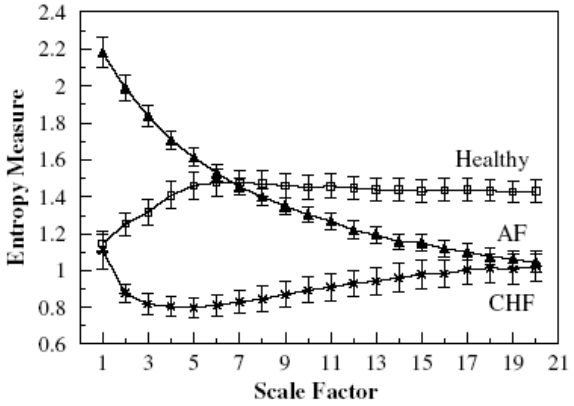
$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, 1 \le j \le \frac{N}{\tau} \tag{3.58}.$$

Za prvu skalu, gdje je $\tau = 1$, niz je isti kao i originalni. Duljina svakog usitnjenog vremenskog niza jednaka je duljini originalnog vremenskog niza podijeljenog s faktorom τ.

Sada se za svaki vremenski niz $y_j^{(\tau)}$ izračuna entropija uzorka SampEn i slijed rezultata se nacrta na grafu kao funkcija od τ. Taj slijed vrijednosti SampEn na različitim skalama naziva se entropijom na više skala. Primjer je dan na slici 3.20.

U svojem članku iz 2005., autori [Costa 2005 (1)] dolaze do vrlo značajnog zaključka: za dovoljno velike vremenske skale, vremenski niz zdravih pacijenata ima najvišu entropijsku vrijednost u odnosu na sve ostale promatrane poremećaje. Time postupak MSE pokazuje da je dinamika zdravog srca najsloženija, što se često nije vidjelo na nižim skalama. Potrebno je uočiti dvije stvari:

1. Vremenski niz RR-intervala je s većim faktorom τ sve kraći, što dakako utječe na preciznost izračuna. Ipak, pokazuje se da je SampEn dovoljno nepristrana mjera za razumno kratke duljine vremenske nizove (oko 100 intervala za najviši τ).
2. Izračun SampEn ovisi o standardnoj devijaciji vremenskog niza. Time na rezultat značajno utječu različiti artefakti i stršeće točke. Ako se oni mogu ukloniti, tj. ako signal nije značajno loš, to se treba provesti prije izračuna MSE. Nelokalne nestacionarnosti kao što su trendovi ne bi se smjeli uklanjati, jer iako utječu na rezultat u nekoj mjeri, time bi se narušila struktura niza.



Slika 3.20. Entropija na više skala, kao mjera entropije na jednoj skali uzeta je SampEn, preuzeto iz [Costa 2002].

Iako izvorni autori koriste mjeru SampEn, nigdje se ne navodi da se za izračun MSE mora koristiti isključivo entropija uzorka. Ono što je poželjno kod SampEn njezina je nepristranost i iskoristivost na malim vremenskim segmentima.

3.5.3.11. Spektralna entropija

Spektralna entropija (engl. *spectral entropy*, kraće: SpectEn) kvantificira spektralnu složenost promatranog niza srčanih otkucaja. Najčešće je spektralna entropija Shannonova entropija vremenskog niza prebačena u frekvencijsku domenu [Rezek 1998].

Pri izračunu značajke SpectEn vremenski niz se može u frekvencijsku domenu prebaciti korištenjem Fourierove transformacije, ili se spektar procjenjuje Burgovim ili nekim drugim postupkom. Izraz za entropiju dan je sa:

$$SpectEn = -\sum_f PSD(f) \log PSD(f) \tag{3.59},$$

pri čemu sumacija ide po svim frekvencijama *f* na kojima je procijenjena spektralna gustoća snage PSD, vidjeti poglavlje 3.3.4.

Heuristički promatrano, spektralna entropija može se promatrati kao mjera nesigurnosti za pojavu događaja na frekvenciji *f*. Za razliku od ShEn u vremenskoj domeni, u frekvencijskoj domeni ne određuje se uvjetna entropija (CE). Autori [Acharya 2004 (1)] i [Asl 2008] koristili su spektralnu entropiju kao jednu od značajki prilikom klasifikacije srčanih aritmija. Upitna je kvaliteta izračuna spektralne entropije na malom broju RR-intervala, budući da tada spektar nije moguće dovoljno precizno procijeniti.

3.5.3.12. Rényijeva entropija

Rényijeva entropija (engl. *Rényi entropy*, kraće : RenEn) je složenija informacijska mjera u odnosu na klasičnu Shannonovu entropiju. Koristi se ako postoji mogućnost negativnih vrijednosti u razdiobi, s kojima ShEn ne zna računati. Jednu od svojih

primjena pronalazi u analizi vremensko-frekvencijskog predstavljanja vremenskog niza RR-intervala [Faust 2004].

Rényijeva entropija dobila je svoj naziv prema izvornom autoru [Rényi 1961]. Ona se izvodi iz njegove predložene teorije srednjih vrijednosti:

$$H = \varphi^{-1}\left(\sum_{k=1}^{N} p_k \varphi(I(p_k))\right) \tag{3.60},$$

pri čemu je s $\varphi(.)$ označen kontinuiran i striktno monotoni podrazred Kolmogorov-Nagumovih funkcija, $\varphi^{-1}(.)$ je inverzna funkcija, dok je $I(p_k)$ bilo koja informacijska mjera. Naprimjer: $I(p_k) = -\log(p_k)$ je Hartleyeva informacijska mjera, pri čemu je p_k vjerojatnost pojave *k*-te vrijednosti u vremenskom nizu. Da bi se zadovoljila ograničenja informacijske mjere za funkciju φ uzimaju se neke od ovih vrijednosti:

$$\varphi(x) = \begin{cases} x & \text{Shannonova entropija} \\ 2^{(1-\alpha)x} & \text{Rényijeva entropija reda } \alpha \end{cases} \tag{3.61}.$$

Pojednostavljujući gornji izraz dobiva se opći izraz za Rényijevu entropiju u obliku [Waheed 2002]:

$$\text{RenEn}_\alpha = \frac{1}{1-\alpha}\log_2\left(\sum_{k=1}^{N} p_k^\alpha\right);\ \alpha > 0,\ \alpha \neq 1 \tag{3.62}.$$

RenEn je viša za srčana oboljenja kao što su I/D kardiomiopatija, CHB, i VF. a niža je za normalan srčani ritam, PVC, AF, SSS, i CHF.

3.5.3.13. Kolmogorovljeva entropija

Kolmogorovljeva entropija, često: K-entropija (engl. *K entropy*), a pojavljuje se i kao KS-entropija (Kolmogorov-Sinai), kvantitativna je mjera koja se koristi za opis dinamičkog sustava iz informacijskog stajališta [Kolmogorov 1958]. Ona se definira na faznom prostoru dinamičkog sustava. Potrebno je napomenuti da se osnovna ideja podjele faznog prostora u ćelije ne razlikuje od one korištene pri opisu korelacijske dimenzije, no ovdje se ponovo navodi radi potpunosti.

Neka je s $I(l,T)$ označena količina informacije koju se dobije slijedeći putanju sustava u vremenskom intervalu *T* s određenom preciznošću *l*. Osnovna ćelija faznog prostora sustava (rekonstrukcijske) dimenzije *m* posjeduje volumen l^m i povezani vremenski interval opažanja $N\tau = T$. Ako se s $p_{i0,i1\ldots iN}$ označi složena vjerojatnost da se varijabla sustava, u oznaci $x(t)$ nađe za $t = 0$ u ćeliji i_0, za $t = \tau$ u ćeliji i_1, a za $t = N\tau$ u ćeliji i_N, informacijski sadržaj bit će jednak

$$I(l,T) = -\sum_{i0,i1\ldots iN} p_{i0,i1\ldots iN} \log_2 p_{i0,i1\ldots iN} \tag{3.63}.$$

Kolmogorovljeva entropija definirana je kao granična vrijednost informacije, kad dimenzije ćelije iščezavaju, a vrijeme promatranja teži beskonačnosti:

$$K = \lim_{l\to 0}\left(\lim_{T\to\infty}\frac{1}{T} I(l,T)\right)\ \text{, bit/s} \tag{3.64}.$$

Važno je naglasiti da K-entropija određuje gubitak informacije koji je nastao u sustavu njegovim odmakom od početne točke kroz neko vrijeme. K-entropija je time povezana s Ljapunovljevim eksponentima pojašnjenima u poglavlju 3.5.3.1. Ona je suma svih pozitivnih LE. Ako postoji barem jedan pozitivan LE, tada će i K entropija biti pozitivna što će sugerirati kaotičan sustav. Ipak, velik iznos K-entropije može se tumačiti kao slučajnost u podacima [Protopescu 2005]. Kolmogorovljeva entropija kao mjera predvidljivosti pouzdano se koristi samo u slučaju vrlo velike količine podataka i u teorijskim razmatranjima bez prisutnosti šuma. Ona je snažno kompromitirana ako postoji stabilan izvor šuma, a za slučajne procese ima beskonačan iznos [Pincus 1994]. Pokazuje se da budući da su vremenski nizovi ograničene, konačne duljine, većina ih ne može imati dobro procijenjenu mjeru K-entropije. Upravo stoga su i uvedene pojednostavljene mjere izračuna entropije kao što su približna entropija, entropija uzorka i entropija na više skala [Pincus 1994, Costa 2002]. Praktično određivanje KS-entropije najčešće se provodi nekim od postupaka procjene spektra LE [Grond 2003].

3.5.3.14. Indeks prostorne popunjenosti

Indeks prostorne popunjenosti (engl. *spatial filling index*, kraće: SFI) je mjera koja kvantificira koncentraciju točaka u faznom prostoru. Postupak je opisan u radu [Faust 2004]. Neka je signal predstavljen s koordinatama točke u *d*-dimenzionalnom faznom prostoru. Dinamičko ponašanje određuje putanju točaka $\bar{x}_n$ u faznom prostoru danu izrazom (3.26).

Oblikuje se matrica $A_d = [\bar{x}_1\ \bar{x}_2 \cdots \bar{x}_M]^T$, pri čemu je $\bar{x}_i, i = 1,\ldots,M$ točka u faznom prostoru dana izrazom (3.25), a $M = N - (d-1)\tau$. N je broj RR-intervala, a τ je broj intervala odgode. Ako se pretpostavi dvodimenzionalni fazni prostor $d = 2$, tada vrijedi $M = N - \tau$ te slijedi $A_2 = \begin{bmatrix} x(1) & x(1+T) \\ x(2) & x(2+T) \\ \cdots & \cdots \\ x(M) & x(N) \end{bmatrix}$. Dalje se određuje normirana matrica B_d koja se dobiva dijeljenjem svakog elementa iz A_d s apsolutno najvećom vrijednosti elemenata matrice A_d, $x_d = \max|x(k)|$, $1 \le k \le N$. Slijedi za matricu B_2 : $B_2 = \frac{A_2}{x_{\max}} = \begin{bmatrix} x(1)/x_{\max} & x(1+T)/x_{\max} \\ x(2)/x_{\max} & x(2+T)/x_{\max} \\ \cdots & \cdots \\ x(M)/x_{\max} & x(N)/x_{\max} \end{bmatrix}$. Za elemente matrice B_d vrijedi $-1 \le b_{ij} \le 1$.

Dvodimenzionalni fazni prostor se nadalje podijeli u $n \times n$ kvadrata (u *d*-dimenzionalnom slučaju govori se o hiperkockama), svaki veličine $R \times R$, $R \in \Re, \frac{2}{R} \in N$. Konstruira se matrica C s elementima c_{ij}, dimenzije $n \times n$ takva da je

c_{ij} broj točaka u faznom prostoru koje padaju u kvadrat $g(i, j), i, j \in \{1, \ldots, n\}$. Matrica *C* naziva se matrica faznog prostora. Općenito, u *d* dimenzija, *C* je *d*-dimenzionalno polje.

Dalje se oblikuje matrica *P*, s elementima $p_{i,j}$, takva da vrijedi $p_{i,j} = \frac{c_{i,j}}{l}$, $l = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}$. Matrica *P* daje vjerojatnost da točka u faznom prostoru padne u kvadrat $g(i, j)$. Matricu *Q* izračuna se tako da se kvadrira svaki element $p_{i,j}$ matrice *P*. Vrijedi $q_{i,j} = {p_{i,j}}^2$. Odredi se suma elemenata matrice *Q* u oznaci *s*: $s = \sum_{i=1}^{n}\sum_{j=1}^{n} q_{i,j}$. Indeks prostorne popunjenosti, u oznaci η definiran je izrazom:

$$\eta = \frac{s}{n^2} \tag{3.65}.$$

Red veličine za η je 10^{-3} i to je veći što je veći broj točaka unutar jednog od kvadrata matrice *C*, odnosno što je veća koncentriranost točaka u faznom prostoru. Budući da poremećaji kao što su atrijalna i ventrikularna fibrilacija imaju visoku raspršenost točaka u faznom dijagramu, za njih je η nizak. Najviši η dobiva se za srčane poremećaje s niskom varijacijom RR-intervala, kao što su CHB, CHF, i I/D kardiomiopatija. U prethodnom radu, koristila se značajka η kao jedna od nelinearnih značajki pri uspješnom razvrstavanju normalnog srčanog ritma, bilo koje aritmije, supraventrikularne aritmije i CHF. Značajka η pokazala se uz ApEn jednom od najprikladnijih značajki [Jović 2011 (1)].

3.5.3.15. Allanov faktor

Allanov faktor (engl. *Allan factor*, kraće: ALF) uveli su u analizu varijabilnosti rada srca autori [Teich 1998]. Ovu nelinearnu statistiku prvi je opisao autor [Allan 1966] koji ju je koristio za procjenu stabilnosti atomskih satova. Allanov faktor je omjer Allanove varijance i dvostruke srednje vrijednosti broja događaja unutar nekog vremena *t*.

$$ALF(t) = \frac{E\left\{\left[N_{i+1}(t) - N_i(t)\right]^2\right\}}{2E\left\{N_{i+1}(t)\right\}} = \frac{\sum_{i=1}^{M-1}\left[N_{i+1}(t) - N_i(t)\right]^2}{2\sum_{i=1}^{M-1} N_{i+1}(t)} \tag{3.66},$$

pri čemu je N_i broj R-vrhova (ili RR-intervala) unutar *i*-tog vremenskog odsječka trajanja *t*, a *M* je ukupan broj odsječaka u analiziranom zapisu. ALF je značajka ovisna o odabranoj vremenskoj skali. Za razliku od obične varijance, Allanova varijanca se definira u terminima varijabilnosti slijednih brojanja. Budući da je djelovanje Allanovog faktora slično derivaciji, ono je dobro za rješavanje problema nestacionarnosti u vremenskom nizu. Vrijednost ALF za homogeni Poissonov proces iznosi $ALF(t) = 1, \forall t$. Bilo koje odstupanje od jedinice ukazuje da točkasti proces nije

Poissonov. Ako je $ALF(t) > 1$, tada je vremenski niz manje uređen od homogenog Poissonovog procesa, a ako je $ALF(t) < 1$ onda je niz uređeniji. Autori [Teich 2001] za ALF uzimaju vremensku skalu $T = 10$, dok je ukupan broj odsječaka *M* velik (≈ 5000).

3.5.3.16. Indeks asimetrije na više skala

Indeks asimetrije na više skala (engl. *multiscale asymmetry index*, kraće: A_I) je mjera koju su relativno nedavno osmislili autori [Costa 2005 (2)]. Mjera se zasniva na opažanju da živi organizmi koriste energiju da bi evoluirali u sve više uređeni hijerarhijski ustroj koji pokazuje sve manju entropiju u usporedbi s okolinom. Samouređenje je povezano s usmjerenosti toka energije u jednom smjeru preko granica sustava te posljedično, s nepovratnosti dotičnih procesa. Budući da je smanjenje sposobnosti za samouređenjem povezano sa starošću i bolesti, gubitak vremenske povratnosti može se smatrati mjerom patologije. Analitički promatrano, nepovratnost vremena ukazuje na nedostatak nepromjenjivosti statističkih svojstava signala pod operacijom obrtanja vremena. Ireverzibilnost vremena je temeljno svojstvo sustava koji nisu u ravnoteži i stoga se može koristiti za opis patoloških srčanih stanja.

Neka je s $\{y_i\}$ označen vremenski niz promjena u vrijednosti RR-intervala, dakle:

$$y_i = x_{i+1} - x_i,\ i = 1,\ldots,N-1 \tag{3.67}.$$

Fiziološki promatrano, ovaj novi niz odražava natjecajući karakter između neuroautonomne stimulacije i stabilnosti sinoatrijalnog čvora.

Radi izlučivanja informacije na više skala, potrebno je grubo usitniti niz $\{y_i\}$:

$$y_\tau(i) = \frac{1}{\tau}\sum_{j=0}^{\tau-1} y_{i+j} \tag{3.68},$$

pri čemu je τ vremenska skala usitnjavanja. Uvodi se pojednostavljena pretpostavka da je svako ubrzanje ili usporenje srčanog ritma neovisno jedno o drugome i da zahtijeva određenu količinu energije *E*. Funkcija gustoće vjerojatnosti ovakvog sustava određena je s: $\rho \propto \exp(-\beta E - \gamma Q)$, pri čemu *Q* označava neuravnoteženi tok topline izvan granica sustava, a β i γ su Lagrangeovi multiplikatori. Operacija obrtanja vremena izvornog niza srčanih otkucaja invertira ubrzanje srčanog ritma u usporenje i obratno. Razlika u promjeni energije između $\rho(y_\tau > 0)$ i $\rho(y_\tau < 0)$ koristi se kao mjera asimetrije obrtanja smjera vremena. Kvantitativno, procjena indeksa asimetrije za konačno dugi vremenski niz srčanih otkucaja dana je izrazom:

$$\hat{A}(\tau) = \frac{\sum_{y(\tau)>0} \Pr(y_\tau)\ln[\Pr(y_\tau)]}{\sum_{y(\tau)} \Pr(y_\tau)\ln[\Pr(y_\tau)]} - \frac{\sum_{y(\tau)<0} \Pr(y_\tau)\ln[\Pr(y_\tau)]}{\sum_{y(\tau)} \Pr(y_\tau)\ln[\Pr(y_\tau)]} \tag{3.69},$$

pri čemu $\Pr(y_\tau)$ označava vjerojatnost pojave vrijednosti y_τ, koju se procjenjuje iz podataka. Za raspon vremenskih skala, definira se indeks asimetrije na više skala kao zbroj vrijednosti indeksa asimetrije na pojedinim skalama:

$$A_I = \sum_{\tau=1}^{L} \hat{A}(\tau) \tag{3.70},$$

pri čemu je *L* promatrani broj skala. Indeks vremenske asimetrije je najveći za mlade i zdrave ispitanike. Zdravi stariji ispitanici imali su značajno niži A_I od mladih, ali još uvijek viši nego kod srčanih bolesnika oboljelih od CHF ili AF. Opaženo je i da umjetno generirani niz srčanih otkucaja također ima niži A_I od niza kod mladih i zdravih ispitanika [Costa 2005 (2)].

3.5.3.17. Mjera središnje težnje

Mjera središnje težnje (engl. *central tendency measure*, kraće: CTM) određuje se na grafu drugog reda razlike. To je kvantitativna mjera koja određuje koliko su točke koncentrirane unutar ograničenog prostora blizu ishodišta u faznom prostoru drugog reda. Mjeru su prvi put definirali autori [Cohen M. 1994].

Neka je s $\{y_i\}$ označen vremenski niz promjena u vrijednosti RR-intervala kao u izrazu (3.67). Točka na grafu drugog reda razlike ima općenito u *m* dimenzija oblik $\bar{y}_n = (y_n \; y_{n+1} \ldots y_{n+m-1})$. Mjera središnje težnje određena je izrazom:

$$CTM = \sum_{i=1}^{N-m} \delta(d(t)) \tag{3.71},$$

pri čemu je:

$$\delta(d(t)) = \begin{cases} 1, & \text{za } \sqrt{y_n^2 + y_{n+1}^2 + \ldots + y_{n+m-1}^2} < r \\ 0, & \text{inace} \end{cases} \tag{3.72}.$$

Ovdje je *N* broj točaka vremenskog niza, a *r* radijus središnjeg područja. Tumačenje ove značajke je prilično jasno. Broje se samo one točke koje su koncentrirane dovoljno blizu ishodištu. Većina istraživača pritom promatra samo graf drugog reda razlike u dvije dimenzije. Time se mogu pratiti promjene srčanog ritma. Što su rjeđe veće promjene ritma, to je veća mjera CTM, jer je većina točaka u tom slučaju blizu ishodištu. Autori [Cohen M. 1996] pokazali su da je moguće za 100 000 točaka dobivenih iz 24-satnog mjerenja EKG-a značajno razlikovati između zdravih osoba i pacijenata oboljelih od CHF. Pacijenti oboljeli od CHF imali su niži CTM na skali $r = 0.1$ s od zdravih osoba, što znači da su promjene ritma kod CHF bile češće. Značajka CTM kasnije je uvedena i u analizu EEG-a gdje se koristi za klasifikaciju osoba oboljelih od shizofrenije [Jeong 2002, Hornero 2006], a primijenjena je uspješno i za otkrivanje poremećaja nedostatka daha prilikom spavanja koristeći podatke dobivene oksimetrijom pulsa [Alvarez 2007].

3.5.3.18. Rekurentni crtež

Rekurentni crtež je jedan od postupaka za mjerenje složenosti vremenskog niza u faznom prostoru [Eckmann 1987]. Koristi se za opis nestacionarnosti u nizu i to tako da kvantizira blizinu između točaka u vremenskom nizu. Točke koje su bliske nazivaju se rekurentnima (ponovno pojavljujućima) i označavaju se s 1 (ili s odgovarajućom bojom), a one koje nisu bliske označavaju se s 0.

Algoritam započinje oblikovanjem točaka niza u vektore dimenzije *m* s odmakom τ prema izrazu 3.25. Rekurentni crtež je simetrična kvadratna matrica A koja se ispunjava s nulama i jedinicama pri čemu je element matrice jednak:

$$a_{ij} = \begin{cases} 1, & d(\bar{x}_i, \bar{x}_j) \le r \\ 0, & inace \end{cases} \quad (3.73),$$

pri čemu je s $d(\cdot)$ označena Euklidska udaljenost između dvaju vektora, a s r radijus hipersfere u *m*-dimenzionalnom faznom prostoru. Dimenzija matrice A iznosi $\dim(A) = N - m + 1$, gdje je *N* broj mjerenja u vremenskom nizu.

Važna značajka matrice A je postojanje kratkih linijskih segmenata jedinica paralelnih s glavnom dijagonalom koji pokazuju da je niz točaka $\bar{x}_i, \bar{x}_{i+1}, \bar{x}_{i+2}, ..., \bar{x}_{i+k}$ blizak nizu točaka $\bar{x}_j, \bar{x}_{j+1}, \bar{x}_{j+2}, ..., \bar{x}_{j+k}$ što znači da postoji determinizam u nizu. Odsustvo ovakvih segmenata sugerira da je niz slučajan.

Iz rekurentnog crteža mogu se izdvojiti razne značajke za opis složenosti BVN. Izbor parametara $m, \tau,$ i r za ispravan izračun ovih značajki diskutiran je u detalje kod autora [Zbilut 2002]. Kao jednostavno pravilo, može se zaključiti da je dovoljno uzeti dovoljno visoku ugradbenu dimenziju faznog prostora m da se bude siguran da je atraktor ugrađen, ako takav postoji. Stoga se preporuča da m bude najmanje $m = 6$ ali poželjno $m = 10$. Odgoda τ je ovisna o vrsti nestacionarnosti, no u pravilu se može uzeti proizvoljna, dakle $\tau = 1$ je prihvatljiva odgoda. Bitan faktor za određivanje radijusa r je da postoji dovoljan broj rekurentnih točaka da ima smisla izračunavati značajke rekurentnog crteža. Najmanji broj rekurentnih točaka trebao bi biti 1%. Kao moguće empirijsko pravilo može se koristiti iznos $r = 0.2\sigma$, kao i za približnu entropiju.

Značajke koje se izlučuju iz rekurentnog crteža uključuju:

1. Mjeru rekurentnosti *REC*:

$$REC = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \sum_{j=1}^{N-m+1} a_{i,j} \quad (3.74).$$

2. Srednju vrijednost duljine kratkih linijskih segmenata paralelnih s dijagonalom:

$$l_{mean} = \sum_{l=l\min}^{l\max} l \cdot N_l \Big/ \sum_{l=l\min}^{l\max} N_l \quad (3.75),$$

pri čemu je s N_l označen broj linija duljine *l*.

3. Determinizam *DET*:

$$DET = \sum_{l=l\min}^{l\max} l \cdot N_l \Big/ \sum_{i=1}^{N-m+1} \sum_{j=1}^{N-m+1} a_{i,j} \quad (3.76).$$

4. Shannonovu entropiju razdiobe dijagonalnih linijskih segmenata rekurentnog crteža:

$$ShEn_R = \sum_{l=l\min}^{l\max} n_l \ln n_l, \quad n_l = N_l \Big/ \sum_{l=l\min}^{l\max} N_l \quad (3.77).$$

5. Laminarnost, kao mjeru ekvivalentnu mjeri DET (3.76), samo s vodoravnim (ili okomitim) linijskim segmentima.

3.5.3.19. Mjera složenosti Lempel-Ziv

Mjera složenosti Lempel-Ziv (engl. *Lempel-Ziv complexity*) služi za opis stupnja složenosti u vremenskom nizu kao i za procjenu koliko je najviše moguće sažeti podatke da bi se održala informacija sadržana u nizu bez gubitaka. Algoritam je namijenjen procjeni složenosti na kratkim vremenskim nizovima, neparametarski je i jednostavan za izračunati [Lempel 1976]. Algoritmom se prebrojavaju različiti podnizovi unutar zadanog niza i stopa njihovog ponavljanja.

Postupak počinje kodiranjem vrijednosti vremenskog niza u binarni kod. Pri tome se vrijednosti niza uspoređuju s medijanom niza koji se postavlja kao vrijednost praga T i slijedi:

$$y_i = \begin{cases} 1, & x_i \geq T \\ 0, & inace \end{cases} \tag{3.78}.$$

Niz se skenira jedan po jedan znak i složenost $c(n)$ se povećava za jedan svaki put kad se otkrije neki novi podniz. Tako npr. niz 011011110001 ima složenost jednaku $c(n) = 6$, budući da se pojavljuju sljedeći različiti podnizovi redom: 0, 1, 10, 11, 110, 001. Obično se računa normirana složenost u iznosu:

$$C(n) = c(n)\frac{\log_2 n}{n} \tag{3.79}.$$

Mjera složenosti Lempel-Ziv, kao mjera harmonične varijabilnosti vremenskog niza, našla je uspješnu primjenu u otkrivanju pojave prelaska VT u VF [Zhang X. 1999] kao i u istraživanjima funkcioniranja mozga i nizu drugih istraživanja [Hornero 2006].

3.6 Analiza obrazaca ritma

Analiza obrazaca ritma (engl. *rhythm pattern analysis*), često i analiza simboličke dinamike (engl. *symbolic dynamics*), promatra ubrzavanje i usporavanje srčanog ritma i simboličkim kodiranjem obrazaca promjena ritma nastoji pružiti bolji uvid u vrstu ritma. Ova grana analize vremenskih nizova potječe još od Hadamarda [Hadamard 1898]. U odnosu na linearnu vremensku, frekvencijsku, vremensko-frekvencijsku i nelinearnu analizu, simbolička analiza obrazaca ritma nije bila predmet značajnijeg broja istraživanja u analizi HRV. Prvi autori koji su uveli analizu simboličke dinamike za opis srčanih poremećaja bili su [Voss 1996]. Oni su istraživali kako se korištenjem dodatnih značajki simboličke dinamike mogu klasificirati pacijenti koji će doživjeti iznenadnu srčanu smrt (SCD). Značajniji članci u području simboličke analize obrazaca ritma su oni autora [Cysarz 2000, Yang 2003, Baumert 2005, Cysarz 2007 i Bogunović 2008].

Glavna ideja kod svih pristupa ove vrste je kodirati srčani ritam određenim postupkom koji će omogućiti da se razlikuje zdrav ritam od bolesnog, često na temelju udaljenosti između kodnih riječi. Za opis složenosti kodiranog niza koriste se i određene

već opisane nelinearne mjere. Kodiranje može također pod određenim uvjetima pružiti i bolji uvid u sam poremećaj ritma, omogućujući fiziološko tumačenje promjene ritmova.

Autori [Cysarz 2000] analizirali su dinamičku informaciju kratkih binarnih sekvenci. Vremenski niz srčanih otkucaja kodirali su binarno i to tako da znamenka 1 predstavlja produljenje, a znamenka 0 skraćenje otkucaja. RR-intervale izlučili su iz 118 24-satnih EKG-a zdravih osoba. Autori su koristili dvije entropijske mjere za tako kodirani signal. Prvo, korištena je ShEn nad vrlo kratkim sekvencama duljine pet, dobivenih iz deset-minutnih intervala. Rezultati su pokazali da je ShEn binarnih sekvenci manja ako je prosječan RR-interval duži. To znači da što je RR-interval duži, to je i manje ravnomjerna raspoređenost obrazaca. Ako je pak RR-interval kraći, tada su binarni obrasci ravnomjernije raspoređeni, odnosno slučajniji. Drugo, pravilnost kratkih binarnih sekvenci duljine pet analizirana je i ApEn-om. ApEn je imao linearnu ovisnost o duljini intervala, s više nepravilnosti koje se pojavljuju pri dužim RR-intervalima. Nepravilnost binarnog niza je pritom određena intuitivno, npr. niz 01010 je pravilniji od niza 01000, budući da se u prvom slijedu izmjenjuju 0 i 1. Vrijednost obje vrste entropija bile su konzistentne barem 48 sati kod ispitanika. Autori su zaključili da kvantifikacija binarnih sekvenci daje svojstva koja se ne mogu pronaći koristeći samo linearne i nelinearne značajke signala.

Autori [Yang 2003] također su proveli analizu HRV koristeći se binarnim kodiranjem vremenskog niza. Kodiranje je provedeno na isti način kao i kod [Cysarz 2000]. Formirane su riječi od osam znakova i tako kroz cijeli signal s pomakom od jednog znaka. Analizirane su mlade i zdrave osobe, stare i zdrave osobe, pacijenti s CHF i pacijenti s AF. Autori su radili usporedbu udaljenosti između riječi prisutnih unutar iste grupe pacijenata kao i unutar zapisa svakog pacijenta. Pokazuje se da je razlika u udaljenostima riječi unutar grupe i unutar pojedinog zapisa značajna svugdje osim za slučaj AF. Svi pacijenti koji imaju AF imaju međusobno podjednaku udaljenost, a i svaki pacijent zasebno ima podjednaku udaljenost riječi između dijelova zapisa. Pacijenti oboljeli od CHF ili AF također imaju više slučajni obrazac od zdravih osoba, ali obrasci nisu posve slučajni. Da bi se odredila udaljenost između riječi, kao i stupanj slučajnosti obrasca, autori koriste mjeru udaljenosti između dvije simboličkog niza definiranu izrazom:

$$D_m(S_1,S_2)=\frac{\sum_{k=1}^{2^m}\left|R_1(w_k)-R_2(w_k)\right|p_1(w_k)p_2(w_k)}{(2^m-1)\sum_{k=1}^{2^m}p_1(w_k)p_2(w_k)} \qquad (3.80),$$

pri čemu su s S_1 i S_2 označeni simbolički nizovi, $p_i(w_k)$ je vjerojatnost pojave, a $R_i(w_k)$ je rang specifične m - bitne riječi u pojedinom nizu. Težinski faktor 2^m-1 drži udaljenost između 0 i 1. Ako se ova mjera primijeni na udaljenost između izvornog niza i umjetno generiranog surogatnog niza (s ispremiješanim fazama u frekvencijskoj domeni) dobivaju se rezultati o stupnju slučajnosti. Autori su također posebno istaknuli da postoje četiri riječi koje se češće pojavljuju kod pacijenata s AF: (00100100), (00110001), (00101000) i (01000100), za što nije ponuđeno fiziološko objašnjenje.

U njihovom istraživanju iz 2005, autori [Baumert 2005] su analizirali udruženu simboličku dinamiku vremenskih nizova srčanih otkucaja i krvnog tlaka. Ideja je bila opisati međudjelovanje srčanih otkucaja i krvnog tlaka kod pacijenata oboljelih od I/D kardiomiopatije. Binarno kodiranje provedeno je s oba niza, a analizirane su udružene samo tri promjene u nizu otkucaja i tlaka. Vrednovala se razdioba simboličkih slijedova kao i DFA značajke α_1 i α_2 udruženog binarnog niza. Autori su pokazali na skupu od 75 zdravih osoba i 75 pacijenata da postoje značajne razlike i u razdiobi i u fraktalnim eksponentima. Zbog regulacije tlaka kod I/D kardiomiopatije, došlo je do gubitka u složenosti dinamike niza srčanih otkucaja.

Autori [Cysarz 2007] spominju da je zdrav srčani ritam složeniji od ritma koji imaju pacijenti oboljeli od CHF, međutim da se vrlo malo zna o složenosti varijacija na razini od otkucaja do otkucaja. Predložili su binarno kodiranje na zapisima od 30 zdravih osoba i 15 pacijenata (CHF). Promatrani su nizovi od osam intervala ubrzanja ili usporenja. Za analizu složenosti korišten je ApEn. Zdravi pacijenti su pokazali veliku količinu pravilnih obrazaca kao i značajnu količinu nepravilnih. Pacijenti koji imaju CHF pokazali su dominantnu učestalost pravilnog izmjenjujućeg niza 10101010 kao i neke nepravilne nizove. Kod zdravih osoba pravilni nizovi ukazuju na djelovanje simpatičke komponente ANS-a, a nepravilni nizovi na parasimpatičku komponentu. Kod CHF pacijenata prisutna je smanjena funkcija obje grane ANS-a.

Autori [Bogunović 2008] primijenili su novi postupak kvalitativne diskretizacije vremenskog niza (engl. *qualitative time-series discretization*, kraće: QTSD) da bi analizirali niz srčanih ritmova. Analizirani su nizovi od 20 zdravih osoba i 20 pacijenata s paroksizmalnom AF (PAF). Postupak QTSD najprije kodira niz srčanih otkucaja u ternarni sustav, u kojem se stagnacija ritma kodira s 0, ubrzanje s 1, a usporenje s 2. Nakon toga oblikuju se parovi (s_i, ns), pri čemu je s_i kodna oznaka (0, 1, ili 2), a ns je oznaka koliko se puta kodna oznaka ponavlja do sljedeće izmjene. Pobroje se svi parovi i te ih se transformira u dinamički broj stanja k putem empirijski određenog izraza:

$$k(s_i, ns) = s_i + 3 \cdot ns - 2 \tag{3.81}.$$

Ovakva definicija vremenskog niza stanja uključuje kvantitativnu informaciju o broju ponavljajućih stanja i kvalitativnu informaciju o vrsti promjene ritma. Nakon što je bio definiran vremenski niz stanja, provelo se mjerenje složenosti niza korištenjem ApEn, zajedničke informacije i Jensen-Shannonove divergencije (engl. *Jensen-Shannon divergence*, kraće: JSD).

Zajednička informacija je mjera informacijske složenosti i definirana je kao:

$$M = \sum_{i,j} \rho_{T,i\to j} \log\left(\frac{\rho_{T,i\to j}}{\rho_{ds,i}\, \rho_{ds,j}} \right) \tag{3.82},$$

pri čemu je $\rho_{T,i\to j}$ vjerojatnost prijelaza iz stanja i u stanje j, a $\rho_{ds,i}$ je vjerojatnost zauzimanja i-tog stanja: $\rho_{ds,i} = N_i / N_K$, gdje je N_i broj stanja i, a N_K je ukupan broj stanja [Fraser 1989(1)]. JSD je prilagodba Kullback-Leiberove divergencije da bi se u

histogramu izbjegle kućice s nula primjeraka [Lin 1991]. JSD je često korištena simetrična mjera sličnosti između signala. Dana je izrazom:

$$JSD = \frac{1}{2}\sum_y q(y)(\log q(y) - \log(avg(q,r))) + r(y)(\log r(y) - \log(avg(q,r)) \quad (3.83),$$

pri čemu je $q(y) = \rho(1)$ i $r(y) = \rho(2)$ i promatraju se svi prijelazi iz dotičnih stanja, dok je $avg(\cdot)$ prosječna vrijednost prijelaza.

Za razlikovanje zdravih pacijenata od pacijenata s AF-om najbolje rezultate dala je zajednička informacija, a dosta dobar rezultat dobiven je i putem JSD. ApEn nije bio toliko uspješan.

4 Izvorno uvedene značajke: abecedna entropija i napredna analiza slijednog trenda

4.1 Abecedna entropija

4.1.1 Teorijska pozadina postupka

Teorija informacijskog sustava, na način kako ju je predložio autor [Lerner 2007], zamišljena je tako da izgradi most između sustavnog matematičkog formalizma i informatičkih tehnologija koje se bave preobrazbom informacije iz jednog oblika u drugi. Cilj te teorije je da se dobiju matematički modeli sustava koji otkrivaju zakone informacije kao i točno određene obrasce po kojima se objekti promatranja ponašaju. U skladu s time, istraživače zanimaju načini kako točno modelirati sustav sa stajališta informacije, što se razlikuje od modeliranja sa stajališta materijala i energije.

Vrednovanje slučajnih podataka sa stajališta informacije u općem slučaju omogućeno je primjenom K-entropije (vidjeti poglavlje 3.5.3.13.). Kolmogorovljeva entropija mjeri stupanj kaotičnosti sustava putem informacijskog opisa faznog prostora. Za izračun entropije fazni prostor se dijeli na višedimenzionalne ćelije određene širine te se zatim određuje kolika je vjerojatnost nalaženja mjerenja u pojedinoj ćeliji. K-entropija određena je izrazom (3.64). Specijalan slučaj K-entropije je Shannonova entropija (ShEn), koja se izračunava izrazom (3.53) i kod koje se promatra prva potencija vjerojatnosti pojave točke u nekoj ćeliji. Također, može se promatrati i neka druga potencija vjerojatnosti pa je tako korelacijska entropija dana kvadratom vjerojatnosti:

$$K_2 = -\sum_{i=1}^{N-d} (p^2(\bar{x}_i) \log p^2(\bar{x}_i)) \tag{4.1},$$

pri čemu sumacija ide po vjerojatnostima pojave svih mjernih vektora $\bar{x}_i$ unutar neke ćelije u faznom dijagramu [Harikrisnhan 2009].

Drugačiji pristup izračunu entropije predložio je filozof Carnap 1977. godine. Carnapova teorija objašnjena je u radu autora [Pudmetzky 2005]. Ideja je da umjesto da se *d*-dimenzionalni fazni prostor R^d podijeli u proizvoljan broj od *k* ćelija s istim volumenom v^d, potrebno je odabrati fiksan broj mjerenja za svaku ćeliju, i to točno po jedno mjerenje. Svaka točka $\bar{x}_i$ u faznom dijagramu se tako pridružuje onom dijelu faznog prostora koji ju sadrži, a taj se dio naziva okolina e_i od točke $\bar{x}_i$. Volumeni v_i okoline bit će različiti od točke do točke. Okoline trebaju zadovoljiti sljedeće uvjete:

1) Svaka okolina e_i je podskup od R^d

2) Svaka okolina e_i ima volumen $v_i > 0$

3) Okoline različitih točaka se ne preklapaju.

4) Gotovo svaka točka (svaka točka s mogućom iznimkom onih točaka koje nemaju zasebnu okolinu) pripada jednoj od k okolina, dakle vrijedi: $\sum_{i=1}^{k} v_i = V^d$, gdje je V^d ukupni volumen faznog prostora R^d.

5) Kartezijeva udaljenost $D = \sqrt{\sum_{k=1}^{d} (x_k^i - x_k^j)^2}$ između dvaju točaka $\bar{x}_i$ i $\bar{x}_j$ treba biti $D > 0$, što znači da dvije točke ne mogu imati iste koordinate.

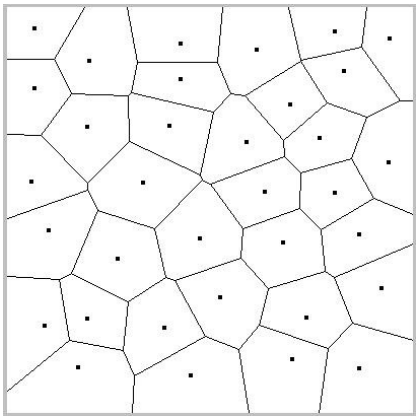
6) Okolina e_i fazne točke $\bar{x}_i$ je skup onih točaka X iz R^d, za koje je udaljenost između X i $\bar{x}_i$ manja nego za bilo koju drugu od $N-1$ točaka u faznom dijagramu.

Može se uočiti da volumen koji ovakva okolina zatvara nije pravokutan u smislu ćelije kakve je zamislio Kolmogorov i na temelju čega se izgradila čitava informacijska teorija. Točnije, u Carnapovom slučaju radi se o konveksnom politopnom obliku (poligonu u više dimenzija). Ovakav prostor kasnije je postao poznat kao Voronojev dijagram (engl. *Voronoi diagram*), u čast istraživaču Georgyju Voronoyju koji je prvi dao matematički opis sličnog prostora [Voronoi 1908]. Podjela prostora R^d u konveksne politope naziva se Voronojeva teselacija, što dolazi od glagola „teselirati", odnosno popločiti na način da nema razmaka između susjednih ploča.

Primjer Voronojevog dijagrama dan je na slici 4.1. Računarska složenost postupka teselacije u dvije dimenzije u općem slučaju je kvadratna $O(n^2)$, gdje je n broj točaka u faznom prostoru. Tek je 1986. predložen brži algoritam zasnovan na praćenju linija (engl. *line sweep*), sa složenosti $O(n \log n)$ koji je omogućio da se mogu konstruirati Voronojevi dijagrami i za razmjerno velik broj točaka (oko 10 000) [Fortune 1986].

Na Voronojevu dijagramu moguće je definirati i entropiju. To je učinio Carnap, koji ju je definirao kao:

$$En_C = \sum_{i=1}^{N} \log_2 v_i \qquad (4.2).$$



Slika 4.1. Voronojeva teselacija faznog prostora.

Za razliku od K-entropije, koja je informacijska mjera definirana na diskretiziranom prostoru, En_C je definirana izravno nad podacima u smislu njihove topologije. Stoga se za En_C koristi i naziv topološka entropija. Iako zanimljiva s teorijskog stajališta, En_C ne koristi se često u praksi, budući da se češće pribjegava diskretizacijskoj entropiji i to najčešće Shannonovog tipa.

Jedan od očitih razloga zašto se više koristi ShEn je brzina izračunavanja. Usprkos brzom algoritmu za izračun teselacije dvodimenzionalnog prostora, ShEn se još uvijek izračunava znatno brže, u $O(n)$ vremenu, budući da je podjela prostora provedena u jednake diskretne ćelije, a ne u konveksne politope.

Ipak, ShEn ima i nekoliko nedostataka u kontekstu numeričkih podataka u faznom prostoru:

1. ShEn definirana je samo nad diskretiziranim vrijednostima, tj. kodnim riječima. Numeričke vrijednosti mjerenja u faznom dijagramu potrebno je najprije pridijeliti određenoj ćeliji da bi se odredila vjerojatnost pojave kodne riječi određene pojedinom ćelijom. Tim se postupkom više mjernih podataka predstave jednom ćelijom, čime se gubi na detaljnosti i preciznosti koja je sadržana u točnim mjerenjima.

2. Prilikom diskretizacije, moguće je da neke ćelije ostanu prazne, što znači da neke kodne riječi nemaju niti jednog primjerka. S obzirom na to da ShEn nije definirana ako neka kodna riječ nema nijedan primjerak, potrebno je umjetno dodati najmanje jedan podatak u svaku praznu ćeliju da bi ShEn imala konačan iznos. Dodavanjem novih podataka stvara se pristranost deformira procjene informacijskog sadržaja ukupnog sustava. Što je praznih ćelija više, to je pristranost veća i odstupanje od točnosti je veće.

3. ShEn ne uzima u obzir dinamiku pojave kodnih znakova što znači da ne opisuje trajektoriju u faznom prostoru. Ona je statička mjera koja daje samo najmanji prosječan broj bitova za kodiranje pojedinog kodnog znaka.

Nedostatci ShEn ukazuju na to da bi bilo dobro pronaći bolju informacijsku mjeru za opis faznog prostora. Carnapova entropija, koju se ovdje razmatra, primjerenija je informacijska mjera za informacijski opis faznog prostora. Za razliku od ShEn, En_C uzima u obzir sama mjerenja time što uračunava volumene koje pojedinačne mjerenja zauzimaju u odnosu na sva ostala susjedna mjerenja. Ipak, En_C ima i dva bitna nedostatka koji će se riješiti u okviru teorije prikazane u ovom radu. To su:

1. En_C, slično kao i ShEn, nije osjetljiva na redoslijed mjerenja u faznom dijagramu. Nije bitno je li se prvo dogodio mjerni vektor $\bar{x}_i$ ili $\bar{x}_j$. Dokaz ovoj tvrdnji slijedi iz izraza (4.2), kod kojeg je očito da se jednaka vrijednost entropije dobije bez obzira na to koji se volumen prvi izračuna. To znači da je En_C primjenjiva na vremenski-neovisne nizove podataka. Budući da je kod BVN najbitnija dinamika kojom se odvijaju događaji, ovaj nedostatak potrebno je riješiti.

2. En_C ne dozvoljava da postoje dva mjerenja s jednakim koordinatama u faznom prostoru, što se vidi iz točke 5 uvjeta na okolinu. Problem kod mjerenja u stvarnim

vremenskim nizovima je konačna preciznost, i to često vrlo mala konačna preciznost, pri kojoj se može dogoditi da se dva mjerenja zaista nađu na istom mjestu.

U ovom radu uvodi se pojednostavljeni pristup Carnapovoj entropiji i to tako da se dimenzija faznog prostora postavlja na vrijednost jedan. To znači da se vrijednosti vremenskog niza promatraju takve kakve jesu, u jednoj vremenskoj dimenziji. Da bi se izračunala En_C, jedina operacija koju je potrebno napraviti je sortiranje podataka u vremenskom nizu prema rastućoj vrijednosti. Ovime se gubi dinamika, ali se omogućuje vrlo jednostavan izračun jednodimenzionalne topološke entropije:

$$En_C{}^{1D} = \sum_{i=1}^{k} \log_2 \frac{|d_i|}{x_k} \tag{4.3},$$

pri čemu je d_i dužina (okolina) točke x_i u jednoj dimenziji, slika 4.2.

Duljina dužine d_i iznosi:

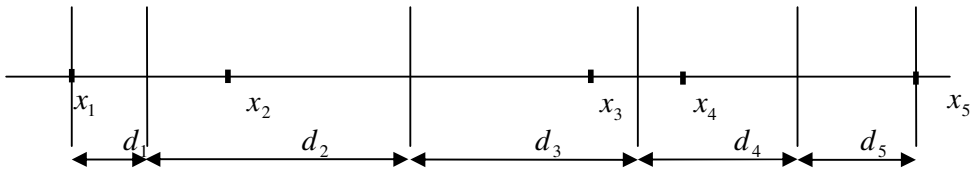
$$|d_i| = \frac{x_i - x_{i-1}}{2} + \frac{x_{i+1} - x_i}{2} = \frac{x_{i+1} - x_{i-1}}{2} \tag{4.4}.$$

Za prvu točku x_1 duljina dužine može se definirati kao: $|d_1| = x_1 + \frac{x_2 - x_1}{2} = \frac{x_1 + x_2}{2}$, a za zadnju točku ona iznosi: $|d_N| = \frac{x_N - x_{N-1}}{2}$. Pri tome se pretpostavlja da su vrijednosti svih točaka x_i pozitivne. Ako vrijednosti nisu pozitivne, tada je potrebno svim mjerenjima u vremenskom nizu dodati pozitivan broj $a \in \Re^+$, i to najmanje takav da vrijedi $x_1 := x_1 + a > 0$.

Iz (4.3) i (4.4) te uzimanjem u obzir graničnih slučajeva slijedi izraz za 1D Carnapovu entropiju niza sortiranih točaka:

$$En_C{}^{1D}(\bar{x}_i) = \log_2 \frac{x_1 + x_2}{2} + \sum_{i=2}^{N-1} \log_2 \frac{x_{i+1} - x_{i-1}}{2} + \log_2 \frac{x_N - x_{N-1}}{2} \tag{4.5}.$$

Ako postoje ista mjerenja u bilo kojim trenutcima u nekom vremenskom nizu, topološki niz u jednoj dimenziji sadržavat će dva ili više mjerenja na istom mjestu. U tom slučaju predlaže se da se doprinos svakog istog mjerenja broji zasebno. To znači da ako za neki *i* vrijedi: $x_i = x_{i+1} = \ldots = x_{i+k}$, pri čemu je *k* broj istih mjerenja, tada se doprinos ukupnoj entropiji računa kao: $k \log \frac{x_{i+k+1} - x_{i-1}}{2}$.



Slika 4.2. Prikaz podjele jednodimenzionalnog faznog prostora pri izračunu 1D Carnapove entropije, primjer za *k* = 5.

4.1.2 Definicija abecedne entropije

Promatrano principijelno, Carnapova entropija kao informacijska mjera opisuje dinamiku mjerenja samo zato što se izračunava u faznom prostoru. Fazni prostor sadrži mjerne vektore čija je prva komponenta trenutno mjerenje, druga komponenta sljedeće mjerenje, a zadnja komponenta mjerenje koje se dogodilo nakon $d-1$ vremenskih trenutaka, gdje je d dimenzija prostora. Ipak, slijed točaka pri izračunu Carnapove entropije nije bitan, kao što se može vidjeti iz izraza (4.2). Ako se iscrta trajektorija sustava u Voronojevu dijagramu tada se u faznom dijagramu prikazuje dinamika mjerenja. Ta dinamika se gubi za potrebe izračuna entropije, budući da se pri izračunu može promatrati samo topologija, ne i trajektorije. U vremenskim nizovima dinamika i svi njezini detalji su izuzetno bitni. Stoga se u ovoj disertaciji predlaže kombinirani pristup, koji će dinamiku modelirati na kvalitativno-kvantitativan način. Izračun En_C će se prilagoditi tako da se omogući kvantizacija dinamike u vremenskom nizu.

4.1.2.1. Delta-modulacija

Za kvalitativno vrednovanje dinamike vremenskog niza za kratke segmente koristit će se analiza vremenskog niza putem delta-promjena (dalje: Δ - modulacija). Kvalitativno promatrano, svaki vremenski niz može se predstaviti uz pomoć Δ - modulacije tako da ga se prikaže kao niz jediničnih promjena. Neka prvo mjerenje u nizu ima vrijednost x_1. Ako se u drugom mjerenju dogodi porast vrijednosti za neki jedinični iznos Δ, tada se govori o pozitivnoj Δ - modulaciji ($+\Delta$), ako dođe do pada vrijednosti, tada se govori o negativnoj Δ - modulaciji ($-\Delta$). U ovom radu koristit će se i treća mogućnost, i to takva kod koje nema zamjetne promjene, što je tzv. nulta Δ - modulacija (0).

Ovdje se predlaže da se za kvantitativni informacijski opis prilikom Δ - modulacija vremenskog niza koristi 1D Carnapova entropija uz jednu izmjenu. Naime, da bi se uzeo u obzir dinamički karakter vremenskog niza, potrebno je definirati slučaj u kojem nema nove informacije, što znači da tada i entropija ostaje ista. To je slučaj nulte Δ - modulacije. Uzimajući u obzir taj slučaj, izraz za Carnapovu entropiju potrebno je normirati na najveću vrijednost ovako:

$$I = En_C{}^{1D}\left(\frac{\bar{x}_i}{x_N}\right) = \log_2\frac{x_1+x_2}{2x_N} + \sum_{i=2}^{N-1}\log_2\frac{x_{i+1}-x_{i-1}}{2x_N} + \log_2\frac{x_N-x_{N-1}}{2x_N} \qquad (4.6).$$

Razlog zašto se ova normalizacija provodi je sljedeći. Ako postoji samo jedno mjerenje x_1 ukupna entropija treba biti jednaka nuli budući da se tek započelo s mjerenjem pa ne postoji ranija informacija o tome kakav je sustav bio. U slučaju da je u drugom mjerenju došlo do nulte Δ - modulacije, informacijski sadržaj vremenskog niza se ne mijenja i dalje ostaje jednak nula:

$I^{x1} = \log_2\frac{x_1}{x_1} = I^{x1+0} = 2\log_2\frac{x_1}{x_1} = 0$. Uz pomoć ovako normiranog izraza (4.6), sad je moguće točno odrediti informacijski sadržaj u slučaju pozitivne i negativne Δ -

modulacije. Ako je došlo do pozitivne Δ - modulacije, što je prikazano na slici 4.3a, tada je informacijski sadržaj jednak:

$$I^{+} = I^{x1+\Delta} = \log_2\left(\frac{x_1 + x_1 + \Delta}{2(x_1 + \Delta)}\right) + \log_2\left(\frac{x_1 + \Delta - x_1}{2(x_1 + \Delta)}\right) = \log_2\left(\frac{x_1 + \Delta/2}{x_1 + \Delta}\right) + \log_2\left(\frac{\Delta/2}{x_1 + \Delta}\right) =$$
$$= I^{x1+\Delta}\left(x_1 + \Delta/2\right) + I^{x1+\Delta}\left(\Delta/2\right) \qquad (4.7).$$

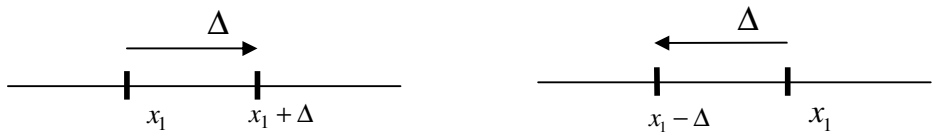
U slučaju negativne Δ - modulacije prikazane na slici 4.3b, informacijski sadržaj jednak je:

$$I^{-} = I^{x1-\Delta} = \log_2\left(\frac{x_1 - \Delta + x_1}{2x_1}\right) + \log_2\left(\frac{x_1 - x_1 + \Delta}{2x_1}\right) = \log_2\left(\frac{x_1 - \Delta/2}{x_1}\right) + \log_2\left(\frac{\Delta/2}{x_1}\right) =$$
$$= I^{x1-\Delta}\left(x_1 - \Delta/2\right) + I^{x1-\Delta}\left(\Delta/2\right) \qquad (4.8).$$

Nakon dvije Δ - modulacije, informacijski sadržaj bit će određen jednim od sljedećih četiriju izraza (samo je za prvi izraz dan potpuni raspis, slika 4.4):

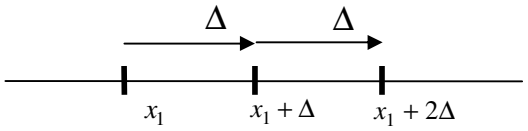
$$I^{++} = I^{x1+2\Delta} = \log_2\left(\frac{x_1 + x_1 + \Delta}{2(x_1 + 2\Delta)}\right) + \log_2\left(\frac{x_1 + 2\Delta - x_1}{2(x_1 + 2\Delta)}\right) + \log_2\left(\frac{x_1 + 2\Delta - x_1 - \Delta}{2(x_1 + 2\Delta)}\right) =$$
$$= \log_2\left(\frac{x_1 + \Delta/2}{x_1 + 2\Delta}\right) + \log_2\left(\frac{\Delta}{x_1 + 2\Delta}\right) + \log_2\left(\frac{\Delta/2}{x_1 + 2\Delta}\right) = I^{x1+2\Delta}\left(x_1 + \Delta/2\right) + I^{x1+2\Delta}(\Delta) + I^{x1+2\Delta}\left(\Delta/2\right)$$
$$I^{+-} = I^{x1+\Delta}\left(x_1 + \Delta/2\right) + 2I^{x1+\Delta}\left(\Delta/2\right)$$
$$I^{-+} = I^{x1}\left(x_1 - \Delta/2\right) + I^{x1}\left(\Delta/2\right)$$
$$I^{--} = I^{x1}\left(x_1 - 3/2\,\Delta\right) + I^{x1}(\Delta) + I^{x1}\left(\Delta/2\right) \qquad (4.9).$$

Informacijski sadržaj nakon tri Δ - modulacije prikazan je u tablici 4.1. Potrebno je zamijetiti da je mjera informacije nakon tri promjene jednaka za slučajeve 4 i 5.



a) Pozitivna Δ - modulacija b) Negativna Δ - modulacija

Slika 4.3. Δ - modulacije vremenskog niza.



Slika 4.4. Dvije pozitivne Δ - modulacije.

Tablica 4.1. Informacijski sadržaj niza nakon tri pozitivne ili negativne Δ - modulacije.

Slučaj	Δ - modulacija	Informacijski sadržaj
1.	I^{+++}	$I^{x1+3\Delta}(x_1+\Delta/2)+2I^{x1+3\Delta}(\Delta)+I^{x1+3\Delta}(\Delta/2)$
2.	I^{++-}	$I^{x1+2\Delta}\left(x_1+\Delta/2\right)+2I^{x1+2\Delta}(\Delta)+I^{x1+2\Delta}\left(\Delta/2\right)$
3.	I^{+-+}	$2I^{x1+\Delta}\left(x_1+\Delta/2\right)+2I^{x1+\Delta}\left(\Delta/2\right)$
4.	I^{+--}	$I^{x1+\Delta}\left(x_1-\Delta/2\right)+2I^{x1+\Delta}(\Delta)+I^{x1+\Delta}\left(\Delta/2\right)$
5.	I^{-++}	$I^{x1+\Delta}\left(x_1-\Delta/2\right)+2I^{x1+\Delta}(\Delta)+I^{x1+\Delta}\left(\Delta/2\right)$
6.	I^{-+-}	$2I^{x1}\left(x_1-\Delta/2\right)+2I^{x1}\left(\Delta/2\right)$
7.	I^{--+}	$I^{x1}\left(x_1-3/2\,\Delta\right)+2I^{x1}(\Delta)+I^{x1}\left(\Delta/2\right)$
8.	I^{---}	$I^{x1}\left(x_1-5/2\,\Delta\right)+2I^{x1}(\Delta)+I^{x1}\left(\Delta/2\right)$

Ako se u obzir uzima i mogućnost nulte Δ - modulacije, tada nakon tri Δ - modulacije postoji ukupno $3^3=27$ trajektorija koje se mogu informacijski opisati. Informacijski sadržaj u tom slučaju dan je u tablici 4.2.

Tablica 4.2. Informacijski sadržaj niza nakon Δ - modulacije koje uključuju i nultu modulaciju.

Slučaj	Δ - modulacija	Informacijski sadržaj
1.	I^{000}	$4I^{x1}(x_1)=0$
2.	I^{00+}	$3I^{x1+\Delta}\left(x_1+\Delta/2\right)+I^{x1+\Delta}(\Delta/2)$
3.	I^{00-}	$3I^{x1}\left(x_1-\Delta/2\right)+I^{x1}(\Delta/2)$
4.	I^{0+0}	$2I^{x1+\Delta}\left(x_1+\Delta/2\right)+2I^{x1+\Delta}(\Delta/2)$
5.	I^{0++}	$2I^{x1+2\Delta}\left(x_1+\Delta/2\right)+I^{x1+2\Delta}(\Delta)+I^{x1+2\Delta}(\Delta/2)$
6.	I^{0+-}	$3I^{x1+\Delta}\left(x_1+\Delta/2\right)+I^{x1+\Delta}(\Delta/2)$
7.	I^{0-0}	$2I^{x1}\left(x_1-\Delta/2\right)+2I^{x1}(\Delta/2)$
8.	I^{0-+}	$I^{x1}\left(x_1-\Delta/2\right)+3I^{x1}(\Delta/2)$
9.	I^{0--}	$I^{x1}\left(x_1-3/2\Delta\right)+I^{x1}(\Delta)+2I^{x1}(\Delta/2)$
10.	I^{+00}	$I^{x1+\Delta}\left(x_1+\Delta/2\right)+3I^{x1+\Delta}(\Delta/2)$
11.	I^{+0+}	$I^{x1+2\Delta}\left(x_1+\Delta/2\right)+2I^{x1+2\Delta}(\Delta)+I^{x1+2\Delta}(\Delta/2)$
12.	I^{+0-}	$2I^{x1+\Delta}\left(x_1+\Delta/2\right)+2I^{x1+\Delta}(\Delta/2)$
13.	I^{++0}	$I^{x1+2\Delta}\left(x_1+\Delta/2\right)+I^{x1+2\Delta}(\Delta)+2I^{x1+2\Delta}(\Delta/2)$
14.	I^{+++}	$I^{x1+3\Delta}(x_1+\Delta/2)+2I^{x1+3\Delta}(\Delta)+I^{x1+3\Delta}(\Delta/2)$
15.	I^{++-}	$I^{x1+2\Delta}(x_1+\Delta/2)+2I^{x1+2\Delta}(\Delta)+I^{x1+2\Delta}(\Delta/2)$
16.	I^{+-0}	$3I^{x1+\Delta}\left(x_1+\Delta/2\right)+I^{x1+\Delta}(\Delta/2)$
17.	I^{+-+}	$2I^{x1+\Delta}(x_1+\Delta/2)+2I^{x1+\Delta}(\Delta/2)$
18.	I^{+--}	$I^{x1+\Delta}(x_1-\Delta/2)+2I^{x1+\Delta}(\Delta)+I^{x1+\Delta}(\Delta/2)$
19.	I^{-00}	$3I^{x1}\left(x_1-\Delta/2\right)+I^{x1}(\Delta/2)$
20.	I^{-0+}	$2I^{x1}\left(x_1-\Delta/2\right)+2I^{x1}(\Delta/2)$

21.	I^{-0-}	$I^{x1}(x_1 - 3/2\Delta) + 2I^{x1}(\Delta) + I^{x1}(\Delta/2)$
22.	I^{-+0}	$I^{x1}(x_1 - \Delta/2) + 3I^{x1}(\Delta/2)$
23.	I^{-++}	$I^{x1+\Delta}(x_1 - \Delta/2) + 2I^{x1+\Delta}(\Delta) + I^{x1+\Delta}(\Delta/2)$
24.	I^{-+-}	$2I^{x1}(x_1 - \Delta/2) + 2I^{x1}(\Delta/2)$
25.	I^{--0}	$2I^{x1}(x_1 - 3/2\Delta) + I^{x1}(\Delta) + I^{x1}(\Delta/2)$
26.	I^{--+}	$I^{x1}(x_1 - 3/2\Delta) + 2I^{x1}(\Delta) + I^{x1}(\Delta/2)$
27.	I^{---}	$I^{x1}(x_1 - 5/2\Delta) + 2I^{x1}(\Delta) + I^{x1}(\Delta/2)$

U tablici 4.2 također se mogu zamijetiti jednake informacijski vrijednosti pojedinih slučajeva. Tako su slučajevi: 2, 6 i 16; 3 i 19; 4, 12 i 17; 7 i 20; 8 i 22; 11 i 15; 18 i 23; te 21 i 26 jednaki. Ukupno od 27 slučajeva njih samo osam imaju jedinstvenu informacijsku vrijednost.

4.1.2.2. Poopćenje na delta-matricu i izgradnja delta-retka

Da bi se svaka vrsta promjene u vremenskom nizu jedinstveno identificirala, nužno je imati nedvosmislenu mjeru informacijskog sadržaja. Da bi se ovaj problem nejedinstvenosti informacijskog sadržaja riješio predlaže se poopćenje Δ - modulacije vremenskog niza u Δ - matricu. Sličan pristup modeliranju dinamike vremenskog niza predložili su autori [Jagnjić 2009]. Autori su odredili postupak modeliranja vremenskog niza koji kvantizira razlike u mjerenjima s konačnim ciljem za što točnijom prognozom ciljne značajke. Ciljna značajka se modelirala putem utežane kombinacije izvornih prediktorskih značajki i izvedenih prediktorskih značajki na temelju razlika vrijednosti u vremenskom nizu.

U ovom radu cilj je definirati nedvosmislenu informacijsku mjeru za vrednovanje dinamike vremenskog niza. Stoga se ovdje kreće od trokutaste Δ - matrice razlika predložene od autora [Jagnić 2009] i nadograđuje ju se informacijskim opisom.

Neka je dan vremenski niz podataka $\{x_1, x_2, \ldots, x_N\}$. Trokutasta matrica razlika u nizu podataka određena je izrazom:

$$M_d = \begin{bmatrix} d_{x1,x2} & d_{x2,x3} & \cdots & d_{x(N-1),xN} \\ \cdot & d_{x1,x3} & \cdots & d_{x(N-2),xN} \\ \vdots & \cdot & \ddots & \vdots \\ \cdot & \cdot & \ldots & d_{x1,xN} \end{bmatrix} \quad (4.10),$$

pri čemu je s $d_{xi,xj}$ označena razlika: $d_{xi,xj} = x_j - x_i$. Matrica razlika sadrži diferencijale do uključivo reda $N-1$. Doljnji trokut u matrici nije bitan.

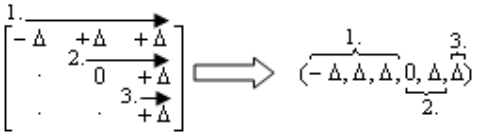
Slično se može promatrati i Δ - modulacija. Ako se promatraju samo tri Δ - modulacije kao što je prikazano u tablici 4.1, tada je moguće svaki od slučajeva raspisati u obliku Δ - matrica dimenzije 3. Prvi redak Δ - matrice sastoji se od tri Δ - modulacije između susjednih mjerenja, zatim slijede dvije Δ - modulacije koje su se dogodile između dvaju mjerenja koje su udaljene međusobno za jedno mjerenje i na kraju, zadnja komponenta Δ - retka je Δ - modulacija između četvrtog i prvog mjerenja. Ovdje su

dani primjeri dviju Δ - matrica za slučajeve kod kojih je informacijski sadržaj nakon tri Δ - modulacije bio jednak, dakle za slučajeve 4 i 5 iz tablice 4.1 (slučajevi 18 i 23 u tablici 4.2):

$$M_\Delta{}^{+--} = \begin{bmatrix} +\Delta & -\Delta & -\Delta \\ \cdot & 0 & -\Delta \\ \cdot & \cdot & -\Delta \end{bmatrix},\ M_\Delta{}^{-++} = \begin{bmatrix} -\Delta & +\Delta & +\Delta \\ \cdot & 0 & +\Delta \\ \cdot & \cdot & +\Delta \end{bmatrix} \qquad (4.11).$$

Sad je moguće iz redaka Δ - matrice oblikovati novi Δ - redak u kojem je sadržana informacija o svim Δ - modulacijama između mjerenja u vremenskom nizu. Izgradnja Δ - retka iz Δ - matrice prikazana je na slici 4.5 (znak + ispred Δ se ispušta).

Kraće će se Δ - redak označavati kao šestorka kod koje se bilježe pozitivne, negativne i nulte Δ - modulacije, npr. Δ - redak sa slike 4.5 kraće se zapisuje kao $(-++0++)$. Postupak proširenja tri Δ - modulacije u Δ - redak moguće je obaviti za sve slučajeve navedene u tablici 4.2. Pritom se ne uzima u obzir početni iznos x_1 pri određivanju informacijskog sadržaja. Bitan je samo Δ - redak, odnosno šest Δ - modulacija. Pokazuje se da je informacijski sadržaj u tom slučaju različit za sve slučajeve. Jedinstveni slučajevi će se odsad nazivati kodni znakovi abecede i označavat će se slovima od A do AA (oznaka „AA“ koristi se za 27. znak abecede). Informacijski sadržaj za sve kodne znakove abecede dan je u tablici 4.3.



Slika 4.5. Prijelaz iz Δ - matrice u Δ - redak.

Tablica 4.3. Jedinstveni informacijski sadržaj kodnih riječi kao rezultat proširenja Δ - modulacija u Δ - redak.

Kodni znak	Δ - modulacija	Δ - redak	Informacijski sadržaj
A	I^{000}	(000000)	$6I^{x1}(x_1) = 0$
B	I^{00+}	(00+0++)	$2I^{x1+3\Delta}(x_1 + \Delta/2) + 3I^{x1+3\Delta}(\Delta) + I^{x1+3\Delta}(\Delta/2)$
C	I^{00-}	(00−0−−)	$I^{x1}(x_1 - 5/2\Delta) + 3I^{x1}(\Delta) + 2I^{x1}(\Delta/2)$
D	I^{0+0}	(0+0+++)	$I^{x1+4\Delta}(x_1 + \Delta/2) + 4I^{x1+4\Delta}(\Delta) + I^{x1+4\Delta}(\Delta/2)$
E	I^{0++}	(0++++++)	$I^{x1+5\Delta}(x_1 + \Delta/2) + 4I^{x1+5\Delta}(\Delta) + I^{x1+5\Delta}(\Delta/2)$
F	I^{0+-}	(0+−+00)	$2I^{x1+\Delta}(x_1 + \Delta/2) + 4I^{x1+\Delta}(\Delta/2)$
G	I^{0-0}	(0−0−−−)	$I^{x1}(x_1 - 7/2\Delta) + 4I^{x1}(\Delta) + I^{x1}(\Delta/2)$
H	I^{0-+}	(0−+−00)	$4I^{x1}(x_1 - \Delta/2) + 2I^{x1}(\Delta/2)$
I	I^{0--}	(0−−−−−)	$I^{x1}(x_1 - 9/2\Delta) + 4I^{x1}(\Delta) + I^{x1}(\Delta/2)$
J	I^{+00}	(+00+0+)	$3I^{x1+3\Delta}(x_1 + 3/2\Delta) + 2I^{x1+3\Delta}(\Delta) + I^{x1+3\Delta}(\Delta/2)$
K	I^{+0+}	(+0++++)	$2I^{x1+5\Delta}(x_1 + \Delta/2) + 3I^{x1+5\Delta}(\Delta) + I^{x1+5\Delta}(\Delta/2)$
L	I^{+0-}	(+0−+−0)	$3I^{x1+\Delta}(x_1 + \Delta/2) + 3I^{x1+\Delta}(\Delta/2)$

M	I^{++0}	$(++0+++)$	$I^{x1+5\Delta}(x_1+3/2\Delta)+4I^{x1+5\Delta}(\Delta)+I^{x1+5\Delta}(\Delta/2)$
N	I^{+++}	$(+++++++)$	$I^{x1+6\Delta}(x_1+3/2\Delta)+4I^{x1+6\Delta}(\Delta)+I^{x1+6\Delta}(\Delta/2)$
O	I^{++-}	$(++-+0+)$	$2I^{x1+3\Delta}(x_1+3/2\Delta)+3I^{x1+3\Delta}(\Delta)+I^{x1+3\Delta}(\Delta/2)$
P	I^{+-0}	$(+-00-0)$	$2I^{x1+\Delta}(x_1+\Delta/2)+3I^{x1+\Delta}(\Delta)+I^{x1+\Delta}(\Delta/2)$
Q	I^{+-+}	$(+-+00+)$	$I^{x1+2\Delta}(x_1+\Delta/2)+4I^{x1+\Delta}(\Delta)+I^{x1+\Delta}(\Delta/2)$
R	I^{+--}	$(+--0--)$	$I^{x1+\Delta}(x_1-5/2\Delta)+4I^{x1+\Delta}(\Delta)+I^{x1+\Delta}(\Delta/2)$
S	I^{-00}	$(-00-0-)$	$3I^{x1-\Delta}(x_1-5/2\Delta)+2I^{x1-\Delta}(\Delta)+I^{x1-\Delta}(\Delta/2)$
T	I^{-0+}	$(-0+-+0)$	$3I^{x1}(x_1-\Delta/2)+3I^{x1}(\Delta/2)$
U	I^{-0-}	$(-0----)$	$I^{x1-\Delta}(x_1-9/2\Delta)+3I^{x1-\Delta}(\Delta)+2I^{x1-\Delta}(\Delta/2)$
V	I^{-+0}	$(-+00+0)$	$I^{x1+\Delta}(x_1-\Delta/2)+3I^{x1+\Delta}(\Delta)+2I^{x1+\Delta}(\Delta/2)$
W	I^{-++}	$(-++0++)$	$I^{x1+3\Delta}(x_1-\Delta/2)+4I^{x1+3\Delta}(\Delta)+I^{x1+3\Delta}(\Delta/2)$
X	I^{-+-}	$(-+-00-)$	$I^{x1}(x_1-3/2\Delta)+4I^{x1}(\Delta)+I^{x1}(\Delta/2)$
Y	I^{--0}	$(--0---)$	$I^{x1-\Delta}(x_1-9/2\Delta)+4I^{x1-\Delta}(\Delta)+I^{x1-\Delta}(\Delta/2)$
Z	I^{--+}	$(--+-0-)$	$I^{x1-\Delta}(x_1-5/2\Delta)+3I^{x1-\Delta}(\Delta)+2I^{x1-\Delta}(\Delta/2)$
AA	I^{---}	$(------)$	$I^{x1-\Delta}(x_1-11/2\Delta)+4I^{x1-\Delta}(\Delta)+I^{x1-\Delta}(\Delta/2)$

Cijela abeceda opisuje 27 različitih mogućnosti promjene u vremenskom nizu tijekom četiri slijedna mjerenja (tri slijedne promjene). Kodni znak sam po sebi daje kvalitativnu informaciju o promjenama u nizu. Kvantitativnu mjeru promjene može se u slučaju promjena uvijek jednakih iznosa (Δ- modulacija) točno izračunati pomoću formula danih u tablici 4.3. Vremenski nizovi u praksi vrlo rijetko sadržavaju promjene mjerenja uvijek istih jediničnih iznosa. Stoga je izračun entropije potrebno prilagoditi za praktičnu primjenu.

4.1.2.3. Algoritam za određivanje kodnog znaka i izračun abecedne entropije općenitog vremenskog niza

---

Algoritam AbEn

1. Neka je zadana četvorka mjerenja iz vremenskog niza $X = (x_1, x_2, x_3, x_4),\ x_i \in \Re$
2. Kodni znak *K* određuje se tako da se odrede predznaci triju razlika:
$$s_1 = sign(x_2 - x_1),\ s_2 = sign(x_3 - x_2),\ s_3 = sign(x_4 - x_3),\ s_i \in \{+, -, 0\},$$
$$K = (s_1\ s_2\ s_3),\quad K \in \{A, B, \ldots, AA\} \qquad (4.12).$$
3. Iz četvorke *X* gradi se šestorka $Y = (y_1, y_2, y_3, y_4, y_5, y_6)$ takva da vrijedi:
$$y_1 = x_2,\ y_2 = x_3,\ y_3 = x_4,\ y_4 = y_3 + x_3 - x_1,\ y_5 = y_4 + x_4 - x_2,\ y_6 = y_5 + x_4 - x_1 \qquad (4.13).$$
Time se pamti informacija dana u prvom, drugom i trećem diferencijalu uz dodavanje apsolutnog iznosa mjerenja, npr. $y_1 = x_1 + (x_2 - x_1) = x_2$.
4. Šestorka *Y* se sortira po rastućem redoslijedu čime se dobiva šestorka
$Z = (z_1, z_2, z_3, z_4, z_5, z_6), z_i \le z_j, \forall i < j$.
Sortiranje se provodi kako bi se omogućilo jednoznačno računanje izmijenjene jednodimenzionalne Carnapove entropije. Ako vrijedi $z_1 \le 0$, tada $z_i := z_i - z_1 + \varepsilon, \varepsilon > 0$. Sada sigurno vrijedi $z_i \in \Re^+, \forall i$ i stoga nema bojazni da će se pokušati izračunati logaritam negativnog broja.
5. Abecedna entropija definirana je izrazom:
$$AbEn = -\left(\log_2 \frac{z_1 + z_2}{2z_6} + \sum_{i=2}^{5} \log_2 \frac{z_{i+1} - z_{i-1}}{2z_6} + \log_2 \frac{z_6 - z_5}{2z_6} \right),\ \text{[bit/niz]} \qquad (4.14)$$
ili u obliku produkata:
$$AbEn = -\left(\log_2 \left((z_1 + z_2)(z_6 - z_5) \prod_{i=2}^{5} (z_{i+1} - z_{i-1}) \right) - 6(1 + \log_2 z_6) \right),\ \text{[bit/niz]} \qquad (4.15).$$
Znak negacije dodaje se ispred logaritama da bi se uvijek dobio pozitivan iznos entropije.
6. Ako postoje neke vrijednosti unutar šestorke *Z* koje su jednake, tada postoji mogućnost dobivanja nule unutar logaritama pa se logaritam ne može izračunati. Može se pokazati da jedini slučajevi koji nisu problematični za izračun entropije su oni kod kojih su samo vrijednosti z_2, z_3 i/ili z_4, z_5 jednake. Samo u tim slučajevima i u slučaju da su sve vrijednosti različite mogu se koristiti izrazi (4.14) i (4.15) za izračun entropije. U svim ostalim slučajevima pojave jednakosti vrijednosti, određeni se logaritmi u izrazu (4.14) uzimaju u obzir više puta, dok se neki drugi zanemare. Tako izraz (4.14) npr. za šestorku $Z = (2, 2, 5, 5, 5, 8)$ iznosi:
$$AbEn = -2\log_2 \frac{2+5}{2 \cdot 8} - 3\log_2 \frac{8-2}{2 \cdot 8} - \log_2 \frac{8-5}{2 \cdot 8} = 9.045\ \text{[bit/niz]}.$$
Na ovaj način izbjegao se nedostatak jednakih točaka koji je bio drugi od dva temeljna nedostatka Carnapove entropije.

Prema klasifikaciji, ovako definirana abecedna entropija pripala bi skupini postupaka simboličke dinamike (analiza obrazaca ritma, poglavlje 3.6), s time da se AbEn može prilagoditi za razvrstavanje. Predložena informacijska teorija razvijena za analizu četiri slijedna mjerenja (tri slijedne promjene) može otkriti samo kratkodjelujuće promjene u vremenskom nizu.

4.1.2.4. Razmatranje abecedne entropije (AbEn)

Shannonova entropija daje najmanji prosječan broj bitova za kodiranje pojedinog znaka abecede. Kodni znak ShEn određen je samo svojom frekvencijom pojavljivanja u određenoj razdiobi. Abecedna entropija izračunava se za svaki znak posebno. Kodni znak kod AbEn daje kvalitativnu informaciju o tome kakva je unutrašnja dinamika tijekom četiri pojedinačna mjerenja. Opis dinamike uključuje prvi, drugi i treći diferencijal. AbEn jednog znaka kvantificira tu dinamiku pa se stoga može koristiti za pronalaženje lokalnih područja u vremenskom nizu koja imaju izraženiju dinamiku. Abecedna entropija je osjetljiva na male promjene od uobičajenih vrijednosti vremenskog niza budući da svaki povećani diferencijal unutar prozora od četiri mjerenja utječe na iznos entropije. Ova vrsta osjetljivosti nije pogodna za sustave kod kojih su promjene česte i velike.

Ako se AbEn određuje na vremenskom segmentu od N mjerenja, $N > 4$, tada je moguće provesti analizu kodne riječi. Kodna riječ sastoji se od kodnih znakova abecede, npr. "AABZXAD" za N=10. Moguće je odrediti udio pojave određenog kodnog znaka u kodnoj riječi. Taj udio jednak je broju pojavljivanja kodnog znaka podijeljenim s ukupnim brojem kodnih znakova u kodnoj riječi.

Za kodnu riječ moguće je izračunati i ShEn uz uvjet da su se svi kodni znakovi abecede pojavili barem jedanput u nekom segmentu.

Za kvantitativni opis vremenskog segmenta mogu se koristiti statističke mjere srednje vrijednosti, varijance, i najveće vrijednosti, s time da ih se primjenjuje na iznose abecednih entropija pojedinačnih znakova. Neka je s $AbEn_i$ označena i-ta vrijednost AbEn nekog vremenskog segmenta. Tada su određene sljedeće mjere:

$$\overline{AbEn} = \frac{1}{N-3}\sum_{i=1}^{N-3} AbEn_i,\ s^2_{AEn} = \frac{1}{N-4}\sum_{i=1}^{N-3}\left(AbEn_i - \overline{AbEn}\right)^2,\ \max(AbEn) = \max_i(AbEn_i) \tag{4.16}$$

Pritom se pretpostavlja da se za vremenski segment duljine N računa ukupno $N-3$ abecedne entropije pojedinih znakova.

4.1.3 Razmatranje primjene abecedne entropije

Abecedna entropija omogućuje analizu promjena koje su se dogodile tijekom četiri slijedna mjerenja u vremenskom nizu. S obzirom na to da srčani ritam može imati

poremećaje koji se očituju u promjenama trajanja otkucaja, AbEn se pokazuje prikladnom za opis ovih promjena.

Teorijski promatrano, svaki kodni znak abecede neće biti koristan za opis nekog poremećaja ritma. Karakteristika svakog pojedinog poremećaja moći će se na najbolji način opisati samo s nekim kodnim znakovima i cjelokupna abeceda barem u teoriji neće biti pokrivena.

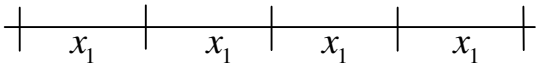
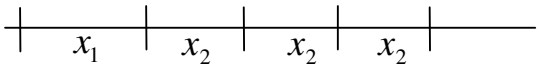
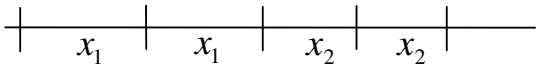
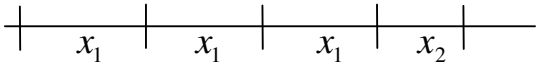
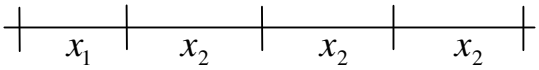
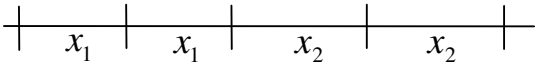
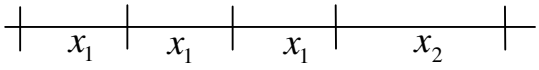
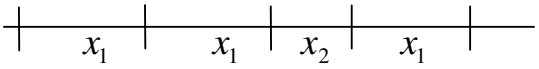
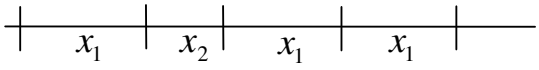
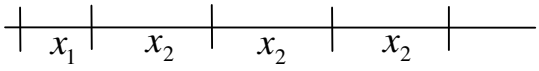
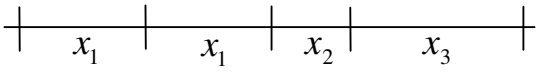
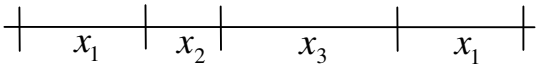
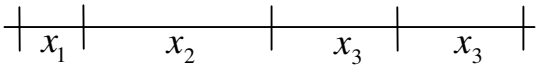
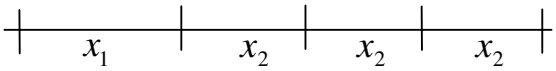
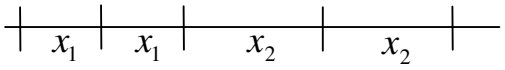
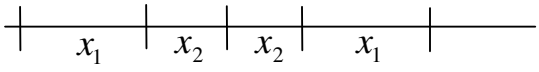
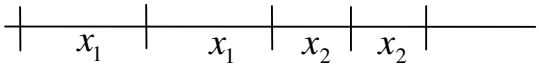
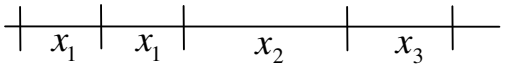
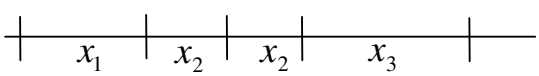
Neki poremećaji ritma protežu se samo kroz jednu promjenu trajanja otkucaja, neki utječu na dvije promjene, a neki na sve tri promjene. Sve takve poremećaje moguće je otkriti pomoću analize AbEn. Postoje i oni poremećaji ritma koji se odnose i na više od tri promjene i koji nisu značajni u praksi (npr. kvadrigeminija) kao i oni poremećaji koji tipično traju mnogo dulje od četiri mjerenja i jesu značajni (npr. supraventrikularna ili ventrikularna tahikardija). AbEn nije namijenjena za otkrivanje takvih vrsta poremećaja.

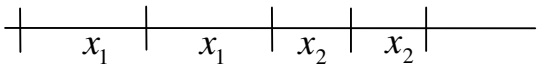
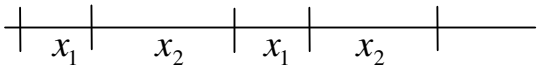
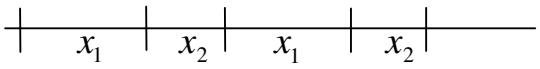
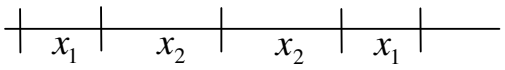
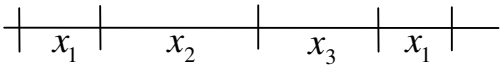
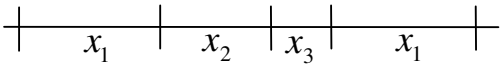
Pretraživanje vremenskog niza radi određivanja AbEn ide slijedno, četiri po četiri mjerenja. Time se neki poremećaj prvo pojavljuje pri kraju prozora od četiri mjerenja, zatim u sredini i na kraju na početku. U slučaju da je poremećaj moguće otkriti samo promatrajući sva četiri mjerenja, tada se on pojavljuje tek kad ga prozor potpuno prekrije, a dotad se može maskirati kao neki drugi poremećaj (kao što je slučaj kod trigeminije). Budući da nije moguće unaprijed znati koji dio kojeg poremećaja se promatra, potrebno je napraviti mapu od svih mogućih abecednih znakova za sve poremećaje ritma tako da se vidi koji znakovi su specifični za pojedine poremećaje. Ta mapa navedena je u tablici 4.4 za one vrste ritmova koji se u teoriji mogu otkriti koristeći analizu AbEn.

Postoji više znakova abecede koji su nespecifični za određivanje poremećaja, jer se mogu pojaviti i kod raznih poremećaja. Tako je npr. znak „C (00-)“ nespecifičan, budući da može biti rezultat normalnog ubrzanja ritma, početka PAC-a, PVC-a, kupleta, bigeminije, trigemenije, itd. Slično, znak „B (00+)“ je također nespecifičan, budući da se može pojaviti kod normalnog usporenja ritma, ali i kod bloka drugog stupnja. Sami pojedinačni znakovi za sebe često nisu specifični ni za jedan poremećaj. Ono što se međutim može uočiti, to je da su kombinacije nekoliko znakova vrlo specifične za pojedine poremećaje. U tablici 4.5 navode se tipične kombinacije znakova koje su specifične za pojedine poremećaje, kao i neke koje su bezopasne i najvjerojatnije su posljedica normalnog srčanog ritma.

Potrebno je napomenuti da se nepravilan ritam atrijalne fibrilacije (AF) može maskirati kao bilo koji znak, stoga ga je teško, ali ne i nemoguće otkriti. Naime, karakteristika AF je velika različitost pojave kodnih znakova. Stoga se kod segmenata koji imaju razne kombinacije znakova koji nisu oni navedeni u tablici 4.5 treba posumnjati u AF. Često razlike između atrijalnih i ventrikularnih ektopičnih otkucaja (PAC i PVC) nisu jasne na temelju samog ritma. Abecedna entropija može međutim dati bolju indikaciju o kojem se poremećaju radi budući da osim samih kodnih znakova računa i iznose promjena. Veće promjene koje postoje pri ventrikularnim ektopičnim otkucajima davat će češće veći iznos abecedne entropije od atrijalnih.

Tablica 4.4. Normalan srčani ritam i poremećaji ritma prikazani kodnim znakovima abecede.

Vrsta ritma	Kodni znakovi	Obrazac ritma
Normalan srčani ritam (bez promjena) ili umjetni pejsmejker	A (000)	
Normalan srčani ritam (ubrzanje)	S (-00)	
	G (0-0)	
	C(00-)	
Normalan srčani ritam (usporenje)	J (+00)	
	D (0+0)	
	B (00+)	
Preuranjena kontrakcija atrija (PAC)	H (0-+)	
	V (-+0)	
	J (+00)	
Preuranjena kontrakcija ventrikula (PVC)	H (0-+)	
	X (-+-)	
	P (+-0)	
	S (-00)	
Kuplet (bez kompenzacije)	D (0+0)	
	T (-0+)	
	G (0-0)	
Kuplet (s kompenzacijom)	F (0+-)	
	T (-0+)	
	G (0-0)	

		
Bigeminija (atrijalna ili ventrikularna)	Q (+-+)	
	X (-+-)	
Atrijalna trigeminija	L (+0-)	
Ventrikularna trigeminija	R (+--)	
	Z (--+)	

Tablica 4.5. Kombinacije kodnih znakova koje ukazuju na pojedinu vrstu ritma.

Kombinacija kodnih znakova	Mogući ritam	Normalno / abnormalno
A	normalan / sinusna tahikardija / sinusna bradikardija	normalno / normalno / normalno ili abnormalno
C, G, S	normalan, jedno ubrzanje	normalno
B, D, J	normalan, jedno usporenje	normalno
I, Y, S	normalan, dva ubrzanja zaredom, sinusna aritimija	normalno
E, M, J	normalan, dva usporenja zaredom, sinusna aritmija	normalno
U, G, S	normalan, prvo usporenje, na kraju ubrzanje, sinusna aritmija	normalno
L, G, S	normalan, drugo usporenje, na kraju ubrzanje, sinusna aritmija	normalno
K, D, J (ili K)	normalan, drugo usporenje, na kraju opet usporenje, sinusna aritmija	normalno
H, V, J	PAC	abnormalno
H, X, P	PVC	abnormalno
D, T, G	Kuplet (bez kompenzacije)	abnormalno
F, T, G	Kuplet (s kompenzacijom)	abnormalno
Q, X, Q (ili V)	Bigeminija (ako je kraj atrijalne bigeminije onda znak V)	abnormalno
L, H, V	Atrijalna trigeminija	abnormalno
R, Z, X	Ventrikularna trigeminija	abnormalno
B, F, Q; P, T, E, itd.	Atrijalna fibrilacija	abnormalno

Ako segmenti imaju više od jednog poremećaja (npr. PAC i kuplet s kompenzacijom), tada će analiza AbEn otkriti više vrsta poremećaja. Ipak, jedan se poremećaj s medicinske strane uvijek smatra dominantnim u nekom segmentu. Pri razvrstavanju vektora značajki u jedan ciljni poremećaj potrebno je imati na umu da će upravo segmenti s više poremećaja biti uzrok najvećih neslaganja pri razvrstavanju.

4.2 Napredna analiza slijednog trenda

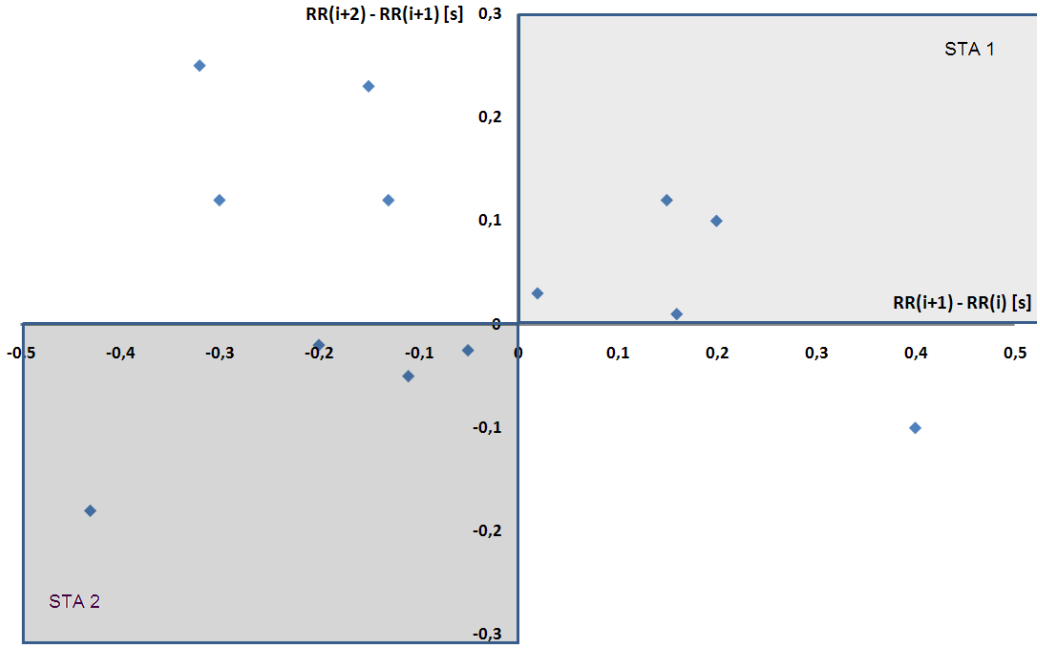
4.2.1 Analiza slijednog trenda

Promatranjem obrazaca usporavanja i ubrzavanja srčanog ritma moguće je doći do zanimljivih informacija o prirodi srčanih poremećaja. Uobičajeni način na koji se ubrzavanje i usporavanje ritma promatra je u kontekstu promjena u dvodimenzionalnom faznom prostoru. Različiti nelinearni postupci koji služe za kvantitativni opis promjena u faznom prostoru drugog reda prikazani su u poglavlju 3.5.3. Analiza slijednog trenda (engl. *sequential trend analysis*, kraće: STA), na način kako su je predložili autori [Schechtman 1992], koristi se za opis poremećaja srca uz pomoć faznog prostora drugog reda razlika.

U faznom prostoru drugog reda razlika, na osi *x* prikazuje se razlika između trajanja slijednih RR-intervala u trenutku *i*+1 i trenutku *i*, dok se na osi *y* istovremeno prikazuje razlika između trajanja RR-intervala u trenutku *i*+2 i trenutku *i*+1. Autori [Schechtman 1992] uveli su STA za razvrstavanje srčanog ritma kod iznenadnog zatajenja srca kod male djece. Pritom definiraju dvije značajke na grafu drugog reda:

1. STA 1 = (Broj točaka u +/+ kvadrantu) / (ukupan broj točaka na dijagramu)
2. STA 2 = (Broj točaka u -/- kvadrantu) / (ukupan broj točaka u dijagramu)

Prvom značajkom kvantiziraju utjecaj parasimpatičkog dijela ANS-a u smislu slijednog produljenja trajanja RR-intervala (tri slijedna produljenja = dvije slijedne razlike produljenja). Drugom značajkom kvantiziraju utjecaj simpatičkog dijela ANS-a u smislu slijednog skraćenja trajanja RR-intervala (tri slijedna skraćenja = dvije slijedne razlike skraćenja). Značajke su predočene na slici 4.6.



Slika 4.6. Analiza slijednog trenda.

Autori [de Carvalho 2002] omogućili su pokretanje STA u okviru sustava Matlab. Osim za opis ritma kod iznenadne srčane smrti, dvije značajke STA koriste i drugi autori prilikom razvrstavanja više vrsta srčanih aritmija [Asl 2008]. Pri razvrstavanju aritmija STA se uvijek koristi u smislu dviju dodatnih nelinearnih značajki, a nikad sama za sebe. Razlog tome je taj što ove dvije značajke nisu dovoljno informativne za točan opis specifičnosti pojedinih aritmija. Postupak analize slijednog trenda unaprijedit će se u ovoj disertaciji tako da omogući detaljniju kvantizaciju pojedinih aritmija u faznom prostoru drugog reda razlika.

4.2.2 Napredna analiza slijednog trenda

Da bi se značajke analize slijednog trenda mogle upotrijebiti za više vrsta aritmija, potrebno je najprije vizualizirati specifičnosti pojedinih aritmija u faznom prostoru drugog reda razlika. Cilj analize u faznom prostoru drugog reda razlike je odrediti područja u prostoru koja najbolje određuju pojedine vrste ritmova i najbolje ih razlikuju od drugih vrsta ritmova. Prilikom konstrukcije prostora moguće je provesti normalizaciju razlika u ovisnosti o srednjoj vrijednosti srčanog ritma na nekom segmentu. Kao uobičajenu vrijednost za normalizaciju može se uzeti brzina od 70 otkucaja/min. Time se rješava problem kada su razlike duljina intervala ovisne o tome kolika je prosječna brzina srčanog ritma. Koordinate točaka $b(x_i, y_i)$u tako definiranom dijagramu dane su kao:

$$x_i = a(RR(i+1) - RR(i))\,[s],$$
$$y_i = a(RR(i+2) - RR(i+1))\,[s], \quad a = \frac{60}{70\overline{RR}_{segm}} \tag{4.17}.$$

U tako određenom prostoru uočava se da pojedine vrste aritmija iscrtavaju svoja vlastita područja.

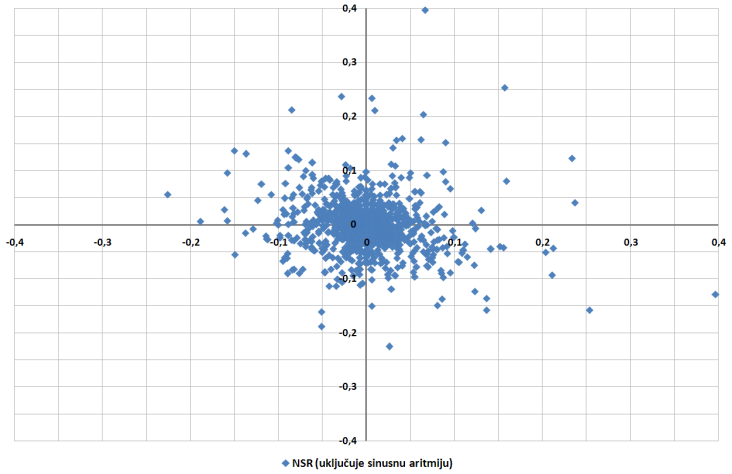
Tablica 4.6. Obrasci ritma analizirani u okviru napredne analize slijednog trenda.

Obrazac ritma	Zapisi iz baza MIT-BIH Arrhythmia (1XX i 2XX) i MIT-BIH Supraventricular Arrhythmia (8XX)	Broj točaka
NSR	800, 802, 803, 805, 806, 807, 808, 809, 810, 812, 822, 824, 827, 840, 844, 846, 862	1569
PVC	800, 801, 802, 803, 805, 806, 808, 810, 811, 812, 820, 821, 822, 824, 840, 841, 843, 847, 848, 849, 851, 852, 853, 855, 856, 859, 860, 861, 862, 863, 864, 865, 868, 869, 870, 871, 105, 106, 107, 109, 114, 116, 201, 202, 214, 223	1090
PAC	800, 801, 806, 807, 809, 811, 812, 820, 821, 823, 824, 825, 827, 829, 840, 842, 843, 845, 846, 847, 848, 849, 853, 855, 857, 859, 861, 863, 864, 865, 100, 108, 114, 118, 200, 202, 209, 213, 219, 220, 223	1078
PVC kuplet	106, 116, 200, 207, 208, 214, 215, 217, 223, 228, 233, 806, 821	119
VBI	800, 801, 804, 841, 854, 856, 857, 859, 866, 868, 870, 879, 881, 887, 890, 891, 892, 106, 119, 200, 207, 213, 217, 223, 228	676
Vođeni ritam	102, 104, 107, 217	612
AFIB	202, 219, 221, 222, 804, 849	2189
BII	217	300
VFL	207	427

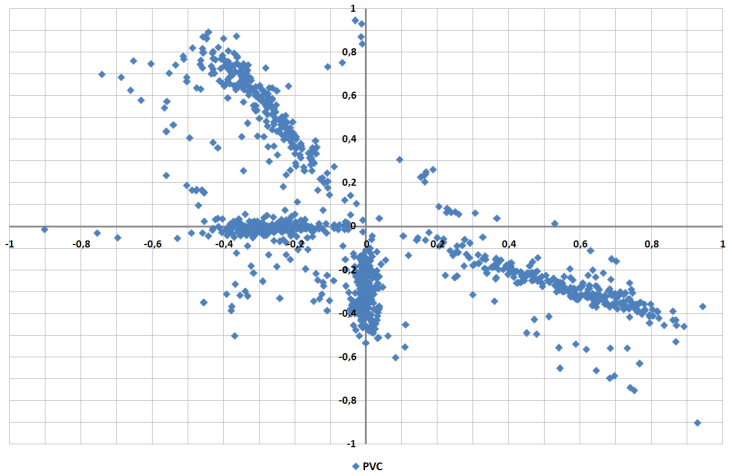
4.2.2.1. Dobavljanje podataka za pojedine poremećaje

Na temelju zapisa dobivenih iz baza MIT-BIH Arrhythmia Database i MIT-BIH Supraventricular Arrhythmia Database dobavljeni su podaci o ukupno devet vrsta obrazaca ritma za iscrtavanje faznog prostora drugog reda razlika. U tablici 4.6. naveden je popis obrazaca ritma i popis zapisa iz baza koji su se koristili za izgradnju faznog prostora. U istoj tablici navodi se i broj pojedinačnih izlučenih točaka na temelju kojih se postavljaju granice za značajke koje će opisivati određeni poremećaj. Navedeni zapisi nisu analizirani čitavi, već su analizirani samo neki dijelovi koji sadrže navedene obrasce ritma. Razlike u trajanjima RR-intervala na određenom segmentu u zapisu dobivene su pomoću radnog okvira HRVFrame (poglavlje 5).

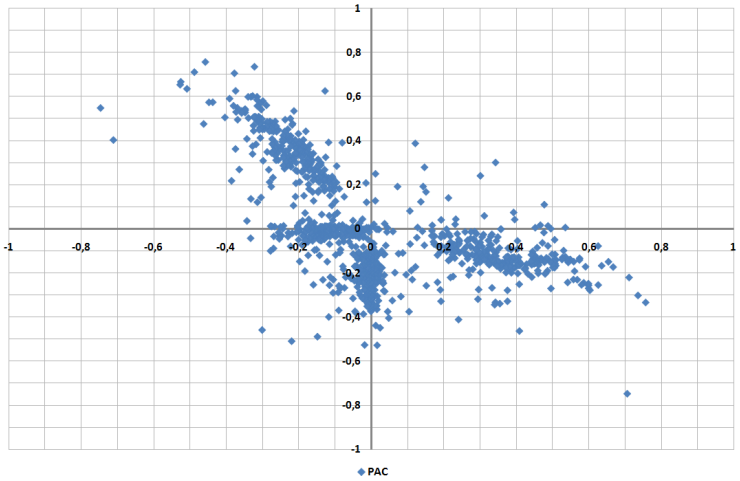
Prikaz tako dobivenih točaka dan je na slikama 4.7 – 4.15. Na slici 4.7 prikazan je karakterističan obrazac normalnog srčanog ritma (NSR) koji može uključivati i sinusnu aritmiju (RSA) kao benigni poremećaj. Iako korisna iz teorijskog stajališta, u praksi se karakteristika NSR bez sinusne aritmije neće koristiti zbog toga što je sinusna aritmija česta pojava i ne smatra se klinički značajnom. Budući da je provedena normalizacija, na slici 4.7 uključeni su i poremećaji sinusne bradikardije (SBR) i sinusne tahikardije (ST). Slika 4.8 prikazuje preuranjenu kontrakciju klijetki (PVC), dok slika 4.9 prikazuje preuranjenu kontrakciju pretklijetki (PAC). Ako se usporede ova dva česta poremećaja ritma, može se uočiti da su poremećaji zaista slični na grafu 2. reda razlike. Na slici 4.10. prikazan je obrazac ventrikularnog kupleta (PVC kuplet). Karakteristika je nešto drugačija u odnosu na PVC i PAC. Posebno je uočljiva koncentracija točaka na osi y koja nije uobičajena za PVC i PAC. Ventrikularna bigeminija (VBI) prikazana na slici 4.11 očituje se vrlo jasnim obrascem koji se dosta razlikuje od ostalih poremećaja. Tu je također prikazan samo središnji dio pojave koju karakterizira izmjena pozitivnih i negativnih razlika, dok se početak i završetak prijelaza u bigeminiju ne prikazuje jer se nužno ne događa unutar nekog vremenskog segmenta. Ritam vođen umjetnim pejsmejkerom (PEJS) prikazan je na slici 4.12. Tu su uočljive vrlo male promjene u iznosima trajanja slijednih RR-intervala. Na slici 4.13 prikazan je obrazac atrijalne fibrilacije (AF). Uočljiva je velika raspršenost točaka i blaga ukošenost raspršenja u smjeru osi $y = -x$. AF prikazana na ovoj slici uključuje i moguće preuranjene kontrakcije ili kuplete na promatranom segmentu, budući da su oni relativno česta pojava. AV-blok drugog stupnja (BII) u omjeru 2:1 prikazan je na slici 4.14. Nije jasno radi li se o Mobitz I ili Mobitz II bloku, ali je jasno da se svaki drugi otkucaj ispušta i da se time trajanja RR-intervala zapravo udvostručuju. Promjene u slijednim razlikama su male. Ventrikularno lepršanje (VFL) prikazano je na slici 4.15. Ako se promjene za vrijeme VFL usporedbe s AF (slika 4.13), uočava se da su dijagrami dosta slični. Jedina bitna razlika koja nije uočljiva sa slike 4.15 je ta da je ritam prilikom ventrikularnog treperenja bitno brži od AF.



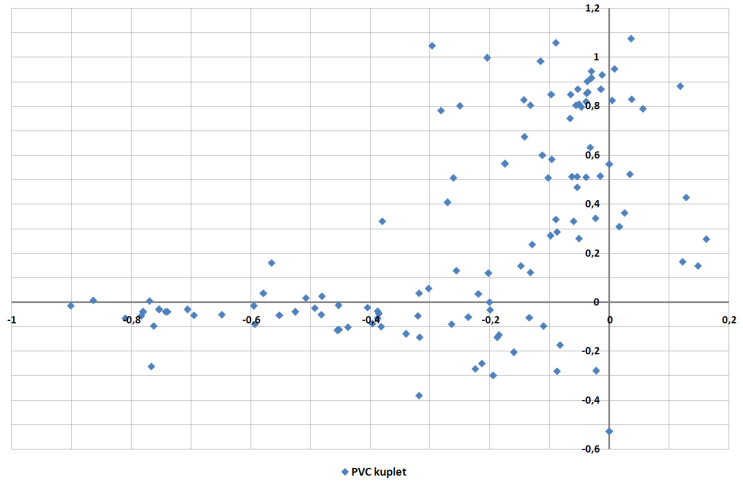
Slika 4.7. Fazni prostor drugog reda razlika za srčani ritam koji uključuje NSR, ST, SBR i RSA.



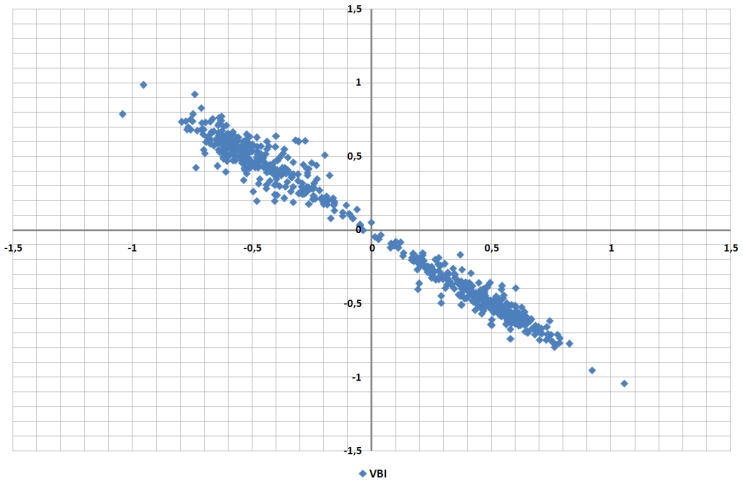
Slika 4.8. Fazni prostor drugog reda razlika za preuranjenu kontrakciju ventrikula (PVC).



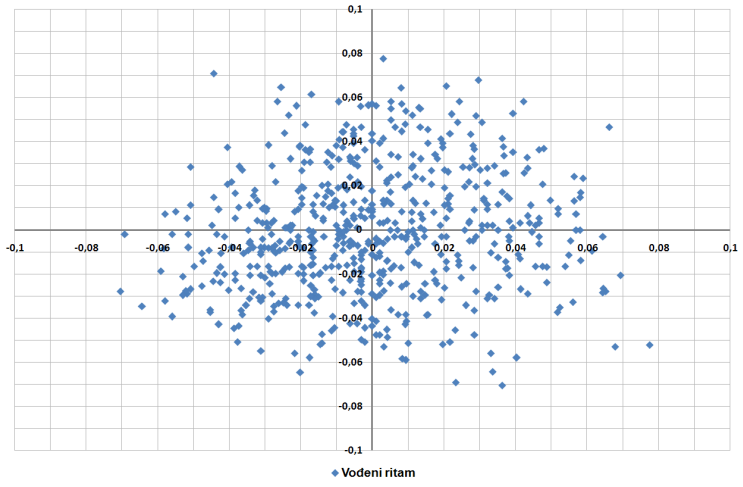
Slika 4.9. Fazni prostor drugog reda razlika za preuranjenu kontrakciju atrija (PAC).



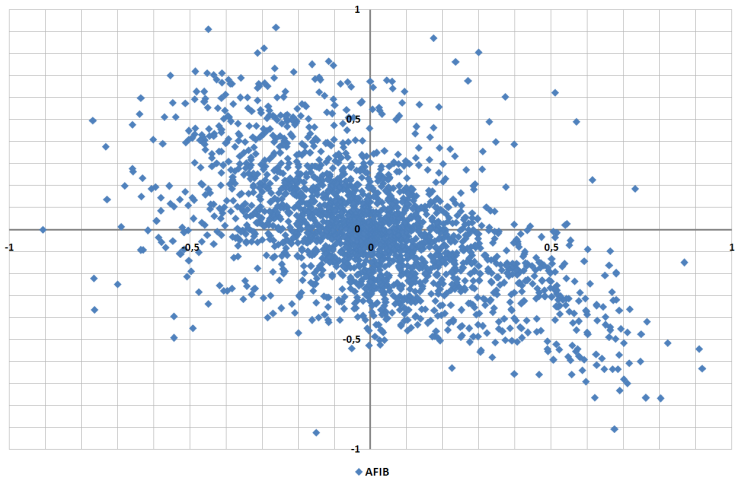
Slika 4.10. Fazni prostor drugog reda razlika za ventrikularni kuplet (PVC kuplet).



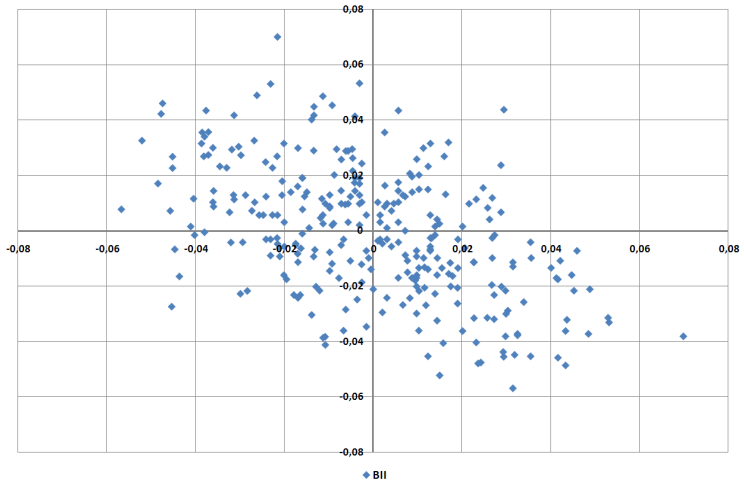
Slika 4.11. Fazni prostor drugog reda razlika za VBI.



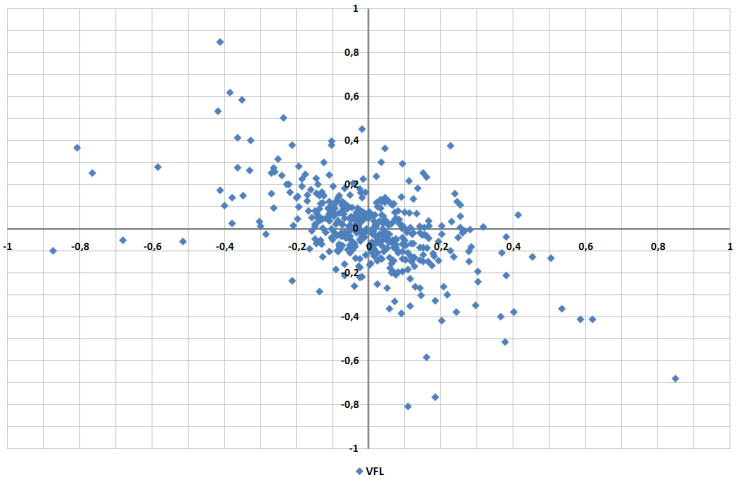
Slika 4.12. Fazni prostor drugog reda razlika za umjetno vođeni ritam (pejsmejker, PEJS).



Slika 4.13. Fazni prostor drugog reda razlika za AF koja može uključivati i PAC i PVC.



Slika 4.14. Fazni prostor drugog reda razlika za AV-blok drugog stupnja (BII), omjera 2:1.



Slika 4.15. Fazni prostor drugog reda razlika za ventrikularno lepršanje (VFL).

4.2.2.2. Postupak dobivanja značajki u naprednoj analizi slijednog trenda

Značajke koje opisuju neki poremećaj trebaju biti što jednostavnije za izračunati i što malobrojnije. Postojeći jednostavni modeli, kao što su recimo oni koje su dali autori [Tsipouras 2005] su uglavnom modeli koje su dali medicinski stručnjaci na temelju literature i vlastitog iskustva. Fazni prostor drugog reda razlika omogućuje relativno jednostavnu definiciju značajki za neke poremećaje na temelju stvarnih podataka o slijednim razlikama. Ipak, potrebno je uočiti da nije lako razlikovati neke poremećaje, kao npr. PAC i PVC.

Na idućoj stranici naveden je algoritam napredne analize slijednog trenda (engl. *advanced sequential trend analysis*, kraće: ASTA), koji pokazuje kako se na temelju točaka u faznom prostoru 2. reda razlika dolazi do kvalitetnih značajki za opis poremećaja.

4.2.2.3. Razmatranje predloženog postupka ASTA

Matematička definicija potpodručja ovisna je o rasporedu točaka koji je dobiven na primjercima za učenje. Matematički i proračunski najjednostavnija definicija za izračun značajke uključuje područje zadano pravokutnikom uz moguću zarotiranost pravokutnika za neki kut φ u odnosu na os x u R^2_{diff}. Svaki takav pravokutnik može se na jedinstven način zadati uz pomoć pravca koji prolazi njegovim središtem i paralelan je duljoj stranici, najvećom dozvoljenom udaljenosti točaka od tog pravca i točkama na pravcu unutar kojih se pravokutnik definira. Nešto složenija, ali i za neke poremećaje preciznija matematička definicija uključuje elipsu za opis potpodručja. Elipsa može također biti zarotirana za neki kut φ s obzirom na os x. U slučaju kad je definirano područje za normalan obrazac nekog BVN, neko područje poremećaja može biti definirano tako da isključuje onaj dio prostora koji pokriva normalan obrazac.

Ako je u prostoru R^2_{diff} neki poremećaj kod kojeg su točke blisko koncentrirane unutar jednog potpodručja, tada se za taj poremećaj definira jedna značajka. Ta značajka bi u teoriji pokrila većinu točaka koje su karakteristične za taj poremećaj. Problem može nastati u slučaju da se potpodručja za neke poremećaje djelomično preklapaju. U tom slučaju ovom metodom nije moguće razlučiti o kojem se poremećaju radi samo na temelju te jedne značajke, već je nužno u analizu uključiti neke dodatne značajke. Primjerice, sa slika 4.12. i 4.14 vidljivo je da se pejsmejker i AV-blok drugog stupnja 2:1 ne razlikuju značajno. Ipak, uvođenjem dodatne značajke srednje vrijednosti trajanja otkucaja uočit će se da ta vrsta bloka daje sporiji ritam.

Ako postoji više potpodručja koja su karakteristična za neki poremećaj, tada se definiraju dvije značajke. Prva, f_i broji sve točke koje se nađu unutar svih definiranih potpodručja za taj poremećaj. Često ta informacija nije dovoljna da bi se zaključilo radi li se ili ne o određenom poremećaju. Dodatnu informaciju dobiva se s drugom značajkom, g_i koja broji povoljne trajektorije za taj poremećaj. Naime, potrebno je da se točke pojavljuju u prostoru R^2_{diff} po određenom redoslijedu koji će sugerirati radi li se

ALGORITAM ASTA:

Neka je s D označen skup poremećaja $D = \{D_1, D_2, \ldots, D_k\}$.

Neka je za svaki $D_i \in D, i = 1...k$ poznat fazni prostor drugog reda razlika R^2_{diff} izgrađen na temelju N_i točaka dobivenih iz skupa za učenje: $T_{ij}(x, y),\ j = 1...N_i$.

Cilj: Naći takav broj značajki $m \geq k$ za što bolji opis i međusobno razlikovanje skupa poremećaja D.

Za svaki $D_i \in D$:

Uoči i definiraj p potpodručja $S = \{S_1, \ldots, S_p\}$ na faznom prostoru drugog reda u kojima se nalazi blisko raspoređeno ukupno $n_i \leq N_i$ točaka, a po potpodručjima

$n_{Sj} \leq n_i,\ j = 1, \ldots, p$

Ako je $p = 1$ (postoji samo jedno takvo potpodručje):

Definiraj samo jednu značajku kao $f_i = n_i$

Inače ako je $p > 1$:

Definiraj prvu značajku $f_i = \sum_{j=1}^{p} n_{Sj} = n_i$

Ako se potpodručja djelomično preklapaju, točka koja se nalazi u dva ili više preklapajućih potpodručja pri izračunu značajke f_i u obzir se uzima samo jednom.

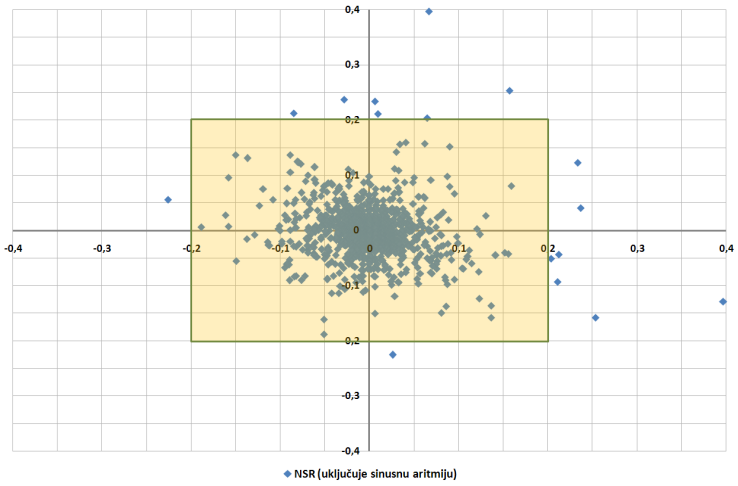
Definiraj drugu značajku $g_i = r_i$, gdje je r_i broj povoljnih trajektorija d_i za poremećaj D_i.

Broj točaka trajektorije d_i u R^2_{diff} za neki poremećaj D_i ovisan je o naravi samog poremećaja i iznosi najmanje dvije točke. Za trajektoriju se kaže da je povoljna ako prolazi kroz $1 \leq n \leq p$ potpodručja. Broj potpodručja n kroz koji trajektorija treba proći da bi ju se smatralo povoljnom ovisi o naravi samog poremećaja D_i.

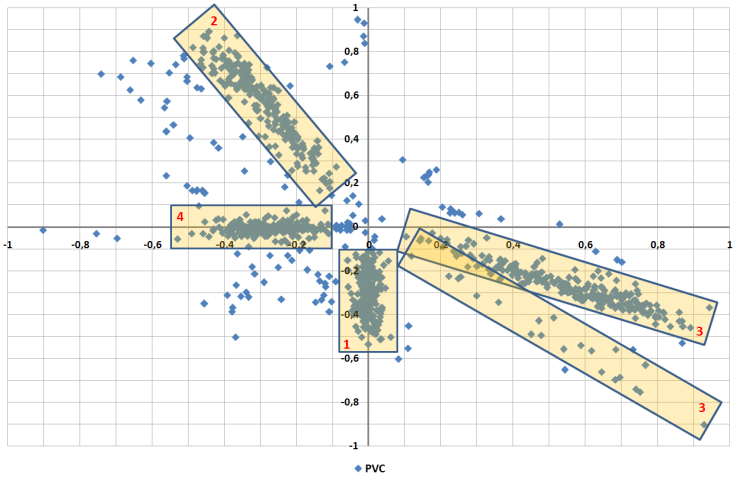
o nekom poremećaju ili ne. Postupkom se dozvoljava da za određeni poremećaj trajektorija neki put promaši potpodručje u kojem bi se trebala naći. Time se povećava robustnost na manja odstupanja od karakterističnog obrasca. Međutim, promašaja ne smije biti puno, inače nema sigurnosti radi li se tu zaista o traženom poremećaju. Ako poremećaj ima *p* potpodručja, povoljna trajektorija je ona koja prolazi kroz barem jedno potpodručje, a moguće i kroz sva potpodručja, ovisno o strogosti definicije.

4.2.2.4. Definicija značajki po promatranim poremećajima

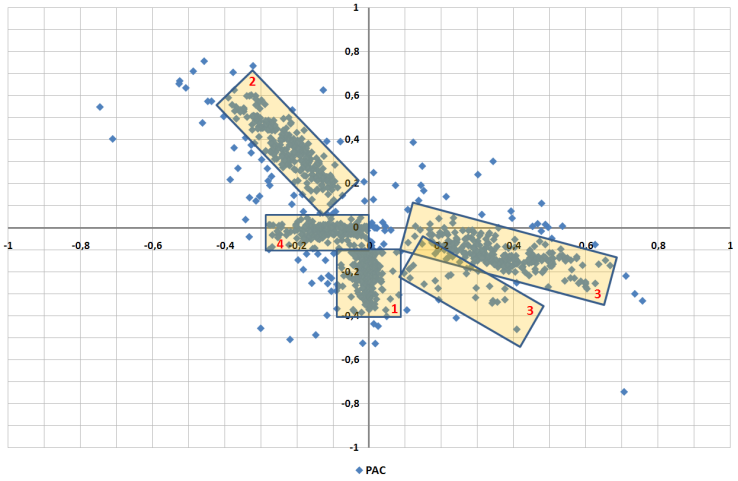
Na slikama 4.16 do 4.24 prikazana je podjela faznog prostora u potpodručja za svaki promatrani poremećaj na temelju predloženog postupka. U slučaju više od jednog potpodručja, svako je označeno rednim brojem prema redoslijedu pojavljivanja u trajektoriji. U tablici 4.7 navedene su matematičke definicije potpodručja kao i ukupan broj potpodručja koji trajektorija treba obići da bi ju se smatralo povoljnom za pridodati u značajku g_i.



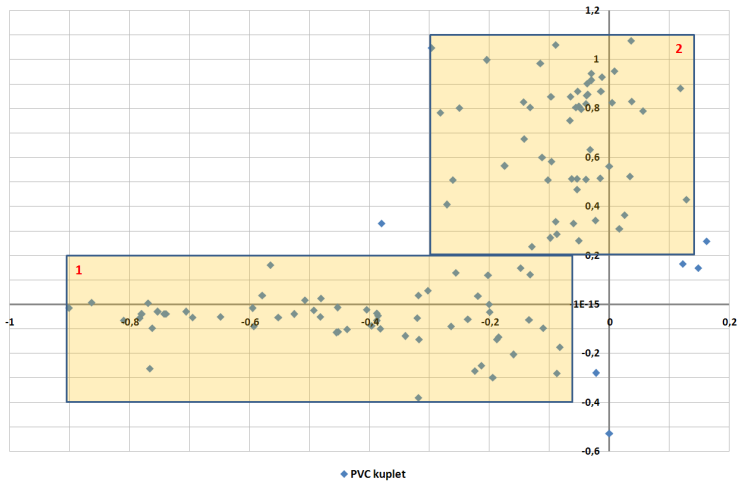
Slika 4.16. Podjela R^2_{diff} u potpodručja za normalan sinusni ritam (NSR), SBR, ST i RSA.



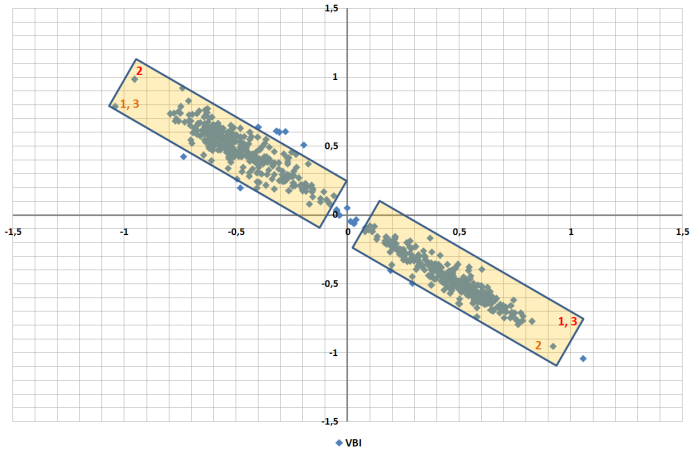
Slika 4.17. Podjela R^2_{diff} u potpodručja za preuranjenu kontrakciju ventrikula (PVC).



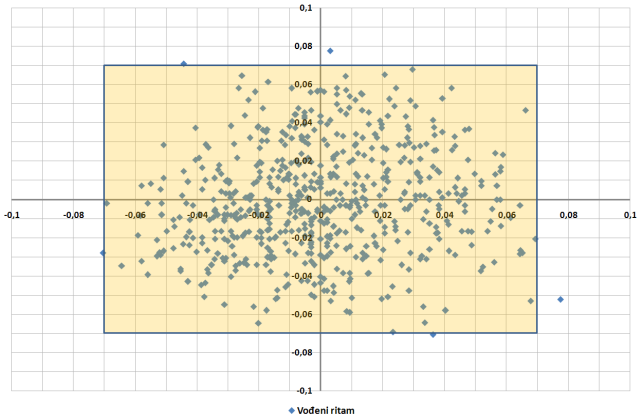
Slika 4.18. Podjela R^2_{diff} u potpodručja za preuranjenu kontrakciju atrija (PAC).



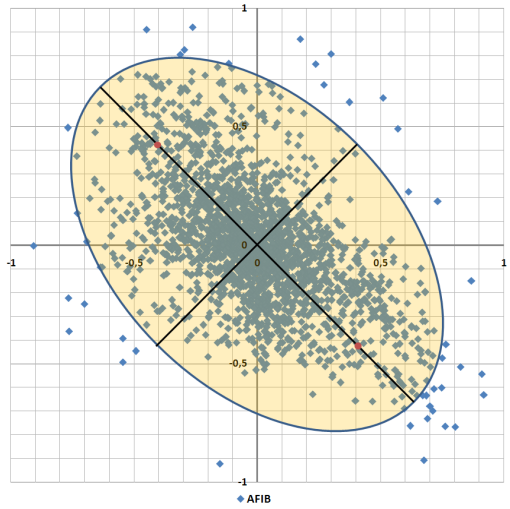
Slika 4.19. Podjela R^2_{diff} u potpodručja za ventrikularni kuplet (PVC kuplet).



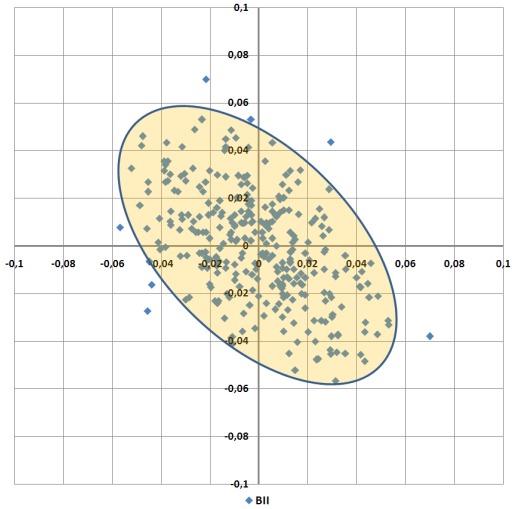
Slika 4.20. Podjela R^2_{diff} u potpodručja za ventrikularnu bigeminiju (VBI).



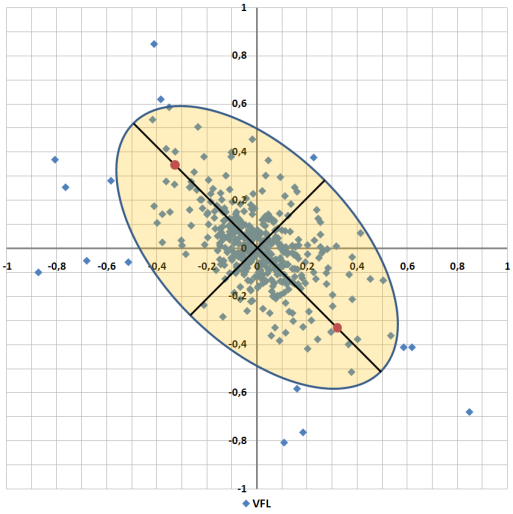
Slika 4.21. Podjela R^2_{diff} u potpodručja za umjetno vođeni ritam (pejsmejker, PEJS).



Slika 4.22. Podjela R^2_{diff} u potpodručja za atrijalnu fibrilaciju (AF). Iako tako nije prikazano, područje ne uključuje točke unutar područja za NSR.



Slika 4.23. Podjela R^2_{diff} u potpodručja za AV-blok drugog stupnja (BII), omjera 2:1.



Slika 4.24. Podjela R^2_{diff} u potpodručja za ventrikularno lepršanje (VFL). Iako tako nije prikazano, područje ne uključuje točke unutar područja za NSR.

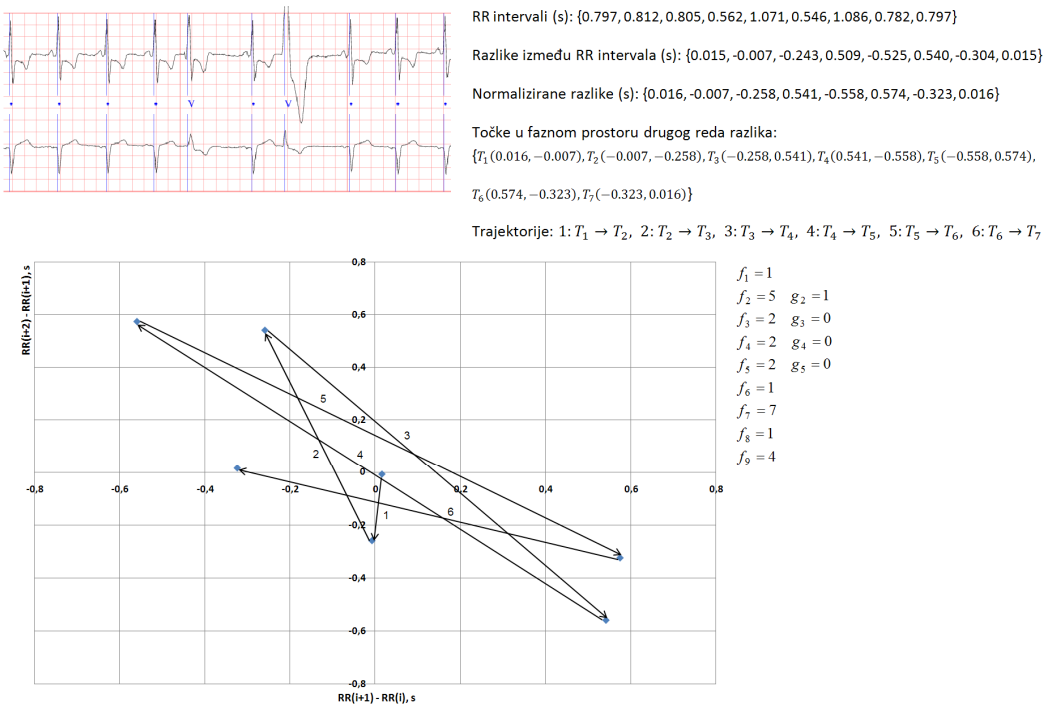
Tablica 4.7. Definicija potpodručja za svaki poremećaj, potrebna trajektorija i značajke napredne analize slijednog trenda.

Obrazac ritma	Definicija potpodručja za točke $T_i(x, y)$	Trajektorija	Značajka
NSR	$\|x\| \le 0.2$ i $\|y\| \le 0.2$	-	f_1
PVC	1. $-0.08 \le x \le 0.08$ i $-0.55 \le y \le -0.1$ 2. $p_1 : y = -2x, -0.5 \le x \le -0.1, d(T_i, p_1) \le 0.1$ 3. $p_2 : y = -\frac{1}{2}x, 0.1 \le x \le 1, d(T_i, p_2) \le 0.1$ $p_3 : y = -x, 0.2 \le x \le 1, d(T_i, p_3) \le 0.1$ 4. $-0.55 \le x \le -0.1$ i $-0.1 \le y \le 0.1$	1->2->3(p_2 ili p_3) -> 4, barem tri od četiri	f_2, g_2
PAC	1. $-0.08 \le x \le 0.08$ i $-0.4 \le y \le -0.1$ 2. $p_4 : y = -1.6x, -0.45 \le x \le -0.05, d(T_i, p_4) \le 0.1$ 3. $p_5 : y = -0.47x + 0.05, 0.1 \le x \le 0.65, d(T_i, p_5) \le 0.1$ $p_6 : y = -x, 0.12 \le x \le 0.4, d(T_i, p_6) \le 0.1$ 4. $-0.3 \le x \le 0.0$ i $-0.1 \le y \le 0.08$	1->2->3(p_5 ili p_6) -> 4, barem tri od četiri	f_3, g_3
PVC kuplet	1. $-0.9 \le x \le -0.07$ i $-0.4 \le y \le 0.2$ 2. $-0.3 \le x \le 0.13$ i $0.2 \le y \le 1.1$	1 -> 2	f_4, g_4
VBI	1 i 3 (ili 2). $p_7 : y = -x, -1 \le x \le -0.05, d(T_i, p_7) \le 0.15$ 2 (ili 1 i 3). $p_8 : y = -x, 0.05 \le x \le 1, d(T_i, p_8) \le 0.15$	1->2->3 (bilo da se počne s p_7 ili p_8	f_5, g_5
PEJS	$\|x\| \le 0.07$ i $\|y\| \le 0.07$	-	f_6
AFIB	$a = 0.9, b = 0.6, S(0,0), F_1(-0.474,0.474), F_2(0.474,-0.474)$, bez $\|x\| \le 0.2$ i $\|y\| \le 0.2$	-	f_7
BII	$a = 0.07, b = 0.04, S(0,0), F_1(-0.041,0.041), F_2(0.041,-0.041)$	-	f_8
VFL	$a = 0.7, b = 0.4, S(0,0), F_1(-0.406,0.406), F_2(0.406,-0.406)$, bez $\|x\| \le 0.2$ i $\|y\| \le 0.2$	-	f_9

Primjer faznog prostora 2. reda razlike za trajektoriju od 7 točaka (9 RR-intervala) dan je na slici 4.25. Riječ je o segmentu na kojem je došlo do dvije preuranjene kontrakcije ventrikula (PVC), s razmakom od jednog normalnog otkucaja. Odsječak EKG-a prikazan je na istoj slici gore, a značajke f_i i g_i izračunate su sa strane.

Važno je uočiti da su značajke f_2 i g_2 za PVC izražene, pogotovo stoga što je uočena jedna trajektorija koja je u skladu s obilježjima PVC-a. Za razliku od toga, u ovom slučaju nije otkriven PAC. Moguće je pitanje zašto je otkrivena samo jedna trajektorija PVC-a, ako postoje dva PVC otkucaja. Razlog tome je taj što je početak drugog PVC otkucaja prekriven s krajem prvoga u smislu da nije došlo do smirenja ritma.

Da je kojim slučajem postojao još jedan normalan otkucaj između, bila bi uočena dva PVC obrasca. Ipak, i ovako je očito da se radi o segmentu s PVC. Ono što se još može primijetiti iz rezultata to je da značajka f_7 za atrijalnu fibrilaciju iznosi 7. To je zato što obrazac točaka na slici 4.25 uistinu može predstavljati AF. Razlika je samo u tome što je osim toga otkriven i PVC, što čini vjerojatnijim da se radi o PVC-u, a ne o AF.



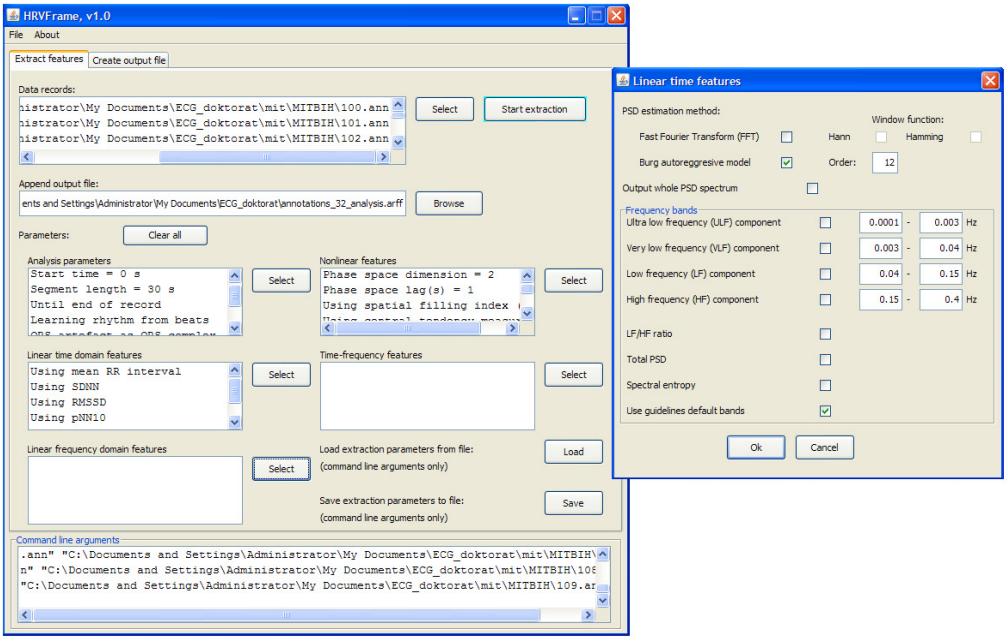
Slika 4.25. Primjer napredne analize slijednog trenda na segmentu od 7 točaka (9 RR-intervala). Označena je i trajektorija u faznom prostoru 2. reda razlika kao i iznosi značajki izračunati na temelju tablice 4.7.

5 Radni okvir za izlučivanje značajki

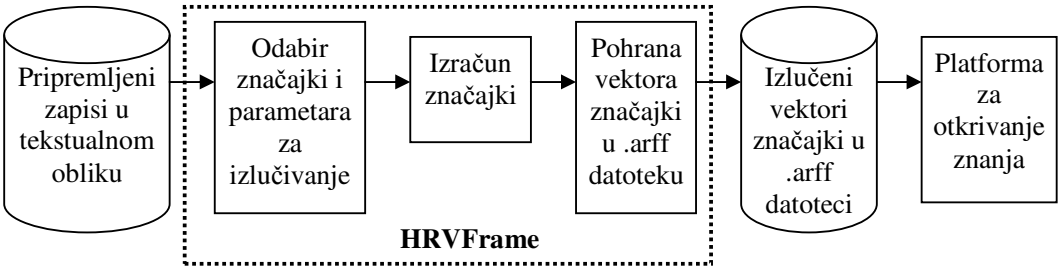
5.1 Pregled radnog okvira

Radni okvir pod nazivom HRVFrame implementiran je u programskom jeziku Java, inačice 1.6, korištenjem razvojnog okruženja Eclipse. HRVFrame je opširan računalni radni okvir namijenjen izlučivanju značajki BVN s trenutnom implementacijom učitavanja ulaznih datoteka predviđenom za srčani ritam. Glavni zadatak radnog okvira je da uzme jednu ili više ulaznih datoteka, izluči značajke i pohrani ih u izlaznu datoteku pripremljenu za dubinsku analizu podataka [Jović 2011 (2)]. Prethodna verzija radnog okvira razvijena u okviru diplomskog rada sadržavala je manji broj značajki s mogućnosti vizualizacije EKG-a, niza otkucaja i faznog prostora [Jović 2007, 2008].

Proces izlučivanja značajki unutar okvira HRVFrame provodi se na pretpripremljenim datotekama. Radni okvir je samostojeći programski produkt u obliku izvršne Java arhive `HRVFrame.jar` koji podržava naredbeno-linijski način rada i grafičko sučelje, slika 5.1. Radni okvir nije integriran niti u jednu platformu za otkrivanje znanja, već se suradnja između izlučivanja značajki i dubinske analize uspostavlja posredno, putem izlaznih datoteka u obliku .arff. Čitav proces analize biomedicinskog vremenskog niza pomoću okvira HRVFrame prikazan je na slici 5.2. HRVFrame je organiziran u tri glavna logička dijela: ulaz, izračun značajki i izlaz. U nastavku se opisuju mogućnosti radnog okvira iz perspektive ova tri dijela.



Slika 5.1. Grafičko sučelje radnog okvira HRVFrame s implementacijom u programskom jeziku Java.



Slika 5.2. Analiza biomedicinskog vremenskog niza uporabom računalnog radnog okvira HRVFrame.

5.2 Ulazni dio radnog okvira

Ulazni dio radnog okvira je osmišljen tako da omogući:
1) učitavanje jednog ili više zapisa biomedicinskog vremenskog niza,
2) izbor značajki i parametara pod kojima se značajke izlučuju,
3) specifikaciju parametara vezanih uz vrstu podataka, način analize i pozicioniranje u zapisu.

5.2.1 Ulazni podaci

Kao ulazni podaci trenutno su dozvoljeni zapisi srčanog ritma, i to kao tekstualne datoteke u obliku ASCII. Struktura zapisa je ista kao ona koju daje alat *rdann* s portala PhysioNet [PhysioNet]. Ulazna datoteka može sadržavati informacije o: 1) vremenima najviše amplitude srčanog otkucaja (R vrh), 2) tipu srčanog otkucaja, 3) dominantnom srčanom ritmu na nekom segmentu, i 4) ostale informacije.

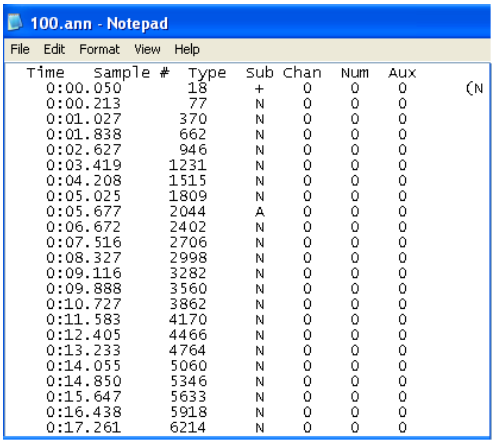
Informacija o vremenu najviše amplitude srčanog otkucaja je jedini nužan dio koji datoteka mora sadržavati. Tip srčanog otkucaja može biti bilo koji podržan u bazama podataka s portala PhysioNet, a isto vrijedi i za oznaku dominantnog srčanog ritma. Korisnik može i sam definirati koji je dominantni tip ritma ili srčane bolesti prisutan u zapisu. Za daljnju analizu radni okvir je ograničen na značajne tipove otkucaja, ritmova i bolesti. Potpuni popis podržanih tipova poremećaja dan je u tablici 5.1.

Ostale informacije mogu uključivati oznaku rednog broja uzorka koji odgovara R-vrhu, specifične informacije vezane uz određenu bazu, komentari i sl. Takve informacije ovaj radni okvir ne uzima u obzir. Primjer valjane ulazne datoteke prikazan je na slici 5.3. Sustav može učitati više datoteka odjednom, koje se analiziraju slijedno.

Zbog modularnosti okvira, učitavanje podataka o nekom drugom biomedicinskom nizu može se jednostavno implementirati dodavanjem odgovarajućeg razreda u Javi za rukovanje dotičnim tipom BVN. Nakon implementacije takvog razreda, značajke bi se mogle računati i nad takvim podacima.

Tablica 5.1. Podržani tipovi srčanih otkucaja, ritmova i bolesti unutar radnog okvira HRVFrame.

Tip poremećaja	Tip poremećaja	Tip poremećaja
Normalan otkucaj / ritam	Atrijalna fibrilacija	Atrijalna bigeminija
Preuranjena kontrakcija atrija (PAC)	Paroksizmalna supraventrikularna tahiaritmija	Atrijalna trigeminija
Preuranjena kontrakcija ventrikula (PVC)	Ventrikularna tahikardija	Ventrikularna bigeminija
Segment sadrži PVC i PAC	Atrijalno lepršanje (*flutter*)	Ventrikularna trigeminija
Kompenzirani kuplet	Ventrikularno lepršanje (*flutter*)	Ventrikularni ektopični otkucaj
Dekompenzirani kuplet	Ventrikularna fibrilacija	Idioventrikularni ritam
Fuzijski otkucaj / ritam	Lutajući atrijalni pejsmejker (WAP)	Sinusna pauza ili zastoj
Blok lijeve grane (LBBB)	AV-blok 2. stupnja 2:1	Sindrom Wolff-Parkinson-White (WPW)
Blok desne grane (RBBB)	Ritam AV-spojnice	Kongestivno zatajenje srca
Sinusna bradikardija	Umjetno vođeni ritam (pejsmejker)	



Slika 5.3. Primjer ulazne datoteke prihvatljive radnom okviru HRVFrame.

5.2.2 Parametri radnoga okvira

Parametri radnog okvira koji su trenutačno dostupni za analizu BVN dani su, uz objašnjenje, u tablicama 5.2. i 5.3. Parametri koje prima program mogu se podijeliti na parametre vezane uz vrstu analize (tablica 5.2) te na značajke i njihove parametre (tablica 5.3). Neki od parametara radnog okvira koji su specifični za analizu srčanog ritma označeni su zvjezdicom u tablici 5.2. Značajke i njihovi parametri navedeni u tablica 5.3. jednako su povezani i uz poglavlje 5.3 u kojem se govori o izračunu značajki. Specifikacija svih navedenih parametara omogućena je jednako tako putem grafičkog sučelja, uz jednostavan odabir opcija stavljanjem kvačice ili upisivanjem traženog parametra u sučelje.

Tablica 5.2. Parametri analize koji se mogu odrediti prilikom pokretanja radnog okvira.

Parametar analize	Objašnjenje
-carff<Naziv_dat>	Stvara izlaznu datoteku naziva <Naziv_dat>.
-aarff<Naziv_dat>	Dodaje vektore značajki u izlaznu datoteku <Naziv_dat>.
-ointervals *	Izlučuje trajanje RR-intervala u zasebnu tekstualnu datoteku.
-rel<Naziv_relac>	Određuje ime relacije u izlaznoj datoteci. Pretpostavljeno ime je HRVdata.
-efr<Naziv_dat>*	Datoteka <Naziv_dat> sadrži prioritete ritmova pri označavanju.
-odifferences *	Izlučuje razlike u trajanju RR-intervala u zasebnu tekstualnu datoteku.
-lfb *	Vrsta ritma određuje se na temelju tipova otkucaja.
-qaq *	Smatra QRS artefakte kao ispravne R-vrhove.
-dnb *	Kod određivanja niza RR-intervala, zanemare se oznake koje nisu otkucaji.
-test *	Razmatra samo vremena R-vrhova i ne označava se ritam, koristi se za testiranje.
-er	Analiza se odvija do kraja zapisa.
-rec	Uključuje naziv zapisa kao atribut u izlaznoj datoteci .arff.
-start*X*	Analiza započinje s *X*-tim rednim brojem mjerenja (ili oznakom) u zapisu.
-st	Koristi se početni redni broj mjerenja kao atribut u izlaznoj datoteci .arff.
-sttime*X*	Analiza započinje u trenutku *X* sekundi nakon početka zapisa.
-tis	Koristi se početno vrijeme kao atribut u izlaznoj datoteci .arff.
-count*X*	Jedan analizirani segment obuhvaća *X* mjerenja.
-co	Koristi se broj mjerenja u jednom segmentu kao atribut u izlaznoj datoteci .arff.
-segm*X*	Jedan analizirani segment traje *X* sekundi.
-seg	Koristi se duljinu segmenta u sekundama kao atribut u izlaznoj datoteci .arff.
-sig<Vrsta_ritma>	Eksplicitno odredi vrstu ritma prisutnu u svim analiziranim segmentima.
-use <Lista_dat>	Koriste se zapisi srčanog ritma dani kao lista datoteka (naziv, odnosno put do datoteke), odvojenih zarezom.

* parametri vezani isključivo uz analizu srčanog ritma

5.2.3 Pokretanje analize

Analizu se pokreće putem grafičkog sučelja ili preko naredbene linije. U slučaju grafičkog sučelja, analiza se pokreće nakon određivanja svih parametara klikom na gumb *Start extraction*. U slučaju naredbene linije, redoslijed navođenja parametara nije bitan, osim što na kraju treba slijediti popis datoteka sa srčanim ritmom. Preporučeni obrazac pokretanja analize korištenjem radnog okvira je sljedeći:

```
java –jar HRVFrame.jar <izlazna_datoteka> <parametri_radnog_okvira>
<značajke_s_parametrima> –use <popis_ulaznih_datoteka>
```

5.3 Izračun značajki

Značajke koje su implementirane u radnom okviru uključuju: linearne vremenske (i to statističke i geometrijske), linearne frekvencijske, vremensko-frekvencijske, i nelinearne značajke (vezane uz fazni prostor, fraktalnost, entropijske i druge). Cjelokupni popis implementiranih značajki dan je u tablici 5.3.

Potrebno je naglasiti da mnoge od implementiranih značajki nisu specifično vezane uz varijabilnost srčanog ritma. Jedine značajke koje su usko vezane uz varijabilnost srčanog ritma označene su zvjezdicom u tablici 5.3. Ipak, sve značajke navedene u tablici su dosad primjenjivane u literaturi za analizu varijabilnosti srčanog ritma, uz iznimku novopredloženih postupaka ASTA i AbEn. Sve značajke implementirane u

radnom okviru imaju ugrađenu najmanju duljinu segmenta za koju se mogu izlučiti. To je omogućeno da se podrži automatsko izlučivanje bez potrebe za informiranim izborom značajki za pojedine duljine segmenata, u skladu sa sustavnim postupkom predloženim u poglavlju 7.

Ukupan broj implementiranih postupaka u radnom okviru trenutačno iznosi 39 (pri čemu jedan postupak može uključivati jednu ili više značajki). Ukupan broj značajki koje je moguće izlučiti je velik i teško ga je procijeniti. Ako se uzme u obzir da se nelinearne fazne značajke (D_2, SFI, CTM) analizira u faznom prostoru određene dimenzije s pet intervala [Jović 2011 (1)], i ako se uzme u obzir da se računa 20 vremenskih skala kod višeskalarne entropije [Costa 2005 (2)], i ako se uzmu u obzir svih 13 značajki koje daje ASTA (poglavlje 4.2.2.4), tada je ukupan broj značajki jednak 97. Abecedna entropija sama za sebe može uvesti do 138 dodatnih značajki.

Tablica 5.3. Značajke implementirane u radnom okviru. Dana je vrsta značajke, parametri radnog okvira u naredbenoj liniji i referenca na poglavlje u disertaciji.

Značajka ili skup značajki	Vrsta značajke	Parametri koje je potrebno odrediti pri izlučivanju	Ref. na poglavlje
Srednja vrijednost RR-intervala *	Lin. vremenska, statistička	-mean	3.2.1.1
SDNN *	Lin. vremenska, statistička	-sdn	3.2.1.2
RMSSD *	Lin. vremenska, statistička	-rms	3.2.1.3
SDSD *	Lin. vremenska, statistička	-sdsd	3.2.1.4
pNN*X* *	Lin. vremenska, statistička	-p*X*: neki od: (-p05, -p10, -p15, -p20, -p30, -p40, -p50)	3.2.1.5
SDANN *	Lin. vremenska, statistička	-sda	3.2.1.6
Fanov faktor	Lin. statistička	-fn, -tscfn*X* (vremenska skala)	3.2.1.7
HTI *	Lin. vremenska, geometrijska	-hti	3.2.1.8
TINN *	Lin. vremenska, geometrijska	-tin	3.2.1.9
ULF *	Lin. frekvencijska	frekvencijski parametri**, -ulf	3.3.4
VLF *	Lin. frekvencijska	frekvencijski parametri**, -vlf	3.3.4
LF *	Lin. frekvencijska	frekvencijski parametri**, -lf	3.3.4
HF *	Lin. frekvencijska	frekvencijski parametri**, -hf	3.3.4
LF/HF *	Lin. frekvencijska	frekvencijski parametri**, -lhf	3.3.4
Total PSD	Lin. frekvencijska	frekvencijski parametri**, -tpsd	3.3.4
Allanov faktor	Nelin. statistička	-al, -tscal*X*	3.5.3.15
Autokorelacijski koeficijent	Nelin. statistička	-acc	3.5.2
SFI	Nelin. fazni prostor	-sfi, -dim*X*, -int*X* (ili -lags*X*,*Y*,...)	3.5.3.14
SD1/SD2 *	Nelin. fazni prostor	-sdr	3.5.3.3
CSI i CVI *	Nelin. fazni prostor	-csi, -cvi	3.5.3.3
Korelacijska dimenzija D_2	Nelin. fazni prostor	-cdi, -dim*X*, -int*X* (ili -lags*X*,*Y*,...)	3.5.3.2
LLE (po Rosenstienu)	Nelin. fazni prostor	-lle, -dim*X*	3.5.3.1
CTM	Nelin. fazni prostor drugog reda razlike	-ctm, -dim*X*, -int*X* (ili -lags*X*,*Y*,...)	3.5.3.17

Ispravljena uvjetna entropija	Nelin. fazni prostor	-cce, -dim*X*, -ksi (finoća podjele prostora)	3.5.3.9
Spektralna entropija	Nelin. entropijska	frekvencijski parametri**, -spc	3.5.3.11
ApEn	Nelin. entropijska	-ape (ApEn1, Apen2, ApEn3, ApEn4), -mape, -mfactAp*X* (faktor *m* za ApEn), -mfAp	3.5.3.7
SampEn	Nelin. entropijska	-sme, -msampe, -mfactSamp*X*, -mfSamp	3.5.3.8
Entropija SampEn na više skala	Nelin. entropijska	-msme, -msscа*X*, -mfactSamp*X*	3.5.3.10
Rényijeva entropija	Nelin. entropijska	-ren, -reo*X* (red Rényijeve entropije)	3.5.3.12
Indeks asimetrije na više skala	Nel. ostalo	-mai, -masca	3.5.3.16
Kompleksnost Lempel-Ziv	Nel. ostalo	-lzc	3.5.3.19
Rekurentni crtež	Nel. ostalo	-rcr	3.5.3.18
Higuchijeva FD	Nelin. fraktalna	-hig, -hfd*X* (kmax za FD)	3.5.3.4
Hurstov eksponent	Nelin. fraktalna	-hur	3.5.3.5
DFA	Nelin. fraktalna	-dfa, -msdfa*X* (najmanja duljina segmenta), -aldfa*X* (granica za alfaL)	3.5.3.6
SD Haarovog valića	Vremensko-frekvencijska	-hw*X*, *X* jedan od {3, 4, 8, 16, 32} (skala)	3.4.1
STA	Nelin. fazni prostor drugog reda razlike	-sea	4.2
ASTA	Nelin. fazni prostor drugog reda razlike	-asta	4.2
Abecedna entropija	Nelin. entropijska	-ae, -aet*X* (prag za promjenu, s); -aeAveLet, -aeVarLet, -aeMaxLet (prosjek, varijanca i maksimum po slovu abecede), -aeAve, -aeVar, -aeMax (prosjek, varijanca i maksimum ukupno po svim slovima), -aeRate (udio pojave pojedinih slova), -aeExi (je li se pojavilo pojedino slovo)	4.1

* značajke vezane isključivo uz analizu srčanog ritma
** frekvencijski parametri: jedan od: {-fft,-burg}; ako fft: {-winHann,-winHamm}, ako burg: {-ord*X*}; može i -psdout (za ispis u zasebnu datoteku procjene PSD na svim frekvencijama)

5.4 Izlazni rezultat

Nakon odabira parametara radnog okvira i značajki te nakon što su vrijednosti značajki izračunate, vektori značajki izlučuju se u izlaznu datoteku. Vektor za vektorom dodaje se na kraj datoteke u procesu izlučivanja značajki po segmentima unutar jednog zapisa i tako za svaki zadani ulazni zapis. Izlazna datoteka ima ekstenziju .arff i može se stvoriti koristeći parametre radnog okvira. Pri tome se navode značajke koje će radni okvir zapisati u .arff datoteku kao atribute u obliku „@attribute <naziv_atributa> <tip_atributa>“. Tip atributa je "numeric" za gotovo sve trenutno izlučivane značajke. Datoteke s ekstenzijom .arff dalje mogu pročitati sustavi za otkrivanje znanja kao što su Weka i RapidMiner.

5.5 Usporedba s postojećim rješenjima

Programski sustav za analizu varijabilnosti rada srca pod nazivom ECGLab ostvaren je u programskom okruženju Matlab [de Carvalho 2002]. Prva inačica ECGLaba sadržavala je rutine za otkrivanje R-vrha u zapisu EKG-a kao i rutine za učitavanje već pripremljenih ASCII datoteka koje sadržavaju RR-intervale. Implementirane značajke u vremenskoj domeni uključuju: srednju vrijednost, SDNN, RMSSD i pNN50. Linearna frekvencijska analiza je također implementirana i uključuje: FFT uz mogućnost korištenja pet različitih funkcija prozora, AR-model, i Lomb-Scargleov periodogram. Nelinearne značajke uključuju Poincaréov crtež (za značajku SD1/SD2) i analizu slijednog trenda (STA). U kasnijem radovima, ista grupa autora implementirala je i dodatne značajke. Te su uključivale vremensko-frekvencijsku analizu signala i to: STFT, CWT skalogram i autoregresijsko modeliranje s vremenskim pomakom [de Carvalho 2003]. Također je implementirana dodatna nelinearna značajka DFA [Leite 2010]. Prednost ECGLaba je njegova integriranost s Matlabom u smislu širokih mogućnosti rukovanja podacima i vizualizacijom. Također, implementacija je vrlo brza. Nedostatci uključuju ovisnost alata o dostupnosti komercijalnog programskog paketa Matlab i njegovim klasifikacijskim mogućnostima te razmjerno malom broju značajki u odnosu na HRVFrame.

ECGLabu srodni sustav KARDIA je nedavno razvijen programski sustav u Matlabu za analizu RR-intervala [Perakakis 2010]. Sustav ima vizualno sučelje i omogućuje analizu nekoliko zapisa odjednom. KARDIA omogućava standardnu vremensku i frekvencijsku analizu zapisa srčanog ritma. Frekvencijska analiza uključuje procjenu PSD korištenjem bilo FFT s funkcijom prozora ili AR-model procijenjen Burgovim algoritmom. KARDIA također može analizirati fazne srčane odgovore (engl. *phase cardiac responses*, kraće: PCR) koristeći teoriju brojanja razlomljenih ciklusa, objašnjenu u [Dinh 1999]. Dodatno, programski sustav omogućuje izračun fraktalnog skaliranja koristeći postupak DFA. U usporedbi s radnim okvirom HRVFrame, KARDIA omogućuje izlučivanje manjeg broja značajki te, kao i ECGLab, ima slične prednosti i nedostatke zbog implementacije u Matlabu.

Projekt BioSig je zanimljiva inicijativa koja ima za zadatak normirati alate za obradu signala i omogućiti lakšu pretvorbu datoteka iz jednog oblika u drugi. Također, cilja se na ponovljivost rezultata i na lakšu usporedbu sa srodnim istraživanjima [Schlögl 2008]. Trenutna implementacija projekta je otvorenog koda, programirana u jezicima C/C++. Iako su neki dijelovi samostojeći, većina alata radi samo s platformom Matlab, što autori priznaju kao nedostatak. Glavno područje primjene ovog softvera je u interakciji mozga i računala (BCI), što uključuje istraživanje svojstava povezanosti dijelova mozga korištenjem podataka na temelju EEG-a. Projekt također omogućuje analizu srčanog ritma u vremenskoj i frekvencijskoj domeni. Autori naglašavaju da implementacije naprednije analize EKG-a i varijabilnosti srčanog ritma ovise o stručnjacima, što znači da se softver može nadograditi u tu svrhu ako će postojati dovoljan interes. Glavna

prednost projekta BioSig je otvorenost koda, dok je glavni nedostatak nedovoljan broj implementiranih općenitih značajki za analizu biomedicinskih vremenskih nizova.

Programski paket pod nazivom HRV Analysis Software (novi naziv je Kubios) [Niskanen 2004] je slobodno-dostupni programski produkt za Windows platformu koji omogućuje analizu varijabilnost srčanog ritma preko vizualnog sučelja. Komponente su bile izvorno napisane u Matlabu, no kasnije su prevedene u jezik C tako da je programski produkt sada samostojeći. Značajke koje program izlučuje uključuju: linearnu vremensku domenu (srednja vrijednost, SDNN, RMSSD, pNN50), linearnu frekvencijsku domenu i to neparametarski FFT i parametarski AR-model te standardna odstupanja SD1/SD2 u faznom prostoru drugog reda. Programski produkt omogućuje unos podataka o RR-intervalima iz ASCII-datoteka, izlaz u oblik .csv kao i izvještaj u obliku spremnom za ispis na pisaču. Ovim sustavom nije omogućeno nadzirano učenje, budući da se mogu učitavati samo vrijednosti RR-intervala, ne i oznake vrste otkucaja ili ritmova. Time je omogućena samo temeljna statistička analiza dobivenih rezultata. Također, nedostatak drugih značajki korištenih u istraživanju HRV je problem u slučaju znanstvenog istraživanja korištenjem ovog programskog produkta. Prednost je grafičko sučelje i jednostavnost uporabe.

U konačnici, većina postojećih računalnih alata ima samo ograničenu znanstvenu vrijednost u analizi BVN, s nedovoljnom specijaliziranosti za specifično područje i to zbog malog broja implementiranih značajki. Korištenje licenciranog softvera (Matlab) ograničava uporabu alata ECGLab, KARDIA i BioSig. Za buduće istraživanje i razvoj, BioSig se čini kao projekt koji najviše obećava u znanstvenom smislu.

6 Dubinska analiza podataka

6.1 Definicija i opće odrednice dubinske analize podataka

Dubinska analiza podataka (engl. *data* mining, kraće: DM, također i: engl. *knowledge discovery in datasets*, kraće: KDD) računalni je način obrade podataka koji podrazumijeva razne netrivijalne postupke koji imaju za cilj pronalaženje razumljivih, ispravnih, novih i potencijalno korisnih obrazaca u skupovima podataka [Fayyad 1996].

Potrebno je uočiti da engleski naziv *data mining* sugerira da je dubinska analiza podataka zapravo iskapanje grumena znanja iz gomila podataka [Seewald]. Osnovu za dubinsku analizu podataka čini konačan skup podataka dobiven iz nekog procesa koji se odvija u stvarnom svijetu. Na temelju podataka o procesu moguće je izraditi model ponašanja nekog određenog dijela procesa koji je nekome od interesa. Dubinska analiza podataka se zbog svoje složenosti treba odvijati na računalu. U teoriji, dubinska analiza podataka ostvariva je i bez računala, no praktično vrijeme potrebno za njegovo provođenje je u tom slučaju predugo. Pri analizi nekog procesa uz pomoć postupaka dubinske analize podataka pretpostavljaju se sljedeći nužni preduvjeti:

1. Proces koji se analizira je malen, ograničen i točno određen dio svijeta.
2. Cilj analize je jasan.
3. Podaci koji su dostupni dovoljno su kvalitetni za opis procesa.
4. Podaci su prikazani u dani u obliku prikladnom za analizu, što najčešće znači u obliku tablice primjeraka (objekata) i mjerenih značajki (varijabli, atributa).

Dubinska analiza podataka obuhvaća tehnike strojnog učenja (engl. *machine learning*, kraće: ML), statistike i baza podataka. To je iterativni proces koji podrazumijeva više koraka. Različita literatura daje finiju ili grublju podjelu procesa analize. Najgrublja podjela je ona u dvije faze i to: faza pripreme podataka i faza modeliranja podataka. Nešto finija podjela je u tri koraka, koji uključuju pripremu podataka, pregled podataka i modeliranje podataka. Na pripremu podataka u slučaju podjele procesa dubinske analize na tri koraka otpada oko 60% ukupnog vremena. Ponekad se u pripremu podataka uključuje i pribavljanje podataka, i u tom slučaju ukupno vrijeme potrošeno na pripremu se približava 90% [Pyle 1999].

U ovom radu primijenit će se postupci dubinske analize podataka na problemu analize BVN. U kontekstu te domene, ovdje se navode i pojašnjavaju oni postupci koji se koriste u okviru predloženog sustavnog postupka analize BVN, koji čine tek mali podskup svih raspoloživih postupaka dubinske analize podataka.

6.1.1 Označavanje i korištena terminologija

U daljnjim poglavljima koristit će se sljedeće oznake i terminologija.

Sa $I = \{I_1, I_2, \ldots, I_n\}$ označavat će se skup primjeraka (objekata), kojih u skupu podataka *S* ima *n*. Sa $A = \{A_1, A_2, \ldots, A_M\}$ označavat će se prediktivni atributi, kojih u skupu podataka ima *M*. Kaže se da je *M* dimenzija skupa podataka, a primjerci *I* su točke u *M* - dimenzionalnom prostoru. U kontekstu dubinske analize na skupu podataka govorit će se o atributima, s time da se podrazumijeva da su atributi izlučene značajke koje se razmatraju u postupku dubinske analize BVN. Jedan ciljni atribut (razred poremećaja) označavat će se s *G*, $G \in A$, dok će se s $C = \{C_1, C_2, \ldots, C_k\}$ označavati skup *k* diskretnih vrijednosti koje može poprimiti ciljni atribut *G* (promatra se *k* poremećaja). Sa $\tau_{Aj} = \{\tau_1, \tau_2, \ldots, \tau_n\}_{Aj}$, označavat će se izmjerene vrijednosti nekog atributa A_j, dok će se s $V_{Aj} = \{V_1, V_2, \ldots, V_d\}_{Aj}$, označavati skup *d* diskretnih vrijednosti koje atribut A_j može poprimiti, ako je taj atribut kategorijskog tipa.

Primjerci za učenje označavat će se s (x_i, y_i), pri čemu je x_i vektor vrijednosti prediktivnih atributa, a y_i je vrijednost ciljnog atributa *G*. Klasifikator koji je izgrađen na temelju skupa za učenje u oznaci $R = \{(x_1, y_1), (x_2, y_2), \ldots, (x_r, y_r)\}, x_i \in I, y_i \in C, r \le n$ daje model *f* skupa podataka *S* . Za dani testni primjerak *x* iz skupa za testiranje *T* predviđanje ciljnog razreda dano je s $y = f(x), y \in C$, dok je stvarna vrijednost jednaka određenom $t \in C$. Pretpostavit će se da je ciljni atribut uvijek samo jedan i to samo kategorijskog tipa.

6.2 Priprema podataka i vrednovanje rezultata

6.2.1 Učenje i testiranje

Početni skup podataka *S* koji je na raspolaganju istraživaču je u većini slučajeva uzorak populacije. Da bi postupak za razvrstavanje mogao izgraditi model koji se dobro ponaša na čitavoj populaciji, potrebno je analizirati uzorak koji stoji na raspolaganju. Na temelju tog uzorka dobar algoritam za razvrstavanje izgradit će model koji će biti točan na čitavoj populaciji, uz pretpostavku da je uzorak reprezentativan dio populacije. Pogreška koju će postupak razvrstavanja počiniti na čitavoj populaciji procjenjuje se na uzorku i to tako da se dio uzorka podijeli na skup za učenje i skup za testiranje. Alternativno, ako postoji dovoljno podataka, skup podataka može se podijeliti na skupove za učenje, validaciju i testiranje.

Skup za učenje služi da postupak razvrstavanja izgradi modele koji će što točnije opisati prirodu populacije. Skup za validaciju služi da se procijeni pogreška različitih modela s različitim parametrima da bi se odabrao onaj najbolji. Procjene pogreške na populaciji obavlja se tako da se odredi pogreška na izdvojenom skupu za testiranje. Skup za testiranje dat će pravu, nepristranu prosudbu pogreške razvrstavanja na populaciji ako je i sam skup za testiranje reprezentativan uzorak populacije i ako ga se ne koristi za učenje niti za validaciju.

Problematika provjere modela izgrađenog nad skupom za učenje na nekom skupom za testiranje je zanimljiv i složen statistički problem. Budući da su rijetko kad istraživaču dostupna dva različita uzorka (skupa podataka) o istoj populaciji, nužno je odlučiti kako najbolje podijeliti dostupni skup podataka na skup za učenje, skup za testiranje i eventualno skup za validaciju. U pravilu, veći je problem ako je skup primjeraka za učenje malen, nego ako je velik. Navode se sljedeća empirijska pravila:

1. Za skupove s više tisuća primjeraka (po razredu, ako se radi o razvrstavanju), koristi se podjela skupa na dva (ili tri) dijela (engl. *holdout procedure*). Pritom ako se skup dijeli u dva dijela, obično se 67% uzima za učenje, a 33% za testiranje [Witten 2005] iako su i druge podjele moguće. Ako se skup dijeli u tri dijela, tada se 50% uzima za učenje, 25% za validaciju, a 25% za testiranje [Hastie 2009].

2. Za skupove između sto i tisuću primjeraka po razredu uvriježeni je postupak unakrsne validacije s k preklopa (engl. *k-fold cross-validation*) [Kohavi 1995]. Alternativno, može se koristiti i postupak unakrsne validacije s deset preklopa i deset ponavljanja [Bouckaert 2003].

4. Za male skupove koristi se uzorkovanje s ponavljanjem (engl. (i dalje): *bootstrap*) [Kohavi 1995], računalno zahtjevan postupak unakrsne validacije s izostavljanjem samo jednog primjerka (engl. *leave-one-out cross-validation,* kraće: LOOCV) [Hastie 2009]. ili postupak unakrsne validacije s dva preklopa i pet ponavljanja (engl. *5x2-fold cross-validation*) [Dietterich 1998].

Pokazuje se da je za postupak unakrsne validacije najbolje uzeti $k = 10$ preklopa [Kohavi 1995]. Koristeći taj postupak, podaci se podijele slučajno u deset jednakih dijelova, postupak učenja se obavlja uvijek nad devet dijelova, a testiranje nad preostalim dijelom. Učenje se ponovi deset puta, tako da svaki dio skupa bude uzet jednom za testiranje. S obzirom na to da se pri procjeni pogreške koristi 90% podataka u skupu, procjena pogreške u populaciji bit će usporedive točnosti kao da se koristio čitav skup podataka. Pokazano je i to da je najčešće unakrsna validacija poželjnija od postupka LOOCV, koji je primjer ekstremnog preklapanja [Kohavi 1995]. Naime, pri tome se samo jedan primjerak ostavlja za testiranje, a uči se na preostalim primjerima i to se ponovi za N preklopa, gdje je N broj primjeraka u skupu podataka. Ipak, za manje od 100 primjeraka, oba postupka daju dobre rezultate. Ako je broj primjeraka zaista malen, *bootstrap* postupak je najpogodniji. Tu se pri učenju izvlači N primjeraka na način da je svaki primjerak moguće izvući više puta ili nijednom. Ako se primjerak nijednom ne izvuče, ulazi u skup za testiranje. U prosjeku, skup za testiranje sadržavat će oko 36.8% primjeraka [Alpaydin 2004].

6.2.2 Pristranost i varijanca algoritama za razvrstavanje, prenaučenost klasifikatora

Prije razmatranja samih algoritama razvrstavanja, potrebno je objasniti obilježja koja su usko vezana uz perfomanse i ograničenja algoritama razvrstavanja da bi se razumjelo

zašto se u okviru ove disertacije razmatra više algoritama razvrstavanja. Ovdje će se ukratko razmotriti pristranost (engl. *bias*) i varijanca (engl. *variance*) te će se objasniti pojam prenaučenosti klasifikatora, dok će se neka druga obilježja razmotriti kroz pojedine algoritme razvrstavanja.

6.2.2.1 Pristranost i varijanca algoritma za razvrstavanje

Autor [Mitchell 1980] definira pristranost kao: „svaku osnovu za odabir jedne generalizacije umjesto druge, osim one koja je u striktnom slaganju s opaženim primjercima za učenje“. Svaki algoritam može se promatrati u kontekstu pristranosti koju uvodi u skup podataka i koja se očituje na modelu. Prema autoru [Fürnkranz 1999], pristranost algoritma može se podijeliti u tri tipa:

1. Pristranost jezika za prikaz podataka ili pristranost ograničenja (engl. *language bias, restriction bias*).
2. Pristranost postupka pretraživanja ili preferencijom (engl. *search bias, preference bias*).
3. Pristranost izbjegavanja prenaučenosti (engl. *overfitting avoidance bias*).

Odabir jezika za prikaz podataka koji koristi neki postupak sam po sebi unosi pristranost u model. Naime, neki koncepti ne mogu biti jasno izraženi u prostoru hipoteza određenog algoritma, te time algoritam unosi pogrešku pri klasifikaciji. Pokazuje se da što je prostor hipoteza koji algoritam može pokriti složeniji, to je i ova vrsta pristranosti manja. Stabla odluke općenito imaju malu pristranost jezika.

Postupak pretraživanja je način na koji se pretražuje prostor hipoteza, a to uključuje algoritam pretraživanja: uspon prema vrhu (engl. *hill climbing*), u širinu, u dubinu, zrakasto pretraživanje (engl. *beam search*) i sl. kao i strategiju pretraživanja: od vrha prema dnu, od dna prema vrhu te heuristiku pretraživanja koja se koristi za vrednovanje nađenih hipoteza. Stabla odluke i indukcijska pravila mogu imati veliku pristranost pretraživanja, iako bolji postupci imaju u pravilu malu pristranost.

Varijanca modela je mjera koja iskazuje koliko su predviđene vrijednosti koje daje klasifikator izgrađen nekim algoritmom raspršene u odnosu na ulazni vektor vrijednosti. Za algoritme koji imaju veliku varijancu, kaže se da su nestabilni. To znači da mala promjena u vrijednosti ulaznog vektora utječe značajno na prediktivnu/klasifikacijsku vrijednost atributa.

6.2.2.2. Pogreška generalizacije i prenaučenost klasifikatora

Sa stajališta parametarskih algoritama razvrstavanja, pokazuje se da vrijede sljedeća opažanja vezana uz pogrešku generalizacije na skupu za testiranje [Hastie 2009], slika 6.1:

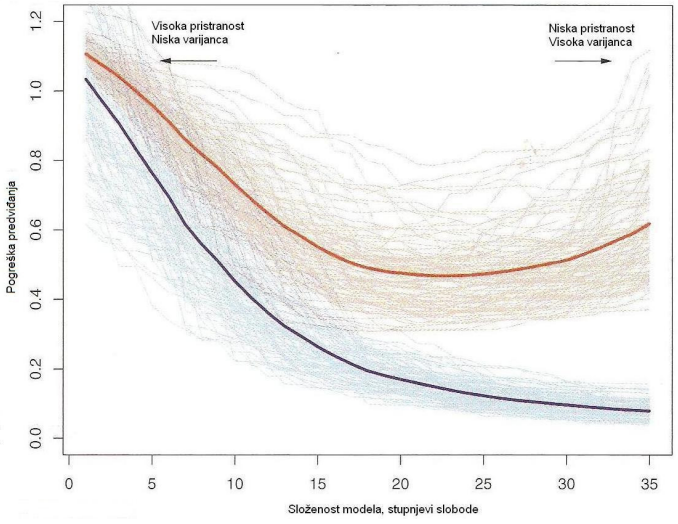
1. Model s premalo parametara (stupnjeva slobode) je netočan zbog prevelike pristranosti (nije dovoljno fleksibilan).
2. Model s previše parametara (stupnjeva slobode) je netočan zbog prevelike varijance (prevelika osjetljivost na uzorak za učenje).

3. Identifikacija modela koji najbolje generalizira na skupu za testiranje zahtijeva identifikaciju odgovarajućeg broja parametara, niti premalog niti prevelikog.

Primjeri nestabilnih parametarskih postupaka s visokom varijancom i niskom pristranosti uključuju: stabla odluke, neuronske mreže, klasifikacijska pravila, dok su stabilni postupci s niskom varijancom, ali visokom pristranosti neparametarski ili malo-parametarski postupci: *k*-najbližih susjeda, linearna diskriminantna analiza (dio diskriminantne analize), linearna regresija, naivni Bayes.

Pristranost algoritma se samo u manjoj mjeri može popraviti izmjenom strukture koju postupak gradi ili odabirom odgovarajućeg postupka pretraživanja. S druge strane, postoji više postupaka za smanjenje varijance odnosno povećanja stabilnosti osnovnog algoritma koji vrlo učinkovito spuštaju pogrešku nastalu zbog varijance. Najpoznatiji primjer postupka za smanjenje varijance je izgradnja ansambala klasifikatora koji većinski glasaju na ciljnom atributu.

Jedan od ciljeva uspješnog algoritma za razvrstavanje je da izbjegne prenaučenost. Pristranost koju postupak za izbjegavanje prenaučenosti pritom uvodi najčešće se zasebno razmatra. Prenaučenost je pojam koji govori o tome da klasifikator predobro nauči skup za učenje što mu daje slabe generalizacijske sposobnosti na skupu za testiranje, vidjeti sliku 6.1. U izvedbi, složenost modela se povećava kako bi obuhvatila što je više moguće karakteristika skupa za učenje. Obuhvaćanje svih detalja skupa za učenje ne znači da će model obuhvatiti i sve karakteristike populacije pa mu pogreška na skupu za testiranje u pravilu raste. Različiti algoritmi za razvrstavanje na različit način nastoje smanjiti ili izbjeći pogrešku zbog prenaučenosti. Primjerice, stabla odluke izbjegavaju prenaučenost podrezivanjem, no time postaju na određeni način pristrana. Neuronske mreže se uči na nekom skupu podataka samo određeni broj iteracija koji se određuje na temelju karakteristika same mreže. Ako je broj iteracija velik, naučit će se predobro struktura skupa podataka za učenje i mreža neće dobro generalizirati.



Slika 6.1. Prosječna pogreška predviđanja na skupovima za učenje (doljnja linija) i za testiranje (gornja linija), prilagođeno iz [Hastie 2009].

6.2.3 Diskretizacija numeričkih atributa

Diskretizacija numeričkih atributa je postupak pretvorbe atributa s kontinuiranim brojčanim vrijednostima u kategoričke atribute. Diskretizacija je često nužan preduvjet da bi pojedini algoritmi strojnog učenja mogli funkcionirati. Iako je jasno da se bilo kojom diskretizacijom gubi na preciznosti opisa stvarnog svijeta koji se promatra, nije jasno gubi li se i na informativnosti i točnosti budući da je moguće da u podacima postoji šum koji uzrokuje pretjeranu raspršenost numeričkih vrijednosti. U svakom slučaju, razlikuje se informirana diskretizacija atributa, koja je ovisna o ciljnom atributu, i neinformirana diskretizacija koja uzima u obzir samo dotični atribut. U okviru ovog doktorata zanimljiva je informirana diskretizacija koja se koristi u postupcima izgradnje stabla odluke i kod sustava zasnovanih na pravilima. Informirana entropijska diskretizacija koju su predložili autori [Fayyad 1993] najčešće je korištena vrsta informirane diskretizacije pa će se ona detaljnije opisati.

Entropijska diskretizacija je rekurzivan algoritam koji dijeli raspon vrijednosti numeričkog atributa uvijek nanovo na dva dijela. Točka podjele odabire se tako da dijelovi imaju što veću homogenost u smislu vrijednosti ciljnog atributa. Kao mjera homogenosti uzima se entropija, dana izrazom:

$$E(S) = -\sum_{i=1}^{k} p_i \log p_i \qquad \text{(bit)} \tag{6.1},$$

pri čemu je S skup primjeraka u tom dijelu, a p_i frekvencija pojave i-tog razreda u tom dijelu. Budući da se traži najhomogenija podjela, odabire se takva točka za koju će utežana suma entropija oba dijela biti najmanja. Entropije dijelova utežane su s brojem primjeraka u dotičnom dijelu. Postupak se ponavlja rekurzivno sve dok se ne zadovolji kriterij za zaustavljanje zasnovan na najmanjoj duljini opisa (engl. *minimum description length*, kraće: MDL).

Izabrana podjela intervala na dva dijela je opravdana, ako je razlika entropija izražena u bitovima između početnog raspona vrijednosti i dva njegova dijela veća od:

$$\frac{\log_2(N-1)}{N} + \frac{\log_2(3^k - 2) - kE + k_1E_1 + k_2E_2}{N} \qquad \text{(bit)} \tag{6.2},$$

pri čemu je N broj primjeraka, k je broj razreda u početnom rasponu vrijednosti, k_1 i k_2 broj razreda u svakom od dijelova, E je entropija u početnom rasponu vrijednosti, a E_1 i E_2 entropije u svakom od dijelova.

6.2.4 Vrednovanje izgrađenih klasifikatora

Za vrednovanje izgrađenih modela klasifikatora mogu se koristiti različite mjere koje procjenjuju kolika je pogreška generalizacije. Kao najosnovnija mjera navodi se ukupna točnost razvrstavanja (ukupna klasifikacijska točnost, engl. *total classification accuracy*, ACC) ili kraće: točnost. Ona se definira kao [Witten 2005]:

$$\text{ACC} = \frac{\text{broj ispravno razvrstanih primjeraka}}{\text{ukupan broj primjeraka}} \quad (6.3).$$

Navodeći samo mjeru točnosti ne uzimaju se u obzir tipovi pogrešaka na skupu podataka niti cijena koju se plaća kad se pogriješi. Iz tih razloga točnost sama po sebi nije dovoljna za usporedbu točnosti rezultata klasifikatora kao niti za primjenu modela koji daju samo točnost razvrstavanja u biomedicinskih istraživanjima [Provost 1997].

Da bi se bolje opisali rezultati modela, potrebno je na neki pregledan način predočiti greške koje je model napravio. Tako se navodi matrica iz koje se može očitati točno kakve su greške počinjene tijekom razvrstavanja. Ta se matrica naziva matrica zabune (engl. *confusion matrix*). Za slučaj razvrstavanja u dva ciljna razreda (npr. bolesni/zdravi) primjer matrice zabune dan je u tablici 6.1:

Tablica 6.1. Primjer matrice zabune u slučaju razvrstavanja u dva ciljna razreda.

		Prognozirani razred	
		bolesni	**zdravi**
Stvarni razred	**bolesni**	45	10
	zdravi	5	55

Tablica zabune se tumači tako da se gleda koliko je točnih pogodaka koje je model dao i koje su vrste grešaka počinjene. U tablici 6.2. prikazani su mogući ishodi pri razvrstavanju:

Tablica 6.2. Mogući ishodi pri razvrstavanju u dva ciljna razreda.

		Prognozirani razred	
		bolesni	**zdravi**
Stvarni razred	**bolesni**	stvarno pozitivan **TP**	lažno negativan **FN**
	zdravi	lažno pozitivan **FP**	stvarno negativan **TN**

Stvarno pozitivni (TP) i stvarno negativni (TN) ishodi smatraju se ispravnim razvrstavanjima, dok se lažno pozitivni (FP) i lažno negativni (FN) smatraju neispravnim razvrstavanjima. Ako je neki primjerak razvrstan kao lažno pozitivan, tada to znači da je postupak razvrstavanja prognozirao da je primjerak pozitivan, dok je on ustvari negativan. Ako je pak neki primjerak razvrstan kao lažno negativan, tada to znači da je postupak razvrstavanja prognozirao da je primjerak negativan, dok je on ustvari pozitivan. Imajući u vidu moguće ishode dane u tablici 6.2, izraz za ukupnu točnost postupka razvrstavanja (6.3) moguće je zapisati kao:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.4).$$

Iz matrice zabune moguće je izdvojiti i druge mjere za vrednovanje rezultata modela osim same točnosti. Tako se u literaturi vezanoj uz područje analize biomedicinskih

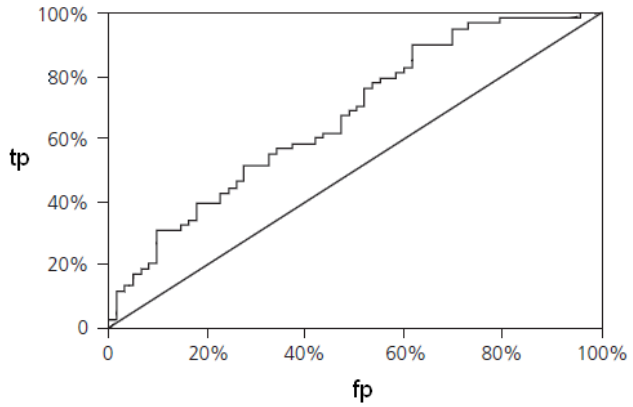
Tablica 6.3. Mjere vrednovanja rezultata razvrstavanja.

Naziv	Engleski naziv	Izračun
Ukupna točnost razvrstavanja	*Total classification accuracy*	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$
Stopa stvarno pozitivnih	*True positive rate,* TP *rate*	$tp = \frac{TP}{TP + FN}$
Stopa lažno pozitivnih	*False positive rate,* FP *rate, fp*	$fp = \frac{FP}{FP + TN} = 1 - SPEC$
Osjetljivost	*Sensitivity*	$SENS = \frac{TP}{TP + FN} = tp$
Specifičnost	*Specificity*	$SPEC = \frac{TN}{FP + TN} = 1 - fp$
Predvidljivost pozitivnih primjeraka	*Positive predictive value*	$PPV = \frac{TP}{TP + FP}$

vremenskih nizova spominju sljedeće poznatije mjere vrednovanja rezultata modela [Fawcett 2003, Witten 2005, Karraz 2006], tablica 6.3.

Od ostalih korištenih mjera vrednovanja modela potrebno je spomenuti dvodimenzionalni grafikon poznat kao krivulja ROC (od engl. *Receiver Operating Characteristic*, ROC *curve*). Krivulja ROC predočava performanse postupka razvrstavanja bez obzira na razdiobu vrijednosti ciljnih razreda i bez obzira na cijenu pogreške. Primjer krivulje ROC dan je na slici 6.2. Na *y*-osi prikazuje se stopa točno pozitivnih *tp*, a na *x*-osi stopa lažno pozitivnih *fp*. Stope se mogu iskazati bilo u obliku decimalnog broja između 0 i 1 bilo kao postotak. Nazubljena linija na slici je krivulja ROC [Witten 2005].

Postupak razvrstavanja je to bolji što je krivulja ROC bliža lijevom gornjem uglu grafa. Kao mjera za opis krivulje ROC navodi se površina pod krivuljom (engl. *Area Under Curve*, kraće: AUC). Mjera AUC ima važno statističko svojstvo: AUC nekog



Slika 6.2. Primjer krivulje ROC, prilagođeno iz [Witten 2005].

postupka razvrstavanja je jednaka vjerojatnosti da će postupak razvrstavanja rangirati slučajno izabran pozitivan primjerak više od slučajno izabranog negativnog primjerka. Ova mjera se izračunava sumacijom po svim trapezoidima ili pravokutnicima od kojih je izgrađena krivulja ROC.

U slučaju kad se provodi razvrstavanje u više ciljnih razreda, kao što je slučaj u ovom radu, tada je potrebno iznaći najbolji način kako da se odrede mjere TP, TN, FP i FN. Intuitivno, ono što je najlakše napraviti je da se problem opet svede na razvrstavanje u dva razreda. Pritom se razmatra sve iz konteksta primjeraka jednog ciljnog razreda *j* u odnosu na primjerke svih ostalih razreda. U tablici 6.4 prikazano je kako se u tom slučaju definiraju mjere TP, TN, FP i FN, na primjeru četiri razreda ciljnog atributa.

Ideja je da svi primjerci ostalih razreda koji se ne nalaze na području koji pokrivaju *j*-ti redak i *j*-ti stupac predstavljaju stvarno negativne (TN), primjerci *j*-tog stupca predstavljaju lažno negativne (FN), a primjerci *j*-tog retka lažno pozitivne (FP). Stvarno pozitivni su samo oni primjerci na dijagonali na mjestu gdje se sijeku *j*-ti redak i *j*-ti stupac. Postupak se ponavlja za svaki razred ciljnog atributa posebno tako da ukupno postoji k mikro-mjera $TP_{Ci}.TN_{Ci}, FP_{Ci}, FN_{Ci}, i = 1,...,k$. Moguće je izračunati i sve mikro-mjere navedene u tablici 6.3 za svaki razred C_i. Konačno, mogu se izračunati i sve mjere iz tablice 6.3 na rezultatima za sve razrede zajedno. Tako se definiraju sljedeće mjere:

$$TP = \sum_{i=1}^{k} TP_{Ci}\ ,\ TN = \sum_{i=1}^{k} TN_{Ci}\ ,\ FP = \sum_{i=1}^{k} FP_{Ci}\ ,\ FN = \sum_{i=1}^{k} FN_{Ci},,$$

$$SENS = \frac{TP}{TP + FN}, SPEC = \frac{TN}{FP + TN}, PPV = \frac{TP}{TP + FP} \qquad (6.5).$$

Ovako definirane mjere nazivaju se mikro-uprosječene mjere i često se koriste kod analize BVN [Sebastiani 2002]. Krivulja ROC se također može prilagoditi za slučaj više ciljnih razreda. To se radi tako da se nacrta *k* krivulja promatranjem svakog razreda zasebno, slično kao i za slučaj ostalih jednostavnijih numeričkih mjera. Ono što se gubi pri ovom postupku je neosjetljivost krivulje ROC na zakrivljenost razdiobe zbog same definicije mjera TP i FP (vidi tablicu 6.4). Međutim, osim ovog manjeg nedostatka, krivulja ROC se i dalje primijenjuje učinkovito u praksi i omogućuje razumnu fleksibilnost pri vrednovanju [Fawcett 2003]. Mjeru AUC moguće je također pouzdano

Tablica 6.4. Definicija mjera TP, TN, FP, i FN u slučaju razvrstavanja u više ciljnih razreda.

		Stvarni razred			
		a	**b**	**c**	**d**
Predviđeni razred	**a**	55	15	20	5
	b	12	45	4	10
	c	0	5	120	15
	d	8	12	10	12

- TP
- TN
- FP
- FN

izračunati u slučaju više razreda promatrajući parove ciljnih razreda na ROC krivulji na način [Hand 2001]:

$$AUC_{total} = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} AUC(C_i, C_j) \qquad (6.6).$$

6.3 Postupci odabira atributa

Računalni postupci za odabir atributa mogu se podijeliti na nekoliko načina. Prva podjela je prema tome je li postupak ispituje značajnost podskupa atributa ili ispituje pojedinačne atribute i rangira ih prema značajnosti [Guyon 2002]. Dobivanje podskupa značajnih atributa postiže se određenim postupkom pretraživanja prostora atributa i računalno je mnogo zahtjevniji postupak od rangiranja. Rangiranje atributa postiže se kvantificiranjem utjecaja pojedinog atributa na ciljni atribut koristeći određenu metriku.

Druga podjela je prema tome koriste li se i na koji način se koriste algoritmi razvrstavanja pri odabiru atributa. Prema toj podjeli, postoje tri glavne skupine postupaka: filterski (engl. *filters*), omotači (engl. *wrappers*) i ugrađeni (engl. *embedded*) [Kohavi 1997]. Kao četvrtu skupinu neki autori izdvajaju i postupke prekrivanja [Liu 1996]. U ovom doktoratu dat će se detaljniji prikaz algoritama za odabir značajki prema ovoj drugoj podjeli s naglaskom na postupke koji se koriste u predloženom sustavnom postupku.

6.3.1 Filterski postupci

Filterski postupci odabiru podskup atributa heuristički za vrijeme pretprocesiranja, u ovisnosti o statističkim značajkama samog skupa podataka, a neovisno od odabranog postupka za klasifikaciju ili predviđanje. Filterski postupci su najbrži u odnosu na ostale postupke odabira atributa. Filterski postupci dijele se prema tome rangiraju li pojedinačne atribute ili vrednuju li skupove atribute. U ovom poglavlju razmatraju se samo postupci koji rangiraju značajke prema nekoj metrici, dok se oni postupci koji vrednuju skupove atributa razmatraju u poglavlju 6.3.4 kao postupci prekrivanja. Pregled poznatijih filterskih postupaka za rangiranje korištenih u dubinskoj analizi podataka dan je u tablici 6.5.

U tablici je dan izraz kojim se metrika izračunava (samo za jednostavne metrike), tip atributa primjene (kategorički, binarni, kontinuirani), i literatura. U nastavku se detaljnije opisuju informacijski dobitak, simetrična nesigurnost i metrika jednog pravila. Ove tri metrike su primjenjive na problemu razvrstavanja koji se razmatra u okviru ove disertacije. Sva tri postupka mogu baratati s ciljnim atributom koji ima više od dva razreda. Također, ovi postupci uzimaju se u obzir zbog velike brzine izvršavanja (npr. zbog sporosti ne koristi se ReliefF postupak) i čestog korištenja s provjereno uspješnim rezultatima na skupovima podataka s razvrstavanjem u više razreda [Forman 2003].

Tablica 6.5. Pregled filterskih metrika korištenih za odabir značajnih atributa rangiranjem.

Naziv metrike	Izračun	Tip atributa: prediktivni/ ciljni	Literatura
Informacijski dobitak (IG)	$IG = E(S) - \sum_{i=1}^{m} \frac{\lvert S_i \rvert}{\lvert S \rvert} E(S_i)$	kat. i kont. / kat.	[Zheng 2004]
Omjer dobitaka (GR)	$GR = \frac{\text{IG}}{-\sum_{i=1}^{m} \frac{\lvert S_i \rvert}{\lvert S \rvert} \log_2 \frac{\lvert S_i \rvert}{\lvert S \rvert}}$	kat. i kont. / kat.	[Grimaldi 2003]
Omjer izgleda	$\log \frac{TP \cdot TN}{FP \cdot FN}$	bin./ bin.	[Forman 2003]
χ - kvadrat		kat./ kat.	[Forman 2003]
ReliefF		kat. i kont./ kat.	[Robnik-Šikonja 2003]
Simetrična nesigurnost (SU)	$SU(A_i, A_j) = 2 \frac{H(A_i) - H(A_i \mid A_j)}{H(A_i) + H(A_j)}$	kat. i kont. / kat.	[Yu L. 2004]
Koeficijent omjera signal-šum	$\frac{\overline{\{\tau_j\}_{Aj}, \tau_{jG} = C_1} - \overline{\{\tau_j\}_{Aj}, \tau_{jG} = C_2}}{s(\{\tau_j\}_{Aj}, \tau_{jG} = C_1) + s(\{\tau_j\}_{Aj}, \tau_{jG} = C_2}$	kont./ bin.	[Golub 1999]
FAST	Procjena površine ispod krivulje ROC	kont./ bin.	[Chen X. 2008]
Jedno pravilo (1Rule)	$1Rule = \sum_{i=1}^{m} \frac{\lvert S_i \rvert}{\lvert S \rvert}$	kat. i kont. / kat.	[Holte 1993]

6.3.1.1 Informacijski dobitak

Informacijski dobitak (engl. *information gain*, kraće: IG) je informacijska mjera koja se koristi prilikom procjene atributa u postupcima razvrstavanja. Mjera je dvostrana, što znači da zanemaruje predznak ispred ocjene atributa i promatra samo apsolutnu vrijednost ocjene.

U slučaju razvrstavanja primjeraka u više ciljnih razreda definirana je entropija skupa podataka *S* izrazom:

$$E(S) = -\sum_{i=1}^{k} p_i \log p_i \qquad (6.7),$$

pri čemu je *k* broj ciljnih razreda u skupu podataka, a p_i vjerojatnost pojave razreda C_i, procijenjena iz skupa podataka. Na osnovi entropije definira se informacijski dobitak nekog atributa A_j, koji služi kao mjera učinkovitosti tog atributa u razvrstavanju primjera:

$$IG = E(S) - \sum_{i=1}^{m} \frac{\lvert S_i \rvert}{\lvert S \rvert} E(S_i) \qquad (6.8).$$

Ovdje je *m* skup diskretnih vrijednosti koje atribut A_j može poprimiti, dok je S_i skup

svih primjeraka za koje je vrijednost jednaka *i*-toj diskretnoj vrijednosti atributa A_j, tj. $S_i = \{I \in S, A_j(I) = V_i\}$

Prvi dio u izrazu (6.8) označava entropiju čitavog skupa podataka, dok težinska suma u drugom dijelu iskazuje očekivanu vrijednost entropije podskupova nastalih grananjem atributa A_j prema njegovim diskretnim vrijednostima. U slučaju da atribut A_j nije kategorički, prije pokretanja izračuna informacijskog dobitka atribut se diskretizira višeintervalnim postupkom, vidi poglavlje 6.2.3.

Informacijski dobitak je dokazano koristan u slučaju kad se kao postupak razvrstavanja koristi stablo odluke C4.5, budući da dotični algoritam razvrstavanja koristi sličnu informacijsku mjeru za odabir najprikladnije značajki pri podijeli u čvoru stabla (omjer dobitaka), poglavlje 6.4.1. Autori [Wasikowski 2010] pokazali su da će informacijski dobitak u prosjeku biti najbolja ili gotovo najbolja metrika za odabrani skup podataka ako postoji neravnoteža u broju primjeraka po razredima ciljne značajke. Također su pokazali da kod bioloških skupova podataka s oko 1000 atributa informacijski dobitak daje vrlo dobre rezultate. Robustnost na neravnotežu u broju primjeraka je dodatni razlog korištenja ove filterske metrike, budući da je ona česta kod razreda u medicinskoj primjeni.

6.3.1.2. Simetrična nesigurnost

Simetrična nesigurnost (engl. *symmetric uncertainty*, kraće: SU) je informacijska mjera slična informacijskom dobitku koja kompenzira slučaj u kojem atribut ima velik broj vrijednosti [Yu L. 2004]. Dana je izrazom:

$$SU(A_i, A_j) = 2\frac{H(A_i) + H(A_j) - H(A_i, A_j)}{H(A_i) + H(A_j)} = 2\frac{MI(A_i, A_j)}{H(A_i) + H(A_j)} \quad (6.9),$$

pri čemu je s $H(A_i)$ označena entropija atributa kategoričkog atributa A_i, $H(A_i) = -\sum_{j=1}^{k} p_j \log p_j$, s $H(A_i, A_j)$ združena entropija dvaju kategoričkih atributa, a s $MI(A_i, A_j)$ označena zajednička informacija dvaju kategoričkih atributa. U slučaju da neki od atributa nije kategorički, prije pokretanja izračuna simetrične nesigurnosti atributi se diskretiziraju višeintervalnim postpukom, poglavlje 6.2.3.

Prilikom izgradnje poretka atributa po značajnosti, koristi se simetrična nesigurnost između pojedinog atributa i ciljnog atributa, dakle:

$$SU(A_i, C) = 2\frac{MI(A_i, C)}{H(A_i) + H(C)} \quad (6.10)$$

i prema [Peng H. 2005]:

$$MI(A_i, C) = -\sum_{j=1}^{|Ai|}\sum_{k=1}^{|C|} p(A_{i,j}, C_k)\log\frac{p(A_{i,j}, C_k)}{p(A_{i,j})p(C_k)} \quad (6.11).$$

Vrijednost simetrične nesigurnosti je uvijek između [0,1] pri čemu vrijednost 1 znači da znajući vrijednosti jednog atributa možemo posve točno predvidjeti vrijednost drugoga, a vrijednost 0 znači da su atributi posve neovisni [Yu L. 2004].

6.3.1.3 Postupak jednog pravila

Postupak jednog pravila (engl. *1-Rule*, kraće: 1Rule) jedan je od najjednostavnijih klasifikatora u strojnom učenju. Razvio ga je autor [Holte 1993]. Ideja je iz skupa svih atributa pronaći onaj atribut koji ima najmanju pogrešku na skupu za učenje pri razvrstavanju primjeraka u ciljni razred. Za numeričke atribute provodi se diskretizacija. Najprije se primjerci sortiraju po vrijednosti atributa i zatim se svrstavaju u kućice definirane mjestima gdje se vrijednost ciljnog atributa mijenja. Ako kućica ima manje od *n* primjeraka većinskog razreda (preporuča se $n = 6$), tada se kućica proširi. Ovakva diskretizacija numeričkih atributa provede se uvijek prije nego što počne usporedba koji atribut ima najmanju pogrešku. Za kategoričke atribute ili za numeričke diskretizirane atribute, pogreška se ocjenjuje podjelom atributa po svim kategoričkim vrijednostima:

$$1Rule = \sum_{i=1}^{m} \frac{|S_i|}{|S|} \tag{6.12},$$

gdje je *m* skup diskretnih vrijednosti koje atribut A_j može poprimiti, dok je $S_i = \{I \in S, A_j(I) = V_i\}$, slično kao i kod informacijskog dobitka. Atributi se rangiraju po stopi pogreške od najmanje prema najvećoj.

6.3.2 Postupci omotača

Postupak omotača koristi određeni klasifikator za procjenu uspješnosti podskupa atributa [Kohavi 1997]. Tipično se uspješnost mjeri ukupnom točnosti razvrstavanja. Za procjenu uspješnosti podskupa atributa na skupu za testiranje može se koristiti unakrsna validacija, izdvajanje skupa za testiranje ili bilo koji drugi postupak naveden u poglavlju 6.2.1. Ako se odabere unakrsna validacija s deset preklopa, tada je za svaki razmatrani podskup atributa potrebno izgraditi deset klasifikatora, što usporava postupak pronalaženja bitnih atributa. Rezultati dobiveni postupcima omotača u pravilu daju točnije rezultate, odnosno bolji odabir značajnih atributa od onih dobivenih filterskim postupcima [Hall 2000, Wasikowski 2010]. S druge strane, glavni problem postupaka omotača je visoka računalna složenost, čime ovi postupci postaju gotovo neuporabljivi na velikim skupovima atributa.

Način odabira podskupa atributa je bitan faktor koji utječe na računalnu složenost i uopće, izvedivost postupka omotača. U pravilu, postoje dva načina odabira skupa atributa: odabir unaprijed (engl. *forward selection*) i uklanjanje unazad (engl. *backward elimination*). Odabir unaprijed započinje s praznim skupom atributa te dodaje i ispituje određeni broj atributa u svakom koraku. Uklanjanje unazad započinje od čitavog skupa atributa i uklanja određeni broj atributa u svakom koraku te promatra utjecaj na točnost klasifikatora. U pravilu, eliminacija unazad daje nešto bolje rezultate u praksi. Glavni

razlog tome je taj da ako se pojavi lokalno najbolje rješenje na kojem se zaustavi odabir podskupa atributa, to nije toliki problem ako još postoji određeni višak atributa. Algoritam razvrstavanja koji će se dalje koristit će ga vjerojatno moći ukloniti. Veći je problem ako odabirom unaprijed još nisu niti uzeti u obzir neki od bitnijih atributa [Witten 2005].

Za pretraživanje prostora atributa bilo unaprijed bilo unazad mogu se koristiti razni heuristički postupci kao što su najbolji prvi (engl. *best-first*), postupak pohlepnog uspona na vrh (engl. *greedy hill climbing*), pretraživanje zrakom, pretraživanje genetskim algoritmom (engl. *genetic search*). Složeniji postupci koji ubrzavaju pretraživanje prostora atributa uključuju utrku (engl. *race*), raspršeno pretraživanje (engl. *scatter search*), pretraživanje zasnovano na shemama (engl. *schemata search*), i druge [Witten 2005, López 2006].

Glavni zadatak svih heurističkih postupaka pretraživanja prostora atributa je da smanje računalnu složenost pretraživanja na što manju moguću polinomijalnu, ali bez garancije pronalaska najboljeg podskupa atributa kao što je to slučaj kod iscrpnog pretraživanja, poglavlje 6.3.4. Korištenje postupka omotača ima jedan bitan problem, a to je da odabir značajnog podskupa atributa dosta zavisi o samom klasifikatoru koji se koristi. Iz razloga računarske složenosti ukupnog postupka, nije poželjno koristiti složeni algoritam za izgradnju klasifikatora. S druge strane, prejednostavni algoritmi možda neće davati dovoljno točna rješenja. U praksi se pokazalo da ima malo takvih algoritama strojnog učenja koji bi bili zahvalni za upotrebu pri izgradnji klasifikatora kod postupka omotača. Jedan od češće korištenih je selektivni naivni Bayesov postupak, što je naivni Bayesov postupak koji se gradi pretraživanjem prostora atributa unaprijed i kome se greška ispituje na skupu za učenje [Witten 2005, Hall 2000].

6.3.3 Ugrađeni postupci

Ugrađeni postupci odabira atributa (engl. *embedded attribute selection*) su takvi postupci koji su specifični za pojedini algoritam strojnog učenja. Oni izabiru bitne atribute tijekom procesa učenja. Za razliku od postupaka omotača koji na izgrađeni klasifikator gledaju kao na crnu kutiju za koji je jedino bitan izlazni odgovor, ugrađeni postupci su nerazdvojivi od samog postupka izgradnje. Do rješenja obično dolaze brže od postupaka omotača jer ne moraju učiti postupak modeliranja svaki puta iznova s nešto drugačijim podskupom atributa – čitav odabir provodi se samo jednom, na skupu za učenje. Većinu algoritama strojnog učenja može se koristiti za takvo filtriranje atributa. Primjerice, algoritam *k*-najbližih susjeda je poznat da ima značajnu degradaciju performansi u slučaju nebitnih atributa. Algoritam C4.5 može se upotrijebiti za dobivanje samo onih atributa koji su značajni, a to su oni koji se pojavljuju u čvorovima stabla, tj. oni koji su odabrani za grananje [Witten 2005].

6.3.4 Postupci prekrivanja

Postupci prekrivanja su takvi postupci koji imaju zadatak naći podskup od *m* atributa od ukupnog broja od *M* atributa, $m \le M$, takvih da ostvare barem jednako dobar rezultat u smislu zadovoljenja zadanog kriterija. Tipičan kriterij koji se želi postići postupcima prekrivanja je što manja pogreška na skupu za testiranje [Liu 1996].

U općenitom slučaju, broj ispitivanja svih mogućih skupova od *m* značajki iz skupa od *M* značajki iznosi:

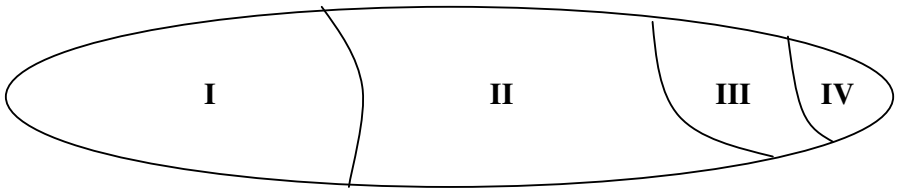
$$\binom{M}{m} = \frac{M!}{m!(M-m)!}. \qquad (6.13).$$

Ako postoji *m* bitnih atributa, tada je ukupna složenost eksponencijalna:

$$\sum_{i=1}^{m}\binom{M}{i} \succ O\left(M^m\right) \qquad (6.14).$$

Jedan od jednostavnijih postupaka je taj da se pronađe takav skup atributa koji će biti taman dovoljan (najmanji) da sve primjerke za učenje potpuno razdvoji. Kod nekih skupova to nije moguće ostvariti, budući da neki put čak i koristeći sve atribute nije moguće potpuno razdvojiti primjerke – i to u slučaju kad postoje takvi primjerci koji imaju jednake vrijednosti svih atributa osim ciljnog. Ipak, kod većine skupova podataka ideja je pronaći najmanji podskup atributa takav da jedinstveno opisuje svaki primjerak. Ovo se može postići iscrpnim pretraživanjem podskupova atributa, međutim takvo pretraživanje je računarski iznimno zahtjevno. Dodatni problem je taj što se pronalaskom takvog podskupa atributa na skupu primjeraka za učenje ne garantira ista uspješnost na skupu za testiranje. Naime, pronalaženjem najmanjeg podskupa atributa za potpuno razdvajanje može kasnije doći do prenaučenosti klasifikatora građenog na temelju takvog skupa atributa.

Drugi način pogleda na postupak prekrivanja je u smislu značajnosti i redundantnosti atributa. Pritom je zadatak pronaći sve atribute koji su jako značajni i one koji su slabo značajni, ali ne redundantni. Sve ostale bi trebalo odbaciti. Na slici 6.3 prikazana je ideja takve podjele atributa, prema autoru [Yu L. 2004]. Do atributa u skupinama III i IV moguće je doći primjenom informacijske teorije zasnovane na Markovljevom pokrivaču (engl. *Markov blanket*). Osnovna ideja je pronaći skup atributa koji uklanjaju potrebu za korištenjem nekog atributa A_j, koji se u tom slučaju smatra redundantnim.



I: nebitni atributi, II: slabo bitni atributi, redundantni, III: slabo bitni atributi, nisu redundantni, IV: jako bitni atributi. Najbolji podskup atributa: III i IV

Slika 6.3. Podjela atributa prema značajnosti i redundantnosti.

Tablica 6.6. Pregled najčešćih postupaka prekrivanja za odabir značajnih atributa.

Vrsta postupka	Naziv postupka	Prediktivni/ ciljni atribut	Literatura
Iscrpno pretraživanje	FOCUS	kat./ kat.	[Almuallim 1994]
Slučajno pretraživanje	Konzistentnost razreda (InCons)	kat./ kat.	[Liu 1996]
Redundantnost	Filter zasnovan na redundantnosti (RBF)	kat./ kat.	[Yu L. 2004]
Korelacijski	Odabir zasnovan na korelaciji (CFS)	kat./ kat.	[Hall 2000]

Postupci prekrivanja prema načinu primjene mogu biti filterski ili postupci omotača. U slučaju kad nije konačni cilj samo pronaći bitne atribute, već se dalje trebaju primijeniti postupci razvrstavanja, češće se koriste filterski postupci koji su neovisni o postupku razvrstavanja. U tablici 6.6 dan je pregled najpoznatijih filterskih postupaka za prekrivanje, zajedno s pripadnom literaturom. Od navedenih postupaka, filter zasnovan na redundantnosti (engl. *redundancy based filter*, kraće: RBF) je jedini koji rangira atribute po značajnosti, dok algoritmi FOCUS, konzistentnost razreda (engl. *class consistency*, kraće: InCons) i odabir značajki zasnovan na korelaciji (engl. *correlation-based feature selection*, kraće: CFS) pronalaze podskupove značajnih atributa. U nastavku se detaljnije opisuje samo postupak konzistentnosti razreda, budući da se on koristi u sustavnom postupku predloženom u ovom radu.

6.3.4.1. Konzistentnost razreda

Poznati postupak prekrivanja zasnovan na vjerojatnosti predložili su autori [Liu 1996]. Autori odabiru na slučajan način podskup atributa koji potom provjeravaju prema jednostavnom kriteriju za konzistentnost između skupa atributa i ciljnog razreda. Stopa nekonzistentnosti za neki podskup od *m* atributa dana je s:

$$InCons(S_m) = \frac{\sum_{i=0}^{J} |D_i| - |M_i|}{N} \tag{6.15},$$

pri čemu je *J* broj skupova primjeraka koji imaju iste vrijednosti na svih *m* atributa, $|D_i|$ je broj primjeraka u *i*-tom takvom skupu primjeraka, $|M_i|$ je broj primjeraka u *i*-tom skupu primjeraka koji imaju najčešću vrijednost ciljnog razreda u odnosu na sve ostale primjerke u *i*-tom skupu, a *N* je ukupan broj primjeraka u skupu podataka. Naprimjer, razmatra se podskup od $m = 5$ atributa od ukupno njih 15 na skupu od 50 primjeraka i za takav podskup pronađe se da postoje $J = 2$ skupa primjeraka s istim vrijednostima na ovih 5 atributa. Neka prvi takav skup ima $|D_1| = 15$ primjeraka, od čega njih $|M_1| = 9$ pripada razredu C_1, 5 razredu C_2 i 1 razredu C_3. Neka drugi takav skup ima $|D_1| = 8$ primjeraka i to 3 u razredu C_1, 1 u razredu C_2 i $|M_2| = 4$ u razredu C_3. Tada ukupna stopa nekonzistentnosti podskupa od 5 atributa iznosi: $InCons(S_5) = \frac{(15-9)+(8-4)}{50} = 0.2$.

Kao parametar algoritma zadaje se prag za stopu nekonzistentnosti koji treba biti zadovoljen da bi se određeni podskup atributa uzeo u obzir. Uobičajeno je uzeti za vrijednost praga onu stopu nekonzistentnosti koju ima ukupan skup od *M* atributa. Time sad problem postaje pronalazak pravog podskupa atributa koji će imati manju stopu nekonzistentnosti od ukupnog skupa. Između više takvih podskupova pamti se samo onaj s najmanjom stopom nekonzistentnosti. Ostaje neriješeno kako pronaći skup od *m* atributa s manjom nekonzistentnosti. Autori predlažu pretraživanje slučajnim odabirom podskupa atributa u svakoj iteraciji ispitivanja nekonzistentnosti. U slučaju velikog broja atributa, potrebno je puno iteracija da bi se pronašao najbolji podskup. Kao parametar postupka može se zadati broj iteracija koje treba provesti ili alternativno, koliki postotak prostora svih mogućih podskupova treba pregledati. Što veći postotak podskupova se pregleda, to je veća šansa pronaći onaj s najmanjom nekonzistentnosti.

6.4 Algoritmi razvrstavanja

U ovom poglavlju daje se prikaz nekoliko poznatih algoritama strojnog učenja primjenjivih na razvrstavanje primjeraka u više ciljnih razreda koji se koriste u okviru ove disertacije. Za svaki algoritam opisan je postupak izgradnje klasifikatora na skupu za učenje. Također su navedeni parametri koje je moguće mijenjati u cilju postizanja točnijih rezultata odnosno manje pogreške generalizacije na skupu za testiranje. Potrebno je naglasiti da je izbor algoritama koji su ovdje opisani proizvoljan. Ipak, u obzir pri odabiru algoritama uzete su sljedeće karakteristike:

1. Da je algoritam poznat i često upotrebljavan postupak strojnog učenja.
2. Primjenjivost algoritma na skup podataka koji se analizira u okviru ove disertacije.
3. Brzina algoritma razvrstavanja treba biti dovoljna za potencijalnu primjenu u stvarnom vremenu, što uključuje i brzinu izgradnje klasifikatora na skupu za učenje, a posebice i brzinu razvrstavanja novih primjeraka na izgrađenim modelima.

6.4.1 Stablo odluke C4.5

Stablo odluke C4.5 je vjerojatno najpoznatiji algoritam strojnog učenja. Razvio ga je autor [Quinlan 1993]. Stablo odluke C4.5 služi za razvrstavanje primjeraka u ciljne razrede. Algoritam pripada u opću skupinu postupaka „podijeli-pa-vladaj“ (engl. *divide-and-conquer*) zajedno s drugim postupcima izgradnje stabla i produkcijskim pravilima. Algoritam gradi stablastu strukturu u kojoj se u svakom čvoru ispituju vrijednosti atributa s ciljem što informiranije podjele skupa primjeraka i što točnijeg razvrstavanja.

C4.5 omogućuje razvrstavanje primjeraka u više ciljnih razreda, pri čemu je dozvoljeno da atributi budu kategorički ili numerički. Također, algoritam zna baratati s primjercima koji sadrže neke nedostajuće vrijednosti i relativno je otporan na šum uz pravilno podešene parametre.

Stablo se gradi od korijena prema listovima, pri čemu se ne podrezuje sve dok nije do kraja izgrađeno. Stablo se smatra do kraja izgrađenim ako su mu listovi čisti, što znači da sadrže sve primjerke istog razreda ili ako na krajnjem čvoru-listu nije moguće više podijeliti primjerke. Primjerke nije moguće dalje dijeliti ako imaju iste vrijednosti svih atributa, a različite ciljne razrede.

Količina informacije u skupu primjeraka R' u nekom čvoru mjeri se u odnosu na vrijednosti ciljnog atributa *G* koristeći Shannonovu mjeru entropije danu izrazom:

$$E(R') = \sum_{i=1}^{|C|} - p(C_i) \log_2 p(C_i) \tag{6.16},$$

pri čemu je s $p(C_i)$ označen udio ciljnog razreda C_i u skupu podataka R'. Najveća vrijednost entropije dana je u slučaju kad su udjeli svih ciljnih razreda jednaki i iznosi $\log_2|C|$, a najmanja iznosi 0, i to kad svi primjerci pripadaju istom ciljnom razredu.

Informacijski dobitak koji je ostvaren podjelom atributa A_j na nekom čvoru iznosi:

$$IG = E(R') - \sum_{i=1}^{m(Aj)} \frac{|R'_i|}{|R'|} E(R'_i) \tag{6.17},$$

pri čemu je s $m(Aj)$ označen broj različitih kategorija atributa A_j. Kategorički atributi se dijele po svim kategorijama, a numerički atributi dijele se na dva dijela pri čemu se ispitaju sve moguće točke prekida između dviju sortiranih vrijednosti atributa. Mjera informacijskog dobitka je pristrana s obzirom na atribute s mnogo vrijednosti, stoga se umjesto nje koristi njena normirana vrijednost, omjer dobitaka (GR), dan izrazom:

$$GR = \frac{\text{IG}}{-\sum_{i=1}^{m(Aj)} \frac{|R'_i|}{|R'|} \log_2 \frac{|R'_i|}{|R'|}} \tag{6.18},$$

Stablo odluke C4.5 se podrezuje nakon izgradnje, primjenjujući postupak naknadnog podrezivanja (engl. *post-pruning*).

Algoritam C4.5 obavlja dva tipa podrezivanja nakon izgradnje, a to su zamjena podstabla i uzdizanje podstabla. Zamjena podstabla provodi se nad unutrašnjim čvorovima stabla koji kao djecu imaju samo listove. Za svaki takav čvor razmatra se opravdanost zamjene sa samo jednim listom prema pesimističnoj procjeni pogreške na skupu za učenje (engl. *pessimistic pruning*):

$$e = \frac{f + \frac{z^2}{N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \tag{6.19},$$

pri čemu se za dani parametar pouzdanosti procjene *c* nađe granica pouzdanosti *z*, dok je $f = E/N$ opažena stopa pogreške pri korištenju većinskog razreda u dotičnom listu, gdje je *E* broj primjeraka koje ne pripadaju većinskom razredu, a *N* je broj svih primjeraka u dotičnom čvoru. Opravdanost zamjene određuje se na temelju mjere pogreške prije zamjene i nakon zamjene. Ako do zamjene dođe, tada se postupak

ponavlja rekurzivno uz strukturu stabla dok god ima zamjena koje uzrokuju da unutarnji čvor ima samo listove.

Uzdizanje podstabla je složenija operacija od zamjene podstabla. Ideja je da se cijelo podstablo koje počinje s nekim čvorom *B* uzdigne za jednu razinu tako da zamijeni čvor *A* koji je prije toga postojao iznad čvora *B*. U praksi se ta operacija obavlja samo ako je dijeljenje čvora *A* na podstablo koje počinje čvorom *B* i ostala podstabla (recimo *C* i *D*) takvo da u čvoru *B* postoji više primjeraka nego u čvoru *C* ili čvoru *D*. Pritom se čvor *A* mora granati na barem jedno podstablo *B* koje nije list da bi se zamijenio s tim podstablom (ili nekim drugim koji nije list). Podstablo se uzdiže također na temelju kriterija procjene pogreške *e*. Nakon podrezivanja u listovima ne smije biti manje primjeraka od korisnički zadanog broja $n_{\min}$, koji određuje osjetljivost stabla na šum. Veći $n_{\min}$ dat će manje i razumljivije stablo, no često manje točno. Osim pesimističnog podrezivanja postoji velik broj načina podrezivanja poznatih u literaturi. Vrlo dobar pregled tih postupaka dan je u disertaciji [Frank 2000].

6.4.2 Algoritam AdaBoost primijenjen na stablo odluke C4.5

AdaBoost je poznati algoritam za pojačavanje (engl. *boosting*) slabih klasifikatora koji su razvili autori [Freund 1995]. AdaBoost spada u skupinu algoritama strojnog učenja koji izgrađuju ansambl klasifikatora u više koraka počevši od praznog skupa klasifikatora. U svakoj iteraciji dodaje se novi klasifikator, a klasifikatori djeluju serijski, uzimajući na neki način u obzir rješenje prethodnog klasifikatora i kontinuirano ga poboljšavaju. Pritom prvi klasifikatori u seriji najčešće daje jednostavni model, dok se kasniji klasifikatori kompliciraju da bi detaljnije opisali specifičnost skupa podataka.

Osnovna ideja algoritma AdaBoost je da se pri učenju iz koraka u korak penaliziraju oni primjerci koji su ispravno razvrstani na skupu za učenje. Algoritam više ne uzima toliko u obzir ispravno razvrstane da bi se fokusirao na one problematičnije. Pritom se gradi više klasifikatora, njih *K*, i to tako da prvi klasifikator uzima jednako u obzir svaki primjerak iz skupa. Sljedeći klasifikator daje prednost onim primjercima koje prvi klasifikator nije točno razvrstao. To se postiže izmjenom težina pridijeljenih primjercima i to tako da se samo ispravno razvrstanima smanji težina za iznos:

$$w_j^{i+1} := w_j^i \frac{e_i}{1-e_i}, \quad e_i = \sum_{j=1}^{r} w_j^i e(x_j),\ e(x_j) = \begin{cases} 1, f_i^{Alg}(x_j) = y_j \\ 0, f_i^{Alg}(x_j) \neq y_j \end{cases} \qquad (6.20),$$

pri čemu je s $f_i^{Alg}(x_j)$ označen *i*-ti izgrađeni klasifikator kojeg gradi određeni algoritam strojnog učenja *Alg* koji se nastoji pojačati. Prilikom odabira primjeraka za novi skup za učenje, primjerci se izvlače iz skupa za učenje s ponavljanjem i to razmjerno svojoj težini.

Svaki sljedeći klasifikator će se sve jače koncentrirati na one primjerke koji su više puta neispravno razvrstani, odnosno češće će takve birati pri učenju. Ako pogreška nekog od klasifikatora prijeđe 50% (bez obzira na broj ciljnih razreda) ili ako je

postignuta pogreška jednaka 0 na skupu za učenje, provede se nešto preinačena izmjena težina [Schapire 1997]. Algoritam završava nakon što se izgradi *K* klasifikatora. Svakom izgrađenom klasifikatoru pridjeljuje se težina u ovisnosti od broja pogrešaka koje je napravio na skupu za učenje. Što je klasifikator napravio manje pogrešaka to će mu težina biti veća, budući da je pokazano pouzdaniji.

U fazi razvrstavanja primjeraka, svaki klasifikator u nizu daje prognozu o ciljnom razredu. Svaki puta kad je neki ciljni razred odabran, težina mu se povećava u ovisnosti od težine samog klasifikatora. Na kraju se glasanje provede na način da se primjerak razvrsta u onaj razred koji ima najveću težinu [Schapire 1997].

6.4.2.1 Primjena na stablo odluke C4.5

Izvorna namjena postupka AdaBoost bila je povećati točnost serije slabih stabala odluke, tzv. panjeva odluke (engl. *decision stumps*) koji se dijele samo na korijenskom čvoru u jednom atributu direktno u listove [Freund 1995]. Odabirom većeg broja panjeva koji se vrlo brzo grade i kombinirajući ih na odgovarajući način, bitno se povećava točnost ukupnog klasifikatora. Ipak, istraživanja su pokazala da se algoritam AdaBoost može primijeniti za povećanje točnosti bilo kojeg poznatog klasifikatora. Može se raspravljati o tome da glavni razlog uspjeha ovog algoritma leži u tome da uspješno smanjuje varijancu klasifikatora, ali i pristranost [Alpaydin 2004, Schapire 1997].

Varijanca algoritma karakteristična je za većinu algoritama za učenje stabala odluke, uključujući i postupak C4.5. Stoga su istraživači primijetili da bi bilo moguće dobiti bolje rezultate razvrstavanja ako se stablo C4.5 dodatno pojača korištenjem postupka AdaBoost. Na više skupova podataka iz stvarnog svijeta, uključujući i skupove biomedicinskih podataka, pokazano je da u pravilu algoritam AdaBoost poboljšava rezultate stabla odluke C4.5 [Quinlan 1996, Schapire 1997, Pramanik 2010].

Algoritam C4.5 može koristiti težine pridijeljene primjercima za učenje izravno prilikom izgradnje stabla odluke. Iz tog razloga nije potrebno provesti uzorkovanje s ponavljanjem prema težinama već se u obzir svaki put uzima čitav skup za učenje, samo s izmijenjenim težinama primjeraka. Stablo odluke C4.5 se gradi koristeći težine na sljedeći način.

Svugdje gdje se pobrojavaju primjerci, umjesto cijelog broja (1 primjerak) uzima se broj s pomičnim zarezom koji govori o težini primjerka (npr. 0.88 primjeraka). Tako npr. pri određivanju informacijskog sadržaja u čvoru $E(R') = \sum_{i=1}^{|C|} - p(C_i) \log_2 p(C_i)$ udio primjeraka koji pripadaju razredu C_i ne računa se više na način: $p(C_i) = N_{C(i)} / N_{UK}$, već kao $p(C_i) = w(x_{C(i)}) \Big/ \sum_{i=1}^{N} w(x_i)$, pri čemu je s $w(x_{C(i)})$ označena suma težina svih primjeraka koji pripadaju razredu C_i.

Izgrađeno stablo u listovima donosi odluku od pripadnosti ciljnom razredu na temelju većinskog udjela težina primjeraka koji pripadaju dotičnom razredu. Pri razvrstavanju testnih primjeraka nema razlike u odnosu na izvorni algoritam C4.5.

6.4.3 Slučajna šuma

Slučajna šuma (engl. *random forest*, kraće: RF) [Breiman 2001] spada u algoritme ansambala baznih klasifikatora, pri čemu svaki bazni klasifikator – slučajno stablo djeluje nezavisno nad određenim ulaznim skupom podataka i daje svoje rješenje, dok ukupni rezultat razvrstavanja nastaje većinskim glasanjem. U problemima razvrstavanja točnost rezultata koju daje RF usporediva je s najboljim algoritmima strojnog učenja, a brzina algoritma je često i bitno veća od drugih slično točnih postupaka.

Algoritam slučajnih šuma koristi sljedeće tehnike kako bi smanjio pogrešku generalizacije:

1. Šuma se gradi s mnogo stabala, pri čemu broj stabala ovisi o broju atributa.
2. Šuma koristi dva izvora slučajnosti: pri uzorkovanju i pri izgradnji stabla.
3. Na kraju čitava šuma većinski glasa za vrijednost ciljnog atributa.

Pristranost algoritma se minimizira tako što se stablo gradi do kraja i ne podrezuje se. Time ne dolazi do uklanjanja nekih dijelova stabala koji možda donose važnu informaciju. Greška zbog varijanca minimizira se uz pomoć *bootstrap* postupka uzorkovanja, slučajnog odabira atributa pri grananju (postupak slučajnog potprostora), i velikog broja stabala u šumi. Na temelju zakona velikih brojeva, Breiman je teorijski i empirijski pokazao da kako broj stabala raste, tako će šuma biti sve točnija, ali i da će u jednom trenutku točnost doći u zasićenje i šuma neće naučiti šum, što znači da je postupak neosjetljiv na prenaučenost [Breiman 2001].

Na početku postupka zadaje se broj stabala u šumi: *k*. Zatim, za svako stablo uzorkuje se slučajno *n* primjeraka postupkom *bootstrap* (uzorkovanje s ponavljanjem). Ovako odabrani *bootstrap* skup bit će skup uzoraka za učenje samo za dotično stablo. Odredi se broj atributa *m* koji će se razmatrati prilikom podjele svakog čvora u stablu. Uobičajeno je $m = \sqrt{M}$, gdje je *M* ukupan broj atributa. Također, postavi se najmanji dozvoljen broj primjeraka u čvoru – listu na $n_{\min} = 2$. Zatim se provede gruba zamjena nedostajućih vrijednosti u atributima. Ako je atribut kategorički, tada se sve nedostajuće vrijednosti zamijene s najčešćom vrijednosti atributa. Ako je atribut numerički, tada se sve nedostajuće vrijednosti zamijene medijanom postojećih vrijednosti. Stablo se gradi sve do listova i ne podrezuje se. Prilikom grananja na svakom čvoru:

1. Ako čvor ima manje od $n_{\min}$, ili je čvor čist u smislu da sadrži primjerke samo jednog ciljnog razreda čvor se ne grana i postaje list stabla.
2. Inače, odabere se slučajno *m* atributa – kandidata za grananje
3. Za svaki atribut – kandidat izmjeri se poboljšanje čistoće koje bi nastalo u slučaju podjele na tom atributu. Odabere se onaj atribut koji ima najveću vrijednost indeksa Gini (najveće smanjenje nečistoće).

4. Ako nema atributa koji bi dalje smanjio nečistoću, tada se uzima ona podjela koja ima najmanju nečistoću. Inače se čvor dijeli na dva čvora djece u slučaju numeričkog atributa ili u svaku od kategorija u slučaju kategoričkog atributa. Pokrene se korak 1 na čvoru djetetu i to tako da se stablo gradi u dubinu sve do listova.

Mjera nečistoće čvora pod nazivom indeks Gini pripada porodici mjera nečistoće koja uključuje i informacijski dobitak, omjer dobitka i druge [Robnik-Šikonja 2004]. Za neki kategorički atribut – kandidat A_l koji se razmatra na čvoru, indeks Gini koji mjeri smanjenje nečistoće iznosi [Breiman 1984]:

$$Gini(A_l) = -\sum_{i=1}^{k} p(C_i)^2 + \sum_{j=1}^{m(Al)} p(v_{Al,j}) \sum_{i=1}^{k} p(C_i \mid v_{Al,j})^2 \qquad (6.21),$$

pri čemu je $\sum_{i=1}^{k} p(C_i)^2$ komponenta koja mjeri čistoću čvora bez daljnje podjele, a $\sum_{j=1}^{m(Al)} p(v_{Al,j}) \sum_{i=1}^{k} p(C_k \mid v_{Al,j})^2$ komponenta koja mjeri čistoću podjele čvora na atributu – kandidatu A_l. Ovdje je $p(C_i)^2$ udio ciljnog razreda C_i u primjercima dostupnima u čvoru, $m(A_l)$ je broj kategorija atributa A_l, $p(v_{Al,j})$ je udio kategorije v_j atributa A_l u primjercima dostupnima u čvoru, a $p(C_i \mid v_{Al,j})$ je udio ciljnog razreda C_i u onim primjercima koji imaju kategoriju v_j atributa A_l. U slučaju da je atribut numerički, postoje dvije mogućnosti. Jedna je diskretizacija numeričkog atributa prije izgradnje šume (poglavlje 6.2.3), a druga je isprobavanje svih mogućih mjesta za podjelu atributa na dva skupa, pri čemu se odabere ona podjela koja je najčišća prema indeksu Gini.

Broj stabala u šumi odabire se u ovisnosti o skupu podataka. U pravilu, što je više atributa prisutno u skupu podataka to je potrebno više stabala. Za stotinjak atributa u skupu algoritam će postići vršnu točnost s otprilike 100 stabala [Breiman 2001].

Pri testiranju, svaki testni primjerak provuče se kroz sva izgrađena stabla od korijena stabla do listova, gdje se nalazi odgovor o ciljnom razredu. Izgrađeni bazni klasifikatori - slučajna stabla većinski glasaju za ciljni razred.

Iako je svako pojedino stablo moguće prikazati, složenost same šume (broj stabala, nepodrezivanje) onemogućuje prikaz pravila koje bi se jednostavno tumačilo. Slučajna šuma se stoga koristi samo za dobivanje velike točnosti razvrstavanja, a ne za probleme gdje je potrebno razumjeti kako je do neke odluke došlo.

Skup primjeraka iz početnog skupa za učenje koje neko stablo nije izabralo za učenje naziva se skup izdvojenih primjeraka (engl. *out-of-bag*, kraće: OOB). Prema teoriji uzorkovanja s ponavljanjem, u prosjeku 36.8% primjeraka iz izvornog skupa za učenje neće biti odabrano za učenje, već će ući u skup OOB. Svako izgrađeno stablo ima svoj skup primjeraka OOB koje služi za unutarnju procjenu pogreške generalizacije šume, određivanje značajnosti pojedinih atributa, finu zamjenu nedostajućih vrijednosti i u druge svrhe, vidi [Breiman 2001].

Zbog dvaju izvora slučajnosti algoritam RF spada u rjeđu skupinu algoritama sa stohastičkom komponentom, što znači da je za objektivnu procjenu rezultata razvrstavanja potrebno nekoliko ponavljanja postupka i usrednjavanje rezultata.

6.4.4 Stroj s potpornim vektorima za razvrstavanje

Stroj s potpornim vektorima (engl. *support vector machine*, kraće: SVM) je vrlo točan algoritam za razvrstavanje koji je razvio autor [Vapnik 1996]. U ovom radu opisat će se osnovni stroj s potpornim vektorima za razvrstavanje [Burges 1998, Hastie 2009]. Također će se ukratko opisati postupak slijedne najmanje optimizacije (engl. *sequential minimal optimization*, kraće: SMO) koji će se koristiti kao jedan od algoritama prilikom razvrstavanja podataka iz BVN [Platt 1998, Keerthi 2001].

SVM je jedan od najuspješnijih postupaka za određivanje decizijske granice između primjeraka dvaju razreda i stoga je najveću primjenu pronašao u raspoznavanju uzoraka [Burges 1998]. Postoji više vrsta algoritama SVM. U slučaju problema razmatranog u ovoj disertaciji, zanimljivo je razmatrati slučaj kad primjerci ne mogu biti linearno razdvojeni u izvornom prostoru i kada se obavlja transformacija podataka u višedimenzionalni prostor.

Osnovna ideja postupka je ta da je u višedimenzionalnom prostoru dobivenim transformacijom primjeraka za učenje (x_i, y_i) korištenjem baznih funkcija $h(x)$ olakšano razvrstavanje. U takvom prostoru granica između razreda (decizijska funkcija) postaje linearna, a preslikana u izvorni prostor ona postaje nelinearna. Decizijska funkcija u višedimenzionalnom prostoru ima oblik:

$$f(x) = h(x)^T \beta + \beta_0 = h(x)^T \sum_{i=1}^{r} \alpha_i y_i h(x_i) + \beta_0 = \sum_{i=1}^{r} \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0 \qquad (6.22),$$

pri čemu su s α_i označeni Lagrangeovi multiplikatori, β je vektor ovisan o funkcijama transformacije i ulaznim podacima, a β_0 je pomak decizijske granice u odnosu na ishodište. Klasifikacijsko pravilo za neki primjerak x u tom slučaju glasi $G(x) = sign[f(x)]$.

Da bi se došlo do parametara α_i i β_0 koji bi dali najtočniju klasifikaciju u smislu najšire margine između primjeraka dvaju razreda, razmatra se optimizacija Lagrangeove dualne funkcije koja ima oblik:

$$\begin{aligned} L_D &= \sum_{i=1}^{r} \alpha_i - \frac{1}{2} \sum_{i=1}^{r} \sum_{k=1}^{r} \alpha_i \alpha_k y_i y_k \langle h(x_i), h(x_k) \rangle \\ &= \sum_{i=1}^{r} \alpha_i - \frac{1}{2} \sum_{i=1}^{r} \sum_{k=1}^{r} \alpha_i \alpha_k y_i y_k K(x_i, x_k), \quad 0 \le \alpha_i < C, \sum_{i=1}^{r} y_i \alpha_i = 0 \end{aligned} \qquad (6.23),$$

pri čemu je s $K(x, x') = \langle h(x), h(x') \rangle$ označena funkcija jezgre kao skalarni produkt vektora baznih funkcija. Optimizacijski problem predstavljen izrazom (6.23) je kvadratni optimizacijski problem s obzirom na Lagrangeove multiplikatore α_i. Problem

je riješen kada su za svaki indeks *i* zadovoljeni sljedeći KKT (Karush-Kuhn-Tucker) uvjeti:

$$\begin{aligned} &\alpha_i = 0 \Leftrightarrow y_i f(x_i) \geq 1 \\ &0 < \alpha_i < C \Leftrightarrow y_i f(x_i) = 1 \\ &\alpha_i = C \Leftrightarrow y_i f(x_i) \leq 1 \end{aligned} \qquad (6.24).$$

Zanimljivo je da se određene oblike baznih funkcija skalarni produkt izračunava brzo i učinkovito. Često korištene funkcije jezgre su sljedeće:

1. Polinomijalna, *d*-tog stupnja: $K(x, x') = \left(1 + \langle x, x' \rangle\right)^d$
2. Radijalna: $K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$
3. Sigmoidalna: $K(x, x') = \tanh\left(\kappa_1 \langle x, x' \rangle + \kappa_2\right)$

Parametar β_0 može se dobiti iz (6.22) rješavajući $y_i f(x_i) = 1$ za bilo koji x_i za koji je $0 < \alpha_i < C$. U višedimenzionalnom prostoru definiranom baznim funkcijama često je moguće potpuno odvajanje razreda. Potporni vektori su oni primjerci koji podupiru najveću marginu između dva razreda. U idealnom slučaju, svi će potporni vektori biti na granici margine, a u realnom slučaju mali broj njih nalazit će se unutar margine, ali s ispravne strane decizijske funkcije.

Kvadratni oblik problema predočen izrazom (6.23) ne može se lagano riješiti zbog veličine matrice jezgre, koja ima broj elemenata jednak kvadratu broja primjeraka za učenje. Za više od 10 000 primjeraka takva matrica ne bi stala u memoriju današnjih računala. Da bi to riješio, autor [Platt 1998] predložio je algoritam SMO, koji razlaže izvorni optimizacijski problem u potprobleme.

Najmanji mogući optimizacijski problem uključuje samo dva Lagrangeova multiplikatora, α_1 i α_2. U svakom koraku SMO odabire dva multiplikatora da budu zajednički optimirani, nalazi optimalnu vrijednost i promijeni klasifikator (vrijednosti β i β_0) tako da reflektira dobivene vrijednosti. Prednost postupka je da se dva multiplikatora mogu optimirati analitički i da algoritam ne mora pamtiti cijelu matricu multiplikatora. Pokazuje se da algoritam brzo konvergira. Računski detalji vezani uz sam algoritam dostupni su u radu [Platt 1998]. Autori [Keerthi 2001] pokazali su da u nekim slučajevi Plattov algoritam ne daje najučinkovitije rješenje te su stoga predložili stabilniju i bržu varijantu postupka koja će se koristiti u ovoj disertaciji.

Algoritam SMO, baš kao i njegova poboljšana verzija, grade decizijsku granicu između dva razreda. U slučaju više razreda, nije jasno kako donijeti odluku o pripadnosti ciljnom razredu. Autor [Abe 2003] objašnjava da se takva odluka može donijeti na tri načina:

1) Izgradnjom *C* klasifikatora, i to po jedan za svaki razred, u usporedbi s primjercima svih ostalih razreda. U tom slučaju bira se onaj razred za koji vrijednost decizijske funkcije $f(x) = x^T \beta + \beta_0$ bude najveća u odnosu na sve ostale.

2) Izgradnjom $C(C-1)/2$ klasifikatora, što znači po jedan klasifikator između svaka dva ciljna razreda. U tom slučaju za neki primjerak bira se onaj razred koji ima najviše pozitivnih odgovora klasifikatora. Ovaj postupak se koristiti u ovom radu.
3) Simultanim razvrstavanjem u više razreda, pri čemu se sve decizijske granice moraju odrediti istovremeno, što rezultira problemom s mnogo više varijabli u odnosu na dva prethodna načina. Ovaj se oblik odluke najčešće ne koristi u praksi.

6.4.5 Algoritam za produkcijska pravila RIPPER

Algoritam RIPPER (engl. *Repeated Incremental Pruning to Produce Error Reduction*) je postupak industrijske snage za izgradnju pravila u obliku ako-onda za razvrstavanje primjeraka [Cohen W. 1995]. Glavne prednosti ovog algoritma u odnosu na slične algoritme s pravilima su velika učinkovitost po pitanju brzine i zauzeća memorije, visoka točnost izgrađenih modela razvrstavanja i jednostavnost izgrađenih pravila.

Opći oblik pravila ako-onda koje generira postupak RIPPER može se prikazati ovako:

$$(A_1 \rho\, v_c(A_1)) \wedge (A_2 \rho\, v_c(A_2)) \wedge \ldots \wedge (A_k \rho\, v_c(A_k)) \Rightarrow razred = C_i\,(prekrivanje \mid broj\ pogresaka) \tag{6.25},$$

pri čemu je s ρ označena neka od mogućih relacija $\rho = \{\leq, =, \geq\}$ između atributa A_i i njegove značajne vrijednosti $v_c(A_i)$, prekrivanje (engl. *coverage*) je broj primjeraka (izražen u težinama) čije vrijednosti atributa zadovoljavaju lijevu stranu pravila, a broj pogrešaka je onaj broj primjeraka (izražen u težinama) koji iako zadovoljavaju lijevu stranu pravila nisu ispravno razvrstani u ciljni razred.

Algoritam izgrađuje više pravila po principu ako-onda-inače, što znači da ako ne vrijedi prvo pravilo za neki primjerak, tada se ispituje sljedeće pravilo i tako sve do zadnjeg pravila koje prekriva sve preostale primjerke. Za svaki razred C_i ciljnog atributa, počevši od onog s najmanje primjeraka u skupu primjeraka za učenje, izvode se sljedeći koraci [Witten 2005]: 1) Izgradnja pravila, 2) Optimizacija, 3) Pometanje i čišćenje.

Pri izgradnji pravila, najprije se izvrši podjela skupa za učenje na skup za rast (engl. *growing set*) i na skup za podrezivanje (engl. *pruning set*), u omjeru 2:1. Zatim se gradi jedno ili više pravila za razvrstavanje u dotični razred C_i i to sljedećim algoritmom:

1. Rast pravila: Pravilu se dodaje takav uvjet $(A_1 \rho\, v_c(A_1))$ koji ima najveće smanjenje nečistoće u smislu najvećeg informacijskog dobitka mjere *G* definirane s:

$$G = p\left(\log\left(\frac{p}{t}\right) - \log\left(\frac{P}{T}\right)\right) \tag{6.26},$$

pri čemu je *p* broj ispravno razvrstanih primjeraka pokrivenih pravilom, *t* je prekrivanje pravila, tj. ukupni broj primjeraka koji zadovoljavaju lijevu stranu pravila, *P* je ukupan broj primjeraka razreda C_i u skupu primjeraka za rast, a *T* je ukupan broj

primjeraka u skupu primjeraka za rast. U slučaju kategoričkih atributa, relacija ρ u uvjetu koji se dodaje je uvijek jednakost „=“. Numerički atributi se ispituju na sva moguća mjesta podjele između svojih vrijednosti tako da se oblikuju dva skupa podataka: jedan manji od točke podjele i drugi veći te se relacija ≤ ili ≥ ustanovljava prema mjeri *G*. Uvjeti se nadodaju pohlepno na prethodne uvjete dok pravilo ne postigne nula pogrešaka s obzirom na svoju prekrivenost.

2. Podrezivanje pravila: Pravilo se podrezuje od krajnje desnog para atribut-vrijednost prema krajnje lijevom dokle god vrijednost $W = \frac{p+1}{t+2}$ pravila na skupu za podrezivanje raste.

3. Koraci 1 i 2 se ponavljaju, što znači da se grade nova pravila za opis razreda c_i sve dok nije ispunjen jedan od tri sljedeća uvjeta:
   a) Svi su primjerci razreda c_i prekriveni nekim pravilom.
   b) Opisna duljina (engl. *description length*, kraće: DL) skupa pravila pronađenih do tada je za 64 bita veća od najmanje opisne duljine pronađene do tada.
   c) Stopa pogreške posljednjeg pravila (jednaka omjeru broja pogreške i prekrivanja) prelazi 50%.

DL je složen izraz kojim se želi kodirati skup primjeraka za danu teoriju, odnosno skup pravila da bi ih se moglo poslati kroz komunikacijski kanal i dekodirati na drugoj strani. Autor [Cohen W. 1995], a na temelju rada o DL-u autora [Quinlan 1995] predlaže da se DL nekog pravila *R* izračunava izrazom:

$$DL_R = \frac{1}{2}\left(\|k\| + k \log_2 \frac{n}{k} + (n-k) \log_2 \frac{n}{n-k} \right) \qquad \text{(bit)} \qquad (6.27),$$

pri čemu je s $\|k\|$ dan broj bitova za slanje *k* uvjeta dotičnog pravila, a *n* je broj mogućih uvjeta za dotično pravilo, dok je cijeli izraz pomnožen s faktorom 0.5 da bi se kompenziralo za moguće redundancije u atributima. Opisna duljina čitavog skupa pravila jednaka je zbroju opisnih duljina svih pojedinačnih pravila.

Nakon što je izgrađen skup pravila u prvom koraku, algoritam RIPPER provodi optimizaciju pravila tako da generira novu varijantu pravila zasnovanu da smanjenom skupu primjeraka za podrezivanje i drugačijoj mjeri vrednovanja za podrezivanje i zatim odabire koja je varijanta pravila bolja. Postupak se ponavlja nekoliko puta dok se ne nađe najbolja varijanta pravila.

Pometanje i čišćenje se koriste za prekrivanje pravilima onih primjeraka koji još uvijek nisu prekriveni kao i za uklanjanje onih pravila koje doprinose povećanju DL-a. Točnost RIPPER-a je usporediva s postupkom C4.5 [Cohen W. 1995].

6.4.6 Naivni Bayesov postupak

Naivni Bayesov postupak pripada probabilističkim postupcima dubinske analize podataka i široj skupini postupaka Bayesovih mreža. Temeljna pretpostavka ovih postupaka je da se znanje koje imamo o nekom događaju u svijetu može opisati s

vjerojatnošću pojave tog događaja. Svaki događaj u modelu je moguć i ima makar malenu i uvijek postojeću vjerojatnost pojave. Ta vjerojatnost se zadaje *a priori* ili se izvodi iz podataka. Postoji nekoliko učinkovitih algoritama temeljenih na Bayesovim mrežama poznatih u literaturi [Webb 2005]. U ovom radu ukratko će se objasniti najjednostavniji, naivni Bayesov postupak kao i njegovo proširenje u vidu više Gaussovih jezgri za točniji opis vrijednosti numeričkih atributa.

Naivni Bayesov model gradi se na pretpostavci o međusobnoj nezavisnosti atributa pri čemu ciljni atribut ovisi samo o pojedinačnim atributima, a ne o njihovim parovima ili složenijim kombinacijama. *A posteriorna* vjerojatnost da će ciljni atribut za neki *j*-ti testni primjerak $x_j \in T$ imati vrijednost razreda C_i dana je Bayesovom formulom:

$$p(C_i \mid x_j) = \frac{p(x_j \mid C_i)\, p(C_i)}{p(x_j)} \tag{6.28},$$

pri čemu je $p(x_j)$ *a priorna* vjerojatnost pojave primjerka s takvim određenim rasporedom vrijednosti svih prediktivnih atributa, $p(x_j \mid C_i)$ je *a priorna* vjerojatnost pojave primjerka x_j uz uvjet da je vrijednost ciljnog atributa jednaka C_i. Obje vjerojatnosti, kao i vjerojatnost pojave ciljnog razreda $p(C_i)$ procjenjuju se na skupu za učenje. Vjerojatnost $p(x_j)$ uz pretpostavku o nezavisnosti prediktivnih atributa jednaka je umnošku vjerojatnosti za vrijednosti pojedinačnih atributa:

$$p(x_j) = p(x_j^{(A1)} = x_{j1} \cap x_j^{(A2)} = x_{j2} \cap \cdots \cap x_j^{(AM)} = x_{jM}) = p(x_j^{(A1)} = x_{j1}) * p(x_j^{(A2)} = x_{j2}) * \ldots * \\ * p(x_j^{(AM)} = x_{jM}) = \prod_k p(x_j^{(k)} = x_{jk}) \tag{6.29},$$

a isto vrijedi i za uvjetne vjerojatnosti:

$$p(x_j \mid C_i) = p(x_j^{(A1)} = x_{j1} \mid C_i) * \ldots * p(x_j^{(AM)} = x_{jM} \mid C_i) = \prod_k p(x_j^{(k)} = x_{jk} \mid C_i)\,. \tag{6.30}.$$

Za svaki kategorički atribut, $p(x_j \mid C_i)$ procjenjuje vjerojatnost da će atribut poprimiti određene vrijednosti iz domene uz dani razred ciljnog atributa i to kao broj puta kad je određena vrijednost atributa bila opažena podijeljeno s ukupnim brojem opažanja. S druge strane, svaki numerički atribut je potrebno modelirati s kontinuiranom razdiobom vjerojatnosti kroz opseg vrijednosti dotičnog atributa. Uobičajena pretpostavka je normalna razdioba vjerojatnosti za numeričke atribute. Numerički atributi su uz to predstavljeni pomoću srednje vrijednosti i standardne devijacije koji su procijenjeni na skupu za učenje pa je njihovu vjerojatnost moguće izračunati kao:

$$p(x_j^{(k)} = x_{jk}, C_i) = G(x_{jk}, \mu_{Ci}, \sigma_{Ci}) = \frac{1}{\sqrt{2\pi}\sigma_{Ci}} \exp\left(-\frac{(x_{jk} - \mu_{Ci})^2}{2{\sigma_{Ci}}^2} \right) \tag{6.31}.$$

Pritom se srednja vrijednost i standardna devijacija računaju samo za one primjerke numeričkog atributa koji imaju vrijednost ciljnog atributa jednaku C_i. U slučaju numeričkih atributa konstantni faktor $p(x_j)$ pri izračunu *a posteriorne* vjerojatnosti u izrazu (6.28) se zanemaruje.

U slučaju da se vrijednosti numeričkog atributa ne ponašaju po normalnoj razdiobi, što je dosta često slučaj kod značajki BVN, tada se funkcija gustoće vjerojatnosti može procijeniti koristeći sumu više Gaussovih jezgri. Procjena gustoće dana je izrazom [John 1995]:

$$p(x_j^{(k)} = x_{jk}, C_i) = \frac{1}{d} \sum_d G(x_{jk}, x_d, \sigma_{Ci}) \qquad (6.32),$$

pri čemu je *d* broj primjeraka atributa $x_j^{(k)}$ koji imaju vrijednost ciljnog atributa jednaku C_i na skupu za učenje, a x_d su konkretne vrijednosti atributa $x_j^{(k)}$ na skupu za učenje. Prednost naivnog Bayesa s Gaussovim jezgrama u odnosu na jednostavni je u tome da bolje modelira one atribute koji nemaju normalnu razdiobu. Nedostatak je u tome da je vrijeme izračunavanja više Gaussovih jezgri duže u odnosu na samo jednu.

7 Sustavni postupak za dubinsku analizu biomedicinskih vremenskih nizova

U ovom poglavlju dan je prijedlog sustavnog računalnog postupka koji će osigurati visoku točnost, pouzdanost i sveobuhvatnost analize biomedicinskih vremenskih nizova. Sustavni postupak se sastoji od tri glavna dijela predočenih na slici 7.1, a to su:

1. Dobavljanje podataka
2. Izlučivanje značajki
3. Dubinska analiza

U nastavku poglavlja detaljnije su objašnjeni pojedini dijelovi postupka i na kraju je dana rasprava o nekim aspektima postupka.

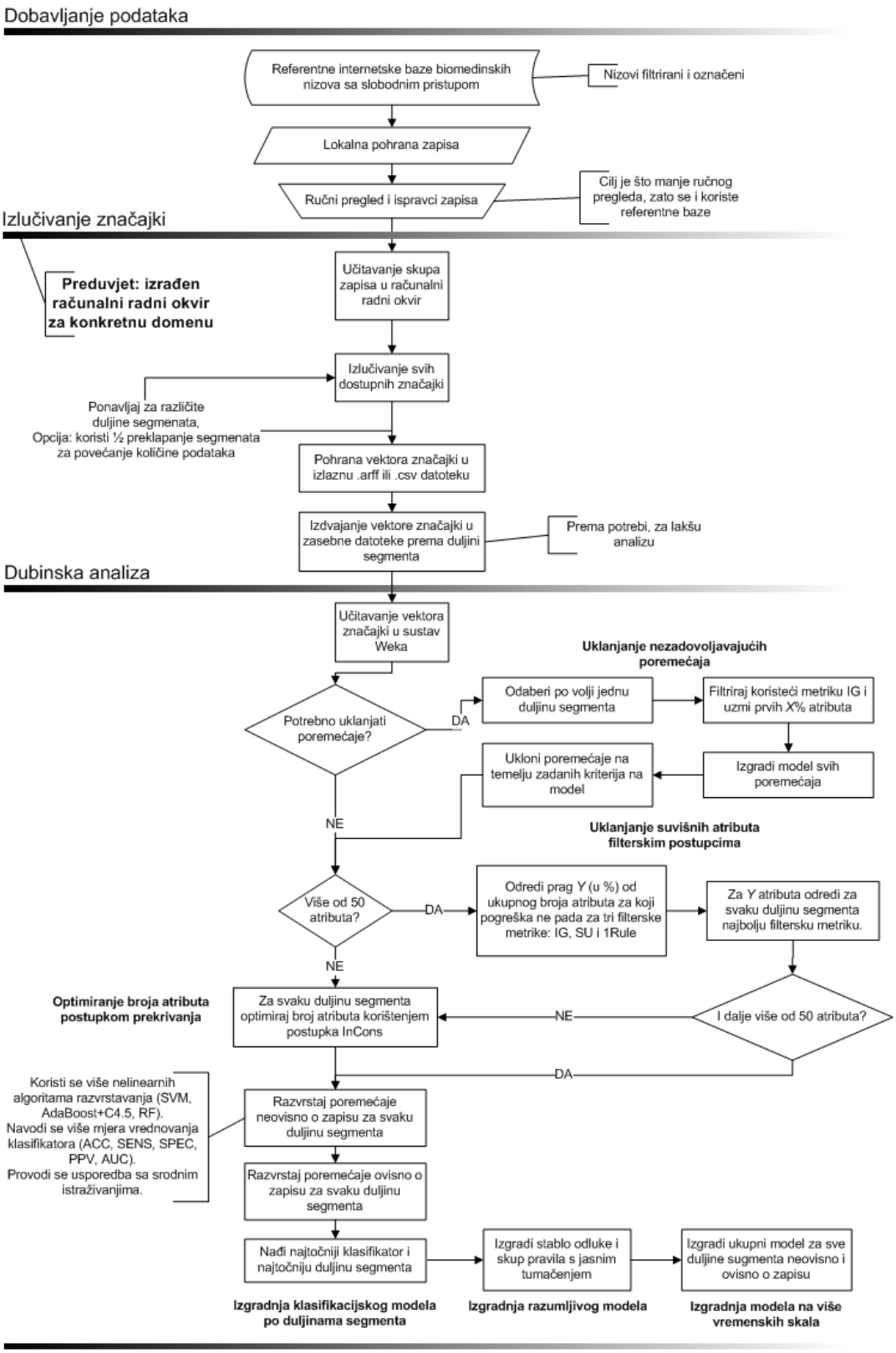
7.1 Dobavljanje podataka

7.1.1 Referentne internetske baze podataka

Izrada točnih modela BVN zahtijeva da su podaci koji se analiziraju pouzdani. Pouzdanost i dostupnost podataka je od velike važnosti da bi se omogućila kvalitetna analiza dostupnih računalnih metoda za klasifikaciju ili predviđanje poremećaja [Jović 2011 (1)]. Jedina garancija da su biomedicinski podaci pouzdani je ta da su podaci pregledani i ispravljeni od strane više nezavisnih stručnjaka u određenom području. Visoko pouzdanih baza koje su besplatno dostupne istraživačima ima malo u svijetu. Najznačajnije takve baze okupljene su na internetskom portalu PhysioNet [PhysioNet]. U ovom postupku preporuča se korištenje pouzdanih, referentnih internetskih baza BVN, no bez ograničenja da portal od kud se podaci preuzimaju mora biti PhysioNet. Razlozi za korištenje referentnih internetskih baza u odnosu na privatne su sljedeći:

1. Veća sigurnost da su podaci kvalitetno pripremljeni.
2. Lakša dostupnost BVN.
3. Omogućena je izravna usporedba rezultata između istraživača.

Baze podataka iz internetske banke podataka PhysioNet su prema kvaliteti podijeljene u tri kategorije [Goldberger 2000]. Prvu kategoriju čine one baze koje su prošle recenziju u više znanstvenih radova i koje su istraživači referencirali u svojim istraživanjima. Također, ove baze su bile pažljivo i detaljno promotrene i ispravljene od strane više liječnika i istraživača te su označeni svi bitni dijelovi u svim zapisima. Istraživanja temeljena na bazama u prvoj kategoriji su najcjenjenija i najčešća. Samo šest baza podataka trenutno dostupnih sa PhysioNeta se nalaze u ovoj kategoriji, od kojih samo jedna baza (*MGH/MF Waveform Database*) sadrži istodobna mjerenja više vremenskih nizova (EKG, arterijski tlak, plućni arterijski tlak, središnji venski tlak, respiratorna impedancija razina CO2).



Slika 7.1. Sustavni postupak za analizu biomedicinskih vremenskih nizova.

Drugu kategoriju čine baze podataka onih istraživača koji su odlučili podijeliti svoje podatke s drugima, a na temelju rezultata objavljenim na konferencijama ili u znanstvenim časopisima. Za razliku od prve kategorije, ove baze nisu detaljno i više puta razmatrane niti ne moraju biti detaljno označavane. Trenutno je ova kategorija u PhysioNetu zastupljena s 20 baza EKG-a, EEG-a, dinamike hoda, EMG-a, respiracije i otpora kože.

U treću kategoriju spadaju sve ostale baze podataka koje su istraživači poslali PhysioNetu i koje nisu prošle priznatu recenziju niti su morale biti označene ili detaljno pregledane. Tu se najčešće radi o bazama podataka pojedinih bolnica koje su dobrovoljno ustupile svoje zapise na korištenje za znanstvena istraživanja. Ukupno je dostupno 27 baza podataka iz treće kategorije: najviše EKG-a, ali i fetalnih EKG-a, EEG-a, EMG-a, respiracije, krvnog tlaka, istodobnih mjerenja više signala, itd.

Sveukupno gledano, najviše je dostupnih zapisa EKG-a, dok je bitno manje dostupnih zapisa EEG-a, EMG-a, respiracije, arterijskog tlaka. Najvrednije u medicinskom smislu su baze podataka koje sadrže istodobna mjerenja više BVN i time omogućuju cjelovitije određivanje poremećaja, no takvih je baza ujedno i najmanje. Postupak predložen u ovoj disertaciji zasad nije prilagođen analizi takvih složenih zapisa, uglavnom zbog manjka kvalitetnih istraživanja.

7.1.2 Lokalna pohrana zapisa

Pri izradi modela poremećaja pretpostavit će se da se BVN zapisuju iz internetskih baza u nekom obliku na lokalno računalo i da se analiziraju s lokalnog računala. Iako analiza s lokalnog računala nije nužan uvjet, iz praktičnih razloga propusnosti podataka i pouzdanosti čitanja podataka u istraživanjima preporuča se da postoji lokalna kopija podataka koja se obrađuje. Također, lokalna pohrana zapisa je obično preduvjet daljnjeg dijela postupka: ručnog pregleda i ispravljanja nedostataka u podacima.

7.1.3 Ručni pregled i ispravci zapisa

Zapise BVN iz baza koje nemaju označene poremećaje (kategorije 2 i 3) potrebno je ručno označiti prije postupka nadziranog učenja i izrade modela. Poželjno je pritom da se analiza što je manje moguće oslanja na podatke iz neoznačenih baza. Ipak, zbog nedostatka visokokvalitetnih baza prve kategorije ponekad je potrebno uzeti u obzir i takve zapise. Označavanje podataka trebali bi provesti stručnjaci ili druge osobe s iskustvom u tom području [Gamberger 2003].

Osim označavanja, podatke je potrebno i prekontrolirati s obzirom na artefakte koji bi mogli naštetiti analizi, a koji nisu bili uklonjeni pri pripremi baze. Tipično, takvi artefakti uključuju lutanje bazne linije zbog pomaka elektrode, nedostatak signala zbog otpadanja elektroda, neispravno otkrivanje karakterističnih točaka u signalu (npr. R-zubaca kod EKG-a) i slično [Clifford 2006]. Od velike je važnosti ukloniti sve manjkavosti u podacima kako bi rezultati modela bili što točniji.

7.2 Izlučivanje značajki pomoću radnog okvira

Izlučivanje značajki slijedi nakon dobavljanja i pripreme ulaznih zapisa. Preduvjet izlučivanju značajki u sklopu sustavnog postupka je izgrađen radni okvir za određenu domenu. Radni okvir treba sadržavati velik broj značajki koje dobro pokrivaju postojeće znanje o određenoj domeni. To znači da je u radnom okviru implementirana većina poznatih značajki primjenjivih na dotičnom problemu, no ne nužno sve, jer bi to bilo teško provedivo. Nužno je, međutim, da radni okvir pokrije sve značajke koje su standardne u dotičnom području, a to su one definirane u medicinskim smjernicama koje pokrivaju to područje.

Pripremljeni zapisi učitavaju se u radni okvir. Prema potrebi, ako su zapisi preveliki da stanu u memoriju računala tada se učitavaju jedan po jedan ili čak po dijelovima. Primjer takvih zapisa su dugotrajni zapisi velikog broja signala EEG-a s visokom frekvencijom uzorkovanja. Zapisi se analiziraju ili cijeli ili po dijelovima od interesa za istraživača. Zapis se analizira po dijelovima u slučaju da su poremećaji lokalizirani.

Nakon što se zapisi učitaju, pristupa se izlučivanju značajki. Izlučivanje značajki se ponavlja u petlji za različite duljine segmenata tako da se pokriju sve moguće vremenske skale za otkrivanje poremećaja. Prema potrebi, ako podataka ima malo, koristi se preklapanje do polovice duljine prethodnog segmenta pri čemu se broj vektora značajki može udvostručiti.

Odabir značajki koje se za neki segment izlučuju ovisi o duljini segmenta. Neke značajke su primjenjive na kraćim vremenskim segmentima, neke na duljim, a neke značajke su gotovo neovisne o duljini segmenta (vidjeti tablicu 3.3). Radni okvir treba imati za svaku značajku ugrađenu najmanju duljinu segmenta za koju bi tu značajku bilo suvislo izlučivati. Suvisla najmanja duljina za koju se izlučuje značajka može se predložiti empirijski kao polovina od najmanje duljine za koju se značajka dosad izlučivala u poznatoj literaturi. Ukoliko postoji teorijska najmanja granica za neku značajku, tada se ona mora uvažiti.

Ako se značajka izlučuje u ovisnosti o nekom parametru, tada je pri izlučivanju potrebno pokriti barem dio spektra tog parametra da bi se detaljno ispitalo ponašanje značajke.

Vektori značajki pohranjuju se u izlaznu datoteku. Izlazna datoteka treba biti prilagođena daljnjoj dubinskoj analizi, stoga se predlaže da datoteka bude u obliku .arff ili alternativno, u obliku .csv radi kompatibilnosti. Osim značajki, bilo bi poželjno u izlaznu datoteku pohraniti naziv zapisa pacijenta kao i početak segmenta i duljinu segmenta koji se analizira.

Vektori značajki pohranjuju se u jednu izlaznu datoteku, a potom se mogu podijeliti i u zasebne datoteke po duljinama segmenata radi lakšeg provođenja daljnje analize.

7.3 Dubinska analiza

Dubinska analiza provodi se nad dobivenom datotekom s vektorima značajki. Za analizu podataka može se koristiti bilo koji dostupan alat sa svim mogućnostima koje se navode u nastavku ovog poglavlja. U okviru ove disertacije korišteno je programsko okruženje Weka [Witten 2005, Hall 2009].

Ciljevi dubinske analize koje pokriva ovdje navedeni sustavni postupak su sljedeći:

1. Izgraditi što točniji klasifikacijski model za svaki razmatrani obrazac BVN i za svaku razmatranu duljinu segmenta i ustanoviti duljinu segmenta s najtočnijim rezultatom.
2. Smanjiti broj atributa koji klasifikator razmatra na najmanju moguću mjeru radi povećanja brzine izgradnje klasifikatora, a da se pritom održi točnost modela.
3. Izgraditi klasifikacijski model koji je razumljiv liječnicima.
4. Izgraditi što točniji klasifikacijski model za svaki razmatrani obrazac BVN koristeći informaciju s više vremenskih skala (više duljina segmenata).

Dubinska analiza podataka provodi se u šest koraka:

1. Uklanjanje poremećaja s preniskom točnosti modela.
2. Uklanjanje suvišnih atributa filterskim postupcima.
3. Optimiranje broja atributa postupkom prekrivanja.
4. Izgradnja klasifikacijskog modela po vremenskim skalama.
5. Izgradnja razumljivog modela.
6. Izgradnja modela na više vremenskih skala.

7.3.1 Uklanjanje poremećaja s preniskom točnosti modela

U slučaju da istraživaču nije unaprijed poznato koje poremećaje ima smisla analizirati korištenjem informacija iz jednog BVN te stoga želi istražiti što veći broj mogućnosti, potrebno je odrediti podskup poremećaja koji imaju potencijal da budu točno opisani. Prvi korak u dubinskoj analizi ima za cilj ukloniti one poremećaje koji više smetaju pri izgradnji točnih modela ostalih poremećaja nego što bi ih imalo smisla uzeti u obzir. Ako istraživač ne želi uklanjati nikakve poremećaje, tada se prvi korak preskače.

Da bi se takvi poremećaji brzo uklonili i da bi se istraživač mogao fokusirati na analizu onih poremećaja koje ima smisla modelirati, provodi se brzi heuristički postupak. Najprije se po volji odabere jedna duljina segmenta na kojoj će se provesti uklanjanje. Predlaže se da ako postoji mnogo atributa, da se tada atributi za tu duljinu segmenta najprije profiltriraju na određeni postotak $X(\%)$ (20%, 30%) od ukupnog skupa atributa koristeći filtersku mjeru informacijskog dobitka (IG) s rangiranjem. Preporuča se da se kao X odabere postotak koji je dovoljno velik da realno procijeni moguće rezultate razvrstavanja, a da je opet dovoljno malen da se model brzo izgradi.

Za učenje i validaciju koristi se postupak unakrsne validacije s 10 preklopa, kao algoritam za razvrstavanje koristi se slučajna šuma s brojem stabala koji približno

odgovara broju atributa. Algoritam slučajnih šuma odabran je zbog brzine izgradnje modela i visoke točnosti.

Za dobiveni model odrede se kriteriji (jedan ili više njih) na temelju kojih će se poremećaji ukloniti (npr. osjetljivost modela za neki poremećaj treba iznositi najmanje 30%, a specifičnost najmanje 60%). Svi poremećaji koji ne zadovolje postavljene kriterije uklanjaju se iz daljnjeg razmatranja. Kriterije postavlja istraživač u suradnji sa stručnjakom [Gamberger 2003]. Ako se procijeni da je to potrebno, postupak uklanjanja poremećaja može se ponoviti više puta. U tom slučaju kriteriji za odbacivanje u svakoj se iteraciji podižu.

7.3.2 Uklanjanje suvišnih atributa filterskim postupcima

Da bi se ostvario cilj #2, odnosno da bi klasifikator izgradio model nad što manjim, a točnim skupom atributa, potrebno je provesti smanjenje broja atributa uklanjanjem onih suvišnih. U okviru ovog postupka, predlaže se da se smanjenju broja atributa pristupi u ovisnosti o broju atributa dobivenih iz radnog okvira. Ako u skupu podataka postoji više od 50 atributa, tada je potrebno provesti uklanjanje atributa korištenjem filterskih metrika i to radi brzine uklanjanja nebitnih atributa. Ako u skupu postoji manje od 50 atributa, tada se može koristiti postupak prekrivanja kako je opisan u poglavlju 7.3.3. U slučaju više od 50 atributa, predlaže se korištenje triju filterskih metrika navedenih u poglavlju 6.3.1: informacijskog gubitka (IG), postupka jednog pravila (1Rule) i simetrične nesigurnosti (SU).

Filterske mjere koriste rangiranje atributa prema nekom kriteriju. Potrebno je provesti detaljnu analizu toga koliko se atributa može ukloniti, a da se točnost razvrstavanja značajno ne smanji u odnosu na ukupan skup atributa. Dakle, potrebno je empirijski odrediti najmanji prag uključenih atributa Y (u % od ukupnog broja atributa) koji ne smanjuje točnost ukupnog skupa atributa. U većini slučajeva, taj prag neće bitno ovisiti o duljini segmenta, već o prirodi značajki uključenih u analizu i to ponajviše na njihovim klasifikacijskim mogućnostima na konkretnom problemu. Atrubuti koje neće imati veliku vrijednost filterske metrike u pravilu nisu korisni u daljnjoj analizi.

Detaljna analiza uključuje ispitivanje podskupova atributa za najmanje jednu, a poželjno za dvije ili više duljina segmenata i to za sve tri filterske metrike. Preporuča se ispitati klasifikacijske mogućnosti skupova od 100%, 50%, 40%, 30%, 20%, 10% i 5% izvornog skupa značajki, no ovisno o broju atributa mogu se ispitati i neki drugi postotci. Podskupovi atributa ispituju se unakrsnom validacijom s 10 preklopa i algoritmom slučajnih šuma za izgradnju modela pri čemu šuma treba sadržavati broj stabala koji je približno jednak broju atributa. Za usporedbu rezultata promatra se samo ukupna točnost razvrstavanja.

Jednom kad je ustanovljeno koji postotak atributa treba koristiti (a to je onaj najmanji koji ne smanjuje značajno rezultat ukupnog skupa atributa), tada još treba odrediti koju je filtersku mjeru najbolje koristiti za takav skup atributa za svaku duljinu segmenta. Radi brzine ispitivanja, preporuča se u tu svrhu korištenje naivnog Bayesovog postupka

s Gaussovim jezgrama koji, iako manje točan od algoritma slučajnih šuma, daje bržu i okvirno točnu procjenu. Postupak ispituje koja je od tri metrike najbolja za svaku od duljina segmenata u smislu ukupne točnosti razvrstavanja. Alternativno, može se za istu procjenu koristiti algoritam slučajnih šuma.

7.3.3 Optimiranje broja atributa postupkom prekrivanja

Dobiveni podskup značajki za svaku duljinu segmenta moguće je optimirati ako on sadrži relativno malo atributa i to koristeći postupak prekrivanja. Preporuča se upotreba postupka konzistentnosti razreda (InCons), poglavlje 6.3.4. Ako najmanji skup atributa dobiven filtriranjem uz pomoć filterskih metrika i dalje broji više od 50 atributa, tada je učinkovito pretraživanje prostora podskupova atributa teško, jer računalo ne može brzo pretražiti niti djelić tako velikog partitivnog skupa. Brojka od 50 atributa uzima se okvirno, s time da su manja odstupanja od tog broja moguća (vidjeti poglavlje 7.4).

Smatra se da je postupak InCons dao bolje rješenje, ako je uspio pronaći takav pravi podskup skupa atributa koji daje jednaku ili veću točnost u odnosu na početni skup, što se može ustanoviti statističkim testom. S obzirom na to je pretraživanje podskupa slučajno, može se dogoditi da za neku duljinu neće biti pronađen podskup koji je bolji od skupa od 50 atributa ni za jednu od tri filterske mjere. U tom slučaju, za konačnu analizu koristi se skup od 50 atributa one filterske mjere koja daje najbolji rezultat uporabom slučajnih šuma.

Za svaku duljinu segmenta provodi se zasebno optimiranje i time se dobivaju konačni podskupovi atributa nad kojima se grade modeli uporabom više algoritama razvrstavanja.

7.3.4 Izgradnja klasifikacijskog modela po duljinama segmenta

Za svaku duljinu segmenta provodi se zasebna analiza postupcima razvrstavanja da bi se odredila uspješnost modela. Specifičnost analize biomedicinskih podataka koju treba uzeti u obzir je biološka individualnost svakog pacijenta. Stoga je prilikom procjene točnosti modela potrebno uzeti u obzir da li se model gradio na različitim segmentima neovisno o zapisu pacijenta ili da li se gradio ovisno o zapisu pacijenta.

Dva tipa modela dakle uključuju: 1) razvrstavanje po segmentima bez obzira na zapis pacijenta, 2) razvrstavanje koje uzima u obzir o kojem se zapisu pacijenta radi.

Prvi služi za ocjenu najveće moguće točnosti koja se može postići za modeliranje poremećaja, budući da će sadržavati neke segmente od istog pacijenta i u skupu za učenje i u skupu za testiranje. U realnim okolnostima prvi tip modela podrazumijeva da: postoji dostupan prethodni zapis pacijenta na temelju kojeg se može lakše opisati daljnje poremećaje dotičnog pacijenta, ili da postoji tako velika baza podataka o konkretnom poremećaju koja pokriva sve moguće biološke raznolikosti koje se mogu dogoditi među pacijentima pa je stoga svejedno o kojem se pacijentu radi.

Drugi tip modela služi za stvarniju ocjenu klasifikacijskog potencijala skupa značajki, budući da se uči na zapisima jednih pacijenata, a testira na zapisima drugih pacijenata.

Za oba tipa modela vrednovanje rezultata provodi se unakrsnom validacijom s 10 preklopa, a modeli se grade korištenjem više algoritama za razvrstavanje. Ustanovljava se koja duljina segmenta daje najtočniji model i to za koji algoritam razvrstavanja. Pritom je potrebno navesti pod kojim su uvjetima (uz koje parametre) klasifikatori postigli najbolji rezultat.

Zbog težine problema preporuča se korištenje algoritama razvrstavanja koji grade nelinearnu decizijsku granicu. Ako neki od algoritama ima stohastičke komponente, potrebno ga je više puta izgraditi i usrednjiti rezultate. Algoritmi se uspoređuju po pitanju postignutih rezultata u smislu mjera vrednovanja pri čemu treba navesti najmanje tri mjere vrednovanja: ukupnu klasifikacijsku točnost (ACC), osjetljivost (SENS) i specifičnost (SPEC), a preporuča se navođenje i mjera predvidljivosti pozitivnih primjeraka (PPV) i površine ispod krivulje ROC (AUC). Mjere SENS, SPEC, PPV i AUC navode se za svaki tip poremećaja, a mogu se, prema potrebi, navesti i mikro-usrednjene vrijednosti mjera ukupno za sve poremećaje.

Usporedba rezultata razvrstavanja sa srodnim istraživanjima provodi se također u ovom koraku. Pritom je poželjno usporediti rezultate s onim istraživanjima koja su rađena na istom skupu podataka i s razvrstavanjem istih vrsti poremećaja, no ukoliko takvih istraživanja nema, tada se rezultati uspoređuju s najsličnijima.

7.3.5 Izgradnja razumljivog modela

Modeli koji postižu najvišu točnost u pravilu nemaju jednostavno tumačenje, budući da su oni često nelinearna kombinacija većeg broja atributa. Cilj je dobiti što točniji model sa zadovoljavajućim tumačenjem za liječnike, što se u više studija naglašava kao vrlo bitan korak u analizi BVN [Cios 2002, Pecchia 2011]. Radi lakšeg tumačenja, u ovom koraku ograničava se izbor postupaka razvrstavanja samo na postupke s pravilima i stabla odluke. Preporučuju se algoritmi C4.5 i RIPPER čiji se modeli pojednostavljuju tako da budu lako razumljivi liječnicima. Što manje atributa sadrže, to će biti razumljiviji, no ne nužno točniji. Kompromis po pitanju razumljivosti i točnosti potrebno je razumno provesti tako da se više puta izgradi model za različite vrijednosti parametara, pri čemu je najvažniji parametar najmanji broj primjeraka koje pravilo (ili list stabla) pokriva.

7.3.6 Izgradnja modela na više vremenskih skala

Informacija dobivena iz više vremenskih segmenata različitih duljina mogla bi biti točnija pri razvrstavanju segmenata nego što je to ona dobivena na segmentu samo jedne duljine. U okviru sustavnog postupka, poželjno je ispitati mogućnost izgradnje modela koji uzima u obzir više vremenskih skala, idealno svih skala na kojima se

provode zasebne analize. Pritom je broj vektora značajki jednak zbroju vektora značajki po segmentima pojedinih duljina što produžava vrijeme izgradnje modela. Da bi se model uspješno sagradio, potrebno je da su već prije provedeni koraci 2 i 3, što znači da je skup atributa optimiran. Za izgradnju modela na svim segmentima može se koristiti algoritam slučajnih šuma s brojem stabala koji odgovara broju atributa i postupak unakrsne validacije za procjenu pogreške generalizacije. Izgradnja modela na više vremenskih skala treba se provesti za oba tipa modela: neovisno o zapisu pacijenta i ovisno o zapisu pacijenta, da bi se ispitao potencijal tako složenog modela.

7.4 Rasprava o postupku

7.4.1 Rasprava o odabiru atributa

Neki elementi predloženog postupka odabira značajki (filterske mjere + optimizacija postupkom konzistentnosti razreda) već su razmatrani u znanstvenoj literaturi. Autori [Zhang K. 2007] koriste utežano glasanje između četiri filterske mjere i zatim postupak omotača da bi odabrali značajke ključne za nadgledanje istrošenosti strojeva. U skupu se nalazilo ukupno 256 značajki. Od filterskih mjera koriste se Pearsonov korelacijski koeficijent, algoritam Relief, Fisherov pokazatelj (engl. *Fisher score*), i koeficijent omjera signal-šum (u radu se naziva odvojivost razreda (engl. *class separability*)). Budući da su ove četiri mjere zasnovane na različitim kriterijima vrednovanja, autori koriste težine da bi odredili koliko je svaka od ovih četiriju mjera jaka. Težine utvrđuju na temelju rezultata prvih 25 značajki svake od mjera na postupku omotača, i to s umjetnom neuronskom mrežom zasnovanoj na radijalnoj funkciji jezgre (RBF). Razlika u odnosu na postupak predložen u ovom radu je ta što autori razmatraju binarni problem razvrstavanja, dok ovdje postoji razvrstavanje u više razreda. Druga razlika je u tome što autori odabiru značajke i razvrstavaju koristeći isti algoritam RBF, dok se u ovom radu razmatraju razni postupci razvrstavanja pa prema tome odabir značajki treba biti neutralan s obzirom na algoritam razvrstavanja.

Autori [Guldogan 2008] koriste glasanje za odabir bitnih značajki pri problemu pronalaženja slika na temelju sadržaja. Glasanje se provodi nad rezultatima triju filterskih mjera: zajedničke informacije, odnosa unutar grupe (engl. *inner-cluster relation*) i Pearsonovog produkt-moment korelacijskog koeficijenta. Ukupni glasovi za svaku značajku određeni su koristeći formulu:

$$v(A_i) = v_{MI}(A_i) + v_{ICR}(A_i) + v_{PPMC}(A_i) \tag{7.1}.$$

U ovoj disertaciji pokušao se sličan princip primijeniti na tri korištene filterske metrike. Za svaku od filterskih mjera provelo se rangiranje atributa i uzeo se podskup od 30% najboljih atributa. Neka su sa $S_{IG}, S_{SU}, S_{1Rule}$ označeni ti podskupovi. Svakoj filterskoj mjeri pridijeljena je odgovarajuća težina koja označava koliko tu mjeru treba uzeti u obzir prilikom odabira podskupa bitnih atributa: $w_{IG}, w_{SU}, w_{1Rule}$. Težinu se

odredilo tako da su se dobiveni podskupovi atributa vrednovali uz pomoć postupka razvrstavanja naivnim Bayesovim klasifikatorom s Gaussovim jezgrama. Neka su s $ACC_{IG}, ACC_{SU}, ACC_{1Rule}$ označene ukupne točnosti razvrstavanja naivnim Bayesom korištenjem postupka unakrsne validacije s deset preklapanja sa podskupovima atributa $S_{IG}, S_{SU}, S_{1Rule}$. Težina pojedine mjere određena je izrazom:

$$w_i = \frac{ACC_i}{\sum_i ACC_i}, i \in \{IG, SU, 1Rule\} \tag{7.2}$$

i vrijedi: $\sum_i w_i = 1$. Konačni rezultat svakog atributa određen je izrazom:

$$R(A_i) = w_{IG} R_{IG}^{(N)}(A_i) + w_{SU} R_{SU}^{(N)}(A_i) + w_{1Rule} R_{1Rule}^{(N)}(A_i) \tag{7.3},$$

pri čemu su $R_i^{(N)}(A_i)$ označeni rezultati pojedinih mjera na atributu A_i, normirani s obzirom na vrijednost najbitnijeg atributa dobivenog dotičnom metrikom. Atributi se rangiraju prema rezultatu.

Rezultati dobiveni ovako utežanom kombinacijom atributa pokazali su se lošijima od najbolje filterske metrike, a boljima od najlošije (poglavlje 8.1). Stoga se primjena utežavanje više filterskih postupaka ne preporuča u okviru sustavnog postupka.

Provedena mjerenja prikazana u tablici 7.1 pokazala su da je moguće pretražiti 1.0×10^{-7} % od ukupnog prostora atributa korištenjem mjere InCons nad 47 atributa i 3300 primjeraka u 9 min. Manje od 1.0×10^{-8} prostora nema smisla pretraživati budući da nije izgledno da se može pronaći podskup atributa koji će dati bolje (ili iste) konačne rezultate. Na temelju mjerenja, došlo se do zaključka da je otprilike 50 atributa neka razumna granica za korištenje mjere InCons za veličine uzorka od nekoliko tisuća primjeraka. U izvornom radu autora [Liu 1996] postupak je isproban na skupovima podataka do 22 atributa što je omogućilo visoku pokrivenost podskupova atributa.

Tablica 7.1. Vrijeme potrebno za provođenje optimizacije skupa atributa korištenjem mjere prekrivanja konzistentnosti razreda (provedeno na četverojezgrenom Q9300 procesoru na 2.5 GHz, 8 GB RAM-a).

Uključeni broj atributa	Broj primjeraka	Pretraženi postotak prostora, %	Potrebno vrijeme
47	3300	1.0×10^{-7}	9 min
	1800	1.0×10^{-7}	4 min
	700	1.0×10^{-7}	1 min
55	3300	1.0×10^{-9}	24 min
	1800	1.0×10^{-9}	11 min
	700	1.0×10^{-9}	2.5 min
60	3300	1.0×10^{-11}	17 min
	1800	1.0×10^{-11}	3.5 min
	700	1.0×10^{-11}	1 min

Jedan od uobičajenih statističkih postupaka za smanjenje dimenzionalnosti skupa podataka je analiza glavnih komponenti (PCA). Taj i slični postupci (analiza faktora, analiza nezavisnih komponenti...) transformiraju podatke na takav način da definiraju manji broj novih atributa (komponenti) u čijem prostoru transformirani podaci zadržavaju većinu informacije iz izvornog prostora. Pritom svaka od komponenti prekriva određeni dodatni postotak varijance ciljnog razreda te ih se u obzir uzima samo prvih nekoliko (do određenog definiranog iznosa varijance). U matematičkom smislu, komponente su određena linearna kombinacija početnih atributa. PCA se često koristi u praksi u području analize BVN [Minhas 2008, Kim 2009, Liang 2010].

U okviru sustavnog postupka predloženog u ovom radu PCA se ne koristi i to iz dva razloga. Prvi je gubitak tumačenja značajki koji je nužna posljedica transformacije podataka u niži prostor. Ipak, PCA bi se mogla koristiti samo za klasifikacijski model bez potrebe za tumačenjem (npr. PCA+slučajna šuma ili PCA+SVM). Drugi razlog je taj što PCA stvara linearnu projekciju podataka (linearnu kombinaciju značajki). Budući da razredi nisu linearno odvojivi, ne mogu se očekivati bolji rezultati korištenjem postupka PCA i sličnih linearnih postupaka [Asl 2008].

7.4.2 Rasprava o algoritmima razvrstavanja

Važno je naglasiti da svaki problem razvrstavanja zahtjeva drugačiji pristup. Stoga nije unaprijed moguće odrediti koja je karakteristika algoritma za učenje od presudne važnosti za uspjeh pri razvrstavanju (pristranost, varijanca, prenaučenost). Upravo zato potrebno je na svakom problemu isprobati više postupaka i doći do odgovora koji je algoritam razvrstavanja najpogodniji za izgradnju modela. S obzirom na složenost problema, klasifikatori koji grade linearne granice (npr. linearna diskriminantna analiza, linearni stroj s potpornim vektorima i sl.) ovdje se ne razmatraju. Već ranije je pokazano da dotični algoritmi nisu dostatni za točan opis poremećaja [Jović 2011 (1)].

8 Vrednovanje i usporedba predloženih postupaka

8.1 Vrednovanje sustavnog postupka

U tablici 8.1 naveden je popis baza podataka, ukupnog broja zapisa pacijenata, kategorije baze, uključenih zapisa i vrste obrazaca ritma koji su uključeni u postupak ispitivanja sustavnog postupka za analizu BVN u domeni srčanog ritma. Svi su zapisi pregledani i ispravljeni po potrebi, a zapisi iz baza kategorije 2 i 3, osim baza s kongestivnim zatajenjem srca (CHF), su označeni s prisutnim obrascima ritma.

Provedene su dvije vrste analiza za vrednovanje sustavnog postupka. U prvoj se analiziraju svi poremećaji ritma pri čemu se koriste prve dvije baze iz tablice 8.1. Druga analiza namijenjena je razlikovanju između zdravih pacijenata, pacijenata s CHF i pacijenata s raznim vrstama aritmije. Tu se koriste sve baze, s time da se za zapise iz prvih dviju baza smatra da su od pacijenata s nekom aritmijom, iz drugih dviju baza da je osoba zdrava, a iz trećih dviju baza da je to zapis pacijenta s CHF.

8.1.1 Razvrstavanje više poremećaja ritma

U prvoj analizi razmatra se za razvrstavanje ukupno 14 obrazaca srčanog ritma koji su zastupljeni kod dovoljnog broja pacijenata u bazama podataka. Pritom se koriste prve dvije baze iz tablice 8.1 budući da one sadrže označene informacije o ritmovima. Baza

Tablica 8.1. Baze podataka, zapisi pacijenata i prisutni obrasci ritma korišteni u vrednovanju sustavnog postupka.

Baza podataka	Kat. baze	Uk. br. zapisa	Zapisi pacijenata	Analizirani obrasci (skraćeni nazivi, vidjeti kazalo oznaka)
MIT-BIH Arrhythmia Database	1	48	Svi u bazi. Prvih pola sata.	NSR, PAC, PVC, PAC i PVC, kuplet, ABI, SVT, VBI, VTR, VT, AFIB, PEJS
MIT-BIH Supraventricular Arrhythmia Database	3	77	Svi u bazi, osim zapisa 877. Prvih pola sata.	NSR, PAC, PVC, PAC i PVC, kuplet, AFL, ABI, ATR, SVT, VBI, VTR, VT, AFIB
Normal Sinus Rhythm RR Interval Database	2	50	Svi osim zapisa: nsr024, nsr048, nsr050, nsr051. Prvih sat vremena.	NSR, PAC, PVC, kuplet, SVT, VBI, VTR.
MIT-BIH Normal Sinus Rhythm Database	3	18	Svi u bazi. Prvih sat vremena.	NSR, PAC, PVC
BIDMC Congestive Heart Failure Database	2	15	Svi u bazi. Prvih sat vremena.	CHF
Congestive Heart Failure RR-interval Database	2	29	Svi u bazi. Prvih sat vremena.	CHF

MIT-BIH Supraventricular Arrhythmia Database izvorno ne sadrži oznake ritmova već samo tipove srčanih otkucaja te je stoga prije ove analize ručno pregledana i izmijenjena tako da ih sadrži. Na taj način proširen je skup zapisa iz kojih se ritmovi mogu učiti. Analizira se 125 zapisa različitih pacijenata, u ukupnom trajanju od 62.5 sata i s ukupnim brojem od preko 280 000 RR-intervala.

Na početku su uklonjeni svi poremećaji koji se pojavljuju u bazama, a koji nisu zastupljeni u značajnom broju zapisa (npr. idioventrikularni ritam, ventrikularno lepršanje, blok drugog reda Mobitz II), ili je za te poremećaje unaprijed poznato iz literature da ih je teško razlikovati od normalnog srčanog ritma koristeći samo informaciju o ritmu (npr. blok lijeve grane, blok desne grane, ritam AV-spojnice). Zapisi iz baza koji su sadržavali većinom one poremećaje koji su uklonjeni (više od 2/3 ukupne duljine zapisa) nisu razmatrani u daljnjoj analizi. Konkretno, odbačeni su zapisi: 109, 111, 124, 207, 212, 230 i 231 iz baze *MIT-BIH Arrhythmia Database*.

Ukupno je izlučeno 230 prediktivnih linearnih i nelinearnih značajki, sve one navedene u tablici 5.3. Nelinearne značajke faznog prostora izlučene su uz pretpostavku ugradbene dimenzije $d = 6$ i tri duljine odgode $\tau = \{1, 2, 3\}$. Analiza je provedena za segmente duljine 10, 15, 20, 25, 30, 40, 50, 75, i 100 sekundi, s ukupno 73 785 izlučenih vektora značajki. Pregled broja vektora značajki po poremećajima za segmente analiziranih duljina dan je u tablici 8.2.

Iz tablice 8.2 uočljivo je da najviše vektora značajki (40.04% ukupno) pripada normalnom srčanom ritmu (NSR), dok su svi poremećaji ritma u manjini. Također, iz tablice se može iščitati da se omjer broja vektora NSR i broj vektora poremećaja smanjuje kako se duljina segmenta povećava: omjer je 10174:10483 za 10 s, ali 404:1505 za 100 s. Razlog tome je što postoji manji broj duljih segmenata kod kojih je ritam cijelo vrijeme normalan nego što ima takvih kraćih segmenata.

Tablica 8.2. Broj vektora značajki za obrasce ritma po duljini segmenta, analizirane baze: *MIT-BIH Arrhythmia Database* i *MIT-BIH Supraventricular Arrhythmia Database*.

	Duljina segmenta, s									
Ritam:	10	15	20	25	30	40	50	75	100	Ukupno/udio, %
NSR	10 174	6062	4121	3049	2390	1576	1144	624	404	29 544/40.04
PAC	1984	1384	1065	888	737	583	473	319	243	7676/10.40
PVC	2801	1940	1466	1191	976	736	581	380	270	10 341/14.02
Kuplet	814	611	527	432	374	299	228	154	113	3552/4.81
PVC i PAC	727	652	563	497	456	369	328	244	191	4027/5.46
AFL	152	112	92	74	65	49	42	29	23	638/0.86
VT	124	118	116	117	114	107	107	102	97	1002/1.36
AFIB	1496	1008	749	594	496	374	288	184	135	5324/7.22
SVT	227	182	174	158	138	121	119	94	83	1296/1.76
VBI	527	436	375	340	304	254	232	178	147	2793/3.79
VTR	401	338	299	258	238	204	175	133	108	2154/2.92
ABI	381	305	272	235	215	176	144	104	79	1911/2.59
ATR	206	198	178	161	150	125	104	77	57	1256/1.70
PEJS	646	427	318	254	210	152	125	80	59	2271/3.08
Ukupno:	20 660	13 773	10 315	8248	6863	5125	4090	2702	2009	73 785

U slučaju da ne postoji sigurnost po pitanju može li se izgraditi zadovoljavajući model za sve analizirane poremećaje, potrebno je najprije pokrenuti postupak uklanjanja nezadovoljavajućih poremećaja. Sustavni postupak preporuča da se za neku proizvoljno odabranu duljinu segmenta i za filtersku metriku IG provede uklanjanje poremećaja koji se po rezultatima razvrstavanje razlikuju od ostalih. Pritom se trebaju postaviti neki najmanji kriteriji u smislu rezultata koje poremećaj mora zadovoljavati da bi ga se ostavilo u razmatranju. Za vrednovanje rezultata koristi se postupak unakrsne validacije s 10 preklopa.

Rezultati u tablici 8.3 navedeni su prilikom korištenja podskupa od 30% (69) najznačajnijih atributa korištenjem metrike IG i postupka slučajnih šuma s 60 stabala za duljinu segmenta od 20 s. Kao kriterije za uzimanje poremećaja u obzir postavljeno je da osjetljivost (SENS) bude najmanje 40%, a predvidljivost pozitivnih primjeraka (PPV) najmanje 50%. Oštriji kriteriji jače bi smanjili skup poremećaja što nije poželjno. Uklonjeni su sljedeći poremećaji: kuplet, PVC i PAC, AFL, VT i SVT, budući da se njihovi modeli nisu pokazali zadovoljavajućima, tablica 8.3.

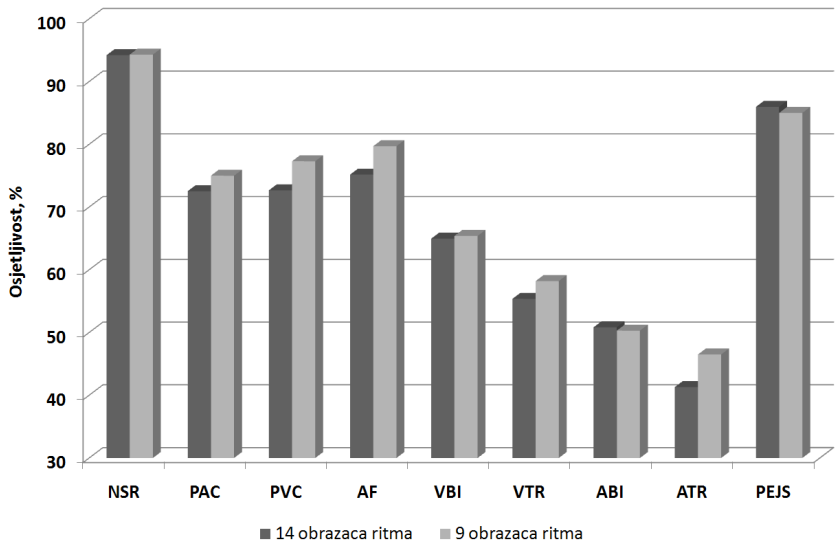
Nakon smanjenja broja poremećaja, ostalo je 9 obrazaca ritma koje se dalje razmatraju: NSR, PAC, PVC, AFIB, VBI, VTR, ABI, ATR, PEJS. Postupak je ponovljen te su izgrađeni okvirni modeli za smanjeni skup obrazaca. Detaljni rezultati navedeni su u tablici 8.4, dok je na slici 8.1 prikazana usporedba u osjetljivosti (SENS) između istih obrazaca pri analizi 14 obrazaca i analizi 9 obrazaca. Iz tablice 8.4. uočljivo je da sad svi poremećaji zadovoljavaju kriterije, a sa slike 8.1. vidi se da su rezultati u slučaju nekih poremećaja (PAC, PVC, AF, VTR, ATR) značajno bolji u odnosu kad se u obzir uzimao veći skup poremećaja. Razlog tome je taj što su uklonjeni neki poremećaji koji su bili međusobno slični, stoga klasifikator više nema toliko dvojbi.

Tablica 8.3. Rezultati (u %) dobiveni za svih 14 vrsta srčanih ritmova korištenjem 30% (69) najznačajnijih atributa filterske metrike IG i slučajne šume s 60 stabala za duljinu segmenta od 20 s. Poremećaji koji se uklanjaju zbog zadanih kriterija označeni su sivo.

Ritam	SENS	SPEC	PPV	AUC	ACC
NSR	94.2	96.6	94.9	98.0	
PAC	72.5	95.3	63.8	95.4	
PVC	72.7	94.2	67.4	94.1	
Kuplet	41.8	97.5	47.0	92.5	
PAC i PVC	39.2	96.9	42.1	90.6	
AFL	7.2	99.6	40.3	95.5	
VT	9.2	99.8	42.3	88.1	74.5±0.2
AF	75.2	94.8	53.1	95.6	
SVT	27.4	99.9	85.6	91.0	
VBI	65.0	98.4	61.1	95.8	
VTR	55.4	99.3	68.9	96.0	
ABI	50.8	99.4	69.2	97.2	
ATR	41.3	99.5	57.9	95.7	
PEJS	86.0	99.5	85.8	99.7	

Tablica 8.4. Rezultati (u %) dobiveni za preostalih 9 vrsta srčanih ritmova korištenjem 30% (69) najznačajnijih atributa filterske metrike IG i slučajne šume s 60 stabala za duljinu segmenta od 20 s. Svi poremećaju zadovoljavaju najmanje kriterije.

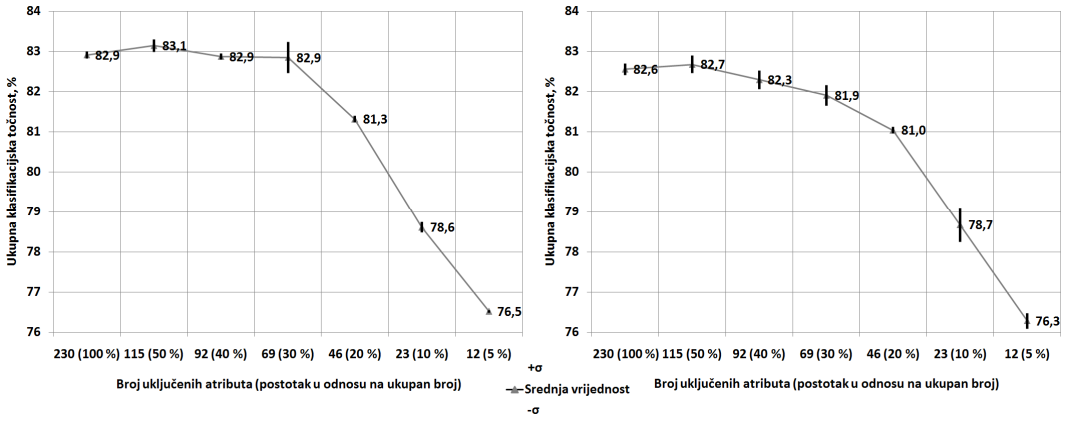
Ritam	SENS	SPEC	PPV	AUC	ACC
NSR	94.3	96.2	95.5	98.0	
PAC	75.0	95.5	69.6	96.0	
PVC	77.3	94.8	74.9	95.7	
AF	79.7	96.5	68.0	97.4	
VBI	65.4	98.5	65.8	97.1	82.9±0.4
VTR	58.2	99.4	78.2	97.0	
ABI	50.3	99.2	68.4	97.7	
ATR	46.5	99.5	66.5	97.3	
PEJS	85.0	99.5	86.0	99.6	



Slika 8.1. Usporedba osjetljivosti (SENS) između modela obrazaca dobivenih u slučaju analize svih 14 obrazaca i analize samo njih 9.

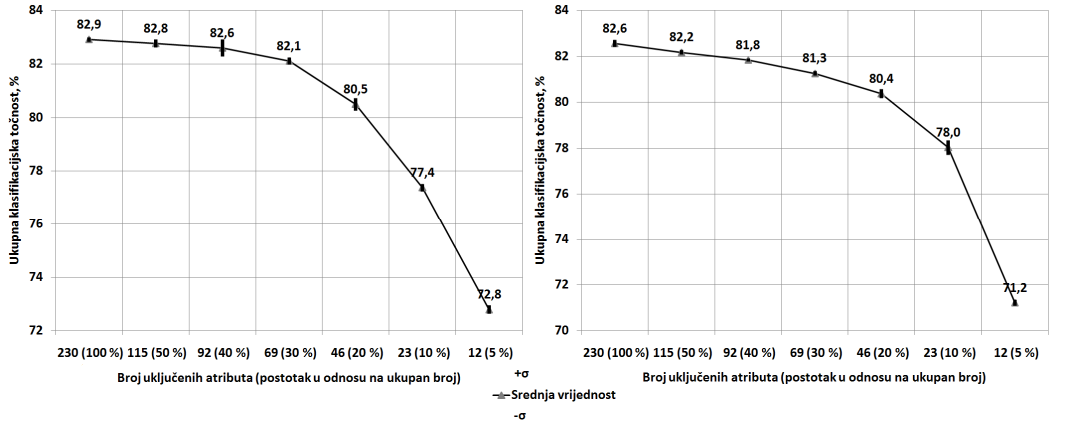
Točnost otkrivanja normalnog ritma nije značajno porasla. Ukupna točnost razvrstavanja je značajno porasla uklanjanjem 5 obrazaca (sa 74.5% na 82.9%).

U nastavku postupka, nakon utvrđivanja konačnog podskupa ispitivanih poremećaja, najprije se provodi sustavna analiza toga koliko se atributa može ukloniti, a da se točnost razvrstavanja bitno ne smanji u odnosu na ukupan skup atributa. Dakle, želi se odrediti parametar $Y(\%)$ sa slike 7.1. Analiza je provedena na segmentima od 20 s i 30 s koristeći 100%, 50%, 40%, 30%, 20%, 10% i 5% izvornog skupa značajki. Pritom su isprobane sve tri filterske mjere za dobivanje rangiranog podskupa najbitnijih atributa i algoritam slučajnih šuma za izgradnju modela. Slučajna šuma je sadržavala broj stabala koji je bio približno jednak broju atributa. Analiza smanjenja skupa značajki prikazana je na slikama 8.2-8.4 i u tablici 8.5. Na slikama 8.2-8.4 prikazana je ukupna klasifikacijska točnost modela, a u tablici 8.5 uspoređene su mjere vrednovanja po poremećajima za čitav skup atributa i za 30% skupa za slučaj filterske mjere IG. Iz slika 8.2-8.4 može se uočiti da se točnost ponekad vrlo malo poboljšava pri korištenju 50%



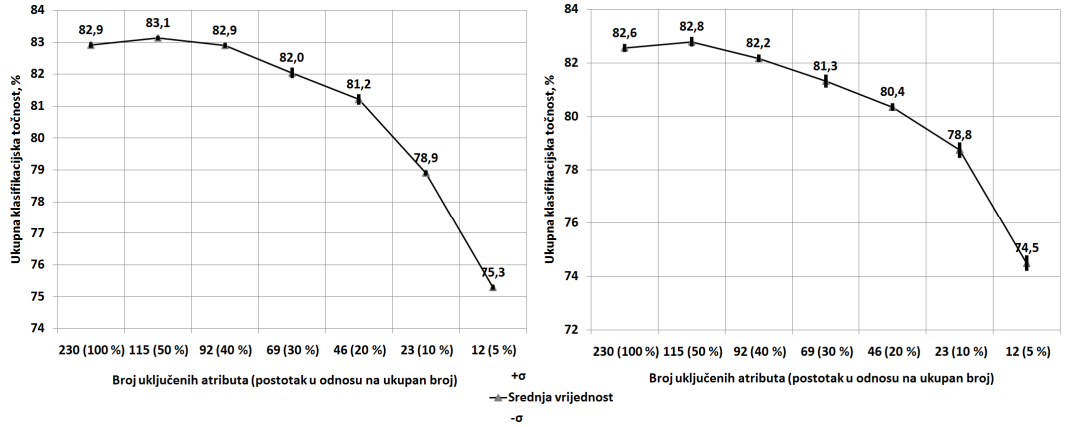
(a) 20 s (b) 30 s

Slika 8.2. Smanjenje dimenzionalnosti skupa atributa korištenjem filterske metrike IG.



(a) 20 s (b) 30 s

Slika 8.3. Smanjenje dimenzionalnosti skupa atributa korištenjem filterske metrike 1Rule.



(a) 20 s (b) 30 s

Slika 8.4. Smanjenje dimenzionalnosti skupa atributa korištenjem filterske metrike SU.

Tablica 8.5. Usporedba rezultata mjera vrednovanja (u %) za puni skup od 230 atributa i za 30% skupa atributa (69 atributa) za filtersku mjeru IG.

(a) 20 s

230 atr.	SENS	SPEC	PPV	AUC	69 atr.	SENS	SPEC	PPV	AUC
NSR	94.9	95.5	95.0	98.1		94.3	96.2	95.5	98.0
PAC	73.4	87.4	70.7	96.1		75.0	95.5	69.6	96.0
PVC	77.7	88.1	75.0	95.7		77.3	94.8	74.9	95.7
AF	81.2	85.9	65.6	97.5		79.7	96.5	68.0	97.4
VBI	65.7	88.6	65.9	97.4		65.4	98.5	65.8	97.1
VTR	57.8	92.6	77.2	97.9		58.2	99.4	78.2	97.0
ABI	45.8	91.7	73.6	98.1		50.3	99.2	68.4	97.7
ATR	49.2	90.4	71.3	98.0		46.5	99.5	66.5	97.3
PEJS	82.1	96.6	90.3	99.7		85.0	99.5	86.0	99.6

(b) 30 s

230 atr.	SENS	SPEC	PPV	AUC	69 atr.	SENS	SPEC	PPV	AUC
NSR	95.3	96.0	94.5	98.4		94.8	96.4	95.0	98.3
PAC	77.3	95.5	50.6	96.3		76.1	95.1	69.6	95.9
PVC	77.2	95.6	78.2	95.8		76.3	95.3	77.2	95.6
AF	82.0	96.1	66.8	97.3		81.3	96.3	67.9	97.0
VBI	64.9	98.1	66.1	97.4		66.3	97.8	62.8	97.0
VTR	61.9	99.3	78.4	97.6		60.1	99.2	76.9	97.1
ABI	48.4	99.3	71.9	97.9		50.9	99.3	73.2	97.2
ATR	56.4	99.3	69.6	98.2		49.5	99.3	66.6	96.9
PEJS	85.1	99.8	94.2	99.7		85.6	99.6	90.0	99.6

skupa atributa u odnosu na puni skup. Razlika je mala (do približno 1% prosječne točnosti) do 30% skupa atributa u odnosu na puni skup, nakon čega točnost skupa atributa značajnije pada. Filterska metrika IG pokazala se nešto bolja u zadržavanju točnosti u odnosu na ostale dvije metrike na ovom problemu. U zaključku, uzima se 30% izvornog skupa atributa (njih ukupno 69) kao prihvatljivo smanjenje dimenzionalnosti u daljnjim analizama. Daljnje sustavno smanjenje broja atributa nažalost nije ostvarivo u ovom slučaju budući da je skup atributa pri kojemu dolazi do smanjenja točnosti prevelik za učinkovitu upotrebu algoritama prekrivanja. Skup od 69 atributa ima prevelik broj podskupova da bi se oni mogli ispitati u razumnom vremenu korištenjem postupka konzistentnosti razreda.

Za svaku duljinu segmenta potrebno je dalje razmotriti koji od tri filterska postupka daje najbolje rješenje. Pokazuje se da u konačnici najbolje rezultate daje onaj filterski postupak koji ima najveću točnost pri korištenju naivnog Bayesovog modela s Gaussovim jezgrama. Ovo je pokazano na temelju ukupne klasifikacijske točnosti modela u tablici 8.6 za periode od 20 s i 30 s za 69 atributa.

Rezultati navedeni u tablici 8.6 preporučaju korištenje filterske mjere IG u daljnjoj analizi za duljine od 20 s i 30 s. Utežana kombinacija filterskih metrika nije dala bolji rezultat od pojedinačnih metrika. Također, rezultati pokazuju da se naivni Bayesov postupak okvirno slaže sa složenijim postupkom slučajnih šuma koji se koristi u izradi konačnog modela. Time je opravdano korištenje naivnog Bayesovog postupka s

Tablica 8.6. Usporedba ukupne točnosti razvrstavanja (u %) pri korištenju filterskih mjera.

20 s	IG	1Rule	SU	Utežana kombinacija
NB+Gauss	74.52	73.45	73.04	-
Konačni model (slučajna šuma)	82.85±0.40	82.11±0.14	82.04±0.17	82.05±0.10
30 s				
NB+Gauss	72.88	72.55	71.50	-
Konačni model (slučajna šuma)	81.91±0.25	81.26±0.10	81.33±0.25	81.58±0.22

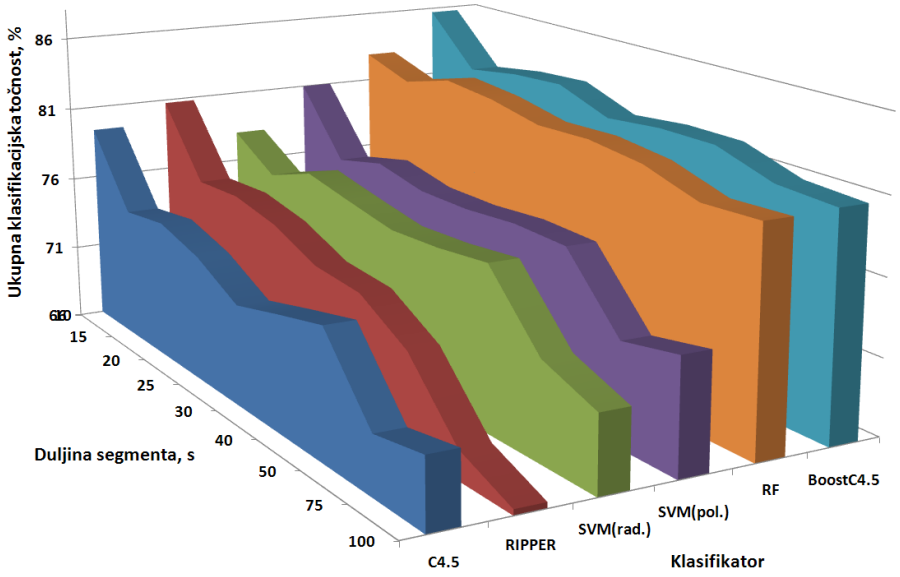
Gaussovim jezgrama u okviru sustavnog postupka analize BVN kao metode za brzu i točnu procjenu kvalitete filterske metrike. Analizom po svim duljinama segmenata ustanovljeno je da za dane podatke filterska mjera IG gotovo uvijek daje bolje rezultate od ostalih dviju mjera (SU i 1Rule), tablica 8.7. Jedina iznimka bila je duljina od 15 s gdje je bolji rezultat dao 1Rule.

U nastavku se za svaku duljinu segmenta grade zasebni modeli koristeći šest algoritama za razvrstavanje: slučajna šuma (RF), C4.5, AdaBoost+C4.5, RIPPER, SVM s polinomnom jezgrom i SVM s radijalnom jezgrom. Potrebno je provesti dvije zasebne procjene točnosti: jednu u slučaju uzimanja u obzir svih dostupnih segmenata neovisno o zapisu pacijenta i drugu, koja uzima u obzir zapis pacijenta te stoga na jednim zapisima uči, a na drugima ispituje. Kod druge procjene, zapisi su slučajno podijeljeni u 10 skupova koji se koriste za unakrsnu validaciju s 10 preklopa s time da je uzeto u obzir da je svaki razred obrasca ritma dovoljno zastupljen u skupu za učenje.

Rezultati razvrstavanja poremećaja za prvi tip modela koji se gradi neovisno o zapisu iz kojih su segmenti stigli prikazani su na slici 8.5 za sve duljine segmenta i za sve algoritme. Algoritmi su prikazani približno poredani po točnosti modela s lijeva prema desno. Uočljivo je da najlošije rezultate daju jednostavniji indukcijski algoritmi C4.5 i RIPPER, dok srednje rezultate daju algoritmi strojeva s potpornim vektorima, pri čemu je nešto bolje koristiti polinom drugog stupnja nego radijalnu jezgru. Parametri algoritama za koje su postignuti rezultati navedeni su u tablici 8.8. Najbolje rezultate dali su algoritmi slučajne šume i AdaBoost+C4.5. Uočava se da je najbolji rezultat kod većine algoritama dobiven za duljinu segmenta od 10 s, koji ujedno ima i najviše

Tablica 8.7. Ukupna klasifikacijska točnost (u %) triju filterskih mjera pri korištenju naivnog Bayesovog modela s Gaussovim jezgrama za svaku duljinu segmenta nad 69 atributa.

Segment	IG	SU	1Rule
10 s	**75.44**	75.19	75.32
15 s	74.58	73.56	**74.64**
20 s	**74.52**	73.04	73.45
25 s	**72.82**	71.68	72.48
30 s	**72.88**	71.50	72.55
40 s	**72.13**	70.31	69.93
50 s	**71.65**	69.23	69.01
75 s	**66.04**	65.80	65.61
100 s	**67.91**	67.11	64.38



Slika 8.5. Ukupna klasifikacijska točnost modela neovisno o zapisu korištenjem 30% najboljih atributa najbolje filterske mjere za sve duljine segmenta po algoritmima.

vektora značajki, a drugi najbolji model je onaj za 20 s (15 s je lošiji). Točnost modela aritmija polako pada s produljenjem segmenta.

U tablici 8.9 prikazani su detaljni rezultati za sve obrasce ritma za duljine segmenata od 10 s i 20 s za algoritme AdaBoost+C4.5 i RF. Detaljni rezultati daju najtočnije predviđanje normalnog ritma (NSR) te konzistentno visoku točnost predviđanja obrazaca AF, PAC, PVC i PEJS, dok su ostali obrasci nešto slabije predvidljivi.

Rezultati razvrstavanja poremećaja za drugu procjenu točnosti koja uzima u obzir točan zapis pacijenta dani su na slici 8.6 za sve duljine segmenta i sve algoritme.

Tablica 8.8. Parametri klasifikatora korišteni za dobivanje rezultata u okviru analize obrazaca ritma.

Algoritam	Parametri
C4.5	Pouzdanost podrezivanja $c = 0.4$, najmanji broj primjeraka u listu $n_{\min} = 2$, uzdizanje podstabala uključeno u podrezivanje
RIPPER	Najmanji broj primjeraka prekriven pravilom $n_{\min} = 2$, broj iteracija optimizacije = 4
SVM s polinomnom jezgrom	Stupanj polinoma = 2, uključeni i članovi nižeg stupnja, parametar složenosti $C = 20$
SVM s radijalnom jezgrom	Širina radijalne jezgre $\gamma = 0.8$, parametar složenosti $C = 20$
Slučajna šuma (RF)	Broj stabala = 60, broj atributa razmotrenih pri svakom čvoru $m = \sqrt{M}$
AdaBoost+C4.5	Broj iteracija postupka uzdizanja = 40, parametri za C4.5 isti kao gore navedeni

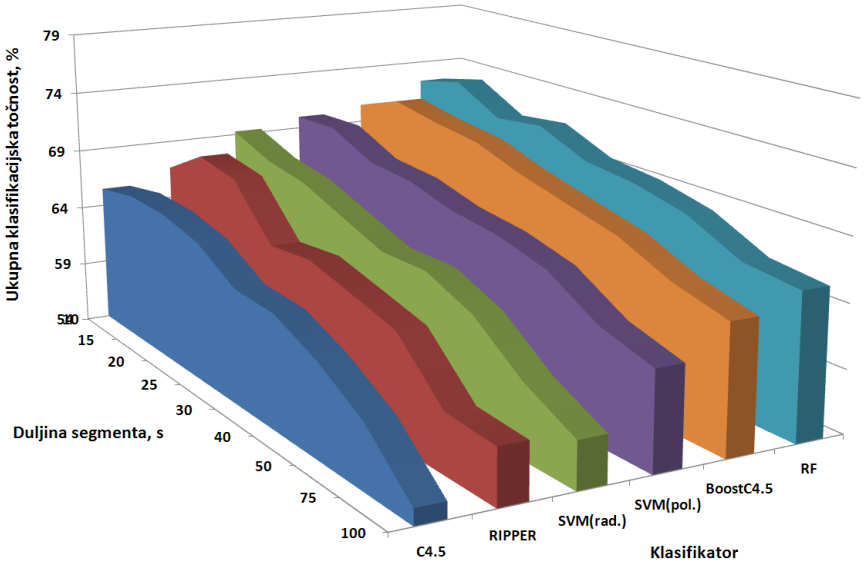
Tablica 8.9. Detaljni rezultati modela (u %) za 9 obrazaca ritma izgrađenih neovisno o zapisu korištenjem 30% najboljih atributa najbolje filterske mjere. Rezultati su dani za klasifikatore AdaBoost+C4.5 i RF.

(a) 10 s

AdaBoost +C4.5	SENS	SPEC	PPV	AUC	RF	SENS	SPEC	PPV	AUC
NSR	94.3	94.2	95.1	97.3		94.9	95.5	95.0	98.1
PAC	68.5	96.6	70.1	95.5		73.4	87.4	70.7	96.1
PVC	77.3	95.3	74.7	95.8		77.7	88.1	75.0	95.7
AF	86.5	97.8	78.0	98.5		81.2	85.9	65.6	97.5
VBI	71.0	99.2	72.3	97.8		65.7	88.6	65.9	97.4
VTR	58.4	99.5	71.3	96.7		57.8	92.6	77.2	97.9
ABI	70.8	99.4	72.5	98.7		45.8	91.7	73.6	98.1
ATR	48.1	99.7	66.0	97.4		49.2	90.4	71.3	98.0
PEJS	87.2	99.5	85.3	99.2		82.1	96.6	90.3	99.7
ACC = 85.7						ACC = 82.9 ± 0.1			

(b) 20 s

AdaBoost +C4.5	SENS	SPEC	PPV	AUC	RF	SENS	SPEC	PPV	AUC
NSR	94.4	96.3	95.7	97.5		94.3	96.1	95.5	98.1
PAC	70.7	96.0	69.6	95.2		72.8	95.9	69.9	95.8
PVC	75.7	94.8	74.7	94.7		78.4	94.6	74.4	95.7
AF	80.8	96.8	70.2	97.1		80.0	96.5	68.0	97.3
VBI	68.0	98.4	65.1	97.1		67.3	98.4	65.1	97.0
VTR	62.2	99.1	70.2	96.3		57.2	99.5	78.7	97.1
ABI	57.6	99.1	66.1	97.4		50.4	99.4	72.6	97.9
ATR	47.2	99.5	65.1	96.6		46.4	99.5	65.5	97.3
PEJS	88.7	99.5	86.5	99.1		85.0	99.5	85.6	99.6
ACC = 82.8						ACC = 82.9 ± 0.1			



Slika 8.6. Ukupna klasifikacijska točnost modela ovisno o zapisu pacijenta korištenjem 30% najboljih atributa najbolje filterske mjere za sve duljine segmenta po algoritmima.

U odnosu na prvi tip modela, rezultati su kod drugoga bitno slabiji, u prosjeku oko 10%. Najtočnije rezultate ponovno su dali klasifikatori AdaBoost+C4.5 i RF, no razlike su ovaj put nešto manje u odnosu na ostale algoritme. Potrebno je uočiti da sad duljina segmenta od 10 s nije više najtočnija, već se najbolji modeli dobivaju uglavnom za duljinu segmenta od 15 s, a točnost polako pada s produljenjem segmenta. U tablici 8.10 prikazani su detaljni rezultati za duljine segmenata od 15 s i 20 s za algoritme AdaBoost+C4.5 i RF. Rezultati dobiveni za neke poremećaje, pogotovo ABI, ATR i PEJS nisu zadovoljavajući. Razlog tome je vjerojatno taj što postoji malo zapisa pacijenata koji sadrže segmente s tim poremećajima. U ukupnoj bazi zapisa postoji samo 14 njih koji sadrže veći broj segmenata s ABI, samo 7 njih s ATR, a tek 4 zapisa zadrže umjetni pejsmejker (PEJS). Za razliku od toga, primjerice poremećaj PVC je sadržan u većem broju segmenata u preko 70 zapisa pa se stoga može izgraditi dosta točan model. Točnost predviđanja normalnog ritma je lošija u prosjeku do 2% u odnosu na model neovisan o zapisu. Ipak, u slučajevima svih obrazaca osjetljivost i ostale mjere padaju što znači da je bitno prethodno poznavati ritam određenog pacijenta da bi klasifikatoru njegov ritam postao predvidljiviji ili je potrebno imati značajno veću bazu zapisa da bi se pokrila sva moguća individualnost poremećaja.

Tablica 8.10. Detaljni rezultati modela (u %) za 9 obrazaca ritma izgrađenih uzimajući u obzir zapise pri učenju korištenjem 30% najboljih atributa najbolje filterske mjere. Rezultati su dani za klasifikatore AdaBoost+C4.5 i RF.

(a) 15 s

AdaBoost +C4.5	SENS	SPEC	PPV	RF	SENS	SPEC	PPV
NSR	91.4	90.3	90.4		92.1	89.7	90.0
PAC	53.4	94.9	56.0		59.5	94.6	57.5
PVC	68.6	92.4	63.3		68.1	92.8	64.6
AF	58.4	95.1	52.7		63.3	94.6	52.0
VBI	36.5	97.6	36.7		39.7	98.0	42.9
VTR	30.5	99.0	47.7		39.3	99.1	57.1
ABI	19.3	98.1	21.2		12.8	98.6	19.4
ATR	18.7	98.7	20.2		11.1	99.5	25.7
PEJS	35.6	98.5	47.4		29.7	98.7	45.7
ACC = 72.0					ACC = 73.2		

(b) 20 s

AdaBoost +C4.5	SENS	SPEC	PPV	RF	SENS	SPEC	PPV
NSR	92.9	91.3	90.4		93.2	90.8	89.9
PAC	57.2	94.2	55.9		59.7	93.4	53.6
PVC	65.9	92.2	62.5		64.6	92.5	63.2
AF	60.3	95.0	53.2		62.3	94.7	52.5
VBI	42.7	97.2	41.0		38.4	97.3	38.8
VTR	42.5	98.6	51.8		43.1	98.9	58.1
ABI	15.1	97.9	18.7		9.2	98.1	13.6
ATR	8.4	99.2	17.2		3.9	99.3	10.0
PEJS	32.1	99.0	55.4		28.3	99.1	54.9
ACC = 71.3					ACC = 71.0		

U tablici 8.11 dano je vrijeme izgradnje modela na ukupnom skupu podataka korištenjem 69 atributa za svaki algoritam. Analiza je provedena na računalu s Intelovim Q9300 procesorom, 2.50 GHz, s 8.0 GB RAM-a. Pokazuje se da algoritmi strojeva s potpornim vektorima (SVM) trebaju jako dugo da izgrade klasifikator u odnosu na ostale algoritme. Promatrajući brzinu izgradnje modela i postignutu točnost, slučajna šuma se pokazuje kao sveukupno najbolji algoritam za analizu BVN, što je također bio zaključak i ranijih, ali ne toliko sveobuhvatnih studija [Jović 2010 (2)].

Važno je usporediti kombinaciju značajki dobivenu metodom predloženom u ovom radu s drugim kombinacijama značajki poznatih iz literature na istom skupu podataka kako bi se utvrdio potencijal metode. U tablici 8.12 navedene su kombinacije značajki koje su se koristile u razvrstavanju raznih srčanih poremećaja i poremećaja ritma i koje

Tablica 8.11. Vrijeme potrebno za izgradnju modela. Ispod duljine segmenta (u sekundama) naveden je broj vektora značajki za koji se gradi model.

Algoritam	10 s (18264)	15 s (11870)	20 s (8675)	25 s (6841)	30 s (5612)	40 s (4104)	50 s (3212)	75 s (2050)	100 s (1483)
Slučajna šuma	35.1 s	19.7 s	14.7 s	12.4 s	9.7 s	5.8 s	5.4 s	3.3 s	2.3 s
C4.5	20.2 s	8.7 s	5.1 s	3.5 s	2.9 s	1.9 s	1.5 s	0.9 s	0.5 s
AdaBoost+C4.5	11.5 min	5.3 min	3.6 min	3.0 min	2.1 min	1.4 min	62.2 s	36.2 s	25.2 s
RIPPER	7.0 min	3.5 min	1.8 min	1.0 min	37.2 s	23.2 s	15.1 s	8.7 s	5.1 s
SVM SMO polinomna jezgra	10.5 h	2.7 h	54.9 min	28.4 min	9.9 min	6.1 min	1.9 min	38.7 s	14.4 s
SVM SMO radijalna jezgra	48.2 min	25.7 min	5.9 min	2.8 min	1.5 min	34.3 s	19.3 s	7.4 s	3.5 s

Tablica 8.12. Kombinacije značajki predložene od strane drugih istraživača.

Rbr.	Rad (broj značajki)	Kombinacija	Istraživani obrasci	Klasifikator i najbolji rezultat
1	[Asl 2008] (14)	Mean, RMSSD, SDSD, pNN5, pNN10, pNN50, LF/HF, SD1/SD2, ApEn3, SpectEn, LLE, DFA, STA1, STA2	NSR, PVC, AF, VF, SSS, BII	SVM(radijalni)+ GDA: ACC=99.16 %
2	[Yaghouby 2009] (9)	Mean, SDNN, pNN50, HTI, LF/HF, SD1/SD2, D2, LLE, SpectEn	LBBB, blok 1. stupnja, SVT, VTR	ANN(MLP) +GDA: ACC=100.0%
3	[Jović 2011 (1)] (23)	SFI, D2, CTM, ApEn1-ApEn4, SDNN, pNN20, RMSSD, HTI; odgoda τ={1, 2, 5, 10, 20}	Zdrav,aritmija, supraventrik. aritmija	Slučajna šuma: ACC=99.6%
4	[Jović 2011 (3)] (16)	Mean, SDNN, RMSSD, SDSD, pNN5, pNN10, pNN20, pNN50, HTI, Total PSD (0-0.4Hz), LF, HF, LF/HF, SD1/SD2, faktor Fano, faktor Allan	NSR, PAC, PVC, VT, VBI, VTR, PEJS, AF, BII	AdaBoost C4.5: ACC = 87.5%
5	[Pecchia 2011] (9)	Mean, SDNN, RMSSD, pNN50, Total PSD (0-0.4Hz), VLF, LF, HF, LF/HF	Zdrav, CHF	CART: SENS=79.3%, SPEC=100.0%
6	-	12 značajki	9 ritmova	-
7	-	23 značajke	9 ritmova	-
8	-	69 značajki	9 ritmova	-

će se ispitati u odnosu na kombinaciju značajki dobivenu u ovom radu. Kombinacije značajke se ispituju za duljine intervala od 15 s i 20 s, s obzirom na to da su prethodni rezultati (vidjeti tablice 8.9 i 8.10), a i [Jović 2011 (3)] sugerirali da su to duljine koje daju najtočnije rezultate pri razvrstavanju aritmija. Pritom se daje samo procjena točnosti kad se analiza provodi neovisno o zapisu pacijenta.

Kombinacije značajki navedene u tablici 8.12 uspoređuju se na segmentima duljine 15 s i 20 s sa kombinacijom od 69 značajki (30% najboljih, kombinacija #8) dobivenih filterskim odabirom atributa korištenjem mjera 1Rule (za 15 s) i IG (za 20 s), kao i s kombinacijama od 23 značajke (10% najboljih, kombinacija #7) i 12 značajki (5% najboljih, kombinacija #6). Za usporedbu se koristi algoritam slučajnih šuma s brojem stabala jednak broju značajki u kombinaciji. Rezultati usporedbe dani su u tablicama 8.13 i 8.14. za sve vrste poremećaja, a ukupna točnost razvrstavanja prikazana je na slici 8.7.

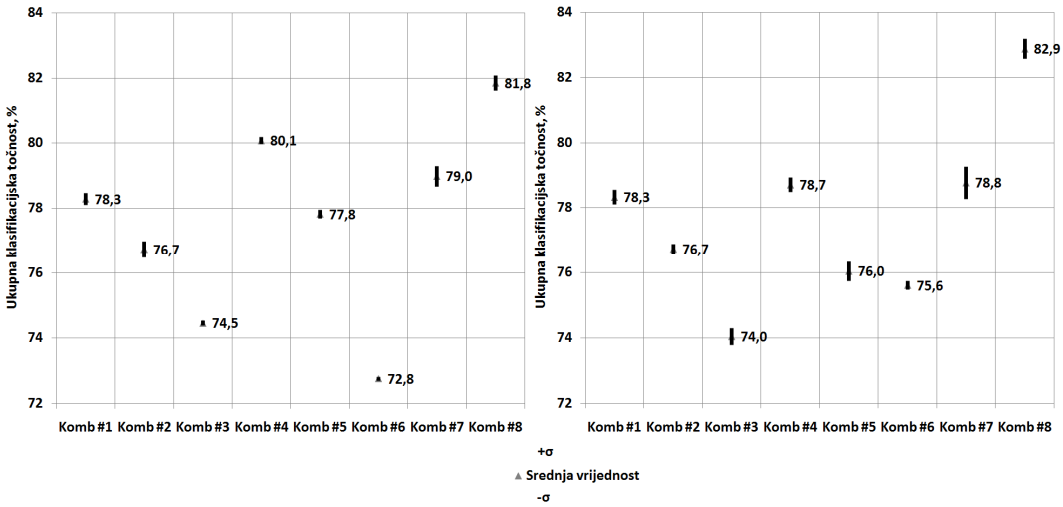
U tablicama 8.13 i 8.14 označena je sivim kombinacija koja daje općenito najbolje rezultate. To je u svim slučajevima kombinacija #8 koja sadržava 69 značajki (prvih 30% ukupnog skupa atributa). Kombinacije značajki #4 i #7 su bliske po pitanju rezultata, a daleko nije ni kombinacija #1. Pri tome je kombinacija #1 jedina od

Tablica 8.13. Usporedba kombinacija značajki na problemu razvrstavanja 9 vrsti srčanog ritma. Navedena je osjetljivost i specifičnost modela za segment duljine 15 s.

SENS, %	Rbr. kombinacije								SPEC, %	Rbr. kombinacije							
Ritam	1	2	3	4	5	6	7	8	Ritam	1	2	3	4	5	6	7	8
NSR	93.4	93.4	92.5	93.8	93.8	92.1	93.6	93.7	NSR	93.1	91.6	92.1	94.0	92.3	87.8	92.2	93.9
PAC	64.5	63.0	58.1	67.2	65.5	61.9	70.1	69.9	PAC	94.7	94.4	93.5	92.0	94.4	93.7	94.8	95.7
PVC	70.0	65.5	64.3	72.6	65.8	67.5	72.3	76.0	PVC	93.2	93.3	91.8	93.5	93.5	93.0	94.5	94.8
AF	68.7	65.8	64.3	73.2	65.8	49.0	70.7	76.5	AF	96.1	96.1	95.8	96.2	96.0	94.9	95.4	96.3
VBI	48.9	45.1	39.9	52.4	49.0	48.2	53.1	68.0	VBI	98.5	98.3	98.2	98.6	98.4	98.1	98.5	98.6
VTR	43.0	36.3	33.5	45.7	39.1	41.5	48.9	57.1	VTR	99.3	99.0	99.0	99.3	99.1	99.1	99.3	99.4
ABI	42.7	38.6	36.1	43.4	44.7	35.6	42.9	49.6	ABI	99.2	99.0	99.2	99.2	99.2	99.3	99.4	99.4
ATR	25.9	24.2	14.5	28.0	25.8	23.2	27.8	39.9	ATR	99.5	99.4	99.5	99.6	99.4	99.5	99.7	99.6
PEJS	76.7	77.1	63.8	83.4	81.2	13.9	50.0	67.1	PEJS	99.2	92.1	98.9	99.4	99.5	98.8	99.3	99.2

Tablica 8.14. Usporedba kombinacija značajki na problemu razvrstavanja 9 vrsti srčanog ritma. Navedena je osjetljivost i specifičnost modela za segment duljine 20 s.

SENS, %	Rbr. kombinacije								SPEC, %	Rbr. kombinacije							
Ritam	1	2	3	4	5	6	7	8	Ritam	1	2	3	4	5	6	7	8
NSR	93.8	93.9	93.5	94.3	94.1	92.7	93.7	94.3	NSR	95.0	93.1	93.4	95.2	93.3	92.1	94.1	96.1
PAC	67.8	65.1	59.3	66.9	65.2	67.1	69.9	74.6	PAC	94.7	94.0	93.4	94.6	93.6	93.8	94.5	95.6
PVC	71.2	67.2	64.2	70.5	63.9	68.0	72.0	77.8	PVC	92.9	93.4	91.2	93.0	93.2	93.3	93.5	94.8
AF	73.1	68.4	67.1	75.2	65.8	66.8	74.0	79.8	AF	96.4	96.1	95.9	96.3	96.0	95.7	96.2	96.5
VBI	50.0	49.9	42.4	53.7	54.0	52.4	56.2	65.7	VBI	98.3	98.0	97.9	98.3	98.1	98.2	98.3	98.5
VTR	46.4	39.8	38.5	42.8	39.9	45.4	50.2	58.2	VTR	99.0	98.8	98.8	98.9	98.7	99.0	99.3	99.5
ABI	40.3	35.2	36.2	39.7	38.1	42.9	44.4	49.8	ABI	98.9	98.8	99.0	99.1	98.9	99.1	99.1	99.3
ATR	27.5	24.7	15.4	31.3	24.7	30.7	36.9	46.8	ATR	99.2	99.2	99.3	99.2	99.1	99.3	99.4	99.5
PEJS	83.1	79.6	69.3	85.3	79.7	47.0	64.2	85.0	PEJS	99.3	99.4	99.2	99.5	99.5	99.0	99.3	99.5



(a) 15 s (b) 20 s

Slika 8.7. Usporedba ukupne točnosti razvrstavanja za kombinacije značajke iz tablice 8.12 na problemu razvrstavanja 9 vrsti srčanog ritma.

kombinacija predloženih od strane drugih autora koja je dovoljno uravnotežena da može postići relativnu visoku točnost razvrstavanja aritmija, no još uvijek bitno manju nego što je postiže kombinacija #8.

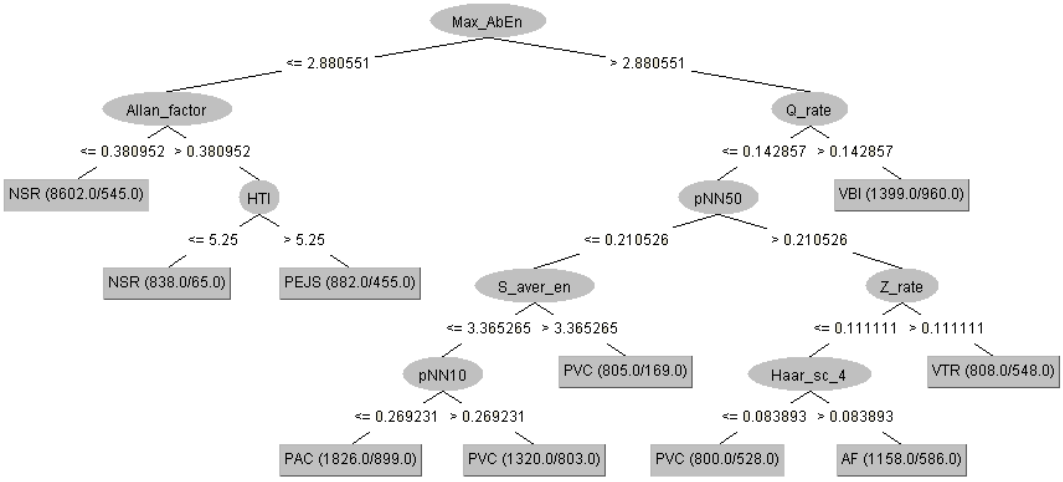
Sadržaj kombinacije #8 dan je u tablici 8.15 za duljinu od 20 s i metriku IG. Dobar dio atributa u toj kombinaciji čine značajke abecedne entropije (za detalje oko tih značajki vidjeti poglavlje 8.3), a kombinacija uključuje i većinu značajki ASTA (vidjeti

Tablica 8.15. Značajke u kombinaciji #8, poredane po relativnoj važnosti pri korištenju metrike IG za duljinu segmenta od 20 s.

Poredak / relativna važnost / naziv značajke			
1./ 0.779 /RMSSD	19./ 0.555 /X_exists	37./ 0.393 /ASTA_f8	55./ 0.304 /REC_lag_1
2./ 0.776 /SDSD	20./ 0.505 /Haar_sc_3	38./ 0.380 /SpectEn	56./ 0.301 /ASTA_g5
3./ 0.775 /X_rate	21./ 0.485 /ApEn_0_2	39./ 0.375 /Haar_sc_4	57./ 0.301/SampEn_02
4./ 0.768 /pNN50	22./ 0.484 /High_freq	40./ 0.372 /P_max_en	58./ 0.295 /R_rate
5./ 0.762 /ASTA_f7	23./ 0.466 /ApEn_015	41./ 0.365 /pNN20	59./ 0.293 /Z_rate
6./ 0.749 /SFI_lag_1	24./ 0.465 /H_max_en	42./ 0.356 /ASTA_f1	60./ 0.290 /S_max_en
7./ 0.743 /Max_AbEn	25./ 0.460 /H_rate	43./ 0.353 /P_aver_en	61./ 0.289 /P_exists
8./ 0.740 /Aver_AbEn	26./ 0.459 /SD1_SD2	44./ 0.348 /Q_max_en	62./ 0.285 /S_aver_en
9./ 0.726 /ASTA_f2	27./ 0.459 /CSI	45./ 0.344 /C_rate	63./ 0.280 /pNN10
10./ 0.689 /pNN40	28./ 0.458 /H_aver_en	46./ 0.344 /P_rate	64./ 0.278 /ASTA_g2
11./ 0.682 /A_rate	29./ 0.458 /ASTA_f5	47./ 0.342 /Q_aver_en	65./ 0.272 /SFI_lag_2
12./ 0.667 /ASTA_f9	30./ 0.440 /ASTA_f6	48./ 0.330 /SampEn025	66./ 0.268 /L-Z_comp
13./ 0.667 /X_max_en	31./ 0.436 /Q_rate	49./ 0.329 /pNN15	67./ 0.266 /R_max_en
14./ 0.657 /SDNN	32./ 0.429 /ApEn_025	50./ 0.318 /Mean	68./ 0.265 /ASTA_g3
15./ 0.630 /X_aver_en	33./ 0.411 /ApEn_0_1	51./ 0.317 /Q_exists	69./ 0.261 /S_rate
16./ 0.602 /AbEn_Var	34./ 0.404 /C_max_en	52./ 0.314 /HTI	
17./ 0.591 /ASTA_f4	35./ 0.398 /C_aver_en	53./ 0.305 /Fano_factor	
18./ 0.559 /pNN30	36./ 0.396 /I_exists	54./ 0.305 /Z_en_vari	

poglavlje 8.2). Osim njih, kombinacija sadrži većinu linearnih vremenskih značajki, kao i neke nelinearne. Važno je naglasiti da se kombinacija #8 ne razlikuje bitno u ovisnosti o duljini segmenta kao niti po odabranoj metrici. To znači da su većina značajki prikazani u tablici 8.15 prisutna i za segmente ostalih duljina kao i za različite metrike, uz tek manje razlike koje se ovdje detaljno ne navode.

Peti korak sustavnog postupka uključuje izgradnju liječnicima razumljivih modela u obliku stabla odluke C4.5 i klasifikacijskih pravila RIPPER na duljini segmenta s najvišom točnosti. Razumljivi model izgrađen je za obje procjene točnosti, neovisno o zapisu pacijenta i u ovisnosti o zapisu pacijenta, oba puta koristeći unakrsnu validaciju s 10 preklopa. Model koji je izgrađen neovisno o zapisu prikazan je na slici 8.8 za period od 10 s.



(a) C4.5

1. (Q_rate >= 0.166667) and (0.001858 <= High_freq <= 0.025493) => signal_type=ABI (367.0/121.0)
2. (Q_rate >= 0.153846) and (X_rate >= 0.375) => signal_type=VBI (435.0/156.0)
3. (Allan_factor >= 0.4) and (High_freq = 0) and (Fano_factor <= 0.133333) => signal_type=PEJS (730.0/263.0)
4. (ASTA_f4 >= 3) and (pNN20 >= 0.357143) and (High_freq <= 1.183055) and (Haar_sc_4 >= 0.113946) => signal_type=AF (628.0/163.0)
5. (pNN50 >= 0.214286) and (HTI >= 6.5) and (0.03406 <= High_freq <= 0.179422) and (Haar_sc_3 >= 0.08448) => signal_type=AF (287.0/47.0)
6. (ASTA_f4 >= 2) and (ASTA_f5 >= 4) and (0.354922 <= High_freq <= 0.80021) => signal_type=AF (428.0/204.0)
7. (Max_AbEn >= 2.853965) and (pNN30 <= 0.1875) and (S_max_en <= 3.320306) and (Haar_sc_4 >= 0.063896) and (STA_minus_quadrant >= 0.3) => signal_type=PAC (505.0/163.0)
8. (Max_AbEn >= 2.82114) and (pNN30 <= 0.222222) and (2.107617 <= P_max_en <= 2.864934) => signal_type=PAC (705.0/317.0)
9. (Max_AbEn >= 2.753112) and (X_max_en <= 3.001568) and (pNN20 <= 0.25) and (H_rate >= 0.083333) and (ASTA_f5 >= 1) => signal_type=PAC (560.0/277.0)
10. (Max_AbEn >= 3.056583) and (ASTA_f5 <= 0) and (R_rate <= 0.076923) => signal_type=PVC (1224.0/213.0)
11. (Max_AbEn >= 3.048495) and (C_max_en >= 3.428303) and (ASTA_f5 <= 1) => signal_type=PVC (247.0/90.0)
12. (Max_AbEn >= 3.025521) and (ASTA_f6 >= 4) and (Mean <= 0.609067) => signal_type=PVC (343.0/141.0)
13. (ASTA_f7 >= 1) and (ASTA_g2 >= 1) and (ASTA_f5 <= 4) => signal_type=PVC (472.0/215.0)
14. => signal_type=NSR (11507.0/2069.0)

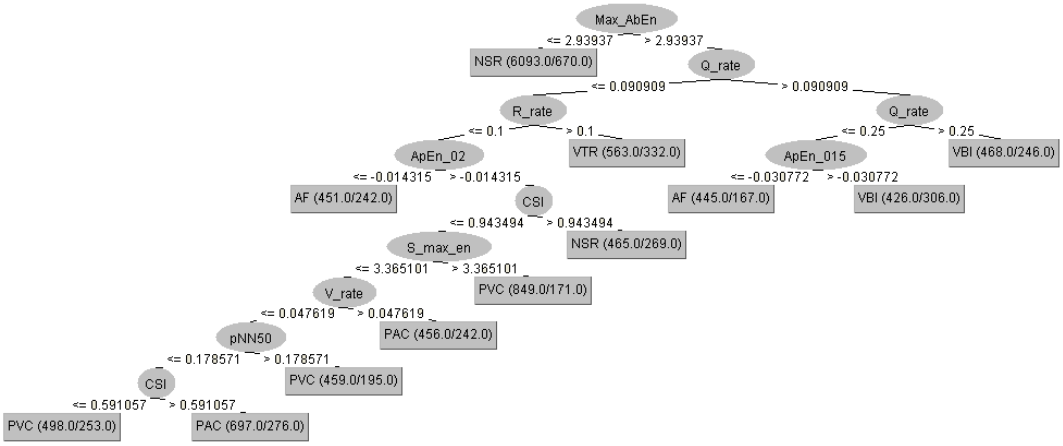
(b) RIPPER

Slika 8.8. Razumljivi model 9 ritmova za klasifikatore C4.5 i RIPPER izgrađen neovisno o zapisu pacijenta za duljinu segmenta od 10 s.

Najbolji razumljivi model za segment od 10 s dobiven je za stablo odluke C4.5 s najmanjim brojem od 800 primjeraka u listovima (ukupno 10 listova), a pokriveno je 7 poremećaja (ABI i ATR nisu mogle biti prikazane s razumljivim modelom). RIPPER je dao najbolji rezultat za 14 pravila koja prekrivaju najmanje 140 primjeraka, pri čemu nije bio obuhvaćen jedino obrazac ATR. Postignuta je ukupna klasifikacijska točnost od 69.4% za C4.5 i 73.4% za RIPPER.

Za duljinu segmenta od 15 s izgrađen je model u ovisnosti o zapisu pacijenta te su dobiveni rezultati dani na slici 8.9. U ovom slučaju, postignuta je ukupna klasifikacijska točnost od 62.9% za C4.5 i 66.0% za RIPPER.

Ako se usporede točnosti postignute razumljivim modelom i one postignute s najboljim postavkama klasifikatora za razvrstavanje, uočljivo je da su točnosti razumljivog modela slabije 5-10% u odnosu na najbolju točnost klasifikatora. Veličina stabla i skupa pravila je pritom drastično smanjena da bi rezultati bili jasni. U tablici 8.16 sumirane su karakteristike klasifikatora pri izgradnji najboljeg klasifikacijskog modela i najboljeg razumljivog modela.



(a) C4.5

1. (R_rate >= 0.117647) and (ASTA_f5 <= 2) => signal_type=VTR (233.0/68.0)
2. (ApEn_015 <= -0.064539) and (HTI >= 8.5) and (RMSSD <= 0.075647) => signal_type=PEJS (482.0/216.0)
3. (Q_rate >= 0.12) and (Q_max_en >= 3.12125) and (X_rate >= 0.307692) => signal_type=VBI (379.0/169.0)
4. (ASTA_f4 >= 4) and (ApEn_015 <= -0.014315) and (SD1_SD2 <= 1.26156) => signal_type=AF (490.0/121.0)
5. (ASTA_f4 >= 4) and (ApEn_015 <= -0.04652) and (pNN40 >= 0.315789) => signal_type=AF (281.0/110.0)
6. (ASTA_f4 >= 3) and (ApEn_015 <= -0.014315) and (pNN50 >= 0.238095) and (X_rate <= 0.190476) and (SDNN >= 0.093587) => signal_type=AF (70.0/26.0)
7. (ASTA_f7 >= 1) and (pNN50 <= 0.16) and (P_max_en <= 2.807901) and (P_aver_en >= 2.206689) => signal_type=PAC (626.0/186.0)
8. (Max_AbEn >= 2.989017) and (X_max_en <= 2.992199) and (pNN30 <= 0.217391) and (S_aver_en <= 3.024966) and (SampEn_02 >= 0.436907) and (H_aver_en >= 2.507017) => signal_type=PAC (205.0/66.0)
9. (Max_AbEn >= 2.721215) and (Aver_AbEn <= 1.84725) and (SD1_SD2 >= 1.049361) and (P_rate <= 0.034483) and (ASTA_f5 <= 2) => signal_type=PAC (308.0/144.0)
10. (Max_AbEn >= 3.255837) and (ASTA_f5 <= 2) and (ASTA_f5 <= 0) => signal_type=PVC (689.0/62.0)
11. (Max_AbEn >= 3.138733) and (ASTA_f5 <= 3) and (P_max_en >= 2.65817) => signal_type=PVC (420.0/130.0)
12. (Max_AbEn >= 3.138733) and (pNN50 <= 0.222222) and (V_rate <= 0.045455) and (AbEn_Var >= 0.83926) and (R_rate <= 0.047619) => signal_type=PVC (301.0/123.0)
13. => signal_type=NSR (7386.0/1858.0)

(b) RIPPER

Slika 8.9. Razumljivi model 9 ritmova za klasifikatore C4.5 i RIPPER izgrađen ovisno o zapisu pacijenta za segmente od 15 s.

Tablica 8.16. Karakteristike klasifikatora izgrađenih postupcima C4.5 i RIPPER i to za slučaj najboljeg klasifikacijskog modela i za slučaj najboljeg razumljivog modela nad skupom od 69 atributa.

Klasifikator	Karakteristika	Najbolji klasifikacijski model		Najbolji razumljivi model	
		Neo.10s	Ov.15 s	Neo.10 s	Ov.15 s
C4.5	ACC, %	79.4	66.2	69.4	62.9
	Broj listova	1337	1023	10	12
	Vrijeme izgradnje, s	12.0	7.3	3.4	2.3
	Min. primjeraka u listovima	2	2	800	400
	Obuhvaćeno obrazaca	9	9	7	6
RIPPER	ACC, %	80.8	68.9	73.4	66.0
	Broj pravila	105	89	14	13
	Vrijeme izgradnje, s	330.8	117.2	58.8	25.5
	Min. primjeraka pokrivenih pravilom	2	2	140	100
	Obuhvaćeno obrazaca	9	9	8	7

Ako se usporede dva algoritma za izgradnju razumljivih pravila, može se ustanoviti da je algoritam C4.5 pogodniji zbog vizualizacije modela u obliku stabla i da ima prednost po pitanju brzine izgradnje modela. RIPPER daje u pravilu nešto točnije rezultate, izgrađuje vrlo jasna pravila te uspijeva obuhvatiti više poremećaja u izgrađenim pravilima.

Konačno, da bi se ustanovilo postoji li mogućnost poboljšanja razvrstavanja poremećaja ako bi se istovremeno promatralo više duljina segmenata (korak 6 sustavnog postupka), provedena je analiza koja uzima u obzir duljine segmenata od 10 s do 50 s, s ukupnim brojem od 62 730 vektora značajki. Pritom je ponovno najprije provedeno smanjenje atributa na 30% izvornog skupa korištenjem filterske mjere IG te se takva kombinacija razvrstavala korištenjem postupka slučajne šume s 60 stabala. Rezultati su prikazani u tablici 8.17 za slučaj analize neovisno o zapisu pacijenta i u tablici 8.18 ovisno o zapisu. Ukupna klasifikacijska točnost modela izgrađenog neovisno o zapisu

Tablica 8.17. Razvrstavanje 9 ritmova korištenjem informacija dobivenih s više različitih duljina segmenata (10, 15, 20, 25, 30, 40 i 50 s) za 30% najbitnijih značajki korištenjem postupka slučajnih šuma. Prikazani su rezultati neovisno o zapisu pacijenta.

Ritam	SENS, %	SPEC, %	PPV, %	AUC, %	ACC, %
NSR	95.9	95.9	95.6	98.8	
PAC	82.8	97.3	80.6	97.9	
PVC	83.6	97.2	85.2	97.6	
AF	85.8	97.2	73.6	98.5	
VBI	81.1	99.0	78.3	98.9	88.30±0.01
VTR	70.3	99.5	83.7	98.7	
ABI	69.8	99.6	84.5	99.2	
ATR	67.9	99.7	81.7	98.9	
PEJS	83.0	99.7	92.5	99.6	

Tablica 8.18. Razvrstavanje 9 ritmova korištenjem informacija dobivenih s više različitih duljina segmenata (10, 15, 20, 25, 30, 40 i 50 s) za 30% najbitnijih značajki korištenjem postupka slučajnih šuma. Prikazani su rezultati ovisno o zapisu pacijenta.

Ritam	SENS, %	SPEC, %	PPV, %	ACC, %
NSR	93.6	90.6	89.7	
PAC	62.1	94.5	60.7	
PVC	68.9	93.4	67.0	
AF	60.7	94.1	49.0	
VBI	43.7	97.5	45.4	72.62±0.04
VTR	42.5	98.6	52.7	
ABI	17.0	98.3	23.7	
ATR	9.8	99.2	20.7	
PEJS	26.1	99.4	61.2	

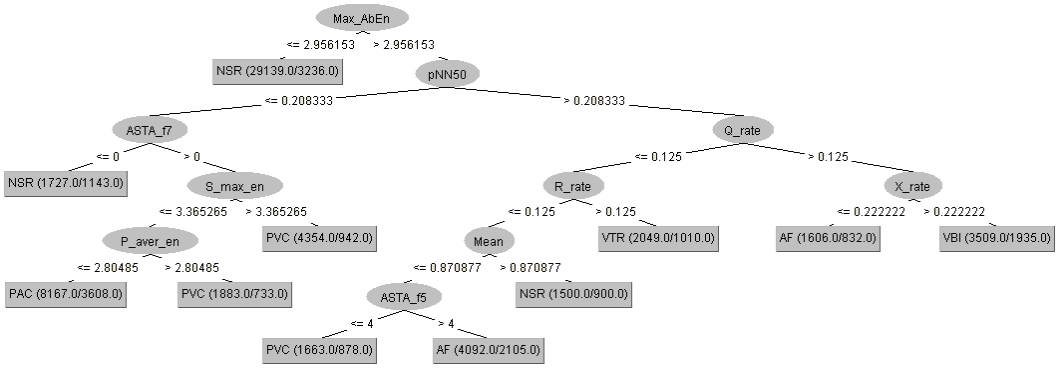
iznosi 88.3%, čime je značajno popravljen rezultat slučajnih šuma na pojedinačnim duljinama segmenta. Vrijeme potrebno za izgradnju modela iznosilo je 3.3 min (usporediti s tablicom 8.11). U slučaju modela izgrađenog ovisno o zapisu pacijenta vršna točnost ostaje nepromijenjena u odnosu na analizu po samo jednom segmentu, što znači da za analizu novih zapisa pacijenata izgradnja modela na više vremenskih skala nije opravdana.

Razumljivi modeli generirani postupcima C4.5 i RIPPER prikazani su na slici 8.10. Klasifikator C4.5 postigao je najbolju točnost od 70.4% sa stablom od 11 listova i najmanje 1500 primjeraka po listu, dok je RIPPER postigao točnost od 71.0% korištenjem skupa od 14 pravila s najmanjim brojem od 400 primjeraka pokrivenih pravilom. Zanimljivo je da za razliku od slučajnih šuma, dobiveni razumljivi modeli uz pomoć C4.5 i RIPPER postupaka za slučaj analize 10-50 s nisu ništa bolji od analize za samo jednu duljinu segmenta te se stoga niti ne bi trebali koristiti u okviru sustavnog postupka.

8.1.2 Razvrstavanje zdravih osoba, pacijenata s aritmijom i pacijenata s CHF

Druga analiza namijenjena je razlikovanju između zdravih osoba, pacijenata s kongestivnim zatajenjem srca i pacijenata sa svim ostalim poremećajima ritma (aritmijama) na temelju varijabilnosti srčanog ritma. Ovakva vrsta analize predstavlja tipičan klasifikacijski problem u biomedicinskim istraživanjima.

Cilj ove analize je utvrditi koliko se točno može razlikovati pacijent s aritmijom od pacijenta s kongestivnim zatajenjem srca te koliko se obje vrste pacijenta razlikuju od osobe koja nema srčanih smetnji. Također, potrebno je pronaći što je moguće kraću duljinu segmenta koja bi davala za liječnike zadovoljavajuće rezultate. U ovom slučaju nije potrebno provoditi smanjenje broja poremećaja sustavnim postupkom budući da postoje jasna tri ciljna razreda: normalan (NORM), aritmija (ARIT) i kongestivno zatajenje srca (CHF).



(a) C4.5

1. (Z_rate >= 0.0678) and (R_rate >= 0.1694) => signal_type=VTR (1547.0/650.0)
2. (0.8034 <= Mean <= 0.8563) and (RMSSD <= 0.0809) and (Haar_sc_3 >= 0.0331) => signal_type=PEJS (2559.0/882.0)
3. (Q_rate >= 0.0926) and (Q_max_en >= 3.2104) and (X_rate >= 0.2917) => signal_type=VBI (1745.0/748.0)
4. (pNN50 >= 0.2105) and (Mean <= 0.8331) and (SD1_SD2 <= 1.3042) and (HTI >= 11.8) => signal_type=AF (2047.0/521.0)
5. (pNN50 >= 0.186) and (Mean <= 0.8536) and (ApEn_02 <= 0.1263) and (pNN20 >= 0.3548) and (X_rate <= 0.1905) and (ASTA_f2 >= 5) => signal_type=AF (967.0/289.0)
6. (ASTA_f7 >= 1) and (pNN40 <= 0.1774) and (P_max_en <= 2.8197) and (P_aver_en >= 2.0082) and (C_aver_en >= 3.5218) and (H_max_en <= 3.0275) => signal_type=PAC (1252.0/128.0)
7. (ASTA_f7 >= 1) and (pNN50 <= 0.1613) and (2.2037 <= S_max_en <= 3.1924) and (STA_plus_quadrant <= 0.14) => signal_type=PAC (1337.0/444.0)
8. (ASTA_f7 >= 1) and (pNN40 <= 0.2029) and (P_max_en <= 2.8657) and (Max_AbEn <= 3.6416) and (pNN40 <= 0.1316) and (pNN10 >= 0.1786) and (Mean >= 0.5921) => signal_type=PAC (1541.0/516.0)
9. (Max_AbEn >= 2.8217) and (pNN40 <= 0.2143) and (S_aver_en <= 3.1507) and (SD1_SD2 >= 1.0544) and (X_max_en <= 2.9533) and (ASTA_f7 >= 2) => signal_type=PAC (2198.0/1078.0)
10. (Max_AbEn >= 3.2318) and (ASTA_f5 <= 1) => signal_type=PVC (5295.0/1027.0)
11. (Max_AbEn >= 3.2159) and (ASTA_f5 <= 3) and (ASTA_g2 >= 1) and (SD1_SD2 >= 1.5832) => signal_type=PVC (858.0/265.0)
12. (Max_AbEn >= 3.1519) and (ASTA_f5 <= 3) and (2.8283 <= P_max_en <= 3.4495) => signal_type=PVC (435.0/171.0)
13. (Max_AbEn >= 3.1351) and (pNN50 <= 0.2222) and (SD1_SD2 >= 1.4815) and (ASTA_f7 <= 7) => signal_type=PVC (1431.0/699.0)
14. => signal_type=NSR (36477.0/9633.0)

(b) RIPPER

Slika 8.10. Razumljivi model 9 obrazaca ritma za klasifikatore C4.5 i RIPPER izgrađen neovisno o zapisu pacijenta za duljine segmenta od 10-50 s.

Ovdje se koriste sve baze navedene u tablici 8.1, ukupno njih šest, što predstavlja rijetkost u analizi BVN. Analizira se ukupno 237 zapisa različitih pacijenata, u ukupnom trajanju od 174.5 sata. Ovo istraživanje pretpostavlja da svi zapisi sadržani u bazama *MIT-BIH Arrhythmia Database* i *MIT-BIH Supraventricular Arrhythmia Database* sadržavaju neku vrstu aritmije. U nekim slučajevima, zapisi mogu većinu vremena sadržavati normalan sinusni ritam, tek ponekad prošaran rijetkim aritmičnim otkucajima. Bez obzira na to radi li se u konkretnom segmentu o normalnom ritmu, benignoj aritmiji ili klinički značajnoj aritmiji, za sve segmente iz svih zapisa iz ovih dviju baza smatra se da pripadaju pacijentu koji ima problem s aritmijom. Slično, neki zapisi iz dviju baza s normalnim sinusnim ritmom sadrže neke segmente s benignim aritmijama. Takvi segmenti se svrstavaju u ovoj analizi pod tip NORM, budući da su to

Tablica 8.19. Broj vektora značajki za tri tipa pacijenta po duljini segmenta, analizirane sve baze navedene u tablici 8.1.

	Duljina segmenta, s						
Tip	200	300	500	750	1000	1500	Ukupno / udio, %
NORM	1875	1182	649	372	259	152	4489 / 35.42
ARIT	2073	1328	738	488	245	121	4993 / 39.40
CHF	1284	807	456	306	202	135	3190 / 25.17
Ukupno:	5232	3317	1843	1166	706	408	12 672

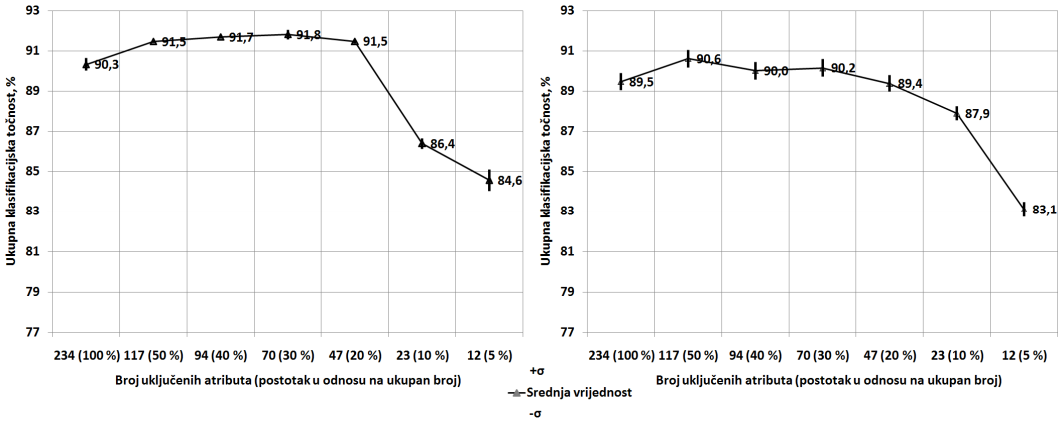
ritmovi prisutni kod osoba koje su liječnici ocijenili zdravima. Konačno, pacijenti iz zadnjih dviju baza koji boluju od CHF povremeno imaju aritmične epizode čime je otežano njihovo razlikovanje u odnosu na pacijente koji boluju samo od aritmije.

Sustavnim postupkom najprije je izlučeno ukupno 234 prediktivne linearne i nelinearne značajke, iste kao i kod analize provedene u poglavlju 8.1.1. Jedina razlika je u tome što su četiri značajke više mogle biti uključene budući da su mogle biti izračunate za sve ovdje analizirane duljine segmenata (ULF, VLF, DFA_alpha_l (ili α_2 iz poglavlja 3.5.3.6), RenEn). Za razliku od analize aritmija, ovdje se izlučivanje značajki provodi za duže segmente: 200, 300, 500, 750, 1000 i 1500 sekundi. Razlog tome je taj što nije realno očekivati dobre rezultate pri razlikovanju srčanih oboljenja na kratkim duljinama segmenata prema dostupnoj literaturi. Također, koristi se preklapanje od 50% duljine segmenta kako bi se udvostručio broj vektora značajki budući da ih nema dovoljno, pogotovo za segmente većih duljina. Ukupno je iz 6 baza izlučeno 12 672 vektora značajki. Popis broja vektora značajki po poremećajima za segmente analiziranih duljina dan je u tablici 8.19. Može se primijetiti da je najviše vektora značajki po tipu ARIT, nešto manje ima NORM i najmanje CHF, što je bilo uvjetovano brojem dostupnih zapisa.

Najprije se provodi sustavna analiza toga koliko se atributa može ukloniti, a da se točnost razvrstavanja bitno ne smanji u odnosu na ukupan skup atributa. Dakle, želi se odrediti parametar $Y(\%)$ sa slike 7.1. Analiza je provedena na segmentima od 300 s i 750 s koristeći 100%, 50%, 40%, 30%, 20%, 10% i 5% izvornog skupa značajki. Pritom su isprobane sve tri filterske mjere za dobivanje rangiranog podskupa najbitnijih atributa i algoritam slučajnih šuma za izgradnju modela. Slučajna šuma je sadržavala broj stabala koji je bio približno jednak broju atributa.

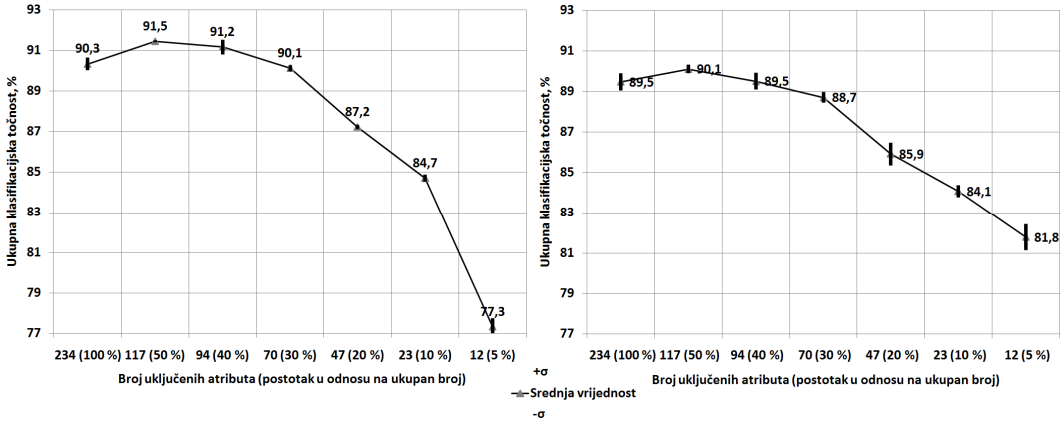
Analiza smanjenja skupa značajki prikazana je na slikama 8.11-8.13 i u tablici 8.20. Na slikama 8.11-8.13 prikazana je ukupna klasifikacijska točnost modela, a u tablici 8.20 uspoređene su mjere vrednovanja po poremećajima za čitav skup atributa i za 20% skupa za slučaj filterske mjere IG.

Analiza pokazuje da su filterske mjere IG i SU održavale točnost sve do uključenih samo 20% izvornog skupa (47 atributa), pri čemu ukupna točnost klasifikacije u svim slučajevima ne pada u odnosu na puni skup, a pada do 1% u odnosu na najviše vrijednosti postignute na smanjenim skupovima. Filterska mjera 1Rule zadržala je točnost samo do 30% izvornog skupa, nakon čega je došlo do značajnog pada rezultata.



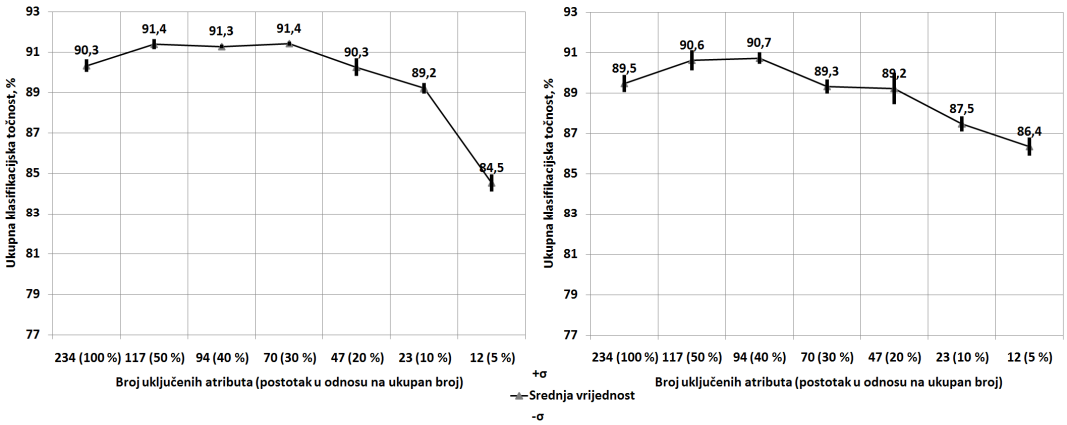
(a) 300 s (b) 750 s

Slika 8.11. Smanjenje dimenzionalnosti skupa atributa korištenjem filterske metrike IG.



(a) 300 s (b) 750 s

Slika 8.12. Smanjenje dimenzionalnosti skupa atributa korištenjem filterske metrike 1Rule.



(a) 300 s (b) 750 s

Slika 8.13. Smanjenje dimenzionalnosti skupa atributa korištenjem filterske metrike SU.

Tablica 8.20. Usporedba rezultata mjera vrednovanja (u %) za puni skup od 234 atributa i za 20% skupa atributa (47 atributa) dobivenih filterskom mjerom IG.

(a) 300 s

234 atr.	SENS	SPEC	PPV	AUC	47 atr.	SENS	SPEC	PPV	AUC
NORM	93.9	92.5	87.4	97.9		94.4	93.3	88.7	98.2
ARIT	92.8	95.8	93.7	98.5		94.3	96.0	94.0	98.8
CHF	81.2	97.0	89.6	97.5		82.5	97.6	91.6	97.6
ACC = 90.34±0.30						ACC = 91.47±0.06			

(b) 750 s

234 atr.	SENS	SPEC	PPV	AUC	47 atr.	SENS	SPEC	PPV	AUC
NORM	91.1	94.1	87.8	98.2		91.9	93.7	87.3	97.6
ARIT	92.5	93.5	91.1	97.9		91.4	93.5	91.0	97.4
CHF	82.6	96.3	88.8	97.4		83.8	96.8	90.3	97.2
ACC = 89.48±0.43						ACC = 89.38±0.41			

Daljnje poboljšanje točnosti i smanjenje broja atributa može se postići za skup od 47 atributa uklanjajući suvišne značajke korištenjem algoritma prekrivanja konzistentnosti razreda. Budući da je skup od 47 atributa velik, potrebno je definirati vrlo mali postotak ukupnog prostora kombinacija atributa za pretraživanje. U ovoj analizi uzima se da se pretražuje 1.0×10^{-7} % od ukupnog prostora. Pretraživanje se provodi za sve duljine segmenata i to za svaku od tri filterske metrike da bi se pronašao najbolji rezultat za tu duljinu. U ovom slučaju metrika 1Rule nema veliku vjerojatnost da će dati bolji podskup od ostalih metrika, ali se svejedno uzima u obzir. Vrednovanje smanjenog skupa atributa dobivenog postupkom konzistentnosti razreda u odnosu na skup od 47 atributa za određenu filterski postupak provodi se algoritmom slučajnih šuma s pet ponavljanja.

U tablici 8.21 navedeni su konačni rezultati o uspješnosti pronalaska boljeg podskupa atributa po duljinama segmenta. U drugom stupcu navodi se podatak je li konzistentnost razreda pronašla bolje rješenje od onog koji daje skup od 47 atributa. U trećem stupcu navodi se filterska metrika koja je korištena za dobivanje skupa od 47 atributa iz kojeg je konzistentnost razreda našla bolje rješenje. Ako konzistentnost razreda nije našla bolje rješenje, ovdje se navodi metrika koja sama daje najbolje rješenje na skupu od 47 atributa. Broj atributa u konačnom podskupu koji se koristi u daljnjoj analizi za svaku

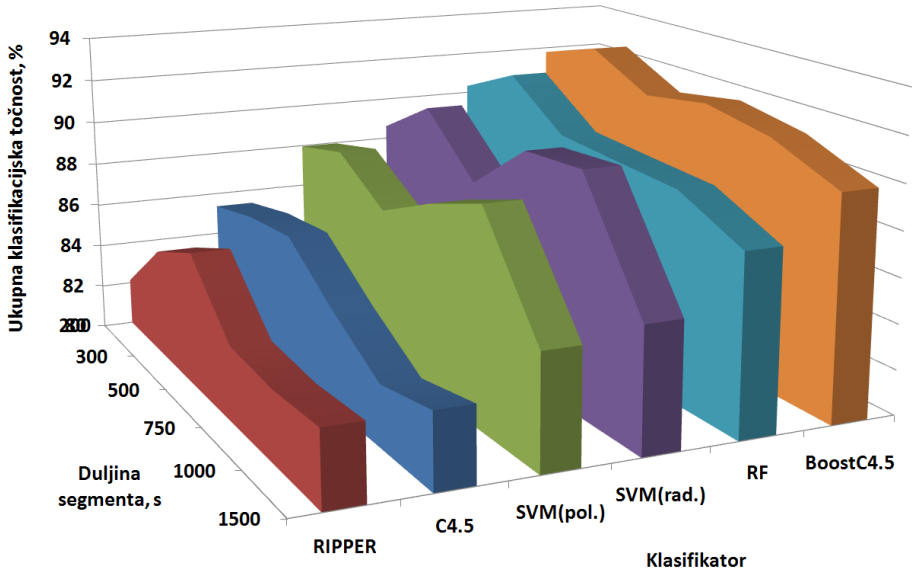
Tablica 8.21. Uspješnost pronalaska boljeg podskupa atributa korištenjem mjere prekrivanja konzistentnost razreda (InCons).

Duljina segmenta, s	InCons uspješna (DA/NE)	Korištena metrika	Br. atr. konačni	RF, ACC, %	RF kon., ACC, %	*p*
200	DA	IG	34	90.48±0.10	90.38±0.13	0.240
300	DA	IG	33	91.48±0.10	91.77±0.21	0.092
500	DA	SU	33	89.34±0.32	90.00±0.43	0.031*
750	NE	IG	47	88.49±0.68	89.45±0.40	0.045*
1000	NE	IG	47	88.67±0.42	89.01±0.63	0.456
1500	DA	SU	33	87.55±0.47	87.74±0.88	0.757

* statistički značajna razlika ($\alpha = 0.05$), parni dvosmjerni *t*-test.

duljinu segmenta navodi se u četvrtom stupcu, dok se u petom i šestom stupcu navode ukupna klasifikacijska točnost modela svih 47 atributa kao i točnost konačnog modela za istu metriku. Ako postupak InCons nije našao bolje rješenje, onda se u petom stupcu navodi točnost drugog najboljeg rješenja. Dobivena značajnost razlike navedena je u sedmom stupcu. Primjetno je da postoji jedna statistički značajna razlika u korist smanjenog broja atributa i to u slučaju duljine segmenta od 500 s za mjeru SU. Za 750 s i 1000 s ni za jednu metriku nije dobiven bolji rezultat postupkom InCons. Za 750 s, postoji statistički značajna razlika između skupa od 47 atributa za mjeru IG i najbolje od svih ostalih mjera (u ovom slučaju skupa dobivenog primjenom InCons na IG), dok za 1000 s mjera SU ne odudara značajno od IG pa je zadržan osnovni skup atributa s mjerom IG (iako je mogla biti uzeta i metrika SU).

U nastavku se za svaku duljinu segmenta grade zasebni modeli koristeći šest algoritama za razvrstavanje: slučajna šuma (RF), C4.5, AdaBoost+C4.5, RIPPER, SVM s polinomnom jezgrom i SVM s radijalnom jezgrom. Potrebno je provesti dvije zasebne procjene točnosti: jednu u slučaju uzimanja u obzir svih dostupnih segmenata neovisno o zapisu pacijenta i drugu, koja uzima u obzir zapis pacijenta te stoga na jednim zapisima uči, a na drugima ispituje. Kod druge procjene, zapisi su slučajno podijeljeni u 10 skupova koji se koriste za unakrsnu validaciju s 10 preklopa. Rezultati razvrstavanja poremećaja za prvu procjenu točnosti koja zanemaruje same zapise iz kojih su segmenti stigli prikazani su na slici 8.14 za sve duljine segmenta i za sve algoritme. Parametri ovdje korištenih klasifikatori isti su kao i u poglavlju 8.1.1 (tablica 8.8). Najboljim klasifikatorom se u ovom slučaju pokazao AdaBoost+C4.5, dok je RF bio drugi po točnosti modela. SVM s radijalnom jezgrom se pokazao boljim od onog s



Slika 8.14. Ukupna klasifikacijska točnost modela neovisno o zapisu korištenjem najboljih kombinacija atributa prema tablici 8.21 za sve duljine segmenta po algoritmima.

polinomnom jezgrom, ali ne toliko točan kao RF. RIPPER je u ovom slučaju ukupno bio najslabiji. Najbolja duljina segmenta za razlikovanje stanja pacijenta pokazala se ona od 300 s (5 min), a visoki rezultat je dobiven i za 750 s. Zanimljivo je primijetiti podbačaj većine klasifikatora za duljinu od 500 s za što nije pronađeno jasno objašnjenje (nije vezano uz filtersku mjeru SU). U tablici 8.22 prikazani su detaljni rezultati za duljine segmenata od 300 i 750 s za algoritme AdaBoost+C4.5 i RF. Postignuta je visoka točnost u otkrivanju stanja pacijenta za 300 s i klasifikator AdaBoost+C4.5, s osjetljivošću između 83% i 96%, specifičnošću između 94% i 98%, te predvidljivošću pozitivnih primjeraka između 90% i 95%. Ostvariva ukupna klasifikacijska točnost korištenjem jedne duljine segmenta je do 93%. Model za 750 s omogućuje nešto veću osjetljivost otkrivanja CHF, dok se smanjuje točnost otkrivanja ARIT i NORM, što je razumljivo iz medicinskog stajališta.

Rezultati razvrstavanja poremećaja za drugu procjenu točnosti koja uzima u obzir same zapise pacijenata iz kojih su segmenti stigli prikazani su na slici 8.15 za sve duljine segmenta i za sve algoritme.

Iz slike se može uočiti da, za razliku od analize provedene neovisno o zapisu pacijenta, ako se uzme u obzir zapis pacijenta tada klasifikacijska točnost uglavnom raste s duljinom segmenta, dosegnuvši blizu 85% za algoritam AdaBoost+C4.5 za 1500 s. Ta činjenica je bitna budući da sugerira da bi uzimanjem duljih segmenata bilo vjerojatno moguće postići i višu točnost. Ograničenje je duljina zapisa u bazi uzeta u obzir u ovoj disertaciji koja ne omogućuje razlikovanje između tri vrste zapisa za period veći od pola sata (1800 s). Klasifikatori su se i ovdje, kao i kod analize ritmova u poglavlju 8.1.1, ponašali drugačije u odnosu na analizu provedenu neovisno o zapisu. Algoritam RF nije dao bolje rezultate u odnosu na algoritme SVM, a algoritmi C4.5 i RIPPER bitno ne zaostaju za jačim algoritmima. Mogući razlog tome je nepostojanje neravnoteže u broju primjeraka ciljnog razreda, dok je neravnoteža postojala kod razvrstavanje ritmova. Time je olakšano razvrstavanje algoritmima osjetljivima na

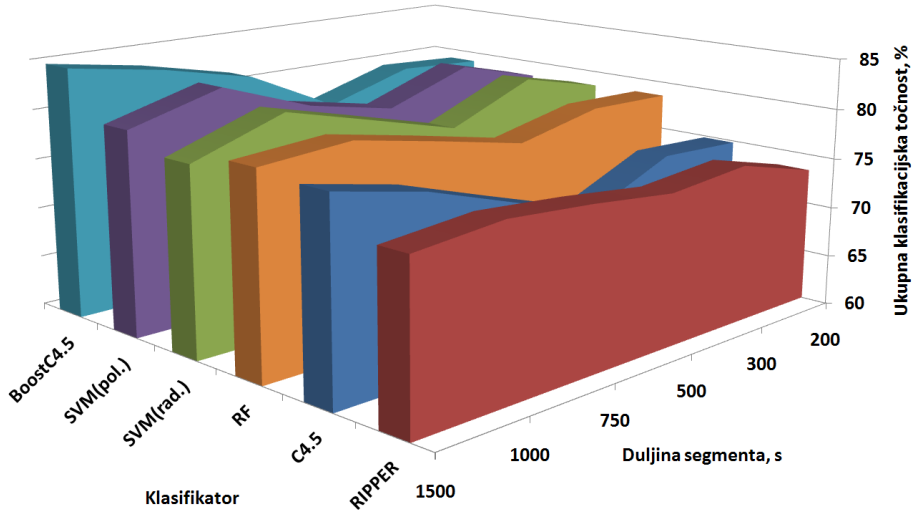
Tablica 8.22. Detaljni rezultati modela (u %) izgrađenih neovisno o zapisu pacijenta pri korištenju smanjenog skupa značajki prema tablici 8.21. Rezultati su dani za klasifikatore AdaBoost+C4.5 i RF za duljine segmenta od 300 s i 750 s.

(a) AdaBoost+C4.5

300 s	SENS	SPEC	PPV	750 s	SENS	SPEC	PPV
NORM	95.3	94.1	89.9		93.0	94.8	89.4
ARIT	96.1	96.8	95.2		93.4	94.2	94.2
CHF	83.3	97.9	92.8		87.6	98.1	98.1
	ACC = 92.70				ACC = 91.77		

(b) RF

300 s	SENS	SPEC	PPV	750 s	SENS	SPEC	PPV
NORM	94.1	93.8	89.4		91.9	93.7	87.3
ARIT	94.8	95.7	93.6		91.4	93.5	91.0
CHF	83.2	97.7	92.2		83.8	96.8	90.3
	ACC = 91.71±0.22				ACC = 89.54±0.45		



Slika 8.15. Ukupna klasifikacijska točnost modela ovisno o zapisu pacijenta korištenjem najboljih kombinacija atributa prema tablici 8.21 za sve duljine segmenta po algoritmima.

neravnotežu kao što su SVM i indukcijski algoritmi zasnovani na stablima i pravilima.

Detaljni rezultati prikazani su u tablici 8.23 za duljine segmenata od 1000 i 1500 s za algoritme AdaBoost+C4.5 i SVM s polinomnom jezgrom. Postignuti rezultati u otkrivanju stanja pacijenta za 1500 s i klasifikator AdaBoost+C4.5 su: osjetljivost između 83% i 96%, specifičnost između 94% i 98%, te predvidljivost pozitivnih primjeraka između 90% i 95%. Model je najslabije osjetljiv na poremećaj CHF, dok je nešto točniji za ARIT i NORM.

Peti korak sustavnog postupka uključuje izgradnju liječnicima razumljivih modela u obliku stabla odluke C4.5 i klasifikacijskih pravila RIPPER. Razumljivi model izgrađen je za obje procjene točnosti, neovisno o zapisu pacijenta i u ovisnosti o zapisu pacijenta, oba puta koristeći unakrsnu validaciju s 10 preklopa. Model koji je izgrađen neovisno o

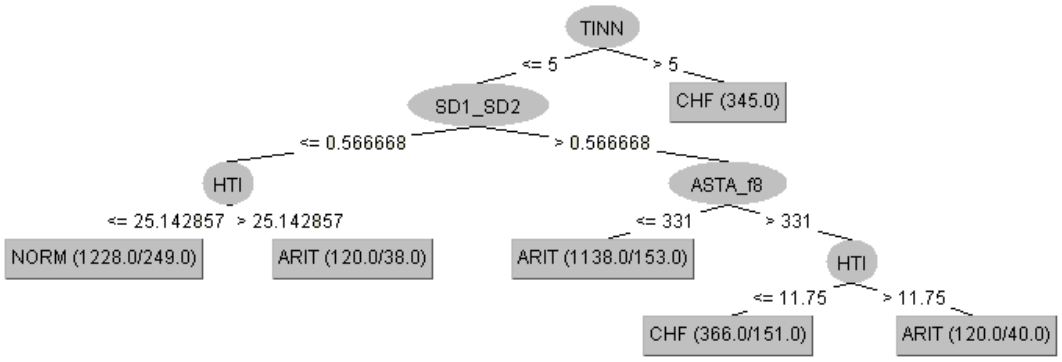
Tablica 8.23. Detaljni rezultati modela (u %) izgrađenih ovisno o zapisu pacijenta pri korištenju smanjenog skupa značajki prema tablici 8.21. Rezultati su dani za klasifikatore AdaBoost+C4.5 i SVM s polinomnom jezgrom za duljine segmenta od 1000 s i 1500 s.

(a) AdaBoost+C4.5

1000 s	SENS	SPEC	PPV	1500 s	SENS	SPEC	PPV
NORM	89.0	89.9	83,9		90.4	90.9	86.1
ARIT	85.7	88.8	80,2		87.3	90.1	78.0
CHF	73.3	95.9	87,6		76.3	96.4	91.2
	ACC = 83.38				ACC = 84.87		

(b) SVM s polinomnom jezgrom

1000 s	SENS	SPEC	PPV	1500 s	SENS	SPEC	PPV
NORM	89.0	87.5	80.7		88.5	88.1	82.2
ARIT	82.0	91.2	83.1		81.4	89.0	75.0
CHF	75.7	95.1	86.0		69.6	93.1	83.2
	ACC = 82.81				ACC = 80.24		



(a) C4.5

(TINN >= 7) => signal_type=CHF (345.0/0.0)
(HTI <= 10.020833) and (pNN15 <= 0.118072) and (CSI <= 1.798564) and (X_rate <= 0.008989) and (L-Z_comp >= 1.31539) => signal_type=CHF (112.0/12.0)
(ASTA_f8 >= 332) and (Low_high_freq <= 1.249181) and (HTI <= 11.914286) and (391 <= ASTA_f6 <= 502) and (X_aver_en <= 2.972766) => signal_type=CHF (162.0/54.0)
(CSI >= 1.765336) and (HTI <= 20.529412) and (Low_high_freq >= 2.39951) => signal_type=NORM (578.0/53.0)
(CSI >= 1.513549) and (HTI <= 25.230769) and (Max_AbEn <= 3.820503) => signal_type=NORM (552.0/119.0)
(X_rate <= 0.009375) and (L-Z_comp <= 1.543938) => signal_type=NORM (193.0/90.0)
=> signal_type=ARIT (1375.0/210.0)

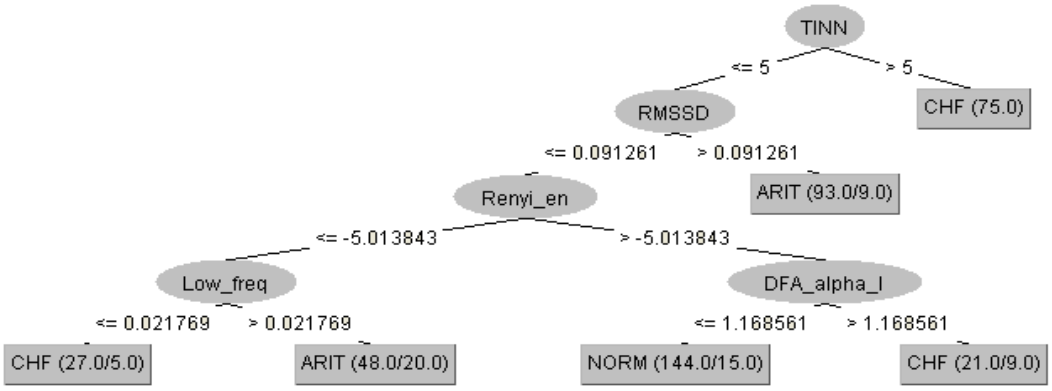
(b) RIPPER

Slika 8.16. Razumljivi model tri stanja srca za klasifikatore C4.5 i RIPPER izgrađen neovisno o zapisu pacijenta za duljinu segmenta od 300 s.

zapisu prikazan je na slici 8.16 za period od 300 s (korištenjem smanjenog skupa od 33 atributa (IG+InCons) prema tablici 8.21). Postignuta je točnost od 80.16% za C4.5 i 82.12% za klasifikator RIPPER, pri čemu je C4.5 izgrađen s najmanjim brojem od 120 primjeraka u listovima, a RIPPER koristi pravila koja prekrivaju najmanje 60 primjeraka. Kod stabla je zanimljivo uočiti granu koja pokriva sve pozitivne primjerke za CHF, a to je ona gdje je značajka TINN > 5. Pokazuje se da je značajka ASTA_f8 također vrlo značajna za otkrivanja ARIT, dok značajka HTI rafinira izbor između CHF i ARIT kao i između NORM i ARIT.

Model koji je izrađen ovisno o zapisu prikazan na slici 8.17 za period od 1500 s (korištenjem smanjenog skupa od 33 atributa (SU+InCons) prema tablici 8.21). C4.5 je postigao ukupnu klasifikacijsku točnost od 78.68% s najmanje 20 primjeraka u listovima, dok pravila izgrađena RIPPER-om postižu točnost od 77.94% i prekrivaju najmanje 10 primjeraka. Zanimljivo je primijetiti da se ovdje pojavljuju neke značajke koje nisu prisutne na kraćim vremenskim segmentima, kao što su Rényijeva entropija, standardna devijacija Haarovog valića, indeks prostorne popunjenosti (SFI) i značajka DFA_alpha_l (ili α_2 iz poglavlja 3.5.3.6) koja se koristi za opis dugoročnih kolebanja ritma.

Konačno, šesti korak sustavnog postupka uključuje izgradnju modela korištenjem informacija s više vremenskih skala. Ovdje se uključuje informacija na temelju vremena segmenata od 200-1500 s (svih 12 672 primjerka). Budući da nije moguće izgraditi



(a) C4.5

1. (RMSSD >= 0.09451) and (TINN <= 5) => signal_type=ARIT (92.0/9.0)
2. (Renyi_en <= -5.013843) and (pNN15 >= 0.239395) and (X_rate <= 0.031831) => signal_type=ARIT (17.0/2.0)
3. (TINN >= 12) => signal_type=CHF (75.0/0.0)
4. (SFI_lag_2 >= 6.827771) and (Haar_sc_4 <= 0.042061) and (RMSSD >= 0.01142) and (pNN20 <= 0.084666) => signal_type=CHF (23.0/0.0)
5. => signal_type=NORM (201.0/50.0)

(b) RIPPER

Slika 8.17. Razumljivi model tri stanja srca za klasifikatore C4.5 i RIPPER izgrađen ovisno o zapisu pacijenta za duljinu segmenta od 1500 s.

zajednički konačni model korištenjem različitih podskupova atributa dobivenih postupkom InCons (tablica 8.21), model će biti izgrađen korištenjem metrike IG za broj od 47 atributa (20% ukupnog skupa), budući da se ta metrika pokazala najčešće najboljom. Konačni model izgrađen je unakrsnom validacijom s 10 preklopa postupcima slučajnih šuma s 50 stabala i 5 ponavljanja i AdaBoost+C4.5 (br. iteracija = 40, C4.5: c=0.4, $n_{\min}$ = 2), a rezultati su dani za slučajeve izgradnje modela neovisno o zapisu pacijenta i ovisno o zapisu pacijenta, u tablicama 8.24 i 8.25. Postignuti su vrlo

Tablica 8.24. Razvrstavanje srčanih stanja korištenjem informacija dobivenih s više različitih duljina segmenata (200, 300, 500, 750, 1000, i 1500 s) koristeći filtersku metriku IG sa skupom od 47 atributa, a model je izgrađen neovisno o zapisu pacijenta.

(a) RF

Stanje	SENS, %	SPEC, %	PPV, %	ACC, %
NORM	96.6	94.9	91.3	94.03±0.06
ARIT	94.9	97.4	96.0	
CHF	89.0	98.6	95.2	

(b) AdaBoost+C4.5

Stanje	SENS, %	SPEC, %	PPV, %	ACC, %
NORM	97.2	96.6	94.0	95.84
ARIT	96.8	98.0	97.0	
CHF	92.4	99.0	96.9	

Tablica 8.25. Razvrstavanje srčanih oboljenja korištenjem informacija dobivenih s više različitih duljina segmenata (200, 300, 500, 750, 1000, i 1500 s) koristeći filtersku metriku IG sa skupom od 47 atributa, a model je izgrađen ovisno o zapisu pacijenta.

(a) RF

Ritam	SENS, %	SPEC, %	PPV, %	ACC, %
NORM	85.3	85.9	76.9	76.88±0.05
ARIT	82.3	84.1	77.1	
CHF	56.6	94.1	76.4	

(b) AdaBoost+C4.5

Ritam	SENS, %	SPEC, %	PPV, %	ACC, %
NORM	84.3	86.6	77.5	77.74
ARIT	84.2	84.3	77.7	
CHF	58.3	94.6	78.3	

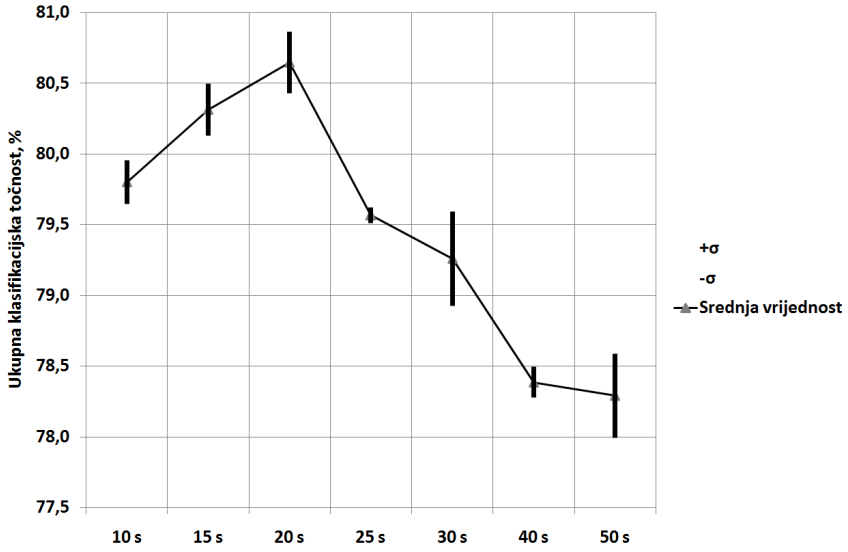
visoki rezultati u slučaju razvrstavanja neovisno o zapisu pacijenta, slično kao i u slučaju ritmova opisanih u poglavlju 8.1.1, dok su rezultati postignuti modelom na problemu razvrstavanja u ovisnosti o zapisu pacijenta relativno loši. Rezultati su u tom slučaju lošiji nego da se uzima samo jedan, duži segment (primjerice onaj od 1000 s ili 1500 s), ali u rangu s rezultatima postignutima za kraće duljine segmenta (kojih je ujedno i najviše pri izgradnji modela, usporediti sliku 8.15 i tablicu 8.25).

8.2 Vrednovanje napredne analize slijednog trenda

Napredna analiza slijednog trenda (ASTA) ispitivana je na istom skupu podataka kao i sustavni postupak u poglavlju 8.1.1, dakle uključene su prve dvije baze iz tablice 8.1. Uz značajke ASTA navedene u tablici 4.7 u analizu je još uključena značajka srednje vrijednosti srčanog ritma budući da informacija o tome nije dostupna iz samih značajki ASTA, a mogla bi biti značajna za razvrstavanje aritmija. Ukupan broj značajki u ovoj analizi je dakle 14 (*f*1-*f*9, *g*2-*g*5, *mean*). Značajke su ispitivane na dva skupa ritmova. Prvi skup od 9 ritmova je onaj dobiven sustavnim postupkom navedenim u poglavlju 8.1.1. Taj skup (skup #1) uključuje ritmove: NSR, PAC, PVC, AF, VBI, VTR, ABI, ATR i PEJS. Drugi skup od također 9 ritmova je onaj za koji su značajke ASTA izvorno definirane, vidjeti poglavlje 4.2. Taj skup uključuje ritmove: NSR, PAC, PVC, Kuplet, VFL, AF, BII, VBI i PEJS (skup #2). Oba skupa ispitivana su korištenjem algoritma slučajnih šuma s 30 stabala, unakrsnom validacijom s 10 preklopa i 3 ponavljanja zbog stohastičnosti algoritma slučajnih šuma.

Na slici 8.18 prikazani su rezultati za skup #1 za vremenske segmente različite duljine. Pokazuje se da se najbolji rezultat dobiva za segment duljine 20 s. Rezultati ostalih mjera vrednovanja za najbolju duljinu segmenta od 20 s prikazani su u tablici 8.26.

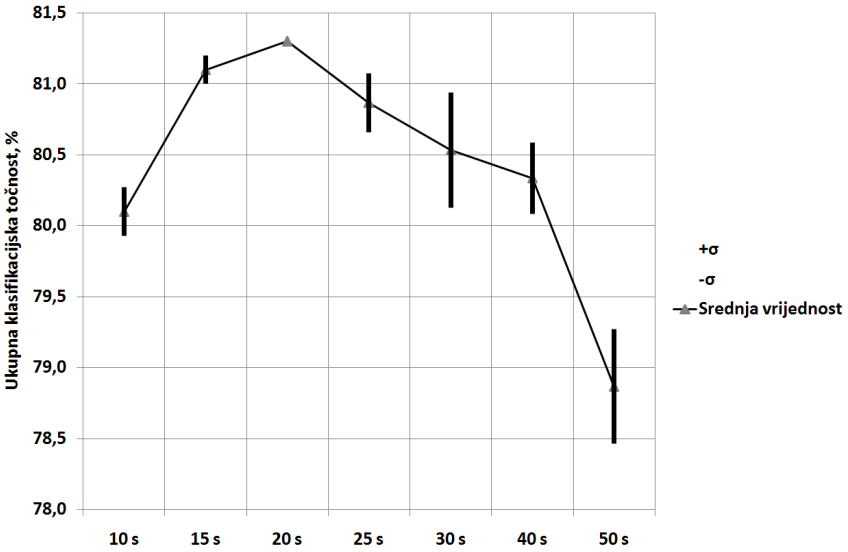
Na slici 8.19 prikazani su rezultati za skup #2 za vremenske segmente različite duljine.



Slika 8.18. Razvrstavanje ritmova iz skupa #1 korištenjem značajki ASTA.

Tablica 8.26. Osjetljivost, specifičnost i površina ispod krivulje pri razvrstavanju ritmova iz skupa #1 korištenjem značajki ASTA za duljinu segmenta od 20 s.

Ritam	SENS, %	SPEC, %	AUC, %
NSR	93.8	95.0	97.6
PAC	69.5	95.0	94.3
PVC	74.1	94.5	94.5
AF	78.4	96.9	96.4
VBI	63.9	98.5	95.3
VTR	53.8	99.1	94.1
ABI	52.6	99.1	94.6
ATR	33.5	99.3	91.9
PEJS	83.0	99.3	99.2



Slika 8.19. Razvrstavanje ritmova iz skupa #2 korištenjem značajki ASTA.

Ponovno je uočljivo da ASTA daje najbolje rezultate za duljinu segmenta od 20 s. U slučaju skupa #2 rezultati su nešto bolji, budući da su izvorno značajke ASTA i definirane za taj skup ritmova. Detaljni rezultati za najbolju duljinu segmenta od 20 s prikazani su u tablici 8.27.

U okviru analize značajki ASTA provedeno je razvrstavanje istovremenim korištenjem segmenata različitih vremenskih duljina. Tako se koriste vektori značajki na duljinama od 10-50s, s ukupnim brojem od 58 663 vektora značajki. Ova analiza je provedena kako bi se ustanovilo postoji li mogućnost poboljšanja razvrstavanja poremećaja ako bi se istovremeno promatralo više duljina segmenata. Uistinu se pokazuje da bi u tom slučaju ukupna točnost razvrstavanja bila nešto povećana, oko 1.7%. Rezultati su dani u tablici 8.28. Uočava se da su rezultati za poremećaje PVC, VBI, VTR i ABI bitno poboljšani (usporediti s tablicom 8.26).

Ovi rezultati ukazuju na to da bi se izlučivanjem značajki ASTA na više duljina segmenata različite duljine povećala točnost ukupnog modela. Preduvjet ovdje postignute točnosti je u tome da su dostupni prethodni podaci o dotičnom pacijentu kao i o tome da se značajke izlučuje na temelju prethodnih 50 sekundi zapisa.

Poredak značajki ASTA po važnosti za razvrstavanje srčanih ritmova iz skupa #1 dan je u tablici 8.29 za tri filterska postupka (IG, 1Rule, SU). Pokazuje se da je značajka $f7$ najbitnija, a slijede ju $f9$, $f2$, $g2$, $f4$ i $f5$. Značajka srednje vrijednosti srčanog ritma za

Tablica 8.27. Osjetljivost, specifičnost i površina ispod krivulje pri razvrstavanju ritmova iz skupa #2 korištenjem značajki ASTA za duljinu segmenta od 20 s.

Ritam	SENS, %	SPEC, %	AUC, %
NSR	93.8	94.6	98.4
PAC	68.6	94.9	96.8
PVC	72.1	94.1	95.9
AF	48.2	97.5	93.4
VBI	42.9	100.0	92.8
VTR	76.7	97.3	98.2
ABI	82.1	100.0	99.9
ATR	66.5	99.0	96.0
PEJS	83.6	99.2	99.2

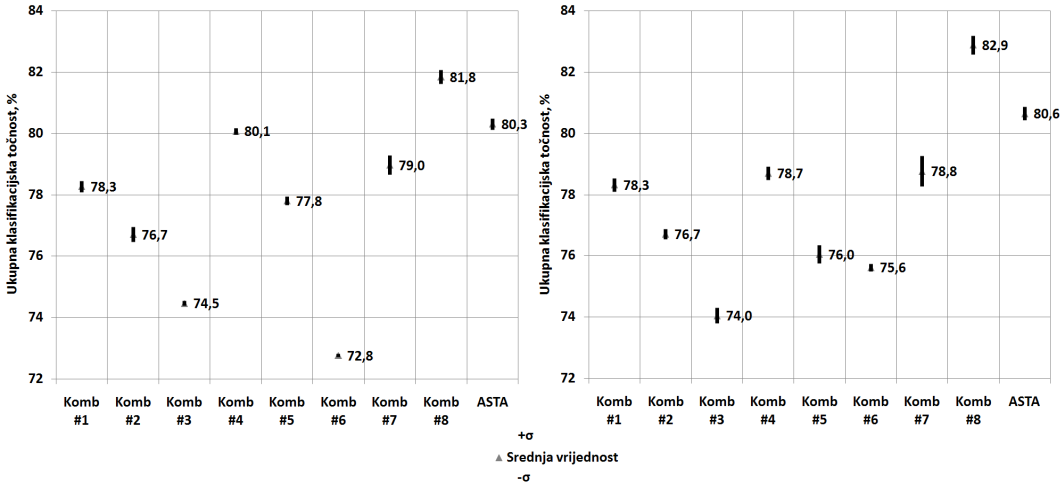
Tablica 8.28. Razvrstavanje ritmova korištenjem informacija dobivenih s više različitih duljina segmenata (10, 15, 20, 25, 30, 40 i 50 s) značajkama ASTA. Rezultati su navedeni za skup #1.

Ritam	SENS, %	SPEC, %	AUC, %	ACC, %
NSR	94.0	94.6	97.9	
PAC	72.2	95.8	95.1	
PVC	76.6	95.4	95.2	
AF	78.3	96.8	96.7	
VBI	67.6	98.6	96.9	82.4±0.1
VTR	58.6	99.2	95.8	
ABI	53.4	99.3	96.0	
ATR	50.7	99.4	96.1	
PEJS	78.9	99.2	98.6	

Tablica 8.29. Poredak značajki ASTA po važnosti za razvrstavanje skupa #1 za duljinu segmenta od 20 s pri korištenju triju filterskih postupaka. Navedeni su i rezultati filterske metrike za svaki atribut.

Poredak	IG	1Rule	SU
1.	0.72 ASTA_f7	63.94 ASTA_f7	0.31 ASTA_f7
2.	0.67 ASTA_f2	62.76 ASTA_f9	0.30 ASTA_f9
3.	0.62 ASTA_f9	59.68 ASTA_g2	0.29 ASTA_f2
4.	0.55 ASTA_f4	58.57 ASTA_f2	0.23 ASTA_f4
5.	0.42 ASTA_f6	58.08 ASTA_f4	0.18 ASTA_f5
6.	0.41 ASTA_f5	56.43 ASTA_g3	0.15 ASTA_g2
7.	0.38 ASTA_f8	54.73 ASTA_f1	0.15 ASTA_g5
8.	0.35 ASTA_f1	54.70 ASTA_f5	0.15 ASTA_f6
9.	0.25 ASTA_g5	53.42 ASTA_g4	0.14 ASTA_f8
10.	0.25 ASTA_g2	53.33 ASTA_f3	0.14 ASTA_g3
11.	0.24 ASTA_g3	53.30 ASTA_g5	0.13 ASTA_f1
12.	0.20 ASTA_f3	52.30 ASTA_f6	0.10 ASTA_g4
13.	0.15 ASTA_g4	51.87 ASTA_f8	0.08 ASTA_f3

sve tri filterske mjere bila je među najslabijim značajkama u usporedbi sa značajkama ASTA te se ne navodi u tablici 8.29. Usporedba kombinacije značajki ASTA s ostalim kombinacijama značajki navedenima u tablici 8.12 dana je na slici 8.20. Na toj slici prikazana je ukupna klasifikacijska točnost razvrstavanja na problemu 9 ritmova iz skupa #1. Pokazuje se da su 13 značajki ASTA-e zajedno s informacijom o srednjoj vrijednosti ritma uspješnije od svih dosad predloženih kombinacija značajki u literaturi, a lošije su jedino od kombinacije 69 značajki (30% ukupnog skupa značajki, komb. #8) dobivenih sustavnim postupkom. Potrebno je istaknuti da značajke ASTA čak nisu ni izvorno definirane za skup poremećaja #1, već za skup #2, što ukazuje na visok potencijal ovih značajki za mnoge probleme pri razvrstavanju srčanih ritmova kao i na valjanost pristupa promatranju poremećaja BVN na temelju faznog prostora drugog reda razlike.



Slika 8.20. Usporedba kombinacije značajki ASTA(+mean) s ostalim kombinacijama značajki na problemu razvrstavanja ritmova iz skupa #1.

8.3 Vrednovanje abecedne entropije

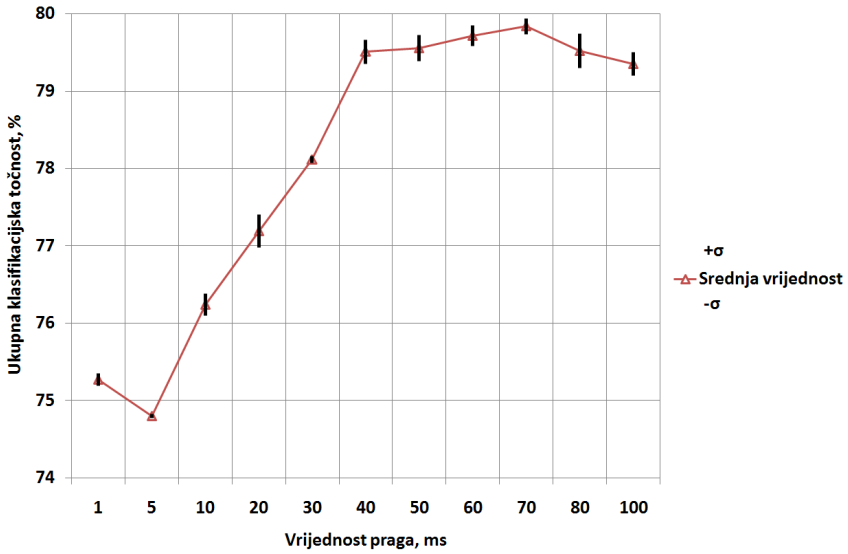
Značajke abecedne entropije (AbEn) izlučene sustavnim postupkom navedene su u tablici 8.30 zajedno s objašnjenjem i kratkim nazivom koji se koristi pri analizi. Ukupno je izlučeno 138 značajki AbEn. Abecedna entropija vrednovana je na skupu podataka iz poglavlja 8.1.1 kao i sustavni postupak i značajke ASTA. Budući da rezultati AbEn izravno ovise o parametru praga koji određuje je li došlo do promjene u signalu ili ne, najprije je potrebno ispitati koja je vrijednost praga najbolja za razvrstavanje.

Ovdje se pravi pretpostavka da ako je rezultat za neku vrijednost praga X bolji nego za vrijednost praga Y za neku duljinu segmenta Z, tada će vrijednost praga X biti bolja od vrijednosti praga Y i za većinu duljina segmenata $W \neq Z$. Analizu AbEn provodi se za skup ritmova dobivenih sustavnim postupkom kao značajne za razvrstavanje, skup #1 (9 ritmova). Za izradu klasifikatora koristi se algoritam slučajne šume s 80 stabala, a rezultati se vrednuju unakrsnom validacijom s 10 preklopa i 3 ponavljanja zbog stohastičnosti algoritma. Na slici 8.21. navodi se ukupna točnost razvrstavanja dobivena s AbEn za skup poremećaja #1 u ovisnosti o vrijednosti praga.

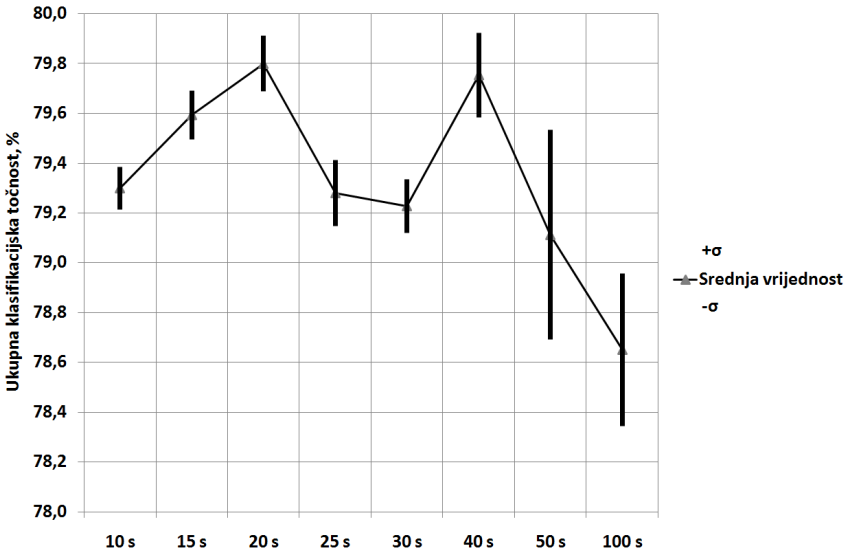
Na temelju slike 8.21 razvidno je da su rezultati dosta ovisni o vrijednosti praga. Očito, dokle god je prag manji od 40 ms sustav uči nepotreban šum koji postoji pri malim i nebitnim promjenama srčanog ritma. Ovo također znači da promjene srčanog ritma manje od nekih 40 ms nisu klinički značajne. Ipak, statistički najbolje je postaviti prag na nekih 60 ili 70 ms. Postavljanjem praga na više od 80 ms dolazi do postepenog pada točnosti, budući da je AbEn postala neosjetljiva na klinički značajne promjene u ritmu. Za 70 ms dobivaju se najbolji rezultati pri razvrstavanju skupa #1. U daljnjoj analizi detaljnije se razmatra prag od 70 ms. Na slici 8.22 dan je rezultat razvrstavanja poremećaja korištenjem praga od 70 ms za sve ispitivane duljine segmenta.

Tablica 8.30. Izlučene značajke abecedne entropije.

Značajka/e	Objašnjenje	Kratak naziv	Broj takvih značajki
Prosječna abecedna entropija	Prosjek AbEn po svim znakovima u segmentu	Aver_AbEn	1
Varijanca abecedne entropije	Varijanca AbEn po svim znakovima u segmentu	AbEn_Var	1
Najveća abecedna entropija	Najveća vrijednost AbEn u segmentu	Max_AbEn	1
Prosječna abecedna entropija po znaku	Prosjek AbEn za svaki znak *X* u segmentu	*X*_aver_en	27
Varijanca abecedne entropije po znaku	Varijanca AbEn za svaki znak *X* u segmentu	*X*_en_var	27
Najveća abecedna entropija po znaku	Najveća vrijednost AbEn za svaki znak *X* u segmentu	*X*_max_en	27
Postojanje znaka	Postoji li znak *X* u segmentu	*X*_exists	27
Udio znaka	Udio pojavljivanja znaka *X* u segmentu u odnosu na sve znakove	*X*_rate	27



Slika 8.21. Rezultati abecedne entropije za skup ritmova #1 u ovisnosti o vrijednosti praga za duljinu segmenta od 20 s.



Slika 8.22. Ovisnost rezultata abecedne entropije o duljini segmenta za iznos praga od 70 ms.

Pokazuje se da najbolji rezultat AbEn postiže za duljinu segmenta od 20 s, iako je vrlo blizak rezultat dobiven i za 40 s. Ono što je značajno zamijetiti to je da ukupna točnost razvrstavanja vrlo malo ovisi o duljini segmenta (primijetiti skalu na osi y na slici 8.22) te ne opada bitno sve do duljine segmenta od 100 s. Slaba ovisnost o duljini segmenta je bitna karakteristika abecedne entropije koju ne iskazuju značajke ASTA pa čak niti ukupni skup značajki analiziran sustavnim postupkom. U tablici 8.31 prikazani su rezultati AbEn na skupu #1 za ostale mjere vrednovanja za duljinu segmenta od 20 s.

Tablica 8.31. Osjetljivost, specifičnost i površina ispod krivulje (u %) pri razvrstavanju ritmova iz skupa #1 korištenjem značajki AbEn za duljinu segmenta od 20 s.

Ritam	SENS	SPEC	AUC
NSR	92.9	93.1	97.0
PAC	70.5	95.7	95.3
PVC	73.5	94.5	93.9
AF	81.3	96.1	97.1
VBI	64.5	98.4	96.3
VTR	62.8	99.1	97.2
ABI	47.1	99.2	97.4
ATR	46.1	99.5	97.7
PEJS	51.9	99.1	95.8

U okviru analize značajki AbEn provedeno je razvrstavanje istovremenim korištenjem segmenata različitih vremenskih duljina. Tako se koriste vektori značajki na duljinama od 10-50 s, s ukupnim brojem od 59 063 vektora značajki. Ova analiza je provedena kako bi se ustanovilo postoji li mogućnost poboljšanja razvrstavanja poremećaja ako bi se istovremeno promatralo više duljina segmenata. Korišten je prag od 70 ms. Uistinu se pokazuje da je u tom slučaju točnost razvrstavanja bitno povećana i doseže visokih 87.6%. Rezultati su dani u tablici 8.32.

Poredak značajki AbEn po važnosti za klasifikaciju dan je u tablici 8.33 za tri filterska postupka (IG, 1Rule, SU). Prikazano je samo prvih 15 najbitnijih značajki. Pokazuje se da su značajke X_rate, X_max_en, AverageAbEn i MaximumAbEn najbitnije. Značajka X_rate opisuje koliki udio u ukupnom broju RR intervala unutar nekog segmenta ima promjena "-+-", dakle obrazac: skraćenje intervala, produljenje intervala, skraćenje intervala; što je specifičnost većine poremećaja, a posebno za PAC i PVC. Najviša postignuta entropija takvog znaka unutar segmenta mogla bi razlikovati PAC od PVC. Prosječna i najveća entropija su očito značajne budući da dobro opisuju varijabilnost vrijednosti unutar segmenta. Prosječna entropija je pritom vjerojatno bolja za otkrivanje radi li se o normalnom segmentu, dok najveća entropija vjerojatno razlikuje između drugih obrazaca ritma (npr. AF može imate velike amplitude promjena ritma, kao i PVC). Od ostalih znakova abecede, značajni su H (0-+), C (00-), P(+-0), Q(+-+), A(000) i S(-00).

Tablica 8.32. Razvrstavanje ritmova korištenjem informacija dobivenih s više različitih duljina segmenata (10, 15, 20, 25, 30, 40 i 50 s) značajkama AbEn za skup #1 s pragom 70 ms.

Ritam	SENS, %	SPEC, %	AUC, %	ACC, %
NSR	94.5	94.7	97.8	
PAC	82.7	97.5	97.7	
PVC	84.1	97.0	96.5	
AF	84.8	97.9	98.2	
VBI	82.0	99.0	98.8	87.6±0.1
VTR	78.1	99.4	99.0	
ABI	72.2	99.6	99.3	
ATR	74.9	99.7	99.4	
PEJS	67.2	99.2	94.9	

Tablica 8.33. Poredak prvih 15 značajki AbEn po važnosti za razvrstavanje skupa #1 za duljinu segmenta od 20 s i praga od 70 ms pri korištenju triju filterskih postupaka. Navedene su relativne važnosti atributa za filterski postupak.

Poredak	IG	1Rule	SU
1.	0.75 Aver_AbEn	62.06 X_rate	0.32 X_exists
2.	0.75 X_rate	60.62 X_max_en	0.32 X_rate
3.	0.75 Max_AbEn	60.11 X_aver_en	0.30 Max_AbEn
4.	0.74 A_rate	59.23 H_aver_en	0.30 Aver_AbEn
5.	0.66 X_max_en	59.23 H_max_en	0.30 X_max_en
6.	0.62 AbEn_Var	59.12 Max_AbEn	0.29 X_aver_en
7.	0.62 X_aver_en	58.63 P_max_en	0.29 A_rate
8.	0.55 X_exists	58.38 P_aver_en	0.26 AbEn_Var
9.	0.48 H_rate	58.35 H_rate	0.25 H_exists
10.	0.47 H_max_en	58.34 X_exists	0.24 Q_rate
11.	0.45 H_aver_en	58.13 C_aver_en	0.22 H_max_en
12.	0.44 Q_rate	57.90 C_max_en	0.22 H_aver_en
13.	0.42 H_exists	57.45 H_exists	0.22 Q_exists
14.	0.41 C_max_en	57.00 P_rate	0.21 H_rate
15.	0.41 C_aver_en	57.00 S_max_en	0.21 Q_max_en

Ako se usporede rezultati dobiveni za značajke ASTA i AbEn na slikama 8.18 i 8.22 može se uočiti da je točnost razvrstavanja bolja za ASTA-u za duljine segmenta od 10, 15, 20 i 25 s, dok je AbEn točnija za veće duljine segmenata. Ipak, ukupno je kombinacija ASTA nešto točnija za najbolju duljinu segmenta od 20 s na skupu #1 i to za oko 0.5%. Ako se uzme u obzir to da kombinacija ASTA (+mean) broji samo 14 značajki, dok AbEn (+mean) koristi čak 138 značajki, onda je jasno da bi se u razvrstavanju aritmija trebalo preferirati koristiti kombinaciju ASTA. Međutim, puni potencijal abecedne entropije dolazi do izražaja kad se koristi razvrstavanje ritmova korištenjem informacije dobivenih s više različitih duljina segmenata. U tom slučaju AbEn prednjači za preko 5% u odnosu na značajke ASTA (usporediti tablice 8.28 i 8.32). AbEn je također u tom slučaju tek nešto slabija od kombinacije najboljih 30% značajki ukupnog skupa značajki (87.6% prema 88.3%), uz ogradu da 30% značajki broji njih 69, dok je značajki AbEn ukupno 138. Nedostatak korištenja informacije s više duljina segmenata su resursi potrebni za izgradnju modela, koji su nekoliko puta veći u smislu memorijskih zahtjeva i vremena izvođenja u odnosu na korištenje informacije samo iz jedne duljine segmenta.

9 Zaključak

U okviru ove disertacije razvijen je računalni radni okvir koji omogućuje sveobuhvatno vrednovanje i usporedbu značajki unutar jednog područja analize biomedicinskih vremenskih nizova (BVN). Okvir se odlikuje implementacijom više od 200 značajki koje uključuju one specifične za analizu srčanog ritma kao i one koje se mogu koristiti za analizu ostalih BVN. Izgrađeni okvir je modularan, prenosiv i lako nadogradiv programski proizvod koji omogućuje: izravno čitanje podataka iz zapisa pacijenata spremljenih u tekstualnom obliku, izlučivanje značajki i pohranjivanje vektora značajki u oblik koristan za dubinsku analizu. Okvir podržava nadzirano učenje kao i izlučivanje značajki bez informacije o ciljnom razredu što je korisno za testiranje. Grafičko korisničko sučelje dozvoljava jednostavno učitavanje većeg broja zapisa i određivanje uvjeta pod kojim se provodi izlučivanje. U računalnom radnom okviru implementirana je većina značajki poznata u području analize srčanog ritma. Ipak, postoji određeni broj značajki koje u ovoj disertaciji nisu spomenute ili koje nisu implementirane u radnom okviru, npr. različite vrste značajki osnovnih valića osim Haarovog [Faust 2004], određene izvedene nenormirane linearne vremenske značajke [Pecchia 2011] i linearne frekvencijske značajke [Kiviniemi 2007]. Namjera radnog okvira je da bude sveobuhvatan i da omogući usporedbu rezultata između istraživača što znači da ga je nužno stalno nadograđivati. U okviru ovog sustavnog postupka, prikazan je samo princip kako bi radni okvir za pojedinu domenu trebao izgledati, dok se dorađivanje okvira za pojedine domene ostavlja za buduća istraživanja. U budućnosti bi se moglo prilagoditi radni okvir za izlučivanje značajke iz više različitih BVN kao i iz složenih zapisa s istovremenim mjerenjima više BVN.

U disertaciji su detaljno opisane sve značajke implementirane u radnom okviru, što uključuje vremenske, frekvencijske, vremensko-frekvencijske i nelinearne značajke. Posebna težnja dana je opisu velikog broja nelinearnih značajki korištenih u posljednje vrijeme u analizi BVN, uključujući značajke faznog prostora, fraktalne, entropijske i druge nelinearne značajke i to na jednoj ili više vremenskih skala.

Jedan od doprinosa disertacije je i razvoj dva nova postupka za izlučivanje značajki iz biomedicinskih vremenskih nizova: abecedne entropije (AbEn) i napredne analize slijednog trenda (ASTA). Uz teorijsku pozadinu i definiciju, u radu je pojašnjena primjena u konkretnoj domeni, s definicijom značajki i raspravom o potencijalnoj uspješnosti primjene. AbEn se temelji na kvalitativno-kvantitativnom modeliranju vremenskog niza, pri čemu se tri slijedne promjene u nizu kodiraju određenim znakom abecede. Stoga AbEn otkriva one poremećaje koji se očituju u kratkoj dinamici. Finija informacija dobiva se kvantificiranjem iznosa promjena za svaki znak abecede kao i za cijeli promatrani segment, s ciljem poboljšane uspješnosti razlikovanja između međusobno sličnih obrazaca u nizu. Pokazano je da se AbEn može uspješno primijeniti u analizi 9 obrazaca srčanog ritma. Značajke postižu točnost od oko 80% na duljini

segmenta od 20 s, a točnost vrlo malo varira s promjenom duljine segmenta. Pokazano je da je prag za priznavanje značajnosti promjene ritma od 40 ms kritičan za točno otkrivanje aritmija, dok se najbolji rezultati postižu za 70 ms. Najveća točnost dobivena korištenjem informacije s više vremenskih skala dosiže 87.6% neovisno o zapisu. Najboljim značajkama pokazale su se srednja entropija segmenta, najviša entropija segmenta te najviša entropija i udio znaka X (niza promjena "-+-") u segmentu.

ASTA se temelji na određivanju područja od interesa za svaki analizirani obrazac unutar faznog prostora 2. reda razlike. Pravilnim određivanjem svakog područja u obliku elipse ili pravokutnika kao i utvrđivanjem kratkih trajektorija ostvaruje se točan i računski nezahtjevan opis obrazaca. ASTA se pokazala usporedivom ili boljom kombinacijom značajki od svih ponuđenih u literaturi za razvrstavanje raznih vrsta srčanih ritmova. Postupak ima potencijal prilagodbe za nove probleme razvrstavanja, čak i one nevezane uz srčani ritam budući da se značajke definiraju u ovisnosti o konkretnim podacima o poremećajima. Jedini nedostatak algoritma ASTA na kojem bi se trebalo poraditi je taj što ne postoji matematički egzaktan postupak za definiranje najboljih potpodručja za pojedini poremećaj, već područja i trajektorije definira istraživač na temelju dostupnih podataka. Najbolja postignuta točnost značajki ASTA na skupu od 9 srčanih ritmova prelazi 80% za duljinu segmenta od 20 s.

Analiza provedena u ovoj disertaciji pokazala je korisnost predloženog sustavnog postupka u analizi poremećaja srčanog ritma i srčanih oboljenja na temelju značajki niza srčanih otkucaja, kao jednog od često analiziranih biomedicinskih vremenskih nizova. Sustavni postupak razvijen je na način najbolje prakse, bez strogih definicija, budući da je područje koje obuhvaća vrlo široko i uključuje raznovrsne BVN. Postupak je ograničen na nadzirano učenje i izgradnju modela poremećaja razvrstavanjem, a ne uključuje druge moguće postupke analize BVN kao što je primjerice nenadzirano učenje.

Korištenjem računalnog radnog okvira unutar sustavnog postupka omogućena je po prvi puta sveobuhvatna direktna usporedba kvalitete različitih značajki varijabilnosti na jednom problemu, što je bitan doprinos disertacije.

Vrednovanje sustavnog postupka provedeno je na dva problema razvrstavanja. Prvi je uključivao razne obrasce srčanog ritma, njih ukupno 14, a u drugom je bilo potrebno razlikovati između zdrave osobe, pacijenta koji ima neku vrstu aritmije i pacijenta koji boluje od kongestivnog zatajenja srca.

Na prvom problemu najprije se uklonilo pet poremećaja koji nisu zadovoljavali kriterije postavljene na točnost modela. Zatim je otkriveno da je moguće smanjiti skup atributa s početnih 230 na njih 69 (30%). Pritom se filterska metrika IG pokazala najuspješnijom u odnosu na metrike SU i 1Rule za gotove sve duljine segmenata. Konačno, provelo se razvrstavanje korištenjem algoritama slučajnih šuma, C4.5, AdaBoost+C4.5, RIPPER, stroja s potpornim vektorima s polinomnom i radijalnom jezgrom za sve duljine segmenata na smanjenom skupu atributa. Postignuta je najbolja točnost od 85.7% algoritmom AdaBoost+C4.5 za duljinu segmenta od 10 s u slučaju

izgradnje modela neovisno o zapisu pacijenta i najbolja točnost od 73.2% algoritmom slučajne šume za duljinu segmenta od 15 s u slučaju ovisno o zapisu. Slučajna šuma pokazala se ukupno najboljim algoritmom uzevši u obzir točnost i brzinu izgradnje modela. Pokazano je da kombinacija koja se sastoji od 30% najboljih atributa postiže točnije rezultate od svih ostalih navedenih u literaturi. Kombinacija sadrži dosta linearnih vremenskih značajki i manji broj nelinearnih značajki, a uključuje i dosta značajki AbEn i ASTA. Najviša osjetljivost pri razvrstavanju obrazaca postignuta je za ritmove NSR, PVC, PAC, AF i PEJS, dok je manja osjetljivost za poremećaje VBI, VTR, ABI i ATR. Razumljivi modeli izgrađeni algoritmima C4.5 i RIPPER postigli su točnost od 69.4% (C4.5) i 73.4% (RIPPER) za slučaj neovisno o zapisu te točnost od 62.9% (C4.5) i 66.0% (RIPPER) za slučaj ovisno o zapisu. Dobiveni modeli mogli bi pomoći liječnicima pri razmatranju o kojoj se aritmiji radi samo na temelju srčanog ritma. Točnost je poboljšana na vrijednost od 88.3% u slučaju razvrstavanja na temelju više vremenskih skala (10-50 s), ali samo u slučaju modela izgrađenih neovisno o zapisu pacijenta.

Na drugom problemu razvrstavanja otkriveno je da se skup atributa može smanjiti na samo 20% (47 atributa) izvornog skupa bez značajnog smanjenja točnosti. Daljnjim optimiranjem korištenjem mjere prekrivanja InCons, skup atributa je uspješno smanjen za četiri od šest duljina segmenata. Razvrstavanje zapisa provedeno je istim postupcima kao i za prvi problem. Najveća točnost postignuta je za duljinu segmenta od 300 s algoritmom AdaBoost+C4.5 od 92.7% u slučaju modela izgrađenog neovisno o zapisu, a isti algoritam dao je točnost od 84.7% za duljinu od 1500 s u slučaju modela izgrađenog ovisno o zapisu što podupire tezu iz literature da je bolje koristiti veće duljine segmenta za otkrivanje razlika između zdrave osobe i osobe oboljele od CHF. Razumljivi modeli izgrađeni algoritmima C4.5 i RIPPER postigli su točnost od 80.2% (C4.5) i 82.1% (RIPPER) za slučaj neovisno o zapisu te točnost od 78.9% (C4.5) i 77.9% (RIPPER) za slučaj ovisno o zapisu.

Jedan od bitnih zaključaka disertacije je taj da je za točnije modele poremećaja koji se rjeđe pojavljuju u zapisima (ABI, ATR, VTR, PEJS) potrebno još podataka. I dok su rezultati za NSR, PAC i PVC približno stabilni pri izgradnji modela ovisno o zapisu pacijenta, osjetljivost otkrivanja ostalih poremećaja bitno pada, što je očiti znak nedostatka referentnih podataka. U tom kontekstu potrebno je promatrati sve rezultate izgrađene neovisno o zapisu pacijenta kao najbolje rezultate koji se mogu postići korištenjem samo analize srčanog ritma u razlikovanju aritmija i srčanih oboljenja.

Iz funkcionalno-implementacijske perspektive promatrano, cjelokupni sustav bi se dao nadograditi integracijom radnog okvira i alata za dubinsku analizu podataka. Nadalje, snimanjem signala srčanog ritma uživo, korištenjem tehnika *streaminga* i predobrade podataka te automatizacijom izgradnje modela korištenjem ovdje predloženog sustavnog postupka u budućnosti bi se mogao izgraditi *on-line* računalni sustav za otkrivanje aritmija i srčanih oboljenja na temelju srčanog ritma koji bi garantirao visoku točnost i ponovljivost rezultata.

Literatura

[Abarbanel 1991] H. D. I. Abarbanel, R. Brown, M. B. Kennel, Lyapunov exponents in chaotic systems: their importance and their evaluation using observed data, Int. J. Mod. Phys. B 5(9):1347–1375, 1991.

[Abe 2003] S. Abe, Analysis of Multiclass Support Vector Machines, in: Proc. Int. Conf. Computational Intelligence for Modelling Control and Automation CIMCA'2003, Vienna, Austria, Feb 12-14, 2003, pp. 385–396, 2003.

[Acharya 2002] R. U. Acharya, C. M. Lim, P. Joseph, Heart rate variability analysis using correlation dimension and detrended fluctuation analysis, ITBM-RBM 23(6):333–339, 2002.

[Acharya 2003] R. U. Acharya, P. S. Bhat, S. S. Iyengar, A. Raod, S. Dua, Classification of heart rate data using artificial neural network and fuzzy equivalence relation, Pattern Recognition 36:61–68, 2003.

[Acharya 2004 (1)] R. U. Acharya, A. Kumar, P. S. Bhat. C. M. Lim, S. S. lyengar, N. Kannathal, S. M. Krishnan, Classification of cardiac abnormalities using heart rate signals, Med. Biol. Eng. Comput. 42:288–293, 2004.

[Acharya 2004 (2)] R. U. Acharya, N. Kannathal, O. W. Sing, L. Y. Ping, T. Chua, Heart rate analysis in normal subjects of various age groups, BioMed. Eng. OnLine 3:24, 2004.

[Acharya 2006] R. U. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, J. S. Suri, Heart rate variability: a review, Med. Biol. Eng. Comput. 44:1031–1051, 2006.

[Adnane 2009] M. Adnane, Z. Jiang, S. Choi, Development of QRS detection algorithm designed for wearable cardiorespiratory system, Comp. Meth. Prog. Biomed. 93:20–31, 2009.

[Allan 1966] D. W. Allan, Statistics of atomic frequency standards, Proceedings of the IEEE 54:221–230, 1966.

[Almuallim 1994] H. Almuallim, T. G. Dietterich, Learning boolean concepts in the presence of many irrelevant features, Artif. Intell. 69(1–2):279–305, 1994.

[Alpaydin 2004] E. Alpaydin, Introduction to machine learning, Cambridge MA, USA: MIT Press, 2004.

[Alvarez 2007] D. Alvarez, R. Hornero, M. Garcia, F. del Campo, C. Zamarron, Improving diagnostic ability of blood oxygen saturation from overnight pulse oximetry in obstructive sleep apnea detection by means of central tendency measure, Artif. Intell. Med. 41(1):13–24, 2007.

[Angelini 2007] L. Angelini, R. Maestri, D. Marinazzo, L. Nitti, M. Pellicoro, G. D. Pinna, S. Stramaglia, S. A. Tupputi, Multiscale analysis of short term heart beat interval, arterial blood pressure, and instantaneous lung volume time series, Artif. Intell. Med. 41(3):237–250, 2007.

[Asl 2008] B. M. Asl, S. K. Setarehdan, M. Mohebbi, Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal, Artif. Intell. Med. 44(1):51–64, 2008.

[Baumert 2005] M. Baumert, V. Baier, S. Truebner, A. Schirdewan, A. Voss, Short- and Long-Term Joint Symbolic Dynamics of Heart Rate and Blood Pressure in Dilated Cardiomyopathy, IEEE Trans. Biomed. Eng. 52(12):2112–2115, 2005.

[Bezerianos 1999] A. Bezerianos, S. Papadimitriou, D. Alexopoulos, Radial basis function neural networks for the characterization of heart rate variability dynamics, Artif. Intell. Med. 15(3):215–234, 1999.

[Boardman 2002] A. Boardman, F. S. Schlindwein, A. P. Rocha, A. Leite, A study on the optimum order of autoregressive models for heart rate variability, Physiol. Meas. 23(2):325–336, 2002.

[Bogunović 2008] N. Bogunović, T. Šmuc, Applicability of Qualitative ECG Processing to Wearable Computing, in: Y.-T. Zhang, ed., Proc. 5th Int. Summer School and Symposium on Medical Devices and Biosensors ISSS-MDBS 2008, Hong Kong, China, Jun 1-3, 2008, Hong Kong, China: IEEE Press, pp. 133–136, 2008.

[Bogunović 2010] N. Bogunović, A. Jović, Processing and Analysis of Biomedical Nonlinear Signals by Data Mining Methods, in: F. R. Leta, A. Conci, eds., Proc. 17th Int. Conf. Systems, Signals and Image Processing IWSSIP 2010, Rio de Janeiro, Brazil, Jun 17-19, 2010, Rio de Janeiro, Brasil: EdUFF Editora da Universidade Federal Fluminense, pp. 276–279, 2010.

[Bos 2002] R. Bos, S. de Waele, P. M. T. Broersen, Autoregressive Spectral Estimation by Application of the Burg Algorithm to Irregularly Sampled Data, IEEE Trans. Instrum. Meas. 51(6):1289–1294, 2002.

[Bouckaert 2003] R. R. Bouckaert, Choosing between two learning algorithms based on calibrated tests, in: T. Fawcett, N. Mishra, eds., Proc. 20th Int. Conf. Mach. Learn., Washington DC, USA, Aug 21-24, 2003, Menlo Park CA, USA: AAAI Press, pp. 51–58.

[Breiman 1984] L. Breiman, Classification and regression trees, Belmont CA, USA: Wadsworth International Group, 1984.

[Breiman 2001] L. Breiman, Random forests, Mach. Learn. 45:5–32, 2001.

[Burges 1998] C. Burges, A tutorial on support vector machines for pattern recognition, Knowledge Discovery and Data Mining 2(2):121–167, 1998.

[Cerutti 2007] S. Cerutti, F. Esposti, M. Ferrario, R. Sassi, M. G. Signorini, Long-term invariant parameters obtained from 24-h Holter recordings: a comparison between different analysis techniques, Chaos 17(1):015108, 2007.

[Ceylan 2009] R. Ceylan, Y. Özbay, B. Karlik, A novel approach for classification of ECG arrhythmias: Type-2 fuzzy clustering neural network, Expert Systems with Applications 36(3)(2):6761–6726, 2009.

[Chattipakorn 2007] N. Chattipakorn, T. Incharoen, N. Kanlop, S. Chattipakorn, Heart rate variability in myocardial infarction and heart failure, International Journal of Cardiology 120:289–296, 2007.

[Chawla 2002] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, JAIR 16:321–357, 2002.

[Chen X. 2008] X. Chen, M. Wasikowski, FAST: A ROC-based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems, Proc. ACM SIGKDD'08, Las Vegas NE, USA: Aug 24-27, 2008, pp. 124–133.

[Chen Z. 2002] Z. Chen, P. C. Ivanov, K. Hu, H. E. Stanley, Effects of nonstationarities on detrended fluctuation analysis, Phys. Rev. E 65:041107, 2002.

[Chua 2008] K. C. Chua, V. Chandran, R. U. Acharya, C. M. Lim, Computer-based analysis of cardiac state using entropies, recurrence plots and Poincare geometry, J. Med. Eng. Technol. 32(4):263–272, 2008.

[Chua 2009] K. C. Chua, V. Chandran, R. U. Acharya, L. C. Min, Cardiac Health Diagnosis Using Higher Order Spectra and Support Vector Machine, The Open Medical Informatics Journal 3:1–8, 2009.

[Cios 2002] K. J. Cios, G. W. Moore, Uniqueness of medical data mining, Artif. Intell. Med. 26(1–2):1–24, 2002.

[Clariá 2008] F. Clariá, M. Vallverdú, R. Baranowski, L. Chojnowska, P. Caminal, Heart rate variability analysis based on time-frequency representation and entropies in hypertrophic cardiomyopathy patients, Physiol. Meas. 29(3):401–416, 2008.

[Clifford 2006] G. D. Clifford, F. Azuaje, P. E. McSharry, Advanced Methods and Tools for ECG Data Analysis, Norwood MA, USA: Artech House, 2006.

[Cohen M. 1994] M. E. Cohen, D. L. Hudson, P. C. Deedwania, Measurement of variability in Holter Tape R-R intervals for patients with congestive heart failure, in: N. F. Sheppard, M. Eden, G. Kantor, eds., Proc. 16th Ann. Int. Conf. IEEE EMBS, Baltimore, MD, USA, Nov 3-6, 1994, IEEE Press, vol. 1, pp. 127–128.

[Cohen M. 1996] M. E. Cohen, D. L. Hudson, P. C. Deedwania, Applying continuous chaotic modeling to cardiac signal analysis, IEEE Eng. Med. Biol. Mag. 15(5):97–102, 1996.

[Cohen W. 1995] W. W. Cohen, Fast Effective Rule Induction, in: A. Prieditis, S. J. Russell, eds., Proc. 12th Int. Conf. Mach. Learn., Tahoe City CA, USA, Jul 9-12, 1995, San Francisco CA, USA: Morgan Kaufmann, pp. 115–123.

[Cooley 1965] J. W. Cooley, J. W. Tukey, An algorithm for the machine calculation of complex Fourier series, Math. Comput. 19:297–301, 1965.

[Costa 2002] M. Costa, A. L. Goldberger, C.-K. Peng, Multiscale Entropy Analysis of Complex Physiologic Time Series, Phys. Rev. Lett. 89(6):068102, 2002.

[Costa 2005 (1)] M. Costa, A. L. Goldberger, C.-K. Peng, Multiscale entropy analysis of biological signals, Phys. Rev. E 71:021906, 2005.

[Costa 2005 (2)] M. Costa, A. L. Goldberger, C.-K. Peng, Broken asymmetry of the human heartbeat: Loss of time irreversibility in aging and disease, Phys. Rev. Lett. 95:198102, 2005.

[Cysarz 2000] D. Cysarz, H. Bettermann, P. van Leeuwen, Entropies of short binary sequences in heart period dynamics, Am. J. Physiol. Heart Circ. Physiol. 278:H2163–H2172, 2000.

[Cysarz 2007] D. Cysarz, S. Lange, P. F. Matthiessen, P. van Leeuwen, Regular heartbeat dynamics are associated with cardiac health, Am. J. Physiol. Regul. Integr. Comp. Physiol. 292:R368–R372, 2007.

[Darrington 2008] J. M. Darrington, L. C. Hool, A new methodology for assessment of the performance of heartbeat classification systems, BMC Medical Informatics and Decision Making 8:7, 2008.

[de Carvalho 2002] J. L. A. de Carvalho, A. F. da Rocha, F. A. O. Nascimento, J. S. Neto, L. F. Junqueira, Development of Matlab Software for Analysis of Heart Rate Variability, in: Proc. 6th Int. Conf. Signal Processing ICSP 2002, Beijing, China, Aug 26-30, 2002, IEEE Press, pp. 1488–1491.

[de Carvalho 2003] J. L. A. de Carvalho, A. F. da Rocha, L. F. Junqueira, J. S. Neto, I. Santos, F. A. O Nascimento, A tool for time-frequency analysis of heart rate variability, Proc. 25th Ann. Int. Conf. IEEE EMBS 2003, Cancun, Mexico, Sep 17-21, 2003, vol. 3, pp. 2574–2577.

[de Lannoy 2008] G. de Lannoy, B. Frenay, M. Verleysen, J. Delbeke, Supervised ECG Delineation Using the Wavelet Transform and Hidden Markov Models, in: J. Vander Sloten, P. Verdonck, M. Nyssen, J. Haueisen, eds., Proc. 4th Eur. Conf. IFMBE, Antwerp, Belgium, Nov 23-27, 2008, vol. 22, part 1, pp. 22–25.

[de Waele 2000] S. de Waele, P. M. T. Broersen, The Burg Algorithm for Segments, IEEE Trans. Signal Process. 48(10):2876–2880, 2000.

[Dietterich 1998] T. G. Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Neural Computation 10:1895–1923, 1998.

[Ding 1993] M. Ding, E. Grebogi, E. Ott, T. Sauer, J. A. Yorke, Estimating correlation dimension from a chaotic time series: when a plateau occurs? Physica D 69:404–424, 1993.

[Dinh 1999] T. Dinh, H. Perrault, P. Calabrese, A. Eberhard, G. Benchetrit, New statistical method for detection and quantification of respiratory sinus arrhythmia, IEEE Trans. Biomed. Eng. 46:1161–1165, 1999.

[Eckmann 1987] J.-P. Eckmann, S. O. Kamphorst, D. Ruelle, Recurrence plots of dynamical systems, Europhys. Lett. 4:973–977, 1987.

[Exarchos 2007] T. P. Exarchos, M. G. Tsipouras, C. P. Exarchos, C. Papaloukas, D. I. Fotiadis, L. K. Michalis, A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree, Artif. Intell. Med. 40(3):187–200, 2007.

[Faust 2004] O. Faust, R. U. Acharya, S. M. Krishnan, L. C. Min, Analysis of cardiac signals using spatial filling index and time-frequency domain, BioMed. Eng. OnLine 3:30, 2004.

[Fawcett 2003] T. Fawcett, ROC Graphs: Notes and Practical Considerations for Researchers, in: HP Labs Tech Report, No. HPL-2003-4, 2003.

[Fayyad 1993] U. M. Fayyad, K. B. Irani, Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, in: R. Bajcsy, ed., Proc. 13th Int. Joint Conf. Artif. Intell., Chambéry, France, Aug 28 - Sep 3, 1993, pp. 1022–1027.

[Fayyad 1996] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Knowledge discovery and data mining: towards a unifying framework, in: E. Simoudis, J. Han, U. Fayyad, eds, Proc. 2nd Int. Conf. Knowledge Discovery & Data Mining KDD'96, Portland OR, USA, Aug 2-4, 1996, Menlo Park CA, USA: AAAI Press, pp. 82–88.

[Forman 2003] G. Forman, An Extensive Empirical Study of Feature Selection Metrics for Text Classification, JAIR 3:1289–1305, 2003.

[Fortune 1986] S. Fortune, A sweepline algorithm for Voronoi diagrams, Proc. 2nd Ann. ACM Symposium on Computational Geometry ACM SIGACT/SIGGRAPH, Yorktown Heights NY, USA, Jun 2-4, 1986, ACM Press, pp. 312–322.

[Frank 2000] E. Frank, Pruning Decision Trees and Lists, Doctoral thesis, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, 2000.

[Fraser 1989] A. M. Fraser, Information and entropy in strange attractors, IEEE Trans. Inf. Theory 35(2):245–262, 1989.

[Freund 1995] Y. Freund, R. E. Schapire, A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting, in: P. M. B. Vitányi, ed., Proc. 2nd Eur. Conf. Computational Learning Theory EuroCOLT '95, Barcelona, Spain, Mar 13-15, 1995, London, UK: Springer-Verlag, pp. 23–37.

[Fürnkranz 1999] J. Fürnkranz, Separate-and-conquer rule learning, Artif. Intell. Rev. 13:3–54, 1999.

[Gamberger 2003] D. Gamberger, N. Lavrač, G. Krstačić, Active subgroup mining: a case study in coronary heart disease risk group detection, Artif. Intell. Med. 28:27–57, 2003.

[Gamero 2002] L. G. Gamero, J. Vila, F. Palacios, Wavelet transform analysis of heart rate variability during myocardial ischaemia, Med. Biol. Eng. Comput. 40:72–78, 2002.

[Garcia 2001] T. B. Garcia, N. E. Holtz, 12-Lead ECG: The Art of Interpretation, Sudbury MA, USA: Jones and Bartlett Publishers, 2001.

[Ge 2002] D. Ge, N. Srinivasan, S. M Krishnan, Cardiac arrhythmia classification using autoregressive modeling, BioMed. Eng. OnLine 1:5, 2002.

[Glass 2001] L. Glass, Synchronization and rhythmic processes in physiology, Nature 410:277–284, 2001.

[Glass 2009] L. Glass, Introduction to Controversial Topics in Nonlinear Science: Is the Normal Heart Rate Chaotic? Chaos 19:028501, 2009.

[Gleick 1987] J. Gleick, Chaos: Making a New Science, New York, USA: Viking Penguin Inc., 1987.

[Goldberger 1988] A. L. Goldberger, D. R. Rigney, J. Mietus, E. M. Antman, S. Greenwald, Nonlinear dynamics in sudden cardiac death syndrome: heartrate oscillations and bifurcations, Experientia 44:983–987, 1988.

[Goldberger 1996] A. L. Goldberger, Non-linear dynamics for clinicians: chaos theory, fractals, and complexity at the bedside, Lancet 347:1312–1314, 1996.

[Goldberger 2000] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components od a New Research Resource for Complex Physiologic Signals, Circulation 101:e215-e220, 2000.

[Goldberger 2002] A. L. Goldberger, L. A. N. Amaral, J. M. Hausdorff, P. C. Ivanov, C.-K. Peng, H. E. Stanley, Fractal dynamics in physiology: Alterations with disease and aging, Proc. Natl. Acad. Sci. USA 99(Suppl 1):2466–2472, 2002.

[Golub 1999] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfeld, E. S. Lander, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science 286:531–537, 1999.

[Grassberger 1983] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D: Nonlinear Phenomena 9(1–2):189–208, 1983.

[Grimaldi 2003] M. Grimaldi, P. Cunningham, A. Kokaram, An Evaluation of Alternative Feature Selection Strategies and Ensemble Techniques for Classifying Music, in: Workshop on Multimedia Discovery and Mining, 14th Eur. Conf. Mach. Learn. ECML 2003, 7th Eur. Conf. Princip. & Pract. KDD, Dubrovnik, Croatia, Sep 22, 2003, vol. 2.

[Grond 2003] F. Grond, H. H. Diebner, S. Sahle, A. Mathias, S. Fischer, O. E. Rossler, A robust, locally interpretable algorithm for Lyapunov exponents, Chaos, Solitons & Fractals 16:841–852, 2003.

[Guldogan 2008] E. Guldogan, M. Gabbouj, Feature selection for content-based image retrieval, SIViP 2:241–250, 2008.

[Gutiérrez 2005] R. M. Gutiérrez, L. A. Sandoval, A Method to Separate Stochastic and Deterministic Information from Electrocardiograms, Phys. Scr. 2005(T118):132, 2005.

[Guyon 2002] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Mach. Learn. 46(1–3), 389–422, 2002.

[Hadamard 1898] J. Hadamard, Les surfaces à courbures opposées et leurs lignes géodésiques, J Math. Pure Appl., 1898.

[Hall 2000] M. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: P. Langley, ed., Proc. 17th Int. Conf. Mach. Learn. ICML2000, Stanford CA, USA, Jun 29 - Jul 2, 2000, San Francisco CA, USA: Morgan Kaufmann, pp. 359–366.

[Hall 2009] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1):10–18, 2009.

[Hand 2001] D. J. Hand, R. J. Tiller, A Simple Generalization of the Area under the ROC Curve to Multiple Class Classification Problems, Mach. Learn. 45(2):171–186, 2001.

[Harikrishnan 2009] K. P. Harikrishnan, R. Misra, G. Ambika, Efficient use of correlation entropy for analysing time series data, Pramana – Journal of Physics 72(2):325–333, 2009.

[Hastie 2009] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd ed., New York, USA: Springer Science+Business Media, LLC, 2009.

[Haykin 1998] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., New Jersey NY, USA: Prentice-Hall, 1998.

[Higuchi 1988] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, Physica D 31:277–283, 1988.

[Ho 1997] K. K. L. Ho, G. B. Moody, C.-K. Peng, J. E. Mietus, M. G. Larson, D. Levy, A. L. Goldberger, Predicting Survival in Heart Failure Case and Control Subjects by Use of Fully

Automated Methods for Deriving Nonlinear and Conventional Indices of Heart Rate Dynamics, Circulation 96:842–848, 1997.

[Holte 1993] R. C. Holte, Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, Mach. Learn. 11(1):63–90, 1993.

[Hornero 2006] R. Hornero, D. Abasolo, N. Jimeno, C. I. Sanchez, J. Poza, M. Aboy, Variability, Regularity, and Complexity of Time Series Generated by Schizophrenic Patients and Control Subjects, IEEE Trans. Biomed. Eng. 53(2):210–218, 2006.

[Hu 1997] Y. H. Hu, S. Palreddy, W. J. Tompkins, A Patient-Adaptable ECG Beat Classifier Using Mixture of Experts Approach, IEEE Trans. Biomed. Eng. 44(9):891–900, 1997.

[Huang 1998] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, H. H. Liu, The empirical mode decomposition and hilbert spectrum for nonlinear and nonstationary time series analysis, Proceedings of the Royal Society A 454(1971):903–995, 1998.

[Huikuri 2000] H. V. Huikuri, T. H. Makikallio, C. K. Peng, A. L. Goldberger, U. Hintze, M. Moller, Fractal correlation properties of R–R interval dynamics and mortality in patients with depressed left ventricular function after an acute myocardial infraction, Circulation 101:47–53, 2000.

[Hurst 1951] H. E. Hurst, Long-term storage capacity of reservoirs, Trans. Am. Soc. Civ. Eng. 116:770–808, 1951.

[Hutchinson 2003] T. P. Hutchinson, Statistics and graphs for heart rate variability: pNN50 or pNN20?, Physiol. Meas. 24(3):N9–N14, 2003.

[Inan 2006] O. T. Inan, L. Giovangrandi, G. T. A. Kovacs, Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval features, IEEE Trans. Biomed. Eng. 53(12):2507–2515, 2006.

[Ivanov 1999] P. C. Ivanov, L. A. Amaral, A. L. Goldberger, S. Havlin, M. G. Rosenblum, Z. R. Struzik, H. E. Stanley, Multifractality in human heartbeat dynamics, Nature 399(6735):461–465, 1999.

[Jagnjić 2009] Ž. Jagnjić, N. Bogunović, I. Pižeta, F. Jović, Time series classification based on qualitative space segmentation, Adv. Eng. Informat. 23(1):116–129, 2009.

[Jeong 2002] J. Jeong, J. C. Gore, B. S. Peterson, A Method for Determinism in Short Time Series, and its Application to Stationary EEG, IEEE Trans. Biomed. Eng. 49(11):1374–1379, 2002.

[John 1995] G. John, P. Langley, Estimating Continuous Distributions in Bayesian Classifiers, in: P. Besnard, S. Hanks, eds., Proc. 11th Conf. Uncertainty in Artif. Intell., Montreal, Quebec, Canada, Aug 18-20, 1995, San Mateo CA, USA: Morgan Kaufmann Publishers, pp. 338–345.

[Jović 2007] A. Jović, N. Bogunović, Feature Extraction for ECG Time-Series Mining Based on Chaos Theory, in: V. Luzar-Stiffler, V. Hljuz Dobrić, eds., Proc. 29th Int. Conf. Information Technology Interfaces ITI 2007, Cavtat/Dubrovnik, Croatia, Jun 25-28, 2007, Zagreb, Croatia: SRCE, pp. 63–68.

[Jović 2008] A. Jović, N. Bogunović, Analysis of ECG Records using ECG Chaos Extractor Platform and Weka System, in: V. Luzar-Stiffler, V. Hljuz Dobrić, Z. Bekić, eds., Proc. 30th Int. Conf. Information Technology Interfaces ITI 2008, Cavtat/Dubrovnik, Croatia, Jun 23-26, 2008, Zagreb, Croatia: SRCE, pp. 347–352.

[Jović 2009] A. Jović, N. Bogunović, Feature Set Extension for Heart Rate Variability Analysis by Using Non-linear, Statistical and Geometric Measures, in: V. Luzar-Stiffler, I. Jarec, Z. Bekić, eds., Proc. 30th Int. Conf. Information Technology Interfaces ITI 2009, Cavtat/Dubrovnik, Croatia, Jun 22-25, 2009, Zagreb, Croatia: SRCE, pp. 35–40.

[Jović 2010 (1)] A. Jović, N. Bogunović, Classification of Biological Signals Based on Nonlinear Features, in: C. J. Debono, M. P. Kazmierkowski, P. Micallef, eds., Proc. 15th IEEE

Mediterranian Electromechanical Conference MELECON 2010, Valletta, Malta, Apr 25-28, 2010, Valletta, Malta: IEEE Press, pp. 1340–1345.

[Jović 2010 (2)] A. Jović, N. Bogunović, Random Forest-Based Classification of Heart Rate Variability Signals by Using Combinations of Linear and Nonlinear Features, in: P. D. Bamidis, N. Pallikarakis, eds., Proc. 12th Mediterranean Conf. on Medical and Biological Engineering and Computing MEDICON 2010, Porto Carras, Chalkidiki, Greece, May 27-30, 2010, Berlin, Germany: Springer, IFMBE Proc. vol. 29, pp. 29–32.

[Jović 2011 (1)] A. Jović, N. Bogunović, Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features, Artif. Intell. Med. 51:175–186, 2011.

[Jović 2011 (2)] A. Jović, N. Bogunović, HRVFrame: Java-Based Framework for Feature Extraction from Cardiac Rhythm, Lecture Notes in Artif. Intell. 6747:96–100, 2011.

[Jović 2011 (3)] A. Jović, N. Bogunović, Evaluating and comparing performance of feature combinations of heart rate variability measures for cardiac rhythm classification, Biomedical Signal Processing and Control, in press, doi:10.1016/j.bspc.2011.10.001

[Kadbi 2006] M. H. Kadbi, J. Hashemi, H. R. Mohseni, A. Maghsoudi, Classification of ECG arrhythmias based on statistical and time-frequency features, in: S. G. Fabri, ed., Proc. IET 3rd Int. Conf. on Advances in Medical, Signal and Information Processing MEDSIP 2006, Glasgow, Scotland, UK, Jul 17-19, 2006, Red Hook NY, USA: Curran Associates, pp. 1–4.

[Kantz 1995] H. Kantz, T. Schreiber, Dimension estimates and physiological data, Chaos 5(1):143–154, 1995.

[Karraz 2006] G. Karraz, G. Magenes, Automatic classification of heartbeats using neural network classifier based on a Bayesian framework, in: Proc. Int. Conf. EMBC 2006, New York, USA, Aug 30 - Sep 3, 2006, IEEE Press, vol. 1, pp. 4016–4019.

[Katz 1988] M. J. Katz, Fractals and analysis of waveforms, Comput. Biol. Med. 18:145–156, 1988.

[Keerthi 2001] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy, Improvements to Platt's SMO Algorithm for SVM Classifier Design, Neural Computation 13(3):637–649, 2001.

[Kim 2009] J. Kim, H. S. Shin, K. Shin, M. Lee, Robust algorithm for arrhythmia classification in ECG using extreme learning machine, BioMed. Eng. OnLine 8:31, 2009.

[Kitlas 2005] A. Kitlas, E. Oczeretko, M. Kowalewski, M. Borowska, M. Urban, Nonlinear dynamics methods in the analysis of the heart rate variability, Roczniki Akademii Medycznej w Białymstoku, Annales Academiae Medicae Bialostocensis 50(Suppl. 2), 2005.

[Kitney 1982] R. I. Kitney, D. Linkens, A. Selman, A. McDonald, The interaction between heart rate and respiration: part II – nonlinear analysis based on computer modelling, Automedica 4:141–153, 1982.

[Kiviniemi 2007] A. M. Kiviniemi, M. P. Tulppo, D. Wichterle, A. H. Hautala, S. Tiinanen, T. Seppanen, T. H. Mäkikallio, H. V. Huikuri, Novel spectral indexes of heart rate variability as predictors of sudden and non-sudden cardiac death after an acute myocardial infarction, Ann. Med. 39:54–62, 2007.

[Kohavi 1995] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: C. S. Mellish, ed., Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI), Montréal, Québec, Canada, Aug 20-25, 1995, San Francisco CA, USA: Morgan Kaufmann, pp. 1137–1143.

[Kohavi 1997] R. Kohavi, G. John, Wrappers for feature selection, Artif. Intell. 97(1–2):273–324, 1997.

[Kolmogorov 1958] A. N. Kolmogorov, A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces, Dokl. Akad. Nauk SSSR 119:861–864, 1958.

[Krstačić 2003] G. Krstačić, A. Krstačić, M. Martinis, E. Vargović, A. Knežević, M. Jembrek-Gostović, A. Šmalcelj, D. Miličić, M. Bergovec, M Gostović, Dynamic Non-Linear Changes in Heart Rate Variability in Patients with Coronary Heart Disease and Arterial Hypertension Treated by Amlodipine Besylate, in: A. Murray, ed., Proc. Computers in Cardiology Conference 2003, Thessaloniki, Greece, Sep 21-24, 2003, pp. 485–488.

[Lagerholm 2000] M. Lagerholm, C. Peterson, G. Braccini, L. Ebendrandt, L. Sornmo, Clustering ECG complexes using Hermite functions and self-organizing maps, IEEE Trans. Biomed. Eng. 47:838–848, 2000.

[Lehnertz 2008] K. Lehnertz, Epilepsy and Nonlinear Dynamics, J. Biol. Phys. 34:253–266, 2008.

[Leite 2010] F. S. Leite, A. F. da Rocha, J. L. A. de Carvalho, Matlab software for detrended fluctuation analysis of heart rate variability, in: Proc. 3rd Int. Conf. Bio-inspired Systems and Signal Processing BIOSIGNALS 2010, Valencia, Spain, Jan 20-23, 2010, INSTICC, pp. 225–229.

[Lempel 1976] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Trans. Inform. Theory 22:75–81, 1976.

[Lerner 2007] V. S. Lerner, Information System Theory and Informational Macrodynamics: Review of the Main Results, IEEE T. Syst. Man Cy. C 37(6):1050–1066, 2007.

[Liang 2010] S.-F. Liang, H.-C. Wang, W.-L. Chang, Combination of EEG Complexity and Spectral Analysis for Epilepsy Diagnosis and Seizure Detection, EURASIP Journal on Advances in Signal Processing 2010:853434, 2010.

[Lin 2006] J. W. Lin, J. J. Hwang, L.Y. Lin, J. L. Lin, Measuring Heart Rate Variability with Wavelet Thresholds and Energy Components in Healthy Subjects and Patients with Congestive Heart Failure, Cardiology 106(4):207–214, 2006.

[Lin 2010] C.-W. Lin, J.-S. Wang, P.-C. Chung, Mining Physiological Conditions from Heart Rate Variability Analysis, IEEE Computat. Intell. Mag. 5(1):50–58, 2010.

[Lisboa 2000] P. J. Lisboa, A. Vellido, H. Wong, Bias reduction in skewed binary classification with Bayesian neural networks, Neural Networks 13:407–410, 2000.

[Liu 1996] H. Liu, R. Setiono, A probabilistic approach to feature selection - A filter solution, in: L. Saitta, ed., Proc. 13th Int. Conf. Mach. Learn. ICML 1996, Bari, Italy, Jul 3-6, 1996, San Francisco CA, USA: Morgan Kaufmann, pp. 319-327.

[López 2006] F. G. López, M. G. Torres, B. M. Batista, J. A. M. Pérez, J. M. Moreno-Vega, Solving feature subset selection problem by a Parallel Scatter Search, Eur. J. Operat. Res. 169(2):477–489, 2006.

[Lu 2008] S. Lu, X. Chen, J. K. Kanters, I. C. Solomon, K. H. Chon, Automatic selection of the threshold value R for approximate entropy, IEEE Trans. Biomed. Eng. 55(8):1966–1972, 2008.

[Mandelbrot 1983] B. B. Mandelbrot, Geometry of nature, San Francisco CA, USA: Freeman, 1983.

[Manis 2007] G. Manis, S. Nikolopoulos, A. Alexandridi, C. Davos, Assessment of the classification capability of prediction and approximation methods for HRV analysis, Comp. Biol. Med. 37:642–654, 2007.

[Melillo 2011] P. Melillo, M. Bracale, L. Pecchia, Nonlinear heart rate variability features for real-life stress detection. Case study: students under stress due to university examination, BioMed. Eng. OnLine 10:96, 2011.

[Mietus 2002] J. E. Mietus, C. K. Peng, I. Henry, R. L. Goldsmith, A. L. Goldberger, The pNNx files: re-examining a widely used heart rate variability measure, Heart 88:378–380, 2002.

[Minhas 2008] F. A. Minhas, M. Arif, Robust electrocardiogram (ECG) beat classification using discrete wavelet transform, Physiol. Meas. 29(5):555, 2008.

[Mitchell 1980] T. M. Mitchell, The need for biases in learning generalizations, Tech. rep. Computer Science Department, Rutgers University, New Brunswick MA, USA, 1980, reprinted in J. W. Shalvik, T. G. Dietterich, eds., Readings in Machine Learning, San Francisco CA, USA: Morgan Kaufmann, 1991.

[Muñoz-Diosdado 2005] A. Muñoz-Diosdado, A non linear analysis of human gait time series based on multifractal analysis and cross correlations, J. Phys.: Conf. Ser 23:87, 2005.

[Niskanen 2004] J.-P. Niskanen, M. P. Tarvainen, P. O. Ranta-aho, P. A. Karjalainen, Software for advanced HRV analysis. Comp. Meth. Prog. Biomed. 76:73–81, 2004.

[Osowski 2001] S. Osowski, T. H. Linh, ECG beat recognition using fuzzy hybrid neural network, IEEE Trans. Biomed. Eng. 48(11):1265–1271, 2001.

[Osowski 2004] S. Osowski, L. T. Hoai, T. Markiewicz, Support vector machine-based expert system for reliable heartbeat recognition, IEEE Trans. Biomed. Eng. 51:582–589, 2004.

[Owis 2002] M. I. Owis, A. H. Abou-Zied, A.-B. M. Youssef, Y. M. Kadah, Study of Features Based on Nonlinear Dynamical Modeling in ECG Arrhythmia Detection and Classification, IEEE Trans. Biomed. Eng. 49(7):733–736, 2002.

[Pan 1985] J. Pan, W. J. Tompkins, Real time QRS detector algorithm, IEEE Trans. Biomed. Eng. 32:230–236, 1985.

[Pecchia 2011] L. Pecchia, P. Melillo, M. Sansone, M. Bracale, Discrimination power of short-term heart rate variability measures for CHF assessment, IEEE Trans. Inf. Technol. Biomed. 15(1):40–46, 2011.

[Peng C. 1995] C.-K. Peng, S. Havlin, H. E. Stanley, A. L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos 5(1):82–87, 1995.

[Peng C. 2009] C.-K. Peng, M. Costa, A. L. Goldberger, Adaptive Data Analysis of Complex Fluctuations in Physiologic Time Series, Advances in Adaptive Data Analysis 1(1):61–70, 2009.

[Peng H. 2005] H. Peng, F. Long, C. Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27(8):1226–1238, 2005.

[Perakakis 2010] P. Perakakis, M. Joffily, M. Taylor, P. Guerra, J. Vila, KARDIA: A Matlab software for the analysis of cardiac interbeat intervals, Comp. Meth. Prog. Biomed. 98:83–89, 2010.

[PhysioNet] PhysioNet: Physiologic signal archives for biomedical research, http://www.physionet.org [2012-01-15]

[Pincus 1994] S. M. Pincus, A. L. Goldberger, Physiological time-series analysis: what does regularity quantify? Am. J. Physiol. 266 (Heart Circ. Physiol. 35):H1643–H1656, 1994.

[Platt 1998] J. C. Platt, Fast Training of Support Vector Machines Using Sequential Minimal Optimization, in: B. Schöllkopf, C. J. C. Burges, A. J. Smola, eds., Advances in Kernel Methods - Support Vector Learning, Cambridge MA, USA: MIT Press, 1998, pp. 185–208.

[Porta 1998] A. Porta, G. Baselli, D. Liberati, N. Montano, C. Cogliati, T. Gnecchi-Ruscone, A. Malliani, S. Cerutti, Measuring regularity by means of a corrected conditional entropy in sympathetic outflow, Biol. Cybern. 78:71–78, 1998.

[Porta 2007] A. Porta, T. Gnecchi-Ruscone, E. Tobaldini, S. Guzzetti, R. Furlan, N. Montano, Progressive decrease of heart period variability entropy-based complexity during graded head-up tilt, J. Appl. Physiol. 103:1143–1149, 2007.

[Portnoff 1980] M. R. Portnoff, Time-frequency representation of digital signals and systems based on short-time Fourier analysis, IEEE Trans. Acoust., Speech, Signal Process. 28(1):55–69, 1980.

[Pramanik 2010] S. Pramanik, U. N. Chowdhury, B. K. Pramanik, N. Huda, A Comparative Study of Bagging, Boosting and C4.5: The Recent Improvements in Decision Tree Learning Algorithm, Asian Journal of Information Technology 9(6):300–306, 2010.

[Press 1992] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, Power Spectrum Estimation Using the FFT, in: Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed., Cambridge, UK: Cambridge University Press, ch. 13.4, pp. 542–551, 1992.

[Pritchard 1995] W. S. Pritchard, D. W. Duke, Measuring chaos in the brain: A tutorial review of EEG dimension estimation, Brain Cogn. 27(3):353–397, 1995.

[Protopopescu 2005] V. Protopopescu, L. M. Hively, Phase-space dissimilarity measures of nonlinear dynamics: Industrial and biomedical applications, Recent Res. Devel. Physics 6:649–688, 2005.

[Provost 1997] F. Provost, T. Fawcett, Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions, in: D. Heckerman, H. Mannila, D. Pregibon, eds., Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, Newport Beach CA, USA, Aug 14-17, 1997, Menlo Park CA, USA: AAAI Press, pp. 43–48.

[Pudmetzky 2005] A. Pudmetzky, Teleonomic Entropy Measuring the Phase-Space of End-Directed System, Appl. Math. Comput. 162(2):695–705, 2005.

[Pyle 1999] D. Pyle, Data Preparation for Data Mining, San Francisco CA, USA: Morgan Kaufmann Publishers, 1999.

[Quinlan 1993] J. R. Quinlan, C4.5: Programs for machine learning, San Francisco CA, USA: Morgan Kaufmann Publishers, 1993.

[Quinlan 1995] J. R. Quinlan, MDL and Categorical Theories (Continued), in: A. Prieditis, S. J. Russell, eds., Proc. 12th Int. Conf. Mach. Learn., Tahoe City CA, USA, Jul 9-12, 1995, San Francisco CA, USA: Morgan Kaufmann, pp. 464–470.

[Quinlan 1996] J. R. Quinlan, Bagging, Boosting, and C4.5, in: AAAI, Proc. 13th Nat. Conf. Artif. Intell., Aug 4-8, 1996, Portland OR, USA, Cambridge MA, USA: MIT Press, pp. 725–730.

[Rényi 1961] A. Rényi, On measures of Entropy and Information, in: J. Neyman, ed., Proc. 4th Berkeley Symposium on Mathematical Statistics and Probabibility, Berkeley CA, USA, Jun 20-30, 1960, University of California Press, 1961, vol. 1, pp. 547–561.

[Rezek 1998] I. A. Rezek, S. J. Roberts, Stochastic Complexity Measures for Physiological Signal Analysis, IEEE Trans. Biomed. Eng. 45(9):1186–1191, 1998.

[Richman 2000] J. S. Richman, J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. Heart Circ. Physiol. 278:2039–2049, 2000.

[Robnik-Šikonja 2003] M. Robnik-Šikonja, I. Kononenko, Theoretical and Empirical Analysis of ReliefF and RReliefF, Mach. Learn. 53(1–2):23–69, 2003.

[Robnik-Šikonja 2004] M. Robnik-Šikonja, Improving Random Forests, in: J. F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi, eds., Proc. 15th Eur. Conf. Mach. Learn. ECML 2004, Pisa, Italy, Sep 20-24, 2004, Berlin, Germany: Springer, pp. 359–370.

[Rodriguez 2008] E. Rodriguez, C. Lerma, J. C. Echeverria, J. Alvarez-Ramirez, ECG scaling properties of cardiac arrhythmias using detrended fluctuation analysis, Physiol. Meas. 29(11):1255–1266, 2008.

[Rosenbaum 1994] D. S. Rosenbaum, L. E. Jackson, J. M. Smith, G. H. Garan, J. N Ruskin, R. J. Cohen, Electrical alternans and vulnerability to ventricular arrhythmia, N. Engl. J. Med. 330:235–241, 1994.

[Rosenstien 1993] M. Rosenstien, J. J. Colins, C. J. de Luca, A practical method for calculating largest Lyapunov exponents from small data sets, Physica D 65:117–134, 1993.

[Schapire 1997] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, Boosting the margin: A new explanation for the effectiveness of voting methods, in: D. H. Fisher, ed., Proc. 14th Int. Conf. Mach. Learn., Nashville TN, USA, Jul 8-12, 1997, San Francisco CA, USA: Morgan Kaufmann, pp. 322–330.

[Schechtman 1992] V. L. Schechtman, S. L. Raetz, R. K. Harper, A. Garfinkel, A. J. Wilson, D. P. Southall, R. M. Harper**,** Dynamic analysis of cardiac R-R intervals in normal infants and in infants who subsequently succumbed to the sudden infant death syndrome**,** Pediatric Res. 31(6):606–612, 1992.

[Schiff 1992] S. J. Schiff, T. Chang, Differentiation of linearly correlated noise from chaos in a biologic system using surrogate data, Biol. Cybern. 67:387–393, 1992.

[Schlögl 2008] A. Schlögl, C. Vidaurre, E. Hofer, T. Wiener, C. Brunner, R. Scherer, F. Chiarugi, Biosig – standardization and quality control in biomedical signal processing using the biosig project, in: P. Encarnação, A. Veloso, eds., Proc. 1st Int. Joint Conf. Biomedical Engineering Systems and Technologies BIOSTEC 2008, Funchal, Madeira, Portugal, Jan 28-31, 2008, IEEE Press, pp. 403–409.

[Sebastiani 2002] F. Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys 34:1–47, 2002.

[Seewald] A. K. Seewald, Seewald solutions, http://alex.seewald.at/kdd.html [2012-1-15]

[Seker 2000] R. Seker, S. Saliu, A. Birand, G. Kudaiberdieva, Validity Test for a Set of Nonlinear Measures for Short Data Length with Reference to Short-Term Heart Rate Variability Signal, Journal of Systems Integration 10:41–53, 2000.

[Small 2000] M. Small, D. Yu, R. G. Harrison, C. Robertson, G. Clegg, M. Holzer, F. Sterz, Deterministic nonlinearity in ventricular fibrillation, Chaos 10:268–277, 2000.

[Soman 2005] T. Soman, P. O. Bobbie, Classification of Arrhythmia Using Machine Learning Techniques, WSEAS Trans. Comp. 4(6):548–552, 2005.

[Stam 2005] C. J. Stam, Nonlinear dynamical analysis of EEG and MEG: review of an emerging field, Clin. Neurophysiol. 116:2266–2301, 2005.

[Syed 2011] Z. Syed, C. M. Stulz, B. M. Scirica, J. V. Guttag, Computationally Generated Cardiac Biomarkers for Risk Stratification After Acute Coronary Syndrome, Science Translational Medicine 3(102):ra95, 2011.

[Talbi 2009] M. L. Talbi, A. Charef, PVC discrimination using the QRS power spectrum and self-organizing maps, Comp. Meth. Prog. Biomed. 94(3):223–231, 2009.

[Task Force 1 1996] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, Heart rate variability guidelines: Standards of measurement, physiological interpretation, and clinical use, Eur. Heart J. 17:354–381, 1996.

[Task Force 2 2008] The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology, ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008, Eur. Heart J. 29(19):2388–2442, 2008.

[Teich 1998] M. C. Teich, Multiresolution Wavelet Analysis of Heart Rate Variability for Heart-Failure and Heart-Transplant Patients, in: H. K. Chang, ed., Proc. 20th Ann. Int. Conf. IEEE/EMBS, Hong Kong, Oct 29 - Nov 1, 1998, IEEE Press, pp. 1136–1141.

[Teich 2001] M. C. Teich, S. B. Lowen, B. M. Jost, K. Vibe-Rheymer, C. Heneghan, Heart-Rate Variability: Measures and Models, in: M. Akay, ed., Nonlinear Biomedical Signal Processing, vol. II, Dynamic Analysis and Modeling, New York, USA: IEEE Press, 2001, ch. 6, pp. 159–213.

[Theiler 1992] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J. D. Farmer, Testing for nonlinearity in time series: the method of surrogate data, Physica D 58:77–94, 1992.

[Thurner 1998] S. Thurner, M. C. Feurstein, M. C. Teich, Multiresolution wavelet analysis of heartbeat intervals discriminates healthy patients from those with cardiac pathology, Phys. Rev. Lett. 80:1544–1547, 1998.

[Toichi 1997] M. Toichi, T. Sugiura, T. Murai, A. Sengoku, A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of R-R interval, J. Auton. Nerv. Syst. 62(1–2):79–84, 1997.

[Tsipouras 2004] M. G. Tsipouras, D. I. Fotiadis, Automatic arrhythmia detection based on time and time–frequency analysis of heart rate variability, Comp. Meth. Prog. Biomed. 74(2):95–108, 2004.

[Tsipouras 2005] M. G. Tsipouras, D. I. Fotiadis, D. Sideris, An arrhythmia classification system based on the RR-interval signal, Artif. Intell. Med. 33(3):237–250, 2005.

[Tulppo 1996] M. P. Tulppo, T. H. Makikallio, T. E. Takala, T. Seppanen, H. V. Huikuri, Quantitative beat-to-beat analysis of heart rate dynamics during exercise, Am. J. Physiol. 71:H244–252, 1996.

[Turcott 1996] R. G. Turcott, M. C. Teich, Fractal character of the electrocardiogram: distinguishing heart-failure and normal patients, Ann. Biomed. Eng. 24:269–293, 1996.

[Übeyli 2008] E. D. Übeyli, Recurrent neural networks with composite features for detection of electrocardiographic changes in partial epileptic patients, Comp. Biol. Med. 38:401–410, 2008.

[Vallverdú 1999] M. Vallverdú, F. Clariá, R. Carvajal, P. Martinez, J. L. Alonso, W. Zareba, X. Vinolas, A. Bayes de Luna, P. Caminal, Heart rate variability characterization: time-frequency representation and nonlinear analysis, in: Proc. Computers in Cardiology Conf. CINC 1999, Hanover, Germany, Sep 26-29, 1999, pp. 257–260.

[Vapnik 1996] V. Vapnik, The Nature of Statistical Learning Theory, New York, USA: Springer, 1996.

[Vetterli 1992] M. Vetterli, C. Herley, Wavelet and filter banks: theory and design, IEEE Trans. Signal Process. 40(9):2207–2232, 1992.

[Voronoi 1908] G. F. Voronoi, Nouvelles applications des paramètres continus à la théorie de formes quadratiques, Journal für die reine und angewandte Mathematik 134:198–287, 1908.

[Voss 1996] A. Voss, J. Kurths, H. J. Kleiner, A. Witt, N. Wessel, P. Saparin, K. J. Osterziel, R. Schurath, R. Dietz, The application of methods of nonlinear dynamics for the improved and predictive recognition of patients threatened by sudden cardiac death, Cardiovasc. Res. 31(3):419–433, 1996.

[Voss 2007] A. Voss, R. Schroeder, S. Truebner, M. Goernig, H. R. Figulla, A. Schirdewan, Comparison of nonlinear methods symbolic dynamics, detrended fluctuation, and Poincare plot analysis in risk stratification in patients with dilated cardiomyopathy, Chaos 17(1):015120, 2007.

[Waheed 2002] K. Waheed, F. M. Salam, A Data-Derived Quadratic Independence Measure for Adaptive Blind Source Recovery in Practical Applications, in: Proc. 45th IEEE Int. Midwest Symposium on Circuits and Systems, Tulsa OK, USA, Aug 4-7, 2002, IEEE Press, pp. 473–476.

[Wang 2001] Y. Wang, Y. S. Zhu, N. V. Thakor, Y. H. Xu, A short-time multifractal approach for arrhythmia detection based on fuzzy neural network, IEEE Trans. Biomed. Eng. 48(9):989–995, 2001.

[Wasikowski 2010] M. Wasikowski, X.-W. Chen, Combating the Small Sample Class Imbalance Problem Using Feature Selection, IEEE Trans. Knowl. Data Eng. 22(10):1388–1400, 2010.

[Webb 2005] G. I. Webb, J. R. Boughton, Z. Wang, Not So Naive Bayes: Aggregating One-Dependence Estimators, Mach. Learn. 58(1):5–24, 2005.

[Weiss 1999] J. N. Weiss, A. Garfinkel, H. S. Karagueuzian, Z. Qu, P.-S. Chen, Chaos and the Transition to Ventricular Fibrillation : A New Approach to Antiarrhythmic Drug Evaluation, Circulation 99:2819–2826, 1999.

[Witten 2005] I. H. Witten, E. Frank, Data mining: practical machine learning tools and techniques, 2nd ed., San Francisco CA, USA: Morgan Kaufmann, 2005.

[Wolf 1985] A. Wolf , J. B. Swift, H. Swinney, J. A.Vastano, Determining Lyapunov exponents from a time series, Physica D 16:285–317, 1985.

[Xia 2009] H. Xia, X. Zhao, J. Bains, D. C. Wortham, A Review of Diagnosis Methods for Heart Rhythm Disorders, in: B. M. Evans III, ed., Proc. 1st Ann. ORNL Biomedical Science & Engineering Conf. BSEC 2009, Oak Ridge TN, USA, Mar 18-19, 2009, pp. 1–4.

[Yaghouby 2009] F. Yaghouby, A. Ayatollahi, R. Soleimani, Classification of Cardiac Abnormalities Using Reduced Features of Heart Rate Variability Signal, World Applied Sciences Journal 6(11):1547–1554, 2009.

[Yang 2003] A. C.-C. Yang, S.-S. Hseu, H.-W. Yien, A. L. Goldberger, C.-K. Peng, Linguistic Analysis of the Human Heartbeat Using Frequency and Rank Order Statistics, Phys. Rev. Lett. 90(10):108103, 2003.

[Yu L. 2004] L. Yu, H. Liu, Redundancy based feature selection for microarray data, in: W. Kim, R. Kohavi, J. Gehrke, W. DuMouchel, eds., Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Seattle WA, USA, Aug 22-25, 2004, ACM Press, pp. 737–742.

[Yu S. 2006] S.-N. Yu, K.-T. Chou, Combining Independent Component Analysis and Backpropagation Neural Network for ECG Beat Classification, in: Proc. 28th Ann. Int. Conf. IEEE EMBS '06, New York, USA, Aug 30 - Sep 3, 2006, pp. 3090–3093.

[Zbilut 2002] J. P. Zbilut, N. Thomasson, C. L. Webber, Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals, Med. Eng. & Phys. 24:53–60, 2002.

[Zeng 1991] X. Zeng, R. Eykholt, R. A. Pielke, Estimating the Lyapunov-Exponent Spectrum from Short Time Series of Low Precision, Phys. Rev. Lett. 66(25):3230–3232, 1991.

[Zhang K. 2007] K. Zhang, A. Ball, F. Gu, Y. Li, A Hybrid Model with a Weighted Voting Scheme for Feature Selection in Machinery Condition Monitoring, in: M. Zhou, ed., Proc. IEEE Int. Conf. Automation Science and Engineering CASE 2007, Scottsdale AZ, USA, Sep 22-25, 2007, IEEE Press, pp. 424–429.

[Zhang X. 1999] X.-S. Zhang, Y.-S. Zhu, N. V. Thakor, Z.-Z. Wang, Detecting Ventricular tachycardia and fibrillation by complexity measure, IEEE Trans. Biomed. Eng. 46:548–555, 1999.

[Zheng 2004] Z. Zheng, X. Wu, R. Srihari, Feature Selection for Text Categorization on Imbalanced Data, ACM SIGKDD Explorations Newsletter 6(1):80–89, 2004.

Popis oznaka

1Rule – postupak jednog pravila (engl. *1-Rule*)
AbEn – abecedna entropija (engl. *alphabet entropy*)
ABI – atrijalna bigeminija (engl. *atrial bigeminy*)
ACC – ukupna točnost razvrstavanja (engl. *total classification accuracy*)
AF – atrijalna fibrilacija (engl. *atrial fibrillation*)
AIC – Akaikeov informacijski kriterij (engl. *Akaike information criterion*)
ALF – Allanov faktor (engl. *Allan factor*)
AMI – akutni infarkt miokarda (srčani udar) (engl. *acute myocardial infarction*)
ANN – umjetna neuronska mreža (engl. *artificial neural network*)
ANS – autonomni živčani sustav (engl. *autonomic nervous system*)
ApEn – približna entropija (engl. *approximate entropy*)
AR-model – autoregresijski model (engl. *autoregressive model*)
ASTA – napredna analiza slijednog trenda (engl. *advanced sequential trend analysis*)
ATR – atrijalna trigeminija (engl. *atrial trigeminy*)
AV – atrioventrikularni (engl. *atrioventricular*)
BCI – sučelje između mozga i računala (engl. *brain-computer interface*)
BII – srčani blok drugog stupnja (engl. *second-degree heart block*)
BVN – biomedicinski vremenski niz (engl. *biomedical time-series*)
CCE – ispravljena uvjetna entropija (engl. *corrected conditional entropy*)
CE – uvjetna entropija (engl. *conditional entropy*)
CFS – odabir značajki zasnovan na korelaciji (engl. *correlation-based feature selection*)
CHB – potpuni srčani blok (srčani blok trećeg stupnja) (engl. *complete heart block*)
CHF – kongestivno zatajenje srca (engl. *congestive heart failure*)
CSI – srčani simpatički indeks (engl. *cardiac sympathetic index*)
CTM – mjera središnje težnje (engl. *central tendency measure*)
CVI – srčani vagalni indeks (engl. *cardiac vagal index*)
CWT – kontinuirana transformacija valićima (engl. *continuous wavelet transform*)
D_2 – korelacijska dimenzija (engl. *correlation dimension*)
DFA – analiza kolebanja s uklonjenim trendom (engl. *detrended fluctuation analysis*)
DL – opisna duljina (engl. *description length*)
DM – dubinska analiza podataka (engl. *data mining*)
DWT – diskretna transformacija valićima (engl. *discrete wavelet transform*)
EEG – elektroencefalogram (engl. *electroencephalogram*, EEG)
EKG – elektrokardiogram (engl. *electrocardiogram*, ECG)
EMG – elektromiogram (engl. *electromyogram*)
En_c – Carnapova entropija (engl. *Carnap entropy*)
FD – fraktalna dimenzija (engl. *fractal dimension*)
FFT – brza Fourierova transformacija (engl. *fast Fourier transform*)
FN – lažno negativni primjerci (engl. *false negatives*)
FP – lažno pozitivni primjerci (engl. *false positives*)
GDA – opća diskriminantna analiza (engl. *general discriminant analysis*)

GR – omjer dobitaka (engl. *gain ratio*)
HE – Hurstov eksponent (engl. *Hurst exponent*)
HF – visoka frekvencija (engl. *high frequency*)
HOS – spektar višeg reda (engl. *higher-order spectrum*)
HRV – varijabilnost srčanog ritma (engl. *heart rate variability*)
I/D kardiomiopatija – ishemijska/dilatacijska kardiomiopatija (engl. *ischemic/dilated cardiomyopathy*)
IG – informacijski dobitak (engl. *information gain*)
InCons – konzistentnost razreda (engl. *class consistency*)
JSD – Jensen-Shannonova divergencija (engl. *Jensen-Shannon divergence*)
KDD – otkrivanje znanja u skupovima podataka (engl. *knowledge discovery in datasets*)
K-entropija – Kolmogorovljeva entropija (engl. *Kolmogorov entropy*)
KKT – uvjeti Karush-Kuhn-Tucker (engl. *Karush-Kuhn-Tucker conditions*)
LBBB – blok lijeve grane (engl. *left bundle branch block*)
LE – Ljapunovljev eksponent (engl. *Lyapunov exponent*)
LF – niska frekvencija (engl. *low frequency*)
LLE – najveći Ljapunovljev eksponent (engl. *largest Lyapunov exponent*)
LOOCV – unakrsna validacija s izostavljanjem jednog primjerka (engl. *leave-one-out cross-validation*)
MFMD – Više značajki / više poremećaja (engl. *multiple features* / *multiple disorders*)
MFSD – Više značajki / jedan poremećaj (engl. *multiple features* / *single disorder*)
ML – strojno učenje (engl. *machine learning*)
MLP – višeslojni perceptron (engl. *multilayer perceptron*)
MSE – entropija na više skala (engl. *multiscale entropy*)
NSR – normalan sinusni ritam (engl. *normal sinus rhythm*)
NYHA – Njujorška udruga za srce (engl. *New York Heart Association*)
OOB – skup izdvojenih primjeraka (engl. *out-of-bag dataset*)
OOT – test najboljeg reda (engl. *optimal order test*)
PAC – preuranjena kontrakcija pretklijetki (engl. *premature atrial contraction*)
PCA – analiza glavnih komponenti (engl. *principal component analysis*)
PCR – fazni srčani odgovori (engl. *phase cardiac responses*)
PEJS - umjetno vođeni ritam (pejsmejker) (engl. *pacemaker*)
PPV – predvidljivost pozitivnih primjeraka (engl. *positive predictive value*)
PSD – spektralna gustoća snage (engl. *power spectral density*)
PVC – preuranjena kontrakcija klijetki (engl. *premature ventricular contraction*)
QTSD – kvalitativna diskretizacija vremenskog niza (engl. *qualitative time-series discretization*)
RBBB – blok desne grane (engl. *right bundle branch block*)
RBF – odabir značajki zasnovan na redundantnosti (engl. *redundancy-based feature selection*); radijalna funkcija jezgre (engl. *radial basis function*)
RenEn – Rényijeva entropija (engl. *Rényi entropy*)
RF – slučajna šuma (engl. *random forest*)
RSA – respiratorna sinusna aritmija (engl. *respiratory sinus arrhythmia*)
SA-čvor – sinoatrijalni čvor (engl. *sinoatrial node*)
SampEn – entropija uzorka (engl. *sample entropy*)

SBR – sinusna bradikardija (engl. *sinus bradycardia*)
SCD – iznenadna srčana smrt (engl. *sudden cardiac death*)
SD – standardna devijacija (engl. *standard deviation*)
SENS – osjetljivost modela (engl. *sensitivity*)
SFI – indeks prostorne popunjenost (engl. *spatial filling index*)
SFMD – Jedna značajka / više poremećaja (engl. *single feature / multiple disorders*)
SFSD – Jedna značajka / jedan poremećaj (engl. *single feature / single disorder*)
ShEn – Shannonova entropija (engl. *Shannon entropy*)
SMO – slijedna najmanja optimizacija (engl. *sequential minimal optimization*)
SPEC – specifičnost modela (engl. *specificity*)
SpectEn – spektralna entropija (engl. *spectral entropy*)
SSS – sindrom bolesnog sinusnog čvora (engl. *sick sinus syndrome*)
ST – sinusna tahikardija (engl. *sinus tachycardia*)
STA – analiza slijednog trenda (engl. *sequential trend analysis*)
STFT – Fourierova transformacija za kratke segmente (engl. *short-time Fourier transform*)
SU – simetrična nesigurnost (engl. *symmetric uncertainty*)
SVM – stroj s potpornim vektorima (engl. *support vector machine*)
SVT – supraventrikularna tahikardija (engl. *supraventricular tachycardia*)
TN – stvarno negativni primjerci (engl. *true negatives*)
TP – stvarno pozitivni primjerci (engl. *true positives*)
ULF – ultra niska frekvencija (engl. *ultra low frequency*)
VBI – ventrikularna bigeminija (engl. *ventricular bigeminy*)
VF – ventrikularna fibrilacija (engl. *ventricular fibrillation*)
VFL – ventrikularno lepršanje (engl. *ventricular flutter*)
VLF – vrlo niska frekvencija (engl. *very low frequency*)
VT – ventrikularna tahikardija (engl. *ventricular tachycardia*)
VTR – ventrikularna trigeminija (engl. *ventricular trigeminy*)
WAP – lutajući atrijalni pejsmejker (engl. *wandering atrial pacemaker*)
WT – transformacija valićima (engl. *wavelet transform*)

Životopis

Alan Jović rođen je u Zagrebu, 24. rujna 1982. godine. Srednju školu, V. gimnaziju završio je u Zagrebu 2001. Iste godine upisao je Fakultet elektrotehnike i računarstva, Sveučilišta u Zagrebu. Diplomirao je na studiju računarstva 2006. godine s temom diplomskog rada pod naslovom: „Ekstrakcija značajki iz elektrokardiografskih signala.“

Na FER-u upisuje poslijediplomski studij 2006. godine. Do ožujka 2007. radio je na Institutu Ruđer Bošković kao stručni suradnik na europskom FP6 projektu pod nazivom: “HEARTFAID“, namijenjenom dijagnostici i pomoći pacijentima oboljelima od kongestivnog zatajenja srca. U sklopu tog projekta razvio je računalnu ontologiju kao dio sustava za pomoć pri odlučivanju, koju je potraživao i američki Stanford i druga sveučilišta.

Od travnja 2007. godine zaposlen je na radnom mjestu asistenta na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave, Fakulteta elektrotehnike i računarstva, Sveučilišta u Zagrebu. Tijekom poslijediplomskog studija, pohađao je dvije ljetne škole: napredni tečaj umjetne inteligencije u Leuvenu, Belgija, 2007. i tečaj formalnih postupaka u oblikovanju računalnih sustava, u Marktoberdorfu, Njemačka, 2008. Sudjelovao je u radu projekta između FER-a i osiguravateljske kuće „CROATIA osiguranje, d.d“ na analizi podataka auto-osiguranika od 2008. do 2009. godine. Profesionalna područja od interesa mu uključuju: dubinsku analizu podataka, predstavljanje znanja u računalnim sustavima i primjenu računarstva u medicini. Autor je 15 radova na konferencijama s međunarodnom recenzijom te šest radova u znanstvenim časopisima, od čega dva rada u CC časopisima. Član je IEEE-a. Služi se engleskim, njemačkim i francuskim jezikom.

Curriculum vitae

Alan Jović was born in Zagreb, Croatia, on September 24th, 1982. He finished his high school education at "V. gimnazija" high school in Zagreb in 2001. He enrolled in the Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Croatia in 2001. He graduated with a degree in computer science in 2006 with graduate thesis entitled: "Feature extraction from electrocardiographic signals".

He enrolled in the postgraduate course of computer science study at FER in 2006. From September 2006 till March 2007 he worked at the "Institut Ruđer Bošković" scientific institute at the position of expert assistant on European FP6 project "HEARTFAID". The project was intended for diagnostics and care of patients suffering from congestive heart failure. During the course of the project he developed a computer ontology as a part of the project's decision support system, which was later requested by Stanford and other universities.

From April 2007 he is employed at the position of research assistant in the Department of Electronics, Microelectronics, Computer and Intelligent Systems, at FER. During the course of his postgraduate study, he attended two summer schools: Advanced Course in Artificial Intelligence, Leuven, Belgium in 2007 and Engineering Methods and Tools for Software Safety and Security, Marktoberdorf, Germany in 2008. He participated in an industrial project between FER and the insurance company "CROATIA osiguranje, d.d" as data analyst for car insurance data, from 2008 till 2009. His professional areas of interest include: data mining, knowledge representation in computer systems, and application of computer science in medicine. He is author or co-author of 15 scientific papers published in proceedings of international conferences and six papers in scientific journals, out of which two are referenced in the Current Contents® database. He is a member of IEEE. He speaks English, German, and French.