Method for classification of fetal phonocardiography signals using empirical mode decomposition and psychoacoustic parameters

Vican, Ivan

University of Zagreb
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Ivan Vican

METHOD FOR CLASSIFICATION OF FETAL PHONOCARDIOGRAPHY SIGNALS USING EMPIRICAL MODE DECOMPOSITION AND PSYCHOACOUSTIC PARAMETERS

DOCTORAL THESIS

Zagreb, 2022

University of Zagreb
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Ivan Vican

METHOD FOR CLASSIFICATION OF FETAL PHONOCARDIOGRAPHY SIGNALS USING EMPIRICAL MODE DECOMPOSITION AND PSYCHOACOUSTIC PARAMETERS

DOCTORAL THESIS

Supervisor:
Professor Kristian Jambrošić, PhD

Zagreb, 2022

Sveučilište u Zagrebu
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Ivan Vican

METODA KLASIFIKACIJE FETALNIH FONOKARDIOLOŠKIH SIGNALA PRIMJENOM EMPIRIJSKE DEKOMPOZICIJE MODOVA I PSIHOAKUSTIČKIH PARAMETARA

DOKTORSKI RAD

Mentor:
prof. dr. sc. Kristian Jambrošić

Zagreb, 2022.

Doctoral thesis was made at the University of Zagreb,
Faculty of Electrical Engineering and Computing,
Department of Electroacoustics

Supervisor: Professor Kristian Jambrošić, PhD

Doctoral thesis contains: 106 pages.

Doctoral thesis number: _____

About the Supervisor

Kristian Jambrošić is a full-time professor at the Department of Electroacoustics of the Faculty of Electrical Engineering and Computing (FER), University of Zagreb, where he obtained his PhD degree in the field of psychoacoustics. His main research activities are in the field of architectural acoustics, acoustic measurements, soundscape research, auralization, perception of sound and noise abatement techniques. He participated in several international and national research projects and has published as author or co-author more than one hundred and forty papers in scientific journals and in conference proceedings. He supervised more than ninety Bachelor and Master thesis and teaches courses on acoustics in Bachelor, Master and Doctoral studies on the Faculty of Electrical Engineering and Computing, as well as on the Academy of Music. He is a member of the executive council of the European Acoustics Association where he served as General Secretary in two terms, and from 2022 he will take over the position of General Secretary of the International Institute of Noise Control Engineering (I-INCE). He has been active in the board of the Acoustical Society of Croatia since its foundation. He is the leader of the research group of the Auralization laboratory of FER.

O mentoru

Kristian Jambrošić je zaposlen kao redoviti profesor na Zavodu za elektroakustiku Fakulteta elektrotehnike i računarstva (FER) Sveučilišta u Zagrebu. Doktorsku disertaciju obranio je na istom fakultetu u području psihoakustike. Njegove glavne istraživačke aktivnosti fokusirane su na područje arhitektonske akustike, akustičkih mjerenja, istraživanja zvučnih okoliša, auralizacije, percepcije zvuka i mjera za smanjenje buke. Sudjelovao je u nekoliko međunarodnih i domaćih istraživačkih projekata, te je kao autor ili koautor objavio više od sto četrdeset radova u znanstvenim časopisima i zbornicima znanstvenih skupova. Mentorirao je više od devedeset preddiplomskih i diplomskih radova, a drži nastavu na preddiplomskom, diplomskom i doktorskom studiju Fakulteta elektrotehnike i računarstva, te na Muzičkoj akademiji. Član je upravnog odbora Europske akustičke asocijacije (EAA) gdje je u dva mandata bio glavnim tajnikom, a od 2022. godine preuzima dužnost glavnog tajnika udruge International Institute of Noise Control Engineering (I-INCE). Aktivan je u upravnom odboru Hrvatskog akustičkog društva od njegovog nastanka. Voditelj je Auralizacijskog laboratorija FER-a.

To my wife and our perfect baby boy.

Acknowledgements

"*It's the questions we can't answer that teach us the most.*
They teach us how to think. If you give a man an answer,
all he gains is a little fact. But give him a question
and he'll look for his own answers."
– Patrick Rothfuss, The Wise Man's Fear (2011)

I would like to use this opportunity to thank the people who have offered unconditional support and understanding during my research.

First, I can only start to express the gratitude to my supervisor, Prof. Kristian Jambrošić, who has been there for me for almost 10 years now. Kristian, your kind words, patience and vast knowledge have shown me the way in finishing research I can be proud of. You have given me all the right answers and questions during critical points, and I am grateful for everything.

To my late mentor, Prof. Hrvoje Domitrović, for not only believing in me during the early years of my studies, but also for showing extreme compassion to a scared kid with chronic tinnitus. Thank you, Professor. I will always miss you.

I also thank my parents Zoran and Maja and my brother Marino, for their words of wisdom and grit in giving me the strength to write this thesis. You have always been my safe place and I will be forever indebted to you.

Finally, to my wife Ivana: you have been my inspiration in completing this research from top to bottom. It would have never happened without you and our Adrian. I dedicate this work to you two.

Abstract

In comparison to other methods used for long-term monitoring of fetal health, fetal phonocardiography has the potential to be more convenient and affordable due to its non-invasive nature and the possibility of implementation on omnipresent devices such as smartphones. Fetal phonocardiography signals can oftentimes be misinterpreted due to various sources of sound in the womb. Therefore, the question remained whether a machine learning model trained for fetal heartbeat detection containing a conventional set of audio features could be improved by introducing features taken from signal representations processed by two methods: empirical mode decomposition (EMD) and pitch shifting. Furthermore, features based on psychoacoustics were proposed as an additional input to the model. In other words, the main goal of this research was to employ EMD and pitch shifting as preprocessing steps, as well as psychoacoustics descriptors such as perceptual linear prediction coefficients, in order to enable the utilization of additional characteristics from the phonocardiography signal.

Features extracted in this fashion were assessed through the analysis of their relevance, usefulness and significance in relation to the performance metrics, such as accuracy or precision. Two raw datasets of audio data were employed as input, one custom recorded and collected through fetal heartbeat acquisition obtained from 8 pregnant women, while the other was taken as an available dataset of simulated fetal heartbeat sounds with different noise levels. Two scenarios with different inputs were introduced in this research.

In Scenario A, the custom dataset was utilized to train a machine learning model from features originating from raw, filtered and EMD-processed versions of the audio signal. The results consistently indicated high ranking of features based on EMD and their ability to improve general detection accuracy once they were introduced to the set of audio features. Namely, the selected subset of combined audio and EMD-based features in comparison to all audio features, improved the detection accuracy by up to 10.28%.

Scenario B contained 6 different cases, incorporating 3 extracted datasets generated from custom raw data and variable segmentation window lengths and 3 extracted datasets acquired through taking the simulated raw dataset with 3 different signal-to-noise ratio values for the fetal heartbeat sound signal. The results of feature selection and ranking methods indicated consistently high relevance of psychoacoustic features, especially in the case where frequency

shifting was used as a preprocessing step. In addition to the random forest models trained with the selected feature subsets, two new classifiers (support vector machine with cubic kernel and bagged trees) were introduced to assess the impact of the entire sets of characteristics added through the proposed preprocessing and feature extraction methods. The analysis showed that the included dimensionality gained through the pitch shifting and EMD in the preprocessing steps with audio and psychoacoustic feature extraction raised the detection accuracy to a higher level, reducing misclassification rate up to nearly 3 times in some instances.

As a final result of this research, the impact of the proposed preprocessing and feature extraction methods in the automatic FPCG classification was shown to be substantial. In addition, the showcased approaches have been demonstrated to be feasible for the implementation in an algorithm for real time usage, further highlighting the possible benefits of the system and its components in the biomedical industry.

KEYWORDS

empirical mode decomposition, psychoacoustics, feature selection, fetal heart sound, machine learning, pitch shifting, fetal phonocardiography

Prošireni sažetak

Metoda klasifikacije fetalnih fonokardioloških signala primjenom empirijske dekompozicije modova i psihoakustičkih parametara

U usporedbi s drugim metodama za dugoročno praćenje zdravlja fetusa, fetalna fonokardiografija ima veliki potencijal - s obzirom na mogućnost implementacije na svakodnevnim uređajima i neinvazivnost, mogla bi postati prilagođenija i jeftinija alternativa. Zbog visoke razine šuma koji proizlazi iz ostalih izvora zvuka u utrobi majke, signali dobiveni navedenom metodom često mogu biti krivo interpretirani. Predložene su dvije metode predobrade zvuka otkucaja srca fetusa: empirijska dekompozicija modova i promjena frekvencije. Uz konvencionalne "audio" značajke, izlučene su i one bazirane na psihoakustici. Također, primjenjene su i metode strojnog učenja u svrhu procjene važnosti pojedinih značajki, kao i evaluaciji kvalitete klasifikacije. Korištena su 2 izvora podataka: podaci snimljeni na 8 trudnica u različitim tjednima trudnoće te simulirani podaci inače primjenjiivi u sličnim istraživanjima. Rezultati pokazuju poboljšanje kvalitete klasifikacije i visoku rangiranost pojedinih značajki iz podskupa predloženih metoda, pogotovo u slučaju kombinacije promjene frekvencije i psihoakustičkih značajki.

Fetalna fonokardiografija je u svojoj suštini prikupljanje, analiza i obrada signala otkucaja srca fetusa. Prednosti navedene metode očituju se u jednostavnosti korištenja (potrebna je jedna sonda) te dostupnosti na uređajima široke primjene, kao što su pametni telefoni. Nažalost, akvizicija akustičkog signala iz utrobe ima više nedostataka: uz jako loš odnos signala i šuma, dolazi do problema u lošem prijenosu zvuka zbog neprilagođenosti akustičke impedancije slojeva kože i mišića na akustičku impedanciju zraka.

Najčešća neinvazivna metoda pri procjeni stanja fetusa uključuje ultrazvučne tehnologije, za koje se ne preporučuju vremenski duži i češći intervali korištenja. Kako je implementacija fetalne fonokardiografije teoretski moguća na svaki uređaj s (kvalitetnim) mikrofonom, postavlja se pitanje mogu li softverske metode poboljšati kvalitetu snimljenog signala te omogućiti pouzdanu procjenu zdravlja fetusa.

S ciljem utvrđvanja postojanja i same pozicije otkucaja fetalnog srca, predloženo je više metoda u svrhu predobrade i izlučivanja značajki signala. Dobivene značajke ključne su za treniranja modela strojnog učenja koji mogu ostvariti superiorne performanse u odnosu na ekspertne sustave, pogotovo u području kompleksnijih zadataka u n-dimenzionalnom prostoru odluke.

Empirijska dekompozicija modova (EMD) predložena je metoda predobrade u svrhu pročišćavanja signala zvuka od šuma. Iterativni proces, zvan prosijavanje, služi za generiranje funkcija intrinzičnih modova koji u idealnom scenariju sadrže "modove oscilacije" signala. Prilikom prosijavanja, na ulaznom signalu računaju se gornja i donja ovojnica signala te se njihova srednja vrijednost oduzima od ulaznog signala. Nakon što su ispunjeni uvjeti za funkciju intrinzičnih modova, ista se sprema te otvara put novom ulaznom signalu za generaciju idućih modova. Ovakav pristup omogućuje odvajanje šuma od korisnog signala, odnosno izlučivanje prvog i drugog zvuka signala fetalnog srca, S1 i S2. EMD, nažalost, pati od problema suboptimalne konvergencije i miješanja modova, tako da je izgledna prisutnost šuma i u funkcijama intrinzičnih modova. Razni korisni signali (npr. S1 i S2) mogu se nalaziti u različitim modovima, dodatno kompliciraući izlaz iz empirijske dekompozicije modova. Unatoč svemu navedenom, EMD i dalje ostaje korisna metoda za dekompoziciju nelinearnih i nestacionarnih biomedicinskih signala.

Psihoakustika je znanstvena disciplina koja se bavi proučavanjem percepcije zvuka, usko povezana s biologijom, fizikom, akustikom, psihologijom te fiziologijom. Jedna od glavnih ideja ovog istraživanja jest iskoristiti mogućnost ljudskog uha da percipira izrazito kompleksne zvukove u raznim neodgovarajućim uvjetima. S obzirom da je ljudsko uho osjetljivije prema višim frekvencijama te da se većina energije zvuka otkucaja srca fetusa nalazi u frekvencijskom rasponu između 20 i 200 Hz, signal je potrebno predobraditi metodom promjene frekvencije signala. S ciljem pomicanja frekvencije signala prema višem dijelu spektra, implementiran je fazni vokoder.

Kao konačna metoda za klasifikaciju zvuka otkucaja, predloženo je strojno učenje. S obzirom da se radi o izrazito složenom problemu klasifikacije u višedimenzionalnom prostoru, odrađeno je izlučivanje značajki signala u domeni konvencionalnih audio značajki i psihoakustičkih parametara. Razne metode iz područja statistike i strojnog učenja iskorištene su za procjenu značaja metoda predloženih u ovom istraživanju.

Ulazni skup podataka ostvaren je snimanjem 8 trudnica u trećem tromjesečju trudnoće, koristeći mjerni mikrofon paralelno s Doppler ultrazvučnim uređajem. Mjerni mikrofon upotrijebljen je za akviziciju signala zvuka otkucaja fetalnog srca, a ultrazvučni uređaj iskorišten je za dobivanje pozicije signala zvuka S1 koja je služila kao oznaka u nadziranom strojnom učenju. Zbog različitih načina rada mikrofona i Doppler uređaja, pojava korisnog signala na jednom instrumentu nije garantirala prisutnost istog na drugom te je predstavljena nekolicina filtera za pročišćavanje ulaznih podataka.

Prikazana su dva scenarija za validaciju predloženih metoda: scenarij A i scenarij B. Dok se prvi scenarij bazira na usporedbi verzije signala obrađene statičkim filtrom s funkcijama intrinzičnih modova dobivenih empirijskom dekompozicijom modova, drugi scenarij uspoređuje statički filtrirani signal s funkcijama intrinzičnih modova i signalom pomaknute frekvencije. Cijela predobrada signala i izlučivanje značajki odrađeni su u programskom sučelju MATLAB, dok su same metode statistike i strojnog učenja implementirane u programskom jeziku Python.

Scenarij A uključuje izlučivanje standardnih audio značajki na 5 verzija signala: neobrađenom signalu, filtriranom signalu te 3 prve funkcije intrinzičnih modova. U svrhu dobivanja primjera za strojno učenje, signal je segmentiran prozorom od 200 ms sa skokom od 100 ms, a svaki je primjer označen s 1 ako sadrži poziciju S1 dobivenu s Doppler uređaja te 0 ako je ne sadrži. Ovakav pristup segmentiranja izgenerirao je 7604 primjera za učenje. Iz svakog od 5 oblika signala izlučeno je 18 statističkih značajki i 9 spektralnih značajki, dajući konačan broj od 135 značajki prije strojnog učenja. Uz samu mjeru poboljšanja točnosti modela, korištene su i razne metode za procjenu važnosti određenih značajki: korelacijska analiza, metoda uzajamne informacije, analiza varijance s jednim promjenjivim faktorom (ANOVA), ugrađena metoda slučajne šume i rekurzivna eliminacija značajki.

Rezultati ukazuju na veću važnost funkcija intrinzičnih modova. Uz konstantnu prisutnost značajki dobivenih kroz verziju signala obrađenu empirijskom dekompozicijom modova, pokazano je da kombinacija značajki iz skupa filtriranog signala te seta funkcija intrinzičnih modova poboljšava točnost klasifikacije do 10,28%.

Scenarij B unaprijedio je istraživanje iz scenarija A na više fronti. Uz predstavljanje dodatnih metoda za procjenu važnosti značajki, uključivanje psihoakustičkih parametara u proces izlučivanja povećalo je broj značajki s originalnih 135 na 212. Konkretno, nakon "izbacivanja" neobrađenog signala te 3. funkcije intrinzičnih modova zbog loših pozicija u rangiranju važnosti, iskorištene su 4 verzije signala: statički filtrirani signal, 2 funkcije intrinzičnih modova te frekvencijski pomaknut signal (za 2 oktave, odnosno pomak od 4x). Uz osnovnih 27 "audio" značajki, izlučeno je i 13 mel-frekvencijskih kepstralnih koeficijenata (MFCC) i 13 perceptivnih linearnih predviđača (PLP). Kako su navedeni psihoakustički parametri izlučivani na manjim prozorima unutar primjera signala, dobiveni 2D tenzori su transformirani u 1D korištenjem 2 statistička funkcionala – srednje vrijednosti i standardne devijacije.

U svrhu dodatne validacije na signalima, predstavljen je i simulirani skup podataka, inače korišten u sličnim istraživanjima u fetalnoj kardiografiji. Korisni dio signala simuliran je generiranjem S1 i S2 valića, dok je pozadinski šum iz raznih izvora unutar utrobe dodan u različitim omjerima. Kako su signali iz skupa podataka pokrivali cijeli raspon snage šuma, za ulazni signal izabrana su 3 primjera s jako niskom razinom odnosa signal-šum: -26,7 dB, -24,4 dB i -22 dB. Za označavanje pozicija S1 signala izabran je signal s najboljim omjerom signala i šuma (-4,4 dB). Također, skupovi podataka dobiveni iz snimljenih signala su proširenи tako da su, uz originalnu širinu prozora od 200 ms, izabrane i 2 nove veličine od 100 i 150 ms. Time je ostvareno 6 skupova podataka za treniranje modela: 3 iz snimljenih signala s različitim veličinama prozora i 3 iz simuliranih signala s različitim omjerima signal-šum.

Rezultati su pokazali superiorni plasman psihoakustičkih značajki u konačnom rasporedu važnosti značajki, pogotovo u slučaju predobrade metodom promjene frekvencije. Navedeno u potpunosti odgovara postavljenoj hipotezi: pomicanje spektra zvuka otkucaja srca fetusa prema višim frekvencijama pobudilo je više nelinearnih psihoakustičkih pojaseva u odnosu na originalnu, nepomaknutu verziju. Nadalje, nesavršena rekonstrukcija signala kao posljedica korištenja faznog vokodera s visokim faktorom promjene frekvencije omogućila je pristup novim, skrivenim informacijama u signalu.

Empirijska dekompozicija modova se također pokazala kao korisna metoda, pogotovo u slučaju snimljenih signala. Primjerice, rekurzivna eliminacija značajki je izabrala 20 od 57 značajki iz skupa funkcija intrinzičnih modova, dok je kao relevantnim procijenila samo 11 od 57 značajki iz skupa statički filtriranog signala. Rezultati na simuliranim signalima demonstrirali su nižu važnost značajki, što se može objasniti umjetno dodanim šumom iz raznih izvora, koji je posljedično doveo do neoptimalne konvergencije procesa prosijavanja.

U svrhu dodatne validacije pristupa, istrenirana su dva različita modela strojnog učenja na podacima: metoda potpornih vektora s kubnim kernelom i ansambl stabala odluke (*Bagged trees*). Rezultati pokazuju da dodavanje značajki iz podskupova funkcija intrinzičnih modova i signala pomaknute frekvencije uvelike povećavaju točnost, preciznost i opoziv istreniranih modela. Primjerice, u slučaju metode potpornih vektora na snimljenim signalima, točnost modela poraste sa 69% u slučaju značajki iz podskupa statički filtriranih signala, pa sve do 76% kada se dodaju i ostale značajke. Situacija je još bolja za simulirane signale (odnos signal-šum od -24,4 dB): točnost poraste s 92% na 97% kada se značajkama statički filtrirane verzije dodaju ostali parametri.

Kao konačan zaključak ovog istraživanja može se navesti veliki potencijal korištenja empirijske dekompozicije modova i promjene frekvencije kao metoda predobrade te psihoakustičkih parametara kao značajki u strojnom učenju primijenjene na zvučnom signalu otkucaja srca fetusa. Ovakav pristup otvara put prema mogućnosti uvođenja softverskih rješenja u realnom vremenu koje bi služile kao dio sustava za dugoročno praćenje stanja fetusa.

KLJUČNE RIJEČI

empirijska dekompozicija modova, psihoakustika, selekcija značajki, zvuk srca fetusa, strojno učenje, promjena frekvencije, fetalna fonokardiografija

Contents

Chapter 1

Introduction

1.1 Background

Fetal heart rate (FHR) monitoring is the most common procedure used to ascertain the health and welfare of the fetus through the assessment of the rhythm and rate of its heartbeat, principally in establishing fast diagnosis in the case of complications during pregnancy [1, 2]. Several techniques [3] are commonly used for non-invasive monitoring: cardiotocography (CTG), Doppler echocardiography (FDE), fetal electrocardiography (FECG), fetal phonocardiography (FPCG), fetal photo-plethysmography (FPPG) and fetal magnetocardiography (FMCG). A comparison of their advantages and disadvantages can be found in Table 1.1 [4, 5].

As the technological and medical breakthroughs in recent years aimed towards improving pre-emptive and personalized healthcare through the Internet of Things (IoT) and wearable technology are gaining an increasing amount of traction [6], studies concerning fetal healthcare are continuously assessing the possibility of introducing remote and/or at-home monitoring, even though the standards of care do not currently allow for that [7]. In any case, remote monitoring has shown benefits in prenatal care, both in high risk [8, 9] and low risk [10] pregnancies. Advances in electronics have also made possible the development of low-cost devices with modern sensoric and communications capabilities [11]. For example, the general and every-day usage of smartphone devices makes them a logical choice for remote monitoring, especially since they have shown their capacity to be exploited as biomedical tools in personalized healthcare monitoring [12, 13, 14].

Regarding clinical fetal health assessment, Doppler-based technologies have remained the most widespread methods for monitoring fetal cardiac activities [15]. The usage of these devices at home has never been recommended [16]: even though the precise correlation between the amount of ultrasonic energy and the fetal well-being was never found, there are a number of studies that have raised some concerns regarding the impact of long-term exposure on animals [17, 18]. Additionally, it has been shown that thermal effects have a stronger impact on possible

fetal damage than non-thermal [19], with the recommendation of keeping the exposure “as low as reasonably achievable”.

The aforementioned facts make fetal phonocardiography a very interesting candidate for remote and long-term monitoring of fetal health as it is non-invasive and simple [20], with the requirement for only a single probe during signal acquisition [21]. Additionally, the increased sensitivity and usage of smartphone-embedded microphones in assessing cardiovascular activity and respiratory and lung health [22] also open the question whether fetal phonocardiography can be employed in extracting fetal heartbeat information.

Table 1.1. *Non-invasive FHR techniques: advantages and limitations*

Technique	Advantages	Limitations	Energy type
FDE	Easy to use, focused source localization	Not suitable for continuous monitoring or home use	Ultrasonic
FPPG	Suitable for long-term recording, harmless	Strong fetal position and probe separation dependency	Optical
CTG	Provides information on fetal heart signal and uterine contractions	Complex design requirements, difficult to interpret	Ultrasonic
FMCG	No problems with impedance boundaries	Large size, complex system design	Magnetic
FECG	Easy to use, suitable for long-term monitoring	Complex design requirements	Electrical
FPCG	Easy to use, suitable for long-term monitoring	Small signal-to-noise ratio, sensor location dependency	Acoustic

1.2 Fetal phonocardiography

First reports of monitoring of the fetal heart sound signal can be traced back to the 17th century [23], with little attention given to the process of listening (also called fetal auscultation) until the early 1800s [24]. The most common instrument for auscultation at the time was the Pinard stethoscope, a horn-like passive device that is placed on the woman’s abdomen on one side while a general practitioner listens from the other side. The device itself is still being used in

developing countries and low-tech environments [25], even though it suffers from various drawbacks, spanning from long training time for practitioners to fairly subjective read-outs.

With the emergence of electronic devices, fetal phonocardiography as a way of capturing and recording fetal heartbeat sounds could be employed to give a more objective assessment of the signal [26]. It can give some critical information regarding the fetal health, as well as detect heart murmurs, extrasystole, split effect, intrauterine growth retardation and other abnormalities that cannot be detected through other techniques [27, 21]. However, the method itself suffers from various drawbacks. The fetal heartbeat signal is very weak as the heart is not fully developed, making the acquisition challenging [27]. Additionally, the acoustic environment of the fetal heart sound incorporates a number of noise sources, reducing the signal-to-noise ratio (SNR) by a substantial amount [28]. The acoustic signal corrupted by noise then moves through a rather complicated arrangement of acoustic layers, each of them attenuating the sound energy by reflecting it on the boundaries with mismatched acoustic impedances [29].

1.3 Fetal heart sound signal

The sounds of a fetal heart are produced by the movement and reshaping of the heart muscles during a cardiac cycle, including the actions of the myocardium and the motion of valve cusps that control the flow of blood [30]. The details on the temporal and spectral content of the fetal heart sounds are well established [31]. A healthy adult heart produces 4 sounds; however the 3rd and 4th heart sounds are considered almost undetectable in a fetus [29]. The two present fetal heart sounds can be found in the lowest segment of the audible spectrum: the largest amount of the fetal heart sound energy is positioned between 20 and 200 Hz [28]. The first heart sound, labeled S1, is created by the asynchronous closure of mitral and tricuspid valves during systole. The second heart sound, labeled S2, is a result of vibrations generated by asynchronous closure of aortic and pulmonary valves during diastole [26, 30].

The range for heart rate in the case of a healthy fetus usually spans from 120 to 180 bpm, with common accelerations and decelerations occurring both in short-term and long-term [32]. As the structural defects of the heart are oftentimes reflected in the sounds and vibrations it produces [33], the specific characteristics of the FPCG signal can be indicative of fetal health. Besides determining FHR from the temporal distance between two adjacent S1 sounds (interbeat interval), various signal characteristics of a fetal heart sound, such as bandwidth and

center frequencies of S1 and S2 sounds [30], can act as important variables in determining cardiovascular conditions.

1.4 Known FPCG analysis methods

***Table 1.2**. Alternative analysis methods for FPCG signals*

Analysis method	Characteristics
Hilbert transform [34]	A linear operator transforming the original signal into an analytic signal, a complex-valued function without negative frequency components. Envelope and instantaneous phase calculations are possible through combination with the original function.
Wavelet transform [35, 36, 37]	A transformation with a window function which can be expanded or compressed to capture both low frequency and high frequency components of the signal. Used for denoising and analysis of non-stationary signals.
Matching pursuit [38]	A greedy algorithm providing sparse signal representation and projecting it over a dictionary in order to find the best match. Used for decomposition and identification of heart murmurs and S1 width.
Multibeat autocorrelation [39]	A process of matching a phonocardiographic signal with the sliding window containing multiple baseline fetal beats. It is robust to noise but dependent on window size, with strong accelerations or decelerations decreasing accuracy.
Cycling frequency spectrum (CFS) analysis [40]	Assuming heart sounds are dominant cyclic components at the heart rate in adjacent fetal cardiac cycles. Detection of basic cycle frequency of the sequence yields FHR.
Eigenvalue decomposition [41, 42]	Requiring multichannel FPCG, where a matrix of eigenvalues and eigenvectors is generated from the data. Extraction of fetal heart signal is possible through channel-to-channel inter-correlations.
Rule-based extraction [43, 44]	Logic blocks generated from a series of expert-based rules concerning fetal heart sound characteristics.

The decomposition of FPCG signals and extraction of important parameters are possible through several analysis methods given in Table 1.2.

The research reported in [39] has shown the comparison of hit rates (ratio of the number of detected beats and the missing ones estimated by baseline) for wavelet transform, matching

pursuit and multibeat autocorrelation. A combination of the aforementioned methods increased the hit rate by 8.2% compared to utilization of a single method, however only in the case of high levels of noise. In another work [40], a simulated dataset was used for the comparison between CFS analysis, one rule-based method and one advanced combination of approaches (wavelet denoising + rule-based system). The CFS analysis outperformed the other methods by 25.2% in the cases of very noisy signals (SNR < 20 dB) but performed worse in the case of high SNR.

1.5 Structure of the thesis

The goals of this thesis are to:

1. Introduce empirical mode decomposition (EMD) and pitch (frequency) shifting (PS) as preprocessing methods for FPCG signal.
2. Extract a set of objective and perceptual (psychoacoustic) features from the frequency filtered signal, as well as signal versions processed by EMD and PS.
3. Employ machine learning principles, including feature ranking and importance, as well as model training and validation, in order to assess the importance of EMD and PS in classification of FPCG signals. Additionally, the impact of features based on psychoacoustics is also evaluated.

The structure of the thesis closely follows the set goals by first explaining the used preprocessing methods, data acquisition, as well as data cleaning, segmentation and feature extraction mechanisms. Afterwards, the thesis moves to first displaying and then discussing the results of employed machine learning processes utilized on two different datasets (one recorded and one simulated). Finally, a comprehensive conclusion is given.

Chapter 2

Methodology

2.1 Empirical mode decomposition

Noisiness, nonstationarity and nonlinearity are innate characteristics of biomedical signals. In comparison to time-frequency based decompositions that require beforehand fixed bases with linearity assumption (such as wavelet or Fourier transform) [45], empirical mode decomposition (EMD) breaks down the nonlinear and non-stationary biomedical signal into intrinsic mode functions (IMFs) based on the local characteristic time scale of the data [46], rendering it suitable for time-frequency domain analysis [45]. The applications of EMD for enabling further insight into biomedical signals are well-known and include works such as the denoising of respiratory signals [47], analysis of thoracic crackles [48] and electroencephalographic (EEG) signals [49]. Furthermore, the method was already shown to be appropriate for application on adult heart sound signals [50, 51]. This makes EMD a reasonable choice for the analysis method in fetal heartbeat detection realized through machine learning, as IMFs can highlight some hidden characteristics of the signal and thus increase the dimensionality in the classification process.

EMD can expose possibly important pieces of information from the signal by decomposing the fetal heartbeat sound signal into "simpler" modes of oscillation. The method is yet to exhibit higher levels of applicability in the context of fetal heart signals: research on detection and classification of essential perinatal parameters through FDE, FECG and CTG has been gaining speed only recently [52, 53, 54], while very limited literature is available in the case of FPCG signals [55, 56, 5].

Figure 2.1 depicts the EMD methodology. The method functions as an iterative process that extracts intrinsic mode functions by recursively subtracting the mean of calculated upper and lower envelopes from the signal representation in the specific iteration [57, 46]. The envelopes are acquired through the interpolation of extrema and contain high frequencies of the IMF candidate, meaning iterative subtraction will leave only lower frequencies, until certain conditions for IMF generation are met. The stored IMF is then subtracted from the original

signal, restarting the process of envelope calculation/subtraction and the generation of a new IMF. That is why every upcoming IMF contains lower frequencies than the IMF generated before it. This approach enables an insight into different modes of oscillation of the biological signal, supporting the benefits of adding features extracted from the IMFs to the dataset [58, 50].

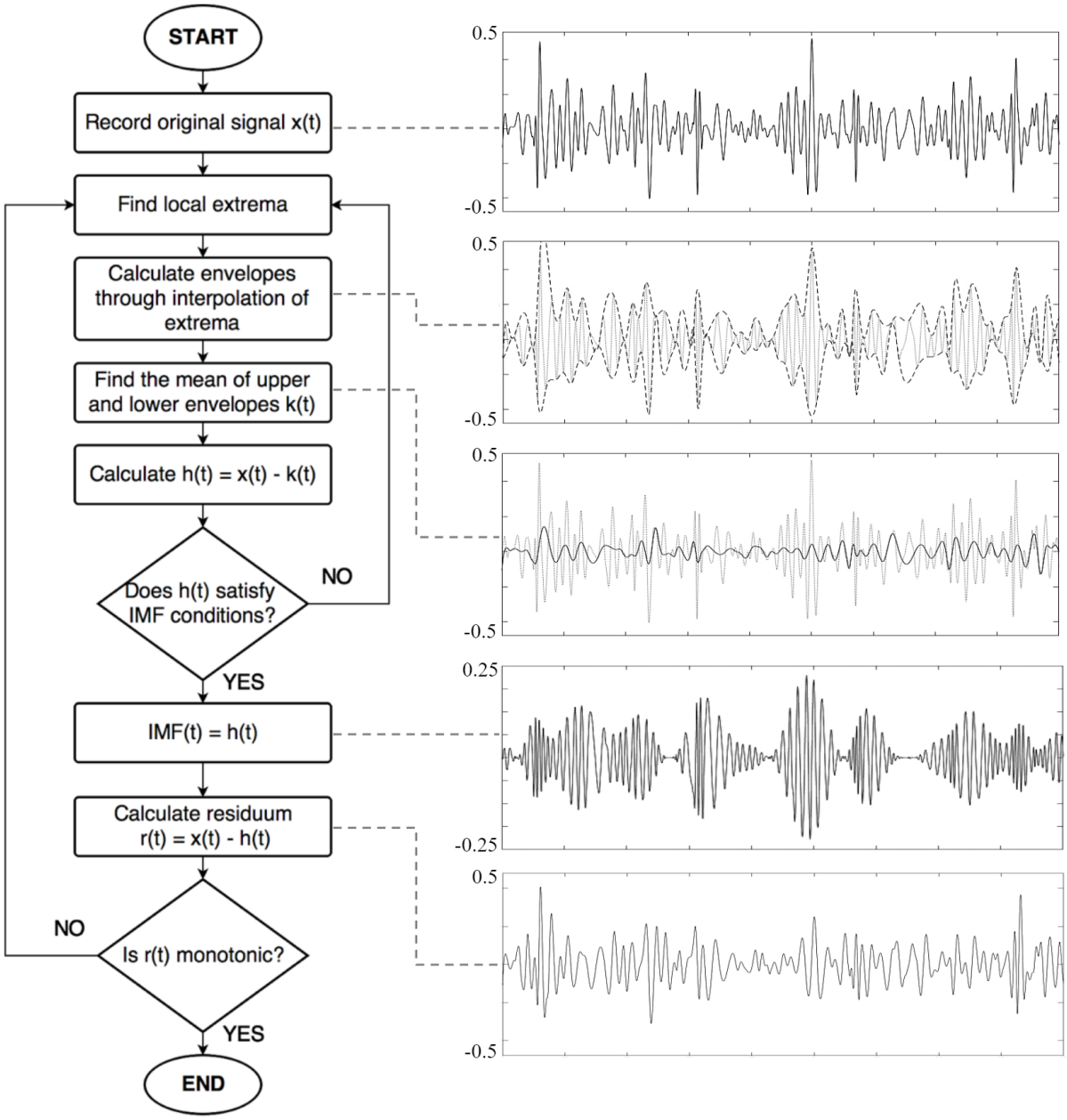

***Figure 2.1**. EMD methodology*

There are two conditions that a sequence resulting from the EMD sifting process needs to satisfy in order to be stored as an IMF:

1. The absolute difference between the number of zero-crossings and extrema needs to be 0 or 1.

2. The mean value of the envelopes defined by local minima and local maxima has to be 0.

As the conditions are quite strict, too many iterations in the sifting process can result in losing relevant amplitude information and problems with computational power. The solution for this was the introduction of some sort of stopping criteria in the IMF generation process to limit the number of iterations, especially if subsequent iterations of the sifting process yield only minor differences [46].

Even though the resulting IMFs exhibit much clearer modes of oscillation and enable the extrapolation of the relevant data needed for signal classification, analysis through EMD can experience problems known as “mode mixing” and “spurious modes” [59]. Put simply, the inherent locality of the method can force extraction of very different modes of oscillation in one IMF and very similar oscillations in several modes, making the sifting process somewhat unstable. For example, this can make one important component of the signal present in three different IMFs but can also produce two distinct signal components in only one IMF. In the case of FPCG signals, this means that noise can remain a very strong factor in all IMFs, if not sifted properly. Several improvements have been proposed to the base method over time: Complex empirical mode decomposition (CEMD) [60], Ensemble Empirical Mode Decomposition (EEMD) [61], Complementary Ensemble Empirical Mode Decomposition (CEEMD) [62], Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) [63] and Improved Complete Ensemble EMD (ICEEMDAN) [64].

Even though EMD doesn’t have a complete and theoretically proven framework, a recent study [65] has presented a mathematical proof for the validity and robustness of the method.

2.2 Psychoacoustics

Psychoacoustics is the interdisciplinary field involving a number of different areas, spanning from acoustics, psychology, physics and biology to physiology and computer science [66]. The core idea revolves around the question of how humans perceive sound, separate from the objective standpoint of acoustic wave propagation. This makes it an important tool in many

disciplines, with applications spanning from speech recognition [67], emotion recognition [68], audio coding and transmission [69], predictive maintenance [70], to marine biology [71] and automotive industry [72]. On the other side, the literature on the usage of psychoacoustics in the detection of biomedical signals is limited, with only a handful of papers available [73, 74, 75].

2.2.1 Human hearing

The peripheral auditory system (Figure 2.2) consists of the outer ear, middle ear and inner ear [76].

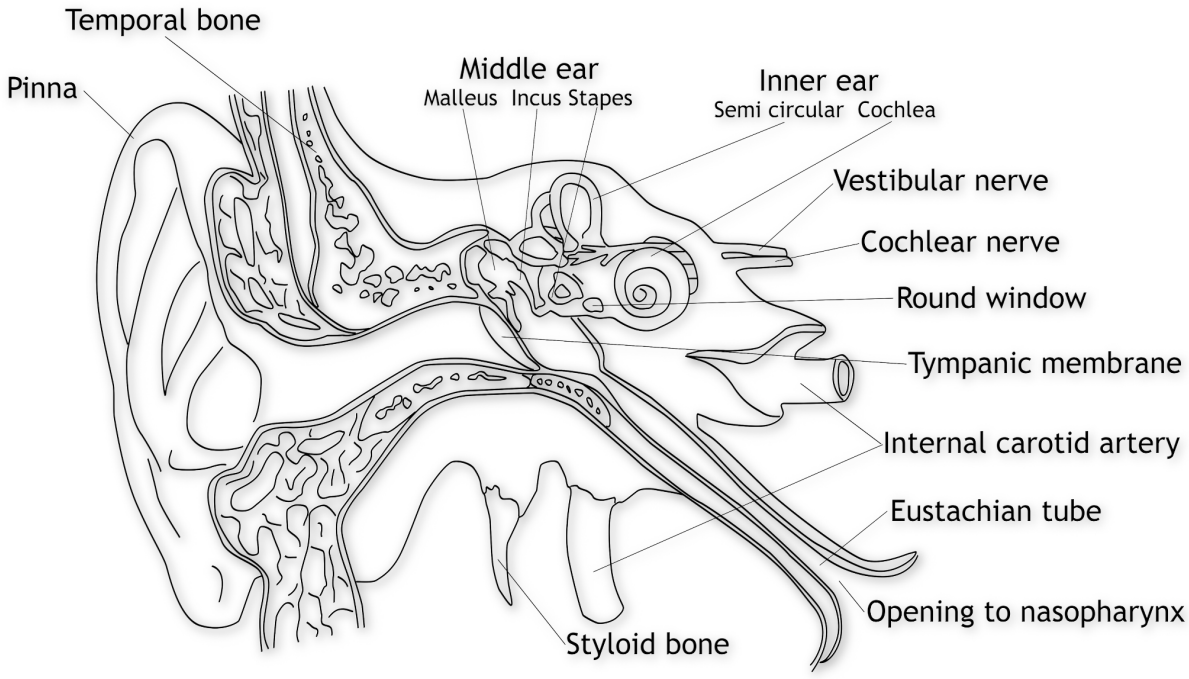

***Figure 2.2**. Human ear anatomy [77]*

The three components incorporate different functions:

1. Outer ear - sound waves are focused through the pinna (visible part of the ear) into the ear canal and towards the eardrum. The distinct shape of the pinna helps localize sounds from different directions, as the sound wave reflects and diffracts differently depending on the angle, both in vertical and horizontal planes [78].

2. Middle ear - three small bones, known as malleus, incus and stapes, are used to further transmit the vibrations captured by the eardrum. This part is critical for matching the

impedances between the air-filled environment as the source of the sound and the extracellular fluids utilized as the conduction medium in the inner ear [79].

3. Inner ear - after the impedance matching, the sound is transmitted through the semicircular canals to the cochlea, a spiral-shaped cavity used for converting acoustic vibrations into electrical neural activity [80]. The basilar membrane found in the cochlea is very important for human hearing, as it separates different frequency components [76].

The specific shape and biological principles of the basilar membrane enables it to be described simply as a bank of overlapping bandpass filters on a nonlinear scale [81, 76], as illustrated in Figure 2.3. Several models, such as the one explained in [82, 83], are used to estimate the perceived frequency, called pitch.

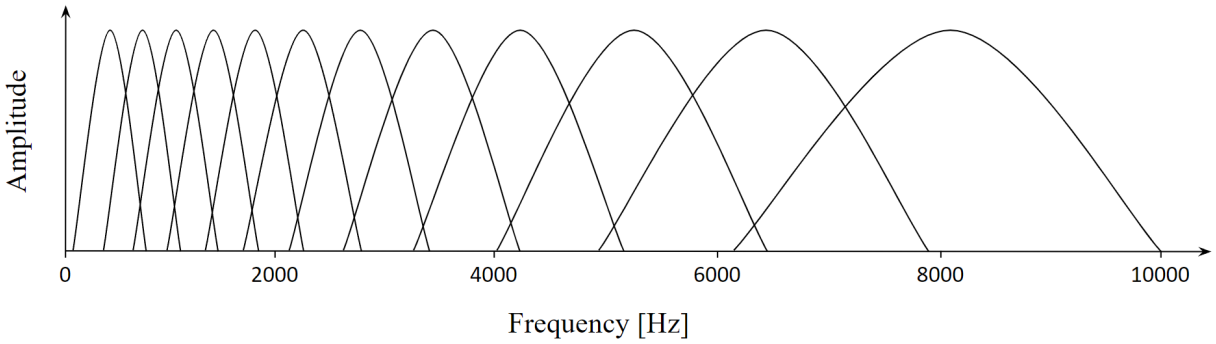

Figure 2.3 *Auditory filters*

Additionally, human hearing also perceives sound intensity in a nonlinear fashion [84], with perceived sound levels (loudness) being dependent not only on the sound intensity, but also frequency and duration [85].

2.2.2 Psychoacoustic feature extraction

In order to extract suitable FPCG signal features that would enable “machine hearing”, as defined by [86], some kind of perceptual modeling of features through signal processing needs to be utilized [87].

Another consideration is a preprocessing method that would make the signal more suitable for psychoacoustic modeling. As a rule of thumb, sensitivity to frequency changes rises towards the higher parts of the spectrum (approximately between 1 and 2 kHz) and starts to deteriorate

after 4 kHz [88]. Since most of the energy of an FPCG signal, as well as biomedical signals in general [89] can be found in the low frequency range, it was considered worthwhile to shift the frequency distribution to higher regions of the spectrum, where psychoacoustics can be utilized in a more feasible manner [90]. Reports on combining pitch (frequency) shifting and psychoacoustics for the classification of biomedical signals have not been found in literature: as FPCG sounds can be perceived given enough preprocessing [91], this research has decided the assess the impact of the aforementioned combination on the quality of the automatic classification process.

2.2.3 Pitch shifting

There are several methods for the implementation of pitch shifting, which can be roughly organized into two categories:

1. Time domain methods, incorporating overlap-add (OLA) and its derivatives Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) [92].

2. Frequency domain methods, including the commonly used phase vocoder and the improvements of the method [93].

Compared to time domain methods that are known for changing the pitch of namely speech signals, phase vocoding can be employed for polyphonic signals and large shifts [94], making it more suitable for the nonlinear and nonstationary FPCG signal.

The working principles of the phase vocoder algorithm are given here, with the graphic flowchart is given in Figure 2.4 [95]:

1. Audio signal is segmented into smaller frames with a high overlap factor. Each of the smaller segments is multiplied by an analysis window, such as Hamming and Blackman windows. This is to achieve better frequency resolution than using a rectangular window.

2. Fast Fourier Transform (FFT) is employed to calculate the magnitude and phase spectrum of the segment. This concludes the analysis stage of the algorithm.

3. During the synthesis stage, the overlap-add procedure of adjacent segments is achieved with different overlap values compared to the analysis phase. This adjustment changes

the phase difference between the successive frames, so instantaneous frequency calculation with unwrapping and accumulation needs to be done in order to avoid discontinuities [96].

4. After the instantaneous frequency is used to adjust the phase, the newly calculated correction can be applied to the original magnitude spectrum and returned to the time domain through Inverse Fast Fourier Transform (IFFT).

5. The phase-adjusted and recalculated segments are overlap-added with hop size values depending on the required pitch shift parameter.

6. The resulting signal is resampled with the pitch shift factor to achieve the same length as the original signal, also effectively shifting the frequency spectrum during the process.

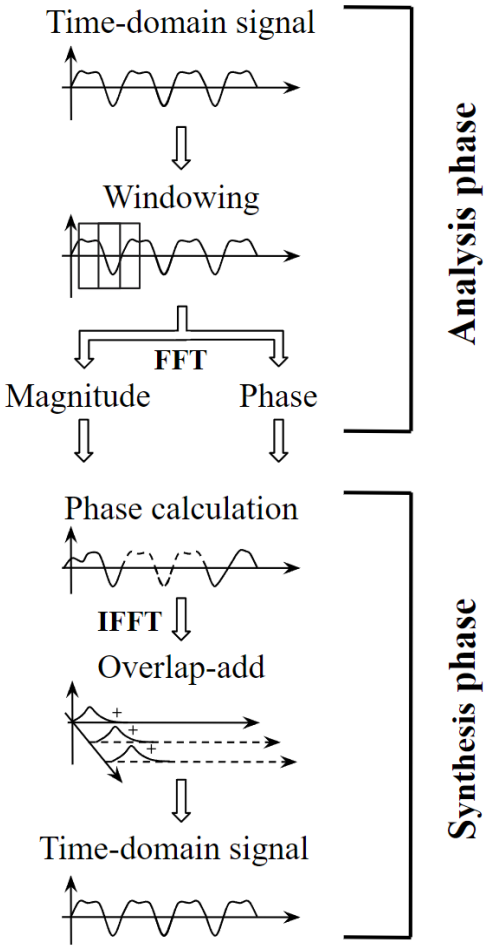

***Figure 2.4**. Pitch shifting algorithm flowchart*

2.3 Machine learning

Machine learning can be seen as a part or a subset of artificial intelligence (AI) [97], which exact definition is a subject of much discussion [98]. One such dictionary-based annotation states that artificial intelligence is "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" [99].

Machine learning algorithms can be seen, in a sense, as a field of computer science concerned with the question of how to construct computer programs that automatically improve with the usage of data/experience [100]. The discipline has the vast potential and application scope in providing insight, prediction and classification benefits in a huge number of industries, including agriculture [101], IoT [102], digital security [103], medicine [104], speech recognition [105] and more. In recent years, it has also become a powerful tool in classification and diagnostics of biomedical signals, involving lung sounds analysis [106], EEG pathology [107], heart disease prediction [108] and chronic kidney disease diagnosis [109]. Fetal heartbeat signal classification is not an exception here, as numerous studies have employed machine learning for the task [110, 111, 112].

2.3.1 Supervised learning

The most common scenario in machine learning is supervised learning, where the learning algorithm receives a number of labeled data points which it then uses for adaptation of internal states of the prediction functions [113]. New data, the one that the algorithm has not used during the training process, can then be classified with more or less success: a properly designed machine learning model that used sufficient amounts of satisfactory distributed data for training will perform well on new data, while a model trained on data of poor quality, non-representative distribution, subpar choice of features or insufficient size will lead to poor performance in classification and prediction on previously unseen data points [114].

Dozens of algorithms are available for machine learning, from simpler ones (namely interpretation-wise) such as logistic regression, decision trees, K-nearest neighbours to more complicated ones like support vector machines and multilayer perceptrons. An illustrative example of how a support vector machine (SVM) model separates the data in two dimensions is given in Figure 2.5.

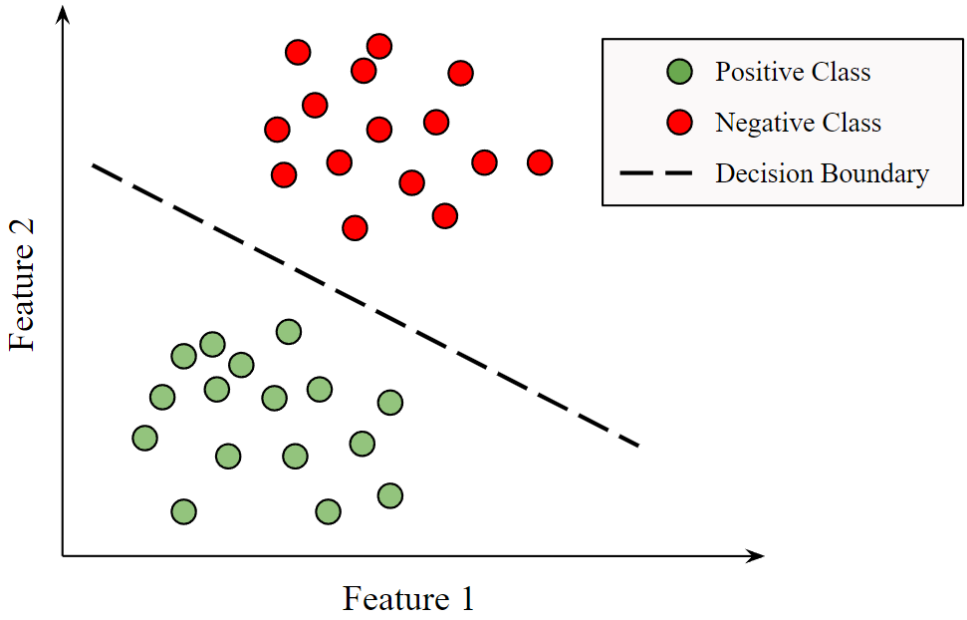

***Figure 2.5**. Decision boundary of an SVM model*

2.3.2 Thesis approach

This study has chosen machine learning principles as a robust approach not only in determining the quality of a potential machine learning model generated from the FPCG data, but also as a powerful way of assessing the importance and impact of the preprocessing methods and feature extraction mechanisms on the classification of FPCG data.

Specifically, data used in this research was organized so it produces two classes: signal windows containing S1 sounds (class 1) and signal windows not containing S1 sounds (class 0). Even though using raw FPCG signal data in machine learning is possible, extraction of meaningful and applicable features is highly encouraged [115]. In the context of FPCG signals, this can include common statistical or audio features as a baseline (explained further in the text) but can also utilize psychoacoustic feature extraction mechanisms. As these signal descriptors can be extracted from different FPCG signal representations (in this case: raw signal, frequency filtered signal, EMD-processed signal and PS-processed signal), the impact of particular features can be seen by including and excluding them from the training process. Additionally, machine learning employs several ways to make univariate and multivariate feature selection and ranking [116, 117]. The most appropriate techniques were used as feature assessment tools in this work.

2.4 Data preparation

The main part of this research was to determine the importance and impact of FPCG signals classification if the data is preprocessed with EMD and pitch shifting, compared to the original recorded data and data preprocessed with a bandpass filter that captures the majority of FPCG signal energy. Two different tests scenarios were considered:

1. Raw data vs bandpass filtered data vs EMD-filtered data - this was achieved through the usage of a custom recorded dataset and feature extraction based on objective audio-based features [118, 56]. A set of features was extracted from the signals and a series of machine learning processes were applied to yield a proper insight into the importance of specific extracted characteristics and through them, the preprocessing methods. This is labeled as Scenario A.

2. Bandpass filtered data vs EMD-filtered data vs PS-processed data - similar to the aforementioned approach in terms of feature organization, but with the additional step of employing psychoacoustic feature extraction on bandpass filtered and pitch shifted signals. Moreover, the original dataset was further expanded and an additional simulated dataset of FPCG signals was introduced into the analysis. This approach was marked as Scenario B.

In both cases, audio-based features, usually consisting of descriptors for objective and perceptive characterization of sound [119, 120], were the baseline used in the generation of the feature set. These features span from simple statistical descriptors in the time domain (e.g. median and standard deviation of amplitude) to more complicated and hardly interpretable expressions in psychoacoustics. Since these approaches have often been applied to extraction of adult heart sound parameters [121, 122], their introduction as an input to the machine learning model was a logical step in evaluating the potential of classifying the presence of fetal heartbeat in an audio recording. Combining audio features extracted from various signals with different preprocessing steps allowed for a meaningful comparison between the feature groups and provide further insight into the impact of a specific method.

In order to alleviate the potential naming confusion for datasets (both raw collected audio and Doppler data time series, as well as matrices of extracted features and labels, were annotated as datasets), the following text would aim to be as clear as possible by introducing the prefix “raw”

for recorded and simulated data consisting of time series signals; and prefix “extracted” for the calculated features vectors and corresponding labels as a preparation for the machine learning processes. This was introduced in strategic places in the following sections to increase clarity.

2.4.1 Custom dataset collection

The custom raw dataset was recorded on 8 pregnant women using a precise electret condenser measurement microphone Behringer ECM8000 for recording FPCG signal and a portable ultrasound Doppler device (Sonotrax Lite by Edan Instruments) with a 2 MHz probe. The latter was employed in parallel with the microphone and utilized for the generation of “ground truth”, i.e. accurate event timings [123]. The amplitude resolution and the sampling frequency taken for both microphone and Doppler recordings were 24 bits and 48 kHz respectively, with each recording having length between 250 and 360 seconds. 4 characteristic positions [124] around the woman's belly button, as shown in Figure 2.6, were used for recording in equal time periods.

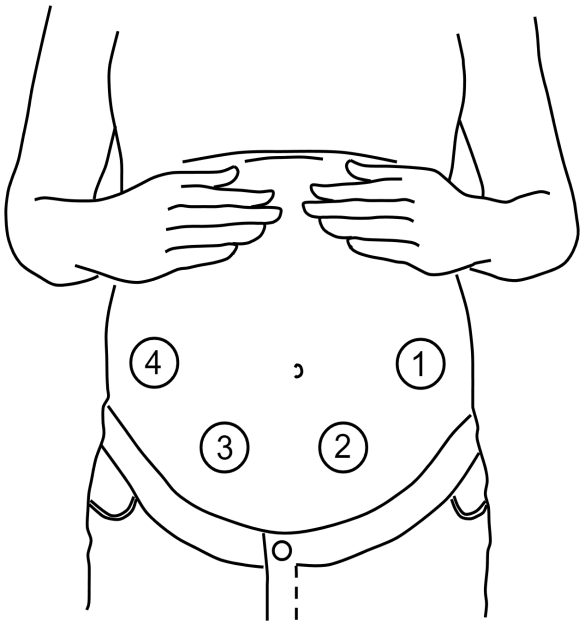

Figure 2.6*. Characteristics positions for placing the microphone diaphragm vertically against the skin. 1 - Left Occiput Posterior (LOP), 2 - Left Occiput Anterior (LOA), 3 - Right Occiput Anterior (ROA), 4 - Right Occiput Posterior (ROP) [125].*

Each recording was captured on a different pregnant woman ranging between the 27th and the 35th week of gestation during regular prenatal appointments. During the recording process, women were positioned on their back on a gynecological examination chair. Only one woman

was diagnosed with an anterior placenta. Every audio recording was made by lightly pressing the microphone diaphragm vertically against the skin without additional coupling mechanisms.

Even though the microphone was fixed in a specific position at the time, the Doppler device was monitored by an obstetrician at all times. The doctor was equipped with a set of professional studio headphones (AKG K240) and listened to the hardware-enabled audible Doppler signal representation in real time. Primarily due to the movement of the fetus, there was a need to continuously adjust the position of the Doppler device so the clearest possible signal could be recorded. The ultrasonic modality of the instrument showed the presence of signal peaks at the beginning of the systole (the S1 sound), making it a good choice for labeling of the audio segments before the machine learning process. Focusrite Scarlett 2i2 sound card was used for both the microphone and Doppler signal acquisition in parallel. A small number of representative audio segments, containing distinct events, such as very clear and prominent S1 sounds and strong fetal kicks, was used to calculate the delay between the two streams. This was achieved by measuring the distances between the peaks of microphone and Doppler signals for the aforementioned cases and was shown to be -61 ms ± 3 ms. The temporal lag of 61 ms was then applied on the microphone signal in order to synchronize the streams in time.

2.4.2 Ground truth label generation and filtering

In order to calculate S1 sound temporal positions from the Doppler signal, a signal processing subroutine was employed on every recording in the custom raw dataset. This is described in Figure 2.7.

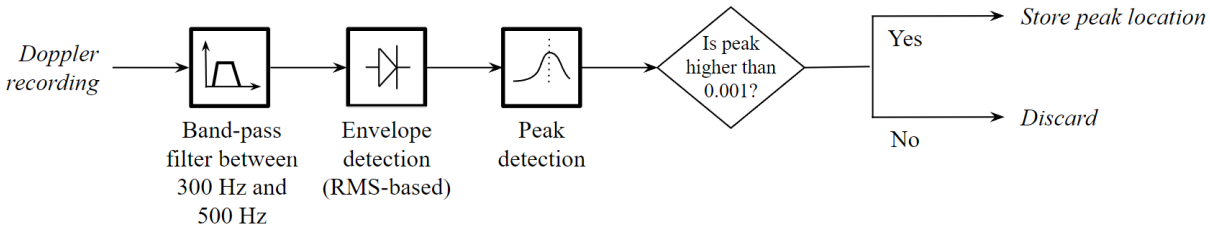

***Figure 2.7**: Extracting S1 locations from the Doppler signal*

After the Doppler signal was bandpassed, the output of the envelope detector based on root mean square (RMS) with 2000 samples in a moving window was differentiated before it was put into a peak detector (*findpeaks()* function in MATLAB) where the proper peak heights were empirically found to be valid if their amplitude was over 0.001 (taking the maximum signal

dynamics to be between -1 and 1). The additional condition was that the minimum distance between peaks cannot be below 333 ms, corresponding to an upper bound of normal FHR, taken as 180 bpm.

2.4.3 Label filtering

The dataset collected this way conformed to real-life scenarios, was sufficiently varied and provided an adequate estimate of the ground truth through the usage of a different data acquisition method. Nevertheless, the objective and subjective assessments of the recordings showed the emergence of two issues in data usability that required mitigation:

1. The presence of clear S1 peaks in the Doppler recordings did not guarantee the presence of the FPCG signal in the audio recording as they have very different working principles.

2. Different ultrasonic reflections and specific positioning of the fetus could “smear” the location of the S1 peak in the Doppler signal, increasing the uncertainty regarding the accuracy of the specific label location.

These ambiguities were alleviated by introducing 2 filtering procedures in the dataset cleaning process:

- First, due to fetal movements, the obstetrician was frequently required to reposition the Doppler device, resulting in a lot of the peaks within the Doppler signal being insufficiently clear. These points were not stored as positive label locations (as they did not satisfy the described peak selection criteria), although the heart sound might have been clearly present at that particular time. In order to avoid mislabeling S1 sounds as negatives, each area that had adjacent positive peaks further apart than 500 ms (<120 bpm) was considered to be missing peaks and removed from the dataset.

- Secondly, movement of the microphone to another position and placement further away from the fetal heart could have become a serious problem in mislabeling. This could have been detected by the characteristic lack of low frequency components, therefore every second of the recording that did not satisfy the condition of having 100x more spectral energy below 100 Hz compared to the energy above 100 Hz was discarded from the dataset.

2.4.4 Details on the test subjects

***Table 2.1**. Details on subjects from the custom dataset*

Rec number	Rec length [s]	Week of gestation	Placenta position	BMI	Subjective description of the sound	Approximated signal-to-noise ratio [dB]
1	249	30	anterior	25.5	Audible FHB in two positions, otherwise muffled and noisy	-9.62
2	258	28	anterior	24.6	Audible FHB in two positions, presence of maternal heartbeat	-1.52
3	263	32	anterior	26.4	Rather faint but somewhat audible FHB throughout the recording, presence of fetal kicks	-12.98
4	249	36	anterior	28.2	Barely audible FHB through most of the recording, presence of maternal heartbeat and constant fetal movement	-5.71
5	243	35	anterior	25.4	Somewhat audible FHB in one position, otherwise containing characteristic lack of low frequency sound implying no FHB presence	-8.36
6	360	32	posterior	24.4	Audible FHB for 30 seconds, after which the fetus moves. Characteristic lack of low frequencies	-9.59
7	312	34	anterior	28.4	Very faint to non-existent FHB sound for 40 seconds. Characteristic lack of low frequencies and a lot of movement	-10.54
8	267	27	anterior	30.0	Very muffled FHB in only a couple of instances, constant fetal movements	-14.8

Table 2.1 gives more details regarding the distribution of the custom raw dataset in respect to different subjects and recordings used in the research.

The quality of the signal was assessed through the approximation of the signal-to-noise ratio by measuring the average energy within 50 ms of the positive label (containing S1 sound and noise) and the average energy of the rest of the signal (containing only noise). As this gave a ratio of a noisy signal (S+N) to the overall noise (N), a simple algebraic adjustment was required to estimate the SNR. In any case, the custom raw dataset was not formed based on the quality of the recordings, but on the reliability of the ground truth found in the data and the variability of the produced FPCG signals. Taking into account the quality of the Doppler signal and recording conditions, those parts of the recordings in which the ground truth was well founded were selected programmatically through the usage of dataset cleaning procedures. In other words, this constructed a dataset with as diverse data as possible by only removing those data points with unclear ground truth (noisy Doppler recordings) and those for which it was meaningless to employ the classification procedure (signals with no fetal heartbeat present due to wrong microphone placement). As a result, a representative dataset was acquired, suitable for the utilization of machine learning principles used for assessment of specific preprocessing and feature extraction methods.

2.4.5 Simulated dataset

As an additional dataset for the validation of preprocessing based on empirical mode decomposition and pitch shifting, as well as psychoacoustic feature extraction, a simulated Fetal PCG Database [27] was employed.

The database consisted of simulated S1 and S2 wavelets corrupted by various amounts of noise that is to be expected within the fetal heartbeat sound environment. The presented noise was a combination of different vibrations generated by maternal body organs, fetal movements, surrounding environments and maternal heart sounds. In addition, white Gaussian noise was also introduced in the dataset. The dataset contained various recordings incorporating the baseline S1 and S2 sounds with variable SNR values, spanning from -26.7 dB to -4.4 dB. The sampling rate of the recordings was 1 kHz and the amplitude resolution was 16 bits.

The labeling of the dataset was achieved by utilizing the recording with the highest SNR of -4.4 dB. Simple bandpass filtering of the recording (cut-off frequencies of 50 and 150 Hz)

showed rather distinct S1 peaks in the recording, so a similar procedure to the label generation in the case of custom dataset was employed: the filtered signal had the differential of its envelope (RMS, moving average of 2000 samples) introduced to the peak detector that stored S1 locations that were at least 333 ms apart. Unlike the peak detection for custom dataset, this one did not include conditions for peak heights due to a more stable and controlled method of the simulated dataset generation.

2.4.6 Implementation details

The entire research was done through the usage of MATLAB programming platform and Jupyter Notebook accessed through Anaconda Python distribution platform. More concretely, MATLAB (version 2020a) was employed for calculating the entirety of data labeling, segmentation, preprocessing and feature extraction, while all feature ranking and selection procedures were done in Python. Regarding model training, most of it was achieved in Python as well, while only the two final classifiers with 10-fold cross-validation introduced in Scenario B were trained and assessed in MATLAB.

Chapter 3

Scenario A

This scenario, originally reported in [56], assessed the impact and importance of EMD-processed FPCG signals compared to raw and statically filtered signal versions. First 7 recordings of the custom recorded dataset were utilized to provide input data, with the 8th recording being originally discarded due to uncertainties concerning the fetal heart sound quality.

3.1 Preprocessing

In order to achieve faster results, the first step in data preprocessing was filtering the recordings with an antialiasing finite impulse response (FIR) filter (Kaiser window with 2400 points and the shape factor of 5) and downsampling them to 2 kHz, essentially limiting the highest available frequency in the spectrum to 1 kHz. The delay introduced by the antialiasing filter was compensated for. Such a downsampled signal with no additional preprocessing steps was used for audio feature extraction and it is referred to as “raw audio” in the further text.

A subsequent preprocessing step was to introduce band-pass filtering in order to only focus on the frequency band that contains the most of the FPCG energy, while removing other frequency components. An 8th order Butterworth filter with cut-off frequencies between 50 and 150 Hz was chosen for the task since it gave the most prominent S1 signal shapes. The signal version filtered in this way was used as an additional input to the feature extraction process. Having a separate set of features extracted from the filtered audio signal without assuming which set of features (the one based on raw signal or the one based on the filtered signal) would contribute more to the dimensionality of the feature set. On one hand, higher-frequency components of the raw audio might have contained predominant noise components that would reduce the quality of the extracted features and the performance of a trained classifier. On the other hand, those components might have also included information useful for prediction. For that reason, following the best practices from feature engineering and machine learning, it was decided for this Scenario to maintain both sets of features and rely on chosen methods for feature ranking

and selection to estimate relevance and usefulness of all available features. This signal obtained by additional band-pass filtering is called “filtered audio” in the following subsections.

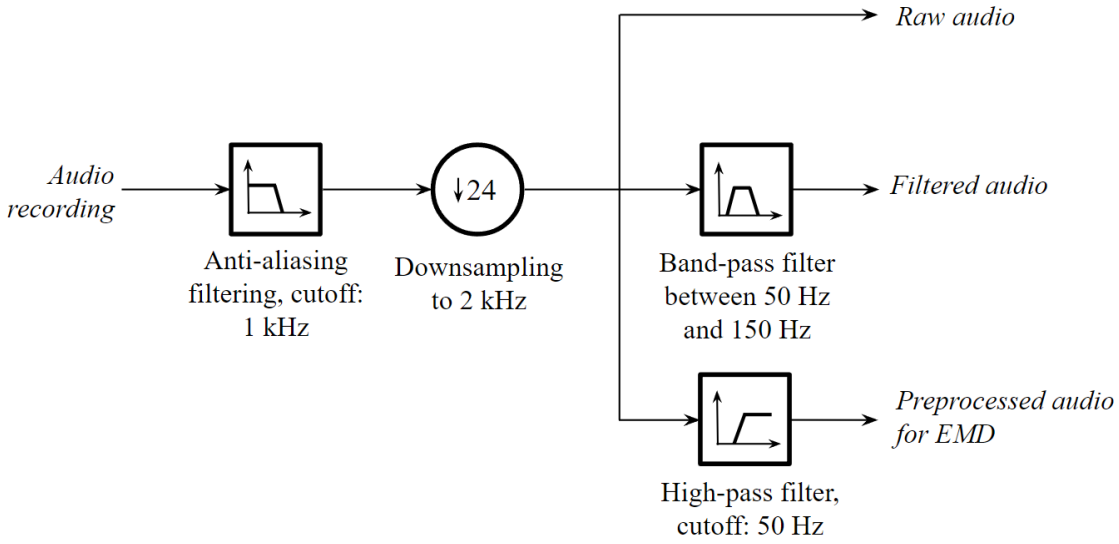

***Figure 3.1**. Preprocessing of the recorded signals in Scenario A*

Regarding the input introduced to the EMD method, visual waveform analysis and subjective listening tests conducted on the raw recordings have shown that the mother's heartbeat was present in some of the recordings, mainly in the lowest part of the spectrum (<50 Hz). It was therefore decided to preprocess the original audio with an 8th order Butterworth high-pass filter (50 Hz cut-off frequency). By utilizing this, the dominant spectrum of the maternal heart sound was suppressed in the FPCG signal, reducing the possibility of IMFs converging to maternal heart sounds instead of relevant fetal heart signals. Even though the higher harmonics of mother’s heartbeat sounds might be present well above the 50 Hz cut-off frequency, the analysis showed that removing frequencies higher than 50 Hz would also take out vital information of fetal heart sounds, which would in the end impede the potential of EMD to discover important data in the FPCG signal.

The whole preprocessing pipeline is depicted in Figure 3.1. MATLAB's built-in version of EMD (function *emd()*, MATLAB version 2020a) was used to calculate the first 3 IMFs from the original audio, with the following parameters:

- sift relative tolerance of 0.2,
- maximum sift iterations count of 100,

- maximum energy ratio of 20,
- cubic spline interpolation.

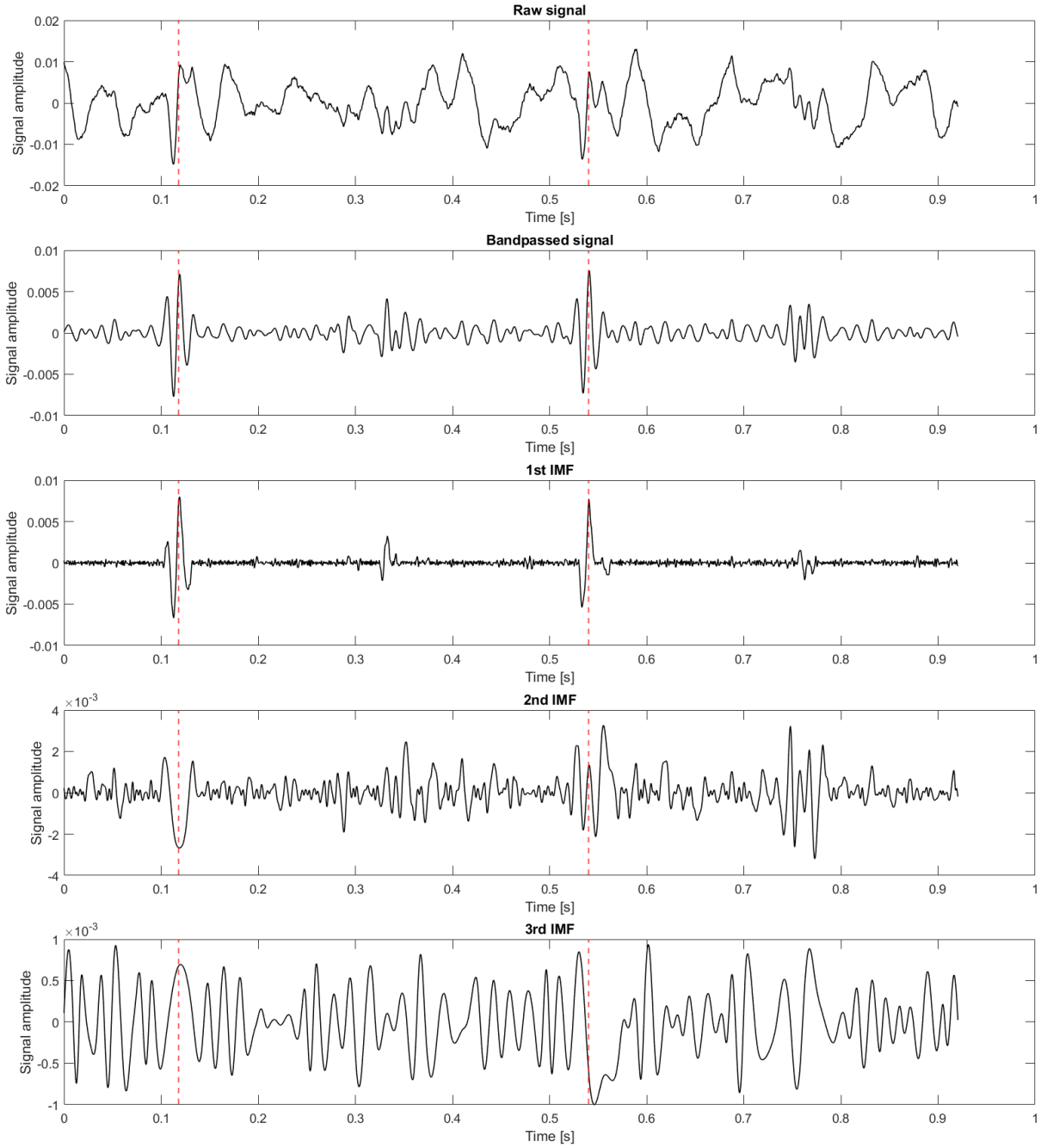

***Figure 3.2**. A representation of the signals used for feature extraction. The red line shows the location of the S1 sound, marked by the Doppler device.*

Spectral analysis of the recordings and the decomposition showed that the first 3 IMFs contained the most significant part of the signal energy above 50 Hz. In some instances during analysis, it was revealed that some of the signals would even meet decomposition stoppage

criteria between the 3rd and the 4th IMF, which would have made the consistent generation of the 4th IMF impossible to achieve through the entire dataset.

The final set included the following 5 signals gained from each of the 7 recordings: 1) “raw audio” (0-1000 Hz), 2) “filtered audio” (50-150 Hz) and 3) the first 3 IMFs (gained from audio with frequencies ranging between 50 and 1000 Hz). Figure 3.2 depicts the aforementioned signals, visually indicating the shift to lower frequencies moving from the first to the last IMF.

3.2 Feature extraction

Feature extraction was done by slicing the signals in 200 ms long windows with a hop of 50 ms (overlapping ratio 1:4). These hyperparameters were chosen empirically, with established window length enabling full encapsulation of S1 sound in the analysis window and selected hop size being adequate for generating a dataset of various positions of S1 within the window. Regarding the potential for precise FHR calculation, a hop size of 50 ms was chosen, as its resolution of 1200 bpm is several times higher than the frequency of S1 events, being 333 to 500 ms apart (i.e. 120-180 bpm). This provided sufficient time localization of S1 events for the purpose of determining the fetal heart rate that could be easily and accurately done through period estimation by averaging time differences between consecutive positive segments or counting detections over a fixed period of time. Nevertheless, if only a small number of detections are available for period estimation in real time applications, time resolution could simply be increased by decreasing the hop size in signal segmentation during prediction/classification stage.

Subsequently, each window was normalized with its root mean square value. This was done to make the classifier invariant to the amplitude, as different amplification levels for new recordings that require classification would influence the final result. Furthermore, various positions of the fetus might have also changed the signal amplitude, so it was decided that the specific amplitude values found in the analysis window were not reliable enough for classification. In other words, if a window of the audio signal contained a fetal heartbeat, it should have been detected without regard to the signal energy.

EMD was applied on each 200 ms window separately in order to calculate IMFs.

If the window's temporal range included the Doppler peak location, it was labeled as a positive data point, otherwise it was labeled negative. Two criteria for ground truth filtering were employed in the process, reducing the size of the extracted dataset considerably. As mentioned before, the reliability of the dataset was imperative for the proper utilization of machine learning processes, so the strict filtering behavior was both expected and desired. The first criteria (inconclusive Doppler peak location) removed the largest bulk of 21577 points, where the ground truth could not be determined with a sufficient degree of certainty. The second criteria (lack of low frequency content) deleted 6160 points. Finally, the points that had the label position very close to the window boundaries (<10% and >90%) were ignored, since their corresponding windows may not have contained the entire S1 signal. This removed 1901 incomplete observations.

The final cleaned dataset consisted of 7604 data points with reliable ground truth, containing 3235 positives and 4369 negatives. Such a dataset exhibited the appropriate variability and size for the purpose of applying methods for features analysis, since it was collected from subjects in different weeks of gestation and with various body mass indices (BMI). Its size was more than 50 times larger than the dimensionality of the feature space, and the distribution of positive and negative labels was rather balanced. The feature set extracted from each signal included 18 statistical features (arithmetical mean, standard deviation, coefficient of variation, maximum, minimum, root mean square, crest factor, 25th percentile, 75th percentile, 90th percentile, interquartile range, skewness, kurtosis, zero crossing rate, median divided by mean, 95th percentile divided by maximum, 5th percentile divided by minimum) [126, 127] and 9 spectral features (centroid, crest factor, decrease, entropy, flatness, kurtosis, skewness, slope and spread) [128, 129]. The spectral features were extracted by first calculating a Fast Fourier Transform (FFT) with the Nyquist frequency of 1000 Hz on the 200 ms window, resulting in 200 FFT points for a single-sided spectrum. Afterwards, the corresponding measure (e.g. centroid) was applied to the calculated spectrum. Considering that there were 5 signals gained from each recording, a total of 135 features were defined. The features were standardized before training the model.

The entire computational pipeline was applicable to work in real time due to filtering schemes and EMD being applied on the segmented 200 ms windows. Stopping criteria set in the EMD method enabled the overall calculation to be 4x faster than real-time on a mid-tier Intel Core i7 CPU from 2017.

3.3 Analysis steps

3.3.1 Correlation analysis

The first step of the exploratory data analysis was the creation and assessment of a correlation matrix. The focus of the analysis was on potential associations between audio and IMF-based features with the aim of better understanding their quantitative nature. Furthermore, the aim was to have a more informed interpretation of the usefulness and relevance of various features in the dataset.

Since all features were continuous numerical variables, focusing on linear associations between them was achieved through the usage of Pearson's product-moment correlation coefficient *r* [130], essentially rendered as normalized measurement of the covariance of two variables.

3.3.2 Feature ranking and selection

The second part of the exploratory analysis was estimating the relevance and usefulness of features in predicting the existence of a fetal S1 sound in the analysis window. The dataset used as an input was composed of audio features and IMF features, with the accompanying label for every window position containing valid data after the filtering process.

As performance of most classification models improves with the removal of highly correlated and/or irrelevant features [131], this research has opted for multiple feature ranking and selection methods to investigate the quality of IMF features in comparison to audio features and to the overall combination of all extracted descriptors. In its simplest form, ranking methods consider predictor characteristics and/or statistical relationship between the predictors and target variables [132, 133]. These approaches usually function as a tool for filtering features in a preprocessing step or as a baseline approach [134, 135, 136]. Univariate rankings based on two different statistical (filter) methods were chosen for the purpose of this research: first of all, the mutual information between each feature and the class label [137]; and secondly, the one-way analysis of variance used to examine whether values of each feature are statistically different for analysis windows that do and do not contain heartbeats [138].

Univariate rankings, however, give only one part of the overall picture in feature selection and relevance. As rankings of individual features obtained by these methods indicate relative

relevance, they may not always align with the measure of how useful a particular feature is when observed in the presence of other available features from the dataset. Two features may thus both be relevant for prediction, but highly correlated and redundant when appearing together. If one is included in the optimal feature subset, the second one becomes less useful [116, 132]. Taking that into account, there was a need to extend the univariate analysis with embedded and wrapper feature selection methods. Embedded methods refer to assessing feature usefulness as an inherent part of the training process with specific machine learning algorithms, while wrappers, in contrast, treat underlying algorithms as black boxes and employ them for testing various subsets of features and assess the relative usefulness of feature subsets [137].

The choice of the embedded feature selection method for the purpose of evaluating features in Scenario A was a random forest ensemble [139]. This was trained separately with audio features, IMF-based features and a combination of all features in order to show relative embedded feature rankings and the overall improvements in the classification accuracy achieved through the addition of IMF-based features. On the other side, recursive feature elimination with 5-fold cross-validation done with the random forest classifier [140] was chosen as an appropriate wrapper method.

3.3.3 Model training

The final step taken within this scenario was a comparison of several fetal heartbeat detectors trained and evaluated on different feature sets: 1) all 135 available features, 2) all 54 audio features, 3) all 81 IMF-based features, and 4) the subset of 48 features selected using the recursive feature elimination in the previous step. The purpose of the comparison was to show the predictive power of each feature set for various classifiers. Three additional classifiers were chosen for providing further insight into the results: a logistic regression model, a support vector machine, and a multi-layer perceptron. Since the provided dataset was rather balanced in terms of the two presented classes (containing and not containing fetal S1 sound), model accuracy (the ratio of correct predictions to total predictions) was chosen as the comparison metric.

3.4 Results

3.4.1 Correlation analysis

The correlation matrix consisted of Pearson's correlation coefficients calculated for each pair of features in the dataset. Correlations between pairs of audio features and correlations between pairs of IMF-based features have not been taken into account in this analysis, as they purely reflect correlations between different statistics on the same signals. Instead, the focus was placed on correlations between mixed pairs of audio and IMF-based features.

The correlation matrix suggested moderate correlations ($|r|$ around 0.5) between most of the statistical features of the filtered audio signal and most of the features of the 1st IMF. The correlations were weak to moderate for the 2nd IMF ($|r|$ between 0.25 and 0.4) and mostly weak between the features of the 3rd IMF and the statistical features of the filtered audio signal.

Similarly, but with lower Pearson's coefficients, weak correlations have been found between most of the statistical features of the raw audio signal and most of the IMF-based features. However, statistics of the filtered audio signal seem to be more correlated with IMFs than the statistics of the raw signal.

Regarding spectral features, weak correlations have also been found between some originating from the filtered audio signal and the 2nd IMF. The exception is a strong correlation ($r = 0.76$) between the spectral slope of the filtered audio signal and the spectral slope of the 2nd IMF. Such correlations between the spectral features of the filtered audio signal and the spectral features of the 1st and 3rd IMFs are less prominent ($|r|$ is between 0.22 and 0.47).

In total, it was observed that a relatively small number of feature pairs have exhibited correlations that would be considered moderate, or stronger. This suggested a high possibility of general improvement of the prediction power if a combination of audio and IMF-based characteristics was taken as a feature set.

3.4.2 Mutual information

Mutual information (MI) is a measure that quantifies how prominent is the reduction of uncertainty about one random variable given knowledge of another random variable. Feature ranking based on mutual information calculated for all pairs of each feature showed that 6 of

the top 10 and 12 of the top 20 features are calculated from IMFs. The first five features taken from the rankings were: (1) the 5th percentile divided by minimum of the filtered audio signal, (2) the 95th percentile divided by maximum of the filtered audio signal, (3) the 95th percentile divided by maximum of the raw audio signal, (4) the 5th percentile divided by minimum of the raw audio signal, and (5) the spectral slope of the 1st IMF. Figure 3.3 shows the mutual information for the top 20 features.

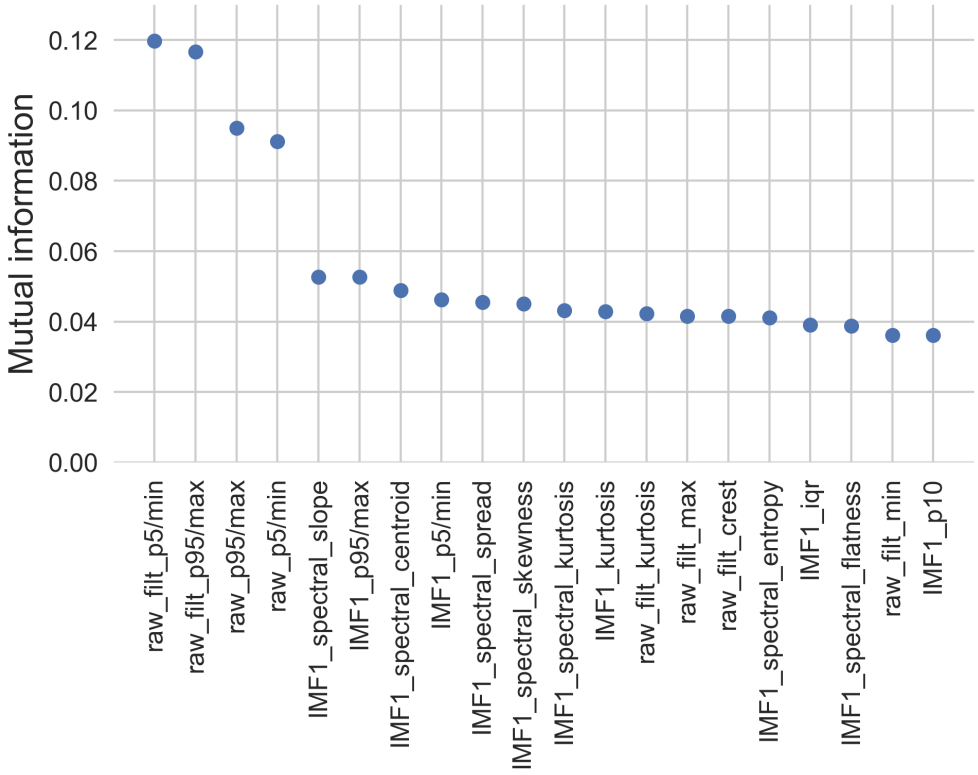

***Figure 3.3**. The top 20 features obtained by univariate ranking based on mutual information.*

3.4.3 ANOVA

One-way analysis of variance (one-way ANOVA) is a robust statistical method used to test for differences among means of two or more independent groups. In the context of feature ranking, ANOVA is computed between each feature and the target vector in order to examine how much the means of feature values differ if grouped by categorical target values.

The negative logarithm of ANOVA's p-values was used as a ranking metric. The results show that 7 of the top 10 and 12 of the top 20 features are calculated from IMFs. The first five best ranked features were: (1) the spectral flatness of the 1st IMF, (2) the kurtosis of the filtered

audio signal, (3) the 5th percentile divided by minimum of the 1st IMF, (4) the spectral entropy of the 1st IMF, and (5) the 10th percentile of the 1st IMF. Figure 3.4 shows the rankings for the top 20 features.

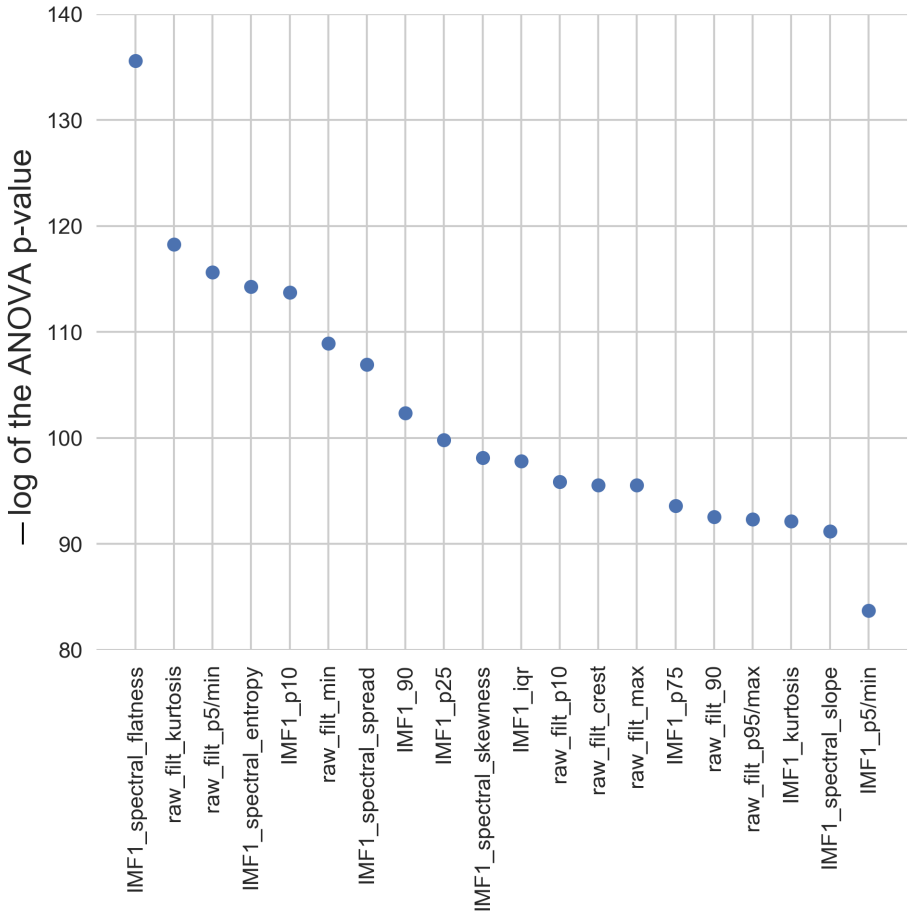

***Figure 3.4**. The top 20 features obtained by univariate ranking based on ANOVA. The y-axis shows the negative common logarithm of p-values.*

3.4.4 Embedded approach

The training process of a random forest classifier was chosen as a feature ranking method embedded in the inherent learning process of the classifier, with the goal of assessing the usefulness of particular features in the context of other available features. Random forest ensemble operates by constructing a multitude of decision trees that inherently choose features in the order of their importance during the training process [141].

100 decision trees were trained using the whole dataset. The overall feature importance obtained from the random forest showed that 6 of the top 10 and 10 of the top 20 features were calculated

from IMFs. The first five most relevant features in decreasing order of importance are: (1) the spectral skewness of the 1st IMF, (2) the spectral kurtosis of the 1st IMF, (3) the 5th percentile divided by minimum of the filtered audio signal, (4) the spectral spread of the raw audio signal, and (5) the spectral centroid of the 1st IMF. Figure 3.5 shows feature importance values obtained through the embedded approach.

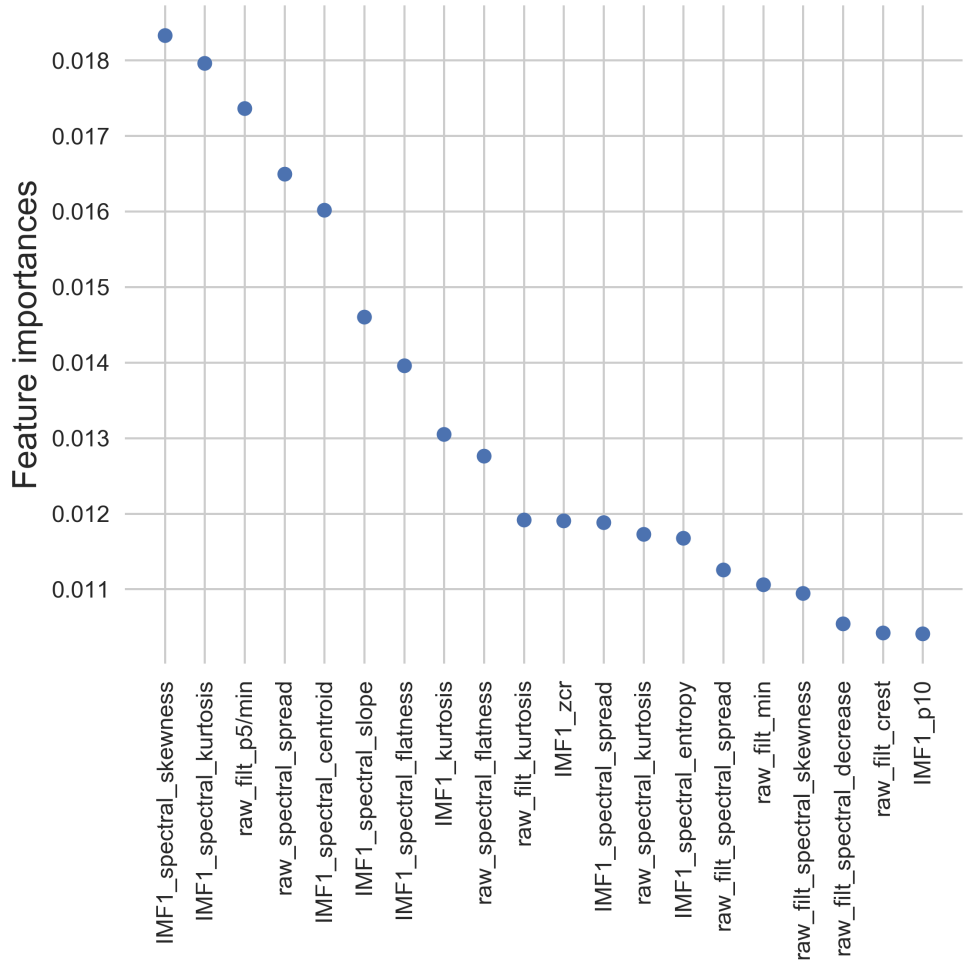

***Figure 3.5**. The top 20 features inherently obtained by the learning algorithm of the random forest classifier.*

The classifier was also used to compare prediction accuracy values between different subsets of available descriptors - features from audio, IMF-based features, and all available features combined. The five-fold cross validation showed that the prediction accuracy was 69.56% for audio features, 70.61% for IMF-based features, and 72.80% for all features.

3.4.5 Recursive feature elimination with cross-validation

The purpose of recursive feature elimination (RFE) with cross-validation was to select a subset of all features that leads to the highest predictive accuracy for the employed classifier. The result for the random forest ensemble with 100 trees contained 48 selected features, among which 22 were based on IMFs. Assessed by five-fold cross-validation, the accuracy improved from 72.80% for all 135 features to 74.13% for the 48 selected features confirming the Hughes phenomenon in machine learning [142].

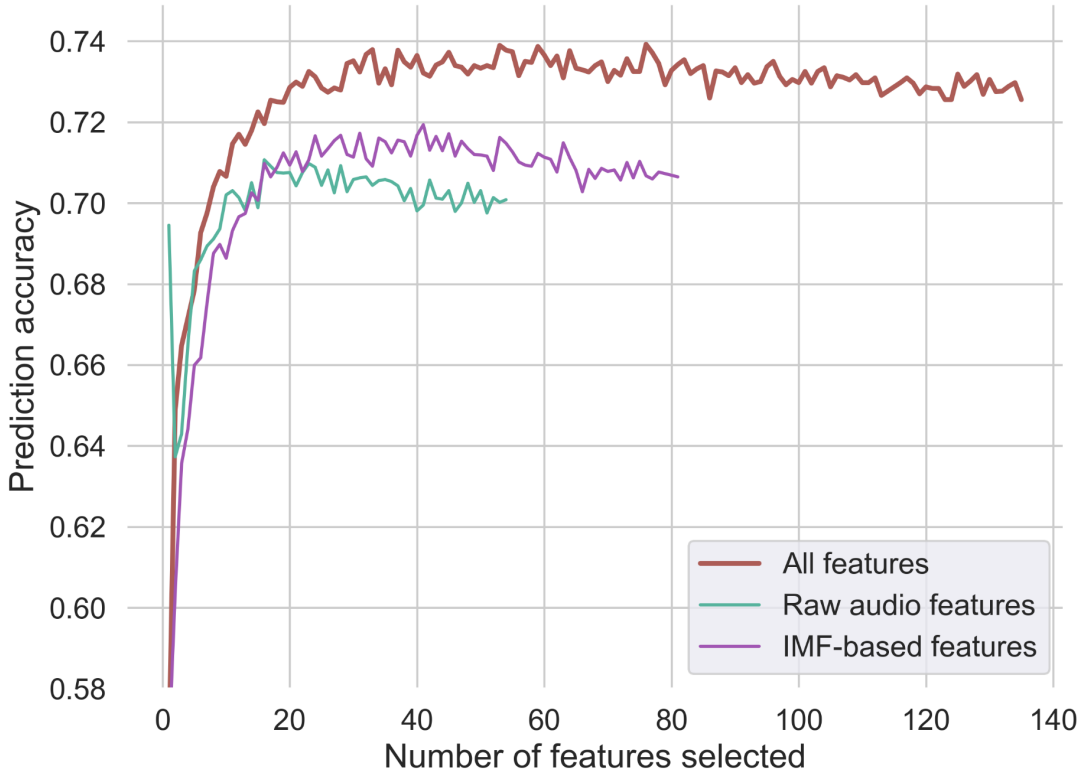

***Figure 3.6**. Prediction accuracies of random forest models trained with different feature subsets selected through recursive feature elimination.*

The prediction accuracy as a function of the number of selected features for 3 subsets (filtered, IMF-based and combined) is shown in Figure 3.6.

Besides running recursive feature elimination on the full set of combined audio and IMF-based features, the same procedure was performed separately on audio features and IMF-features. The random forest ensemble trained only on audio features reached the maximum accuracy of 70.84% for 28 selected audio features, while the same classifier trained only on IMF-based features reached the maximum accuracy of 72.04% for 33 selected IMF-based features. Since

the classifier trained on the combined set of features reached 74.13%, it was demonstrated that adding IMF-based features to the set of conventional audio features improved the predictive accuracy of the random forest by 3.26% before feature selection and by 4.57% with the selected features.

3.4.6 Comparison of trained models with different feature sets

In order to compare the predictive power of all features, their corresponding subsets and the selected features, the following classification models were used: a random forest ensemble with 100 trees, a logistic regression model with the L2-regularization, a linear support vector machine with the L2-regularization, and a multi-layer perceptron (MLP) containing three fully connected hidden layers with 1200, 300, and 150 neurons respectively. Hyperparameters of the classifiers were tuned manually before training the models. Table 3.1 shows mean accuracies for all these classification models after running 5-fold cross-validation for all four feature sets. The results showed that the predictive power of combined audio and IMF-based features was higher than that of particular feature subsets for every classifier trained. Moreover, the smaller feature set selected by the recursive elimination algorithm resulted with the highest accuracies for each classification model.

***Table 3.1**. Mean prediction accuracies obtained with 5-fold cross-validation for all combinations of feature sets and classification models.*

Classifier type	All features	Audio features	IMF-based features	Selected features
Random forest	72.80%	69.56%	70.61%	74.13%
Logistic Regression	66.87%	64.53%	64.82%	68.45%
Linear SVM	66.83%	64.56%	65.00%	68.31%
MLP	62.95%	60.84%	59.97%	71.12%

Chapter 4

Scenario B

This scenario consisted of a more extensive analysis of preprocessing methods and feature extraction mechanisms: given the small importance of audio features calculated from the raw signal, as demonstrated in Scenario A, the methodology was adapted to include the assessment of audio and psychoacoustic features extracted from the bandpass filtered, EMD-processed and PS-processed versions of the FPCG signal.

4.1 Preprocessing

Data preprocessing was done in a similar fashion to the one stated in the Scenario A section, however downsampling factor with moving from 24x (from 48 kHz to 2 kHz) to 12x (from 48 kHz to 4 kHz). This was performed to make a more suitable signal representation for pitch shifting, as including more data points in the analysis and synthesis stages might provide a less noisy output.

Bandpass filtering with an 8th order Butterworth filter between 50 and 150 Hz was again applied to the audio signal. Due to poor ranking of the raw signal representation during Scenario A, it was decided to remove it completely: as these features were regularly ranked lower than the bandpass signal and IMF-based features, it was safe to deduce that the noise contained in the higher parts of the spectrum (from 150 Hz to 1000 Hz) impedes the predictive power of the raw FPCG signal features. This could have included all sorts of noises observed in the aforementioned frequencies: besides the high frequency content of the womb environment that was superimposed on an FPCG signal, external noises generated by e.g. obstetrician changing the location of the Doppler device and repositioning of the mother in the examination chair could have also reduced the relevance of features gained from raw audio.

4.1.1 Pitch shifting parameters

PS-processed signals were introduced instead of the raw signals, using the same bandpass filter setup used for the raw filtered version as an input to the pitch shifting procedure. It was assessed that the versions strongly shifting the signal towards high frequencies would be the most

suitable ones for introducing psychoacoustics-based features, as the excitation of the highest number of critical bands might produce features that rely on psychoacoustic principles in the optimal way. For example, a popular psychoacoustic frequency scale (Bark scale) was proposed in [143], containing 24 critical bands that approximate the cochlear function. Signal bandpassed with filter between 50 and 150 Hz would only show excitation in two of the first bark bands [144], while a potentially pitch shifted signal of three octaves (pitch shift factor of 8) would "push" the spectrum to higher frequency values, so it covers frequencies from 400 to 1200 Hz. This expansion of the frequency range would have the newly constructed signal excite 6 bark bands, which implies a much higher level of separability in the psychoacoustic domain. Subjective listening tests have found that pitch shifting a signal 2 octaves up (factor of 4) provided a good compromise between achieving a substantial move from low to mid-frequency range and acceptable levels of phase artifacts in the signal [145]. A phase vocoder algorithm was employed for signal processing on a 4 kHz sampling rate, window size of 256 samples, an analysis overlap factor of 16 and a synthesis overlap factor of 4.

4.1.2 Revisiting IMF properties

Low ranking of the 3rd IMF showed its lack of usefulness in the classification process. The main reason for this could be found in a strong indication that the valuable signal has already converged to the first and second IMFs, making the 3rd IMF only contain potential noise and irrelevant/redundant information. Due to this, it was completely removed from the analysis in this scenario.

It was empirically determined that different modalities of recorded and simulated datasets would require separate preprocessing setups for the EMD process: as the method is data-driven, different distributions of noise could severely influence the sifting process. Experimentations with various filter cut-off frequencies (in increments of 50 Hz) and visual assessment of the waveform in the cases of two datasets have shown that an adequate frequency range for the custom dataset should have remained the same as in the Scenario A (50-1000 Hz), while a flatter shape of the noise distribution in regard to frequency has shown that it required additional filtering, so a bandpass filter with cut-off frequencies of 50 and 250 Hz was chosen.

4.1.3 Preprocessing pipeline

Signal representations for the 4 signals in the cases of custom and simulated datasets are given in Figures 4.1 and 4.2, while the preprocessing pipeline for filtered, PS-processed and EMD-processed versions of the signal is shown in Figure 4.3.

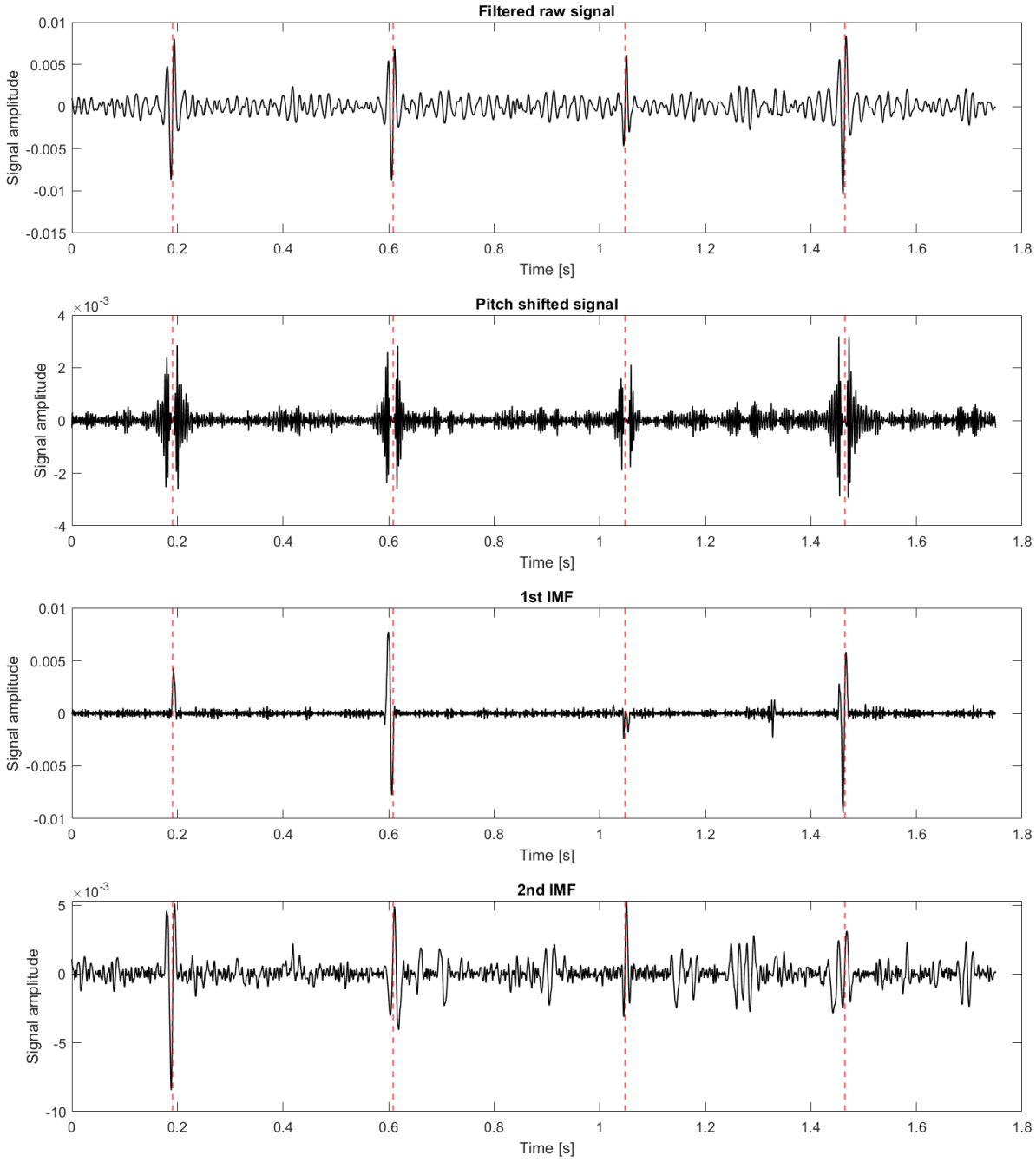

***Figure 4.1**. A representation of custom signals used for feature extraction. The red line shows the location of the S1 sound, marked by the Doppler device.*

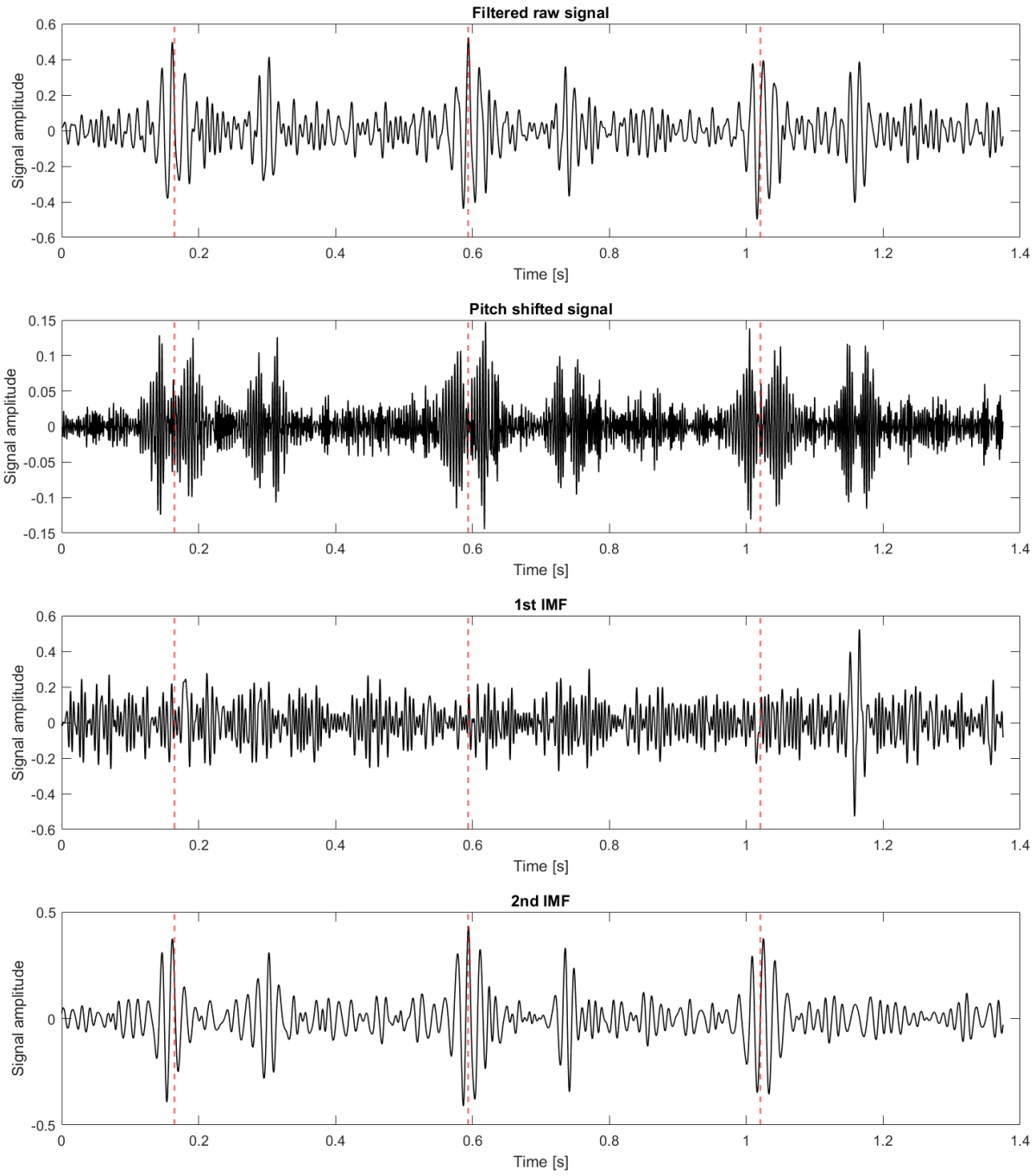

Figure 4.2. *A representation of simulated signals used for feature extraction. The red line shows the location of the S1 sound, marked by the Doppler device.*

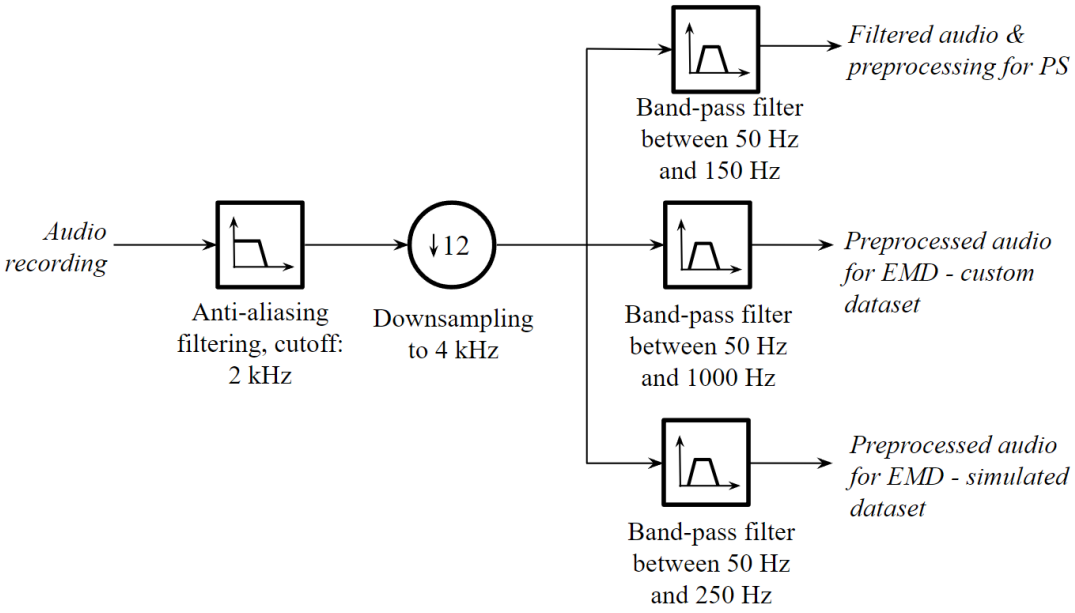

***Figure 4.3**. Preprocessing of the recorded signals in Scenario B*

4.2 Feature extraction

In addition to features described and used in Scenario A (18 statistical and 9 spectral features), Scenario B introduced a set of psychoacoustic features.

4.2.1 Psychoacoustic features

Two versions of psychoacoustic features were utilized for FPCG signal analysis: mel-frequency cepstral coefficients (MFCCs) [146] and perceptual linear prediction (PLP) coefficients [147]. Rastamat framework in MATLAB [148] was used for the calculation in both cases.

MFCCs are popular features in speech recognition [149], music similarity detection [150] and emotion recognition [151]. The algorithm for the calculation of MFCCs is given here:

1. The sound is pre-emphasized (boosting higher frequency magnitude) and segmented into small subwindows (25-50 ms) with even smaller hop lengths (10-20 ms). A window function (such as Hamming [152]) is applied to the segment for better spectral resolution.

2. FFT is utilized to extract the power spectrum of each subwindow.

3. Power spectrum values are mapped on the mel scale [82] through multiplication with non-uniformly placed triangular windows.
4. The newly mapped power spectrum values are logarithmized for better conformity to human hearing principles.
5. Calculated values are decorrelated with Inverse Discrete Cosine Transform (IDCT).
6. Step 5 is repeated for different orders of IDCT, yielding a requested number of MFCCs (usually 13).
7. Liftering of the final cepstral coefficients can be utilized to yield better final performance. This functions as a suppression method for slow variations in the log-power spectrum [153].

MFCC calculation flowchart can be found in Figure 4.4.

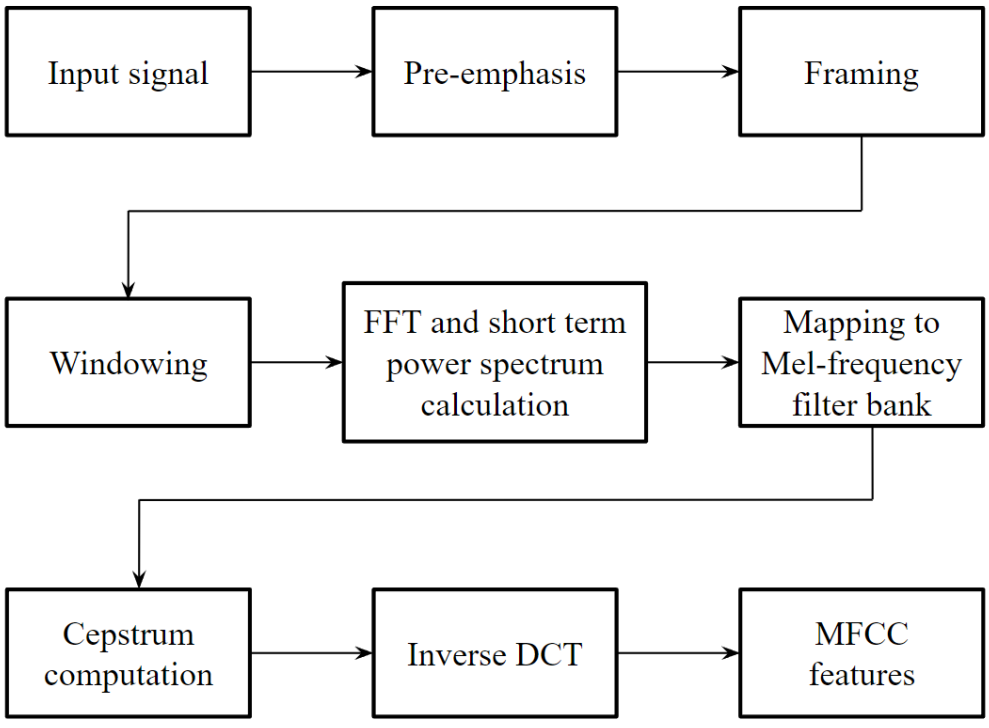

***Figure 4.4**. Calculation flowchart for Mel Frequency Cepstral Coefficients*

The parameters for MFCC calculation were:

- Window size of 25 ms,
- Hop size of 10 ms,
- Hamming analysis window,
- 13 coefficients to return,
- Frequency range from 0 to 2000 Hz,
- Exponential liftering with exponent of 0.6,
- Input signal preemphasis with filter coefficients values of 1 and -0.97.

On the other side, PLP coefficients are a psychoacoustic extension on the concept of linear predictive coding, a technique that uses previous signal values in a time series to predict new values, with applications in speech and audio processing [154]. Linear predictive coefficients can be seen as weights of a finite impulse response (FIR) filter, predicting a new sample from a set of previous ones. The estimation of the coefficients is usually achieved with autoregressive estimation methods, such as Levinson-Durbin [155] and Burg algorithms [156].

The perceptual "extension" for linear prediction coefficients can be found in 3 points:

1. Utilization of the critical-band spectral resolution by warping the spectrum into the Bark frequencies, not unlike the mel scale mapping used in MFCCs.
2. Pre-emphasizing the signal with the equal loudness curve, considering the non-equal sensitivity of human hearing.
3. Mapping the signal amplitude to perceived loudness by the cubic-root conversion [157].

IDCT is then utilized to yield the autocorrelation function, with the autoregression method finally used for outputting PLP coefficients. The block diagram for the calculation of these coefficients is depicted in Figure 4.5.

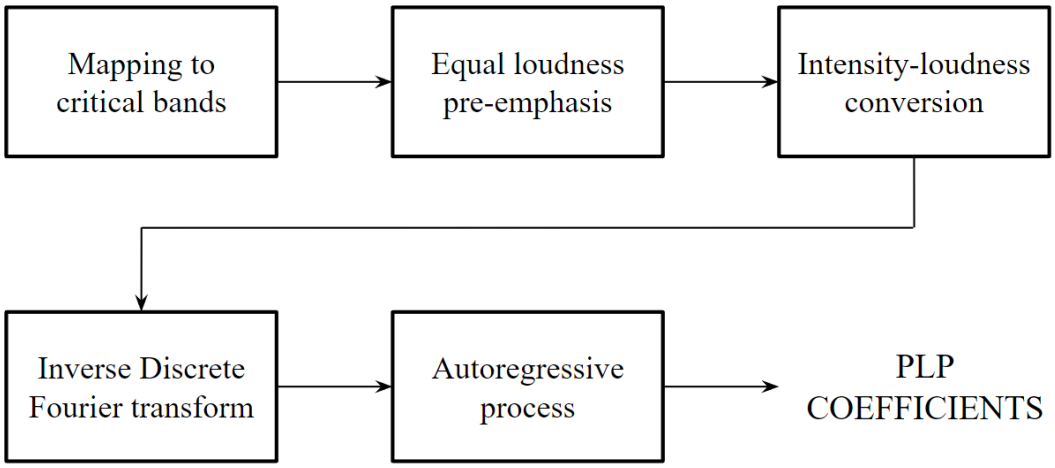

***Figure 4.5**. Calculation flowchart for Perceptual Linear Prediction coefficients*

The parameters for PLP calculation were:

- Window size of 25 ms,
- Hop size of 10 ms,
- Hamming analysis window,
- 13 coefficients to return,
- Frequency range from 0 to 2000 Hz,
- Levinson-Durbin autoregression.

Since every subwindow returned 13 MFCCs and 13 PLPs, for an input window of 200 ms, a feature matrix of 13x18 values was produced. Flattening of the feature matrices was achieved by employing mean and standard deviation as functionals. Basically, these statistical functions were used to summarize the 18 values for each coefficient into one value: for example, one value for the 1st PLP coefficient could be found in each of the 18 subwindows, so taking a mean from the set of numbers yielded only one quantity, making a feature that can be labeled as "mean of the 1st PLP coefficient". This reduced the feature space, made the specific position of a subwindow in the overall window irrelevant and increased the robustness of the feature representation. The total of 54 psychoacoustic-based features were calculated: 13 mean values for PLP coefficients, 13 mean values for MFCCs, 13 standard deviation values for PLP coefficients and 13 standard deviation values for MFCCs.

4.2.2 Overall feature vectors

The generation of feature vectors was similar to the approach done in Scenario A, but further expanded so it included psychoacoustic descriptors.

A total of 212 features were extracted:

1. 79 bandpass-processed features - 27 audio features + 52 psychoacoustic features,
2. 79 PS-processed features - 27 audio features + 52 psychoacoustic features,
3. 54 EMD-processed features - 27 audio features for IMF1 and 27 audio features for IMF2.

The application of psychoacoustic features on IMFs was avoided due to the fact that the data-driven sifting process could have produced IMFs that did not have perceptual value in the strictest sense.

Having more preprocessing and feature extraction subroutines did not increase the computation costs drastically: compared to Scenario A, the entire computation time increased by 80%, making the process still 2.2x faster than real time, using the same hardware as before. Optimization steps can be included in future work, especially considering that large amounts of data being constantly recalculated due to a high overlap factor can be reused in a more adequate fashion.

4.2.3 Dataset considerations

Even though the segmentation of the dataset was done similarly as in Scenario A, there were considerable extensions introduced in this Scenario. Firstly, it was discovered that the pitch shifting procedure induced temporal expansion of the S1 sound (this can be observed in Figures 4.1 and 4.2). In order to validate that the phenomenon does not rate PS-processed signals better in the case of 200 ms window size (introduced in Scenario A), additional datasets were constructed from the custom recorded data: one with 150 ms window size and one with 100 ms window size. In both of these cases, the overlap factor was chosen to be 4. Finally, the 8th recording (originally discarded in Scenario A) was used in the dataset.

Regarding the utilization of the simulated dataset, 3 cases with extremely low SNR values (-22 dB, -24.4 dB and -26.7 dB) were chosen from the simulated dataset with 200 ms window size.

It has to be noted that the amount of noise added to the simulated FPCG signals was randomized for different noise components (such as environmental noise or vibrations from the maternal organs) for all three cases in the original research, making all 3 cases rather different regarding the contents of the recordings.

Secondly, it was empirically determined that the EMD sifting process would perform better given more data besides the window size of 200 ms. Therefore, the input into the EMD was expanded with one second of concurrent data both on the left and right side of the window. After the extraction of IMFs, the central 200 ms of data were cropped and used as generated IMFs, making the final EMD-processed window the same size as before.

The details for the 6 dataset cases used in this Scenario are given in the Table 4.1 below.

***Table 4.1**. Distribution of custom and simulated dataset cases*

Dataset case	Observations num	Positive examples	Negative examples	Pos/neg ratio
Custom, 200 ms window	9346	4043	5303	43:57
Custom, 150 ms window	14039	4005	10034	29:71
Custom, 100 ms window	23553	3987	19566	17:83
Simulated, all 3 cases	6989	3732	3257	53:47

The overall percentage of negative examples for custom dataset cases increased with the decrease of the window size used for segmentation. This was expected, since the smaller window size reduces the chance of a window containing S1 sound labels, while the fixed overlap ratio also shortens the hop size, taking more negative examples between adjacent S1 labels in the process.

4.3 Results

The results for Scenario B are divided into 7 subsections, mostly encapsulating the outcomes of ranking processes for 6 of the extracted datasets (custom and simulated).

4.3.1 Correlation analysis

The correlation matrix consisting of Pearson's correlation coefficients was calculated once again for each of the feature pairs in the dataset. Since Scenario A already demonstrated the results of the correlation analysis in the cases of filter-based and IMF-based features, the focus of the analysis here was to assess the correlation of the distinct pairs of the same functionals applied to a signal with bandpass filtering and pitch shifting, such as standard deviation of the 5th PLP coefficient for both the filtered and shifted signal versions.

The aggregated results for custom datasets showed moderate to strong correlations for feature pairs in the cases of statistical features ($|r|$ between 0.4 and 0.8) but also exhibited fairly low correlations for the spectral features ($|r|$ below 0.6) except for the spectral centroid ($r = 0.93$) and spectral entropy ($r = 0.79$). Regarding psychoacoustic features, very high correlations ($|r|$ above 0.9) were only observed for the means of the first 3 PLP coefficients with somewhat weaker correlations ($|r|$ below 0.3 and 0.8) for all other PLP features. In the case of MFCCs, correlations were moderate ($|r|$ around 0.6) for mean functionals, but very low ($|r|$ below 0.2) for standard deviation statistics.

Regarding simulated datasets, the results exhibited much lower correlation values: besides a couple of statistical and spectral features, only the means of the first 2 PLP coefficients showed moderate correlation ($|r|$ around 0.6 or somewhat higher), with every other feature pair manifesting weak correlation ($|r|$ below 0.5).

On the whole, the rather weak correlations between features of filtered and shifted signal origin were fairly moderate, suggesting high levels of nonlinearity induced into the analysis and synthesis process of the phase vocoding during high pitch shifting factors (4x in this case). Signal shapes shown in Figures 4.1 and 4.2 definitely confirm this. Psychoacoustic features have shown mostly moderate correlations, something that was expected due to the nonlinear nature of the frequency mapping used in both the MFCC and PLP calculations.

4.3.2 Mutual information

As described in the results for Scenario A, mutual information is a measure that describes how much information one random variable tells about the other, i.e. how much information is shared between the variables: in this case, the independent variable (feature) and the target variable (label).

4.3.2.1 Custom datasets

The MI ranking results for extracted custom dataset with 200 ms window length showed that 6 out of 10 top features and 11 out of top 20 features were taken from the features with pitch shifted input. An interesting point was the number of "signal shape" features (e.g. the 5th percentile divided by maximum) that are ranked high and insinuate closer distance between the probability distributions of the features and target variables. This seemed to be the case for all three subsets of features, including filtered, pitch shifted and IMF-based. The first five features with the highest MI in decreasing order were: (1) the 5th percentile divided by minimum of the filtered audio signal, (2) the 95th percentile divided by maximum of the filtered audio signal, (3) the 95th percentile divided by maximum of the shifted audio signal, (4) the 95th percentile divided by maximum of the 2nd IMF, and (5) the 5th percentile divided by minimum of the shifted signal. Figure 4.6 shows the mutual information for the top 20 features.

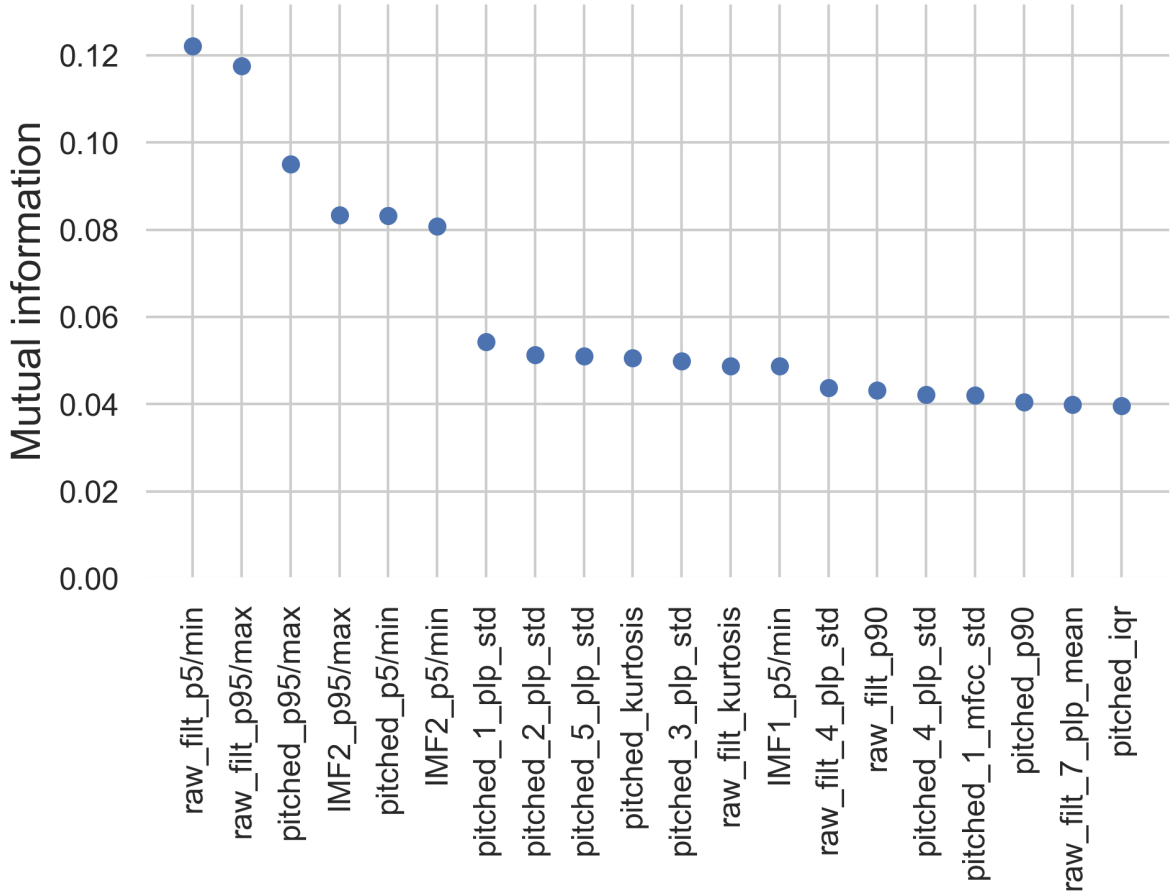

***Figure 4.6**. The top 20 features obtained by univariate ranking based on mutual information. Custom dataset with 200 ms window used as an input.*

The MI ranking results for the second case, containing the extracted custom dataset with 150 ms window length showed similar behavior compared to the latter case and had 5 out of 10 top features and 10 out of top 20 features were taken from the features with pitch shifted input. The univariate analysis does not take into account any redundancy concerns, such as the relations between same statistical functionals taken from different signal representations. In any case, shapes of the filtered and IMF-based signal versions, as well as PLP coefficients taken from the pitched signal, suggested an existing correlation with the target variable. The first five features with the highest MI in decreasing order were: (1) the 5th percentile divided by minimum of the filtered audio signal, (2) the 95th percentile divided by maximum of the filtered audio signal, (3) the 5th percentile divided by minimum of the 2nd IMF, (4) the 5th percentile divided by minimum of the shifted signal, and (5) the 95th percentile divided by maximum of the shifted signal. Figure 4.7 shows the mutual information for the top 20 features.

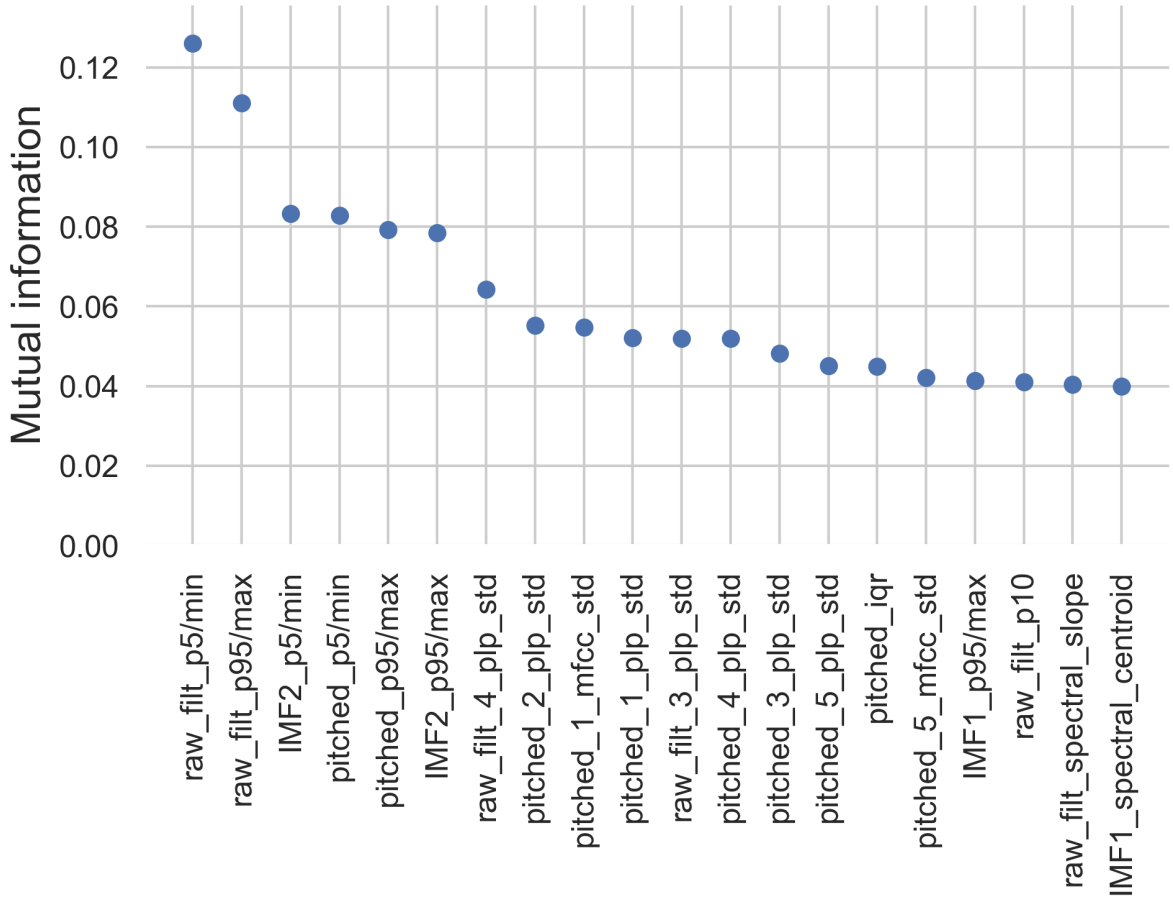

Figure 4.7*. The top 20 features obtained by univariate ranking based on mutual information. Custom dataset with 150 ms window used as an input.*

The MI ranking was also done for the last of the extracted custom datasets, with 100 ms window length. The results are rather related to the two previous cases, however with PLP features (from both inputs that employ them) showing more prominence. The first five features were: (1) the 5th percentile divided by minimum of the filtered audio signal, (2) the 95th percentile divided by maximum of the filtered audio signal, (3) the 5th percentile divided by minimum of the 2nd IMF, (4) the 95th percentile divided by maximum of the second IMF, and (5) standard deviation of the 4th PLP coefficient of the filtered audio signal. Figure 4.8 displays the MI rankings for the described cases.

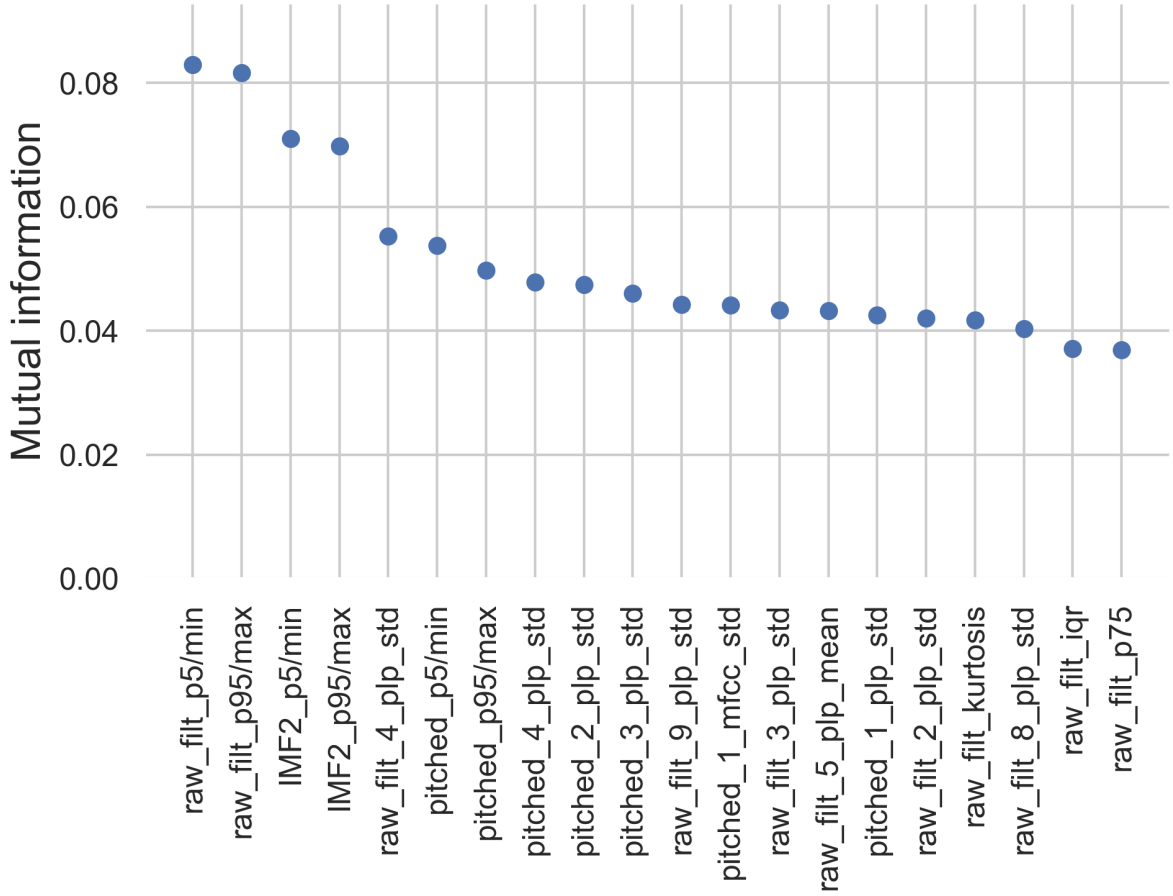

***Figure 4.8**. The top 20 features obtained by univariate ranking based on mutual information. Custom dataset with 100 ms window used as an input.*

4.3.2.2 Simulated datasets

Results for the MI ranking with the simulated datasets revealed a somewhat different situation compared to the custom datasets. For the least severe case of the 3 datasets under analysis in terms of noise (SNR = -22 dB) it was shown that 7 out of 10 top features and even 14 out of

top 20 features are taken from the subset with pitch shifted input. The rankings also demonstrated a heavy reduction of the percentile-based features and the increase in perceptually based feature ranking, implying higher relevance of psychoacoustic modeling, especially in the case of pitch shifted signals. The first five features with the highest MI in decreasing order were: (1) mean of the 4th MFCC of the shifted audio signal, (2) mean of the 4th PLP coefficient of the shifted audio signal, (3) mean of the 3rd MFCC of the shifted audio signal, (4) spectral decrease of the filtered audio signal, and (5) zero crossing rate of the shifted audio signal. This is all depicted in Figure 4.9.

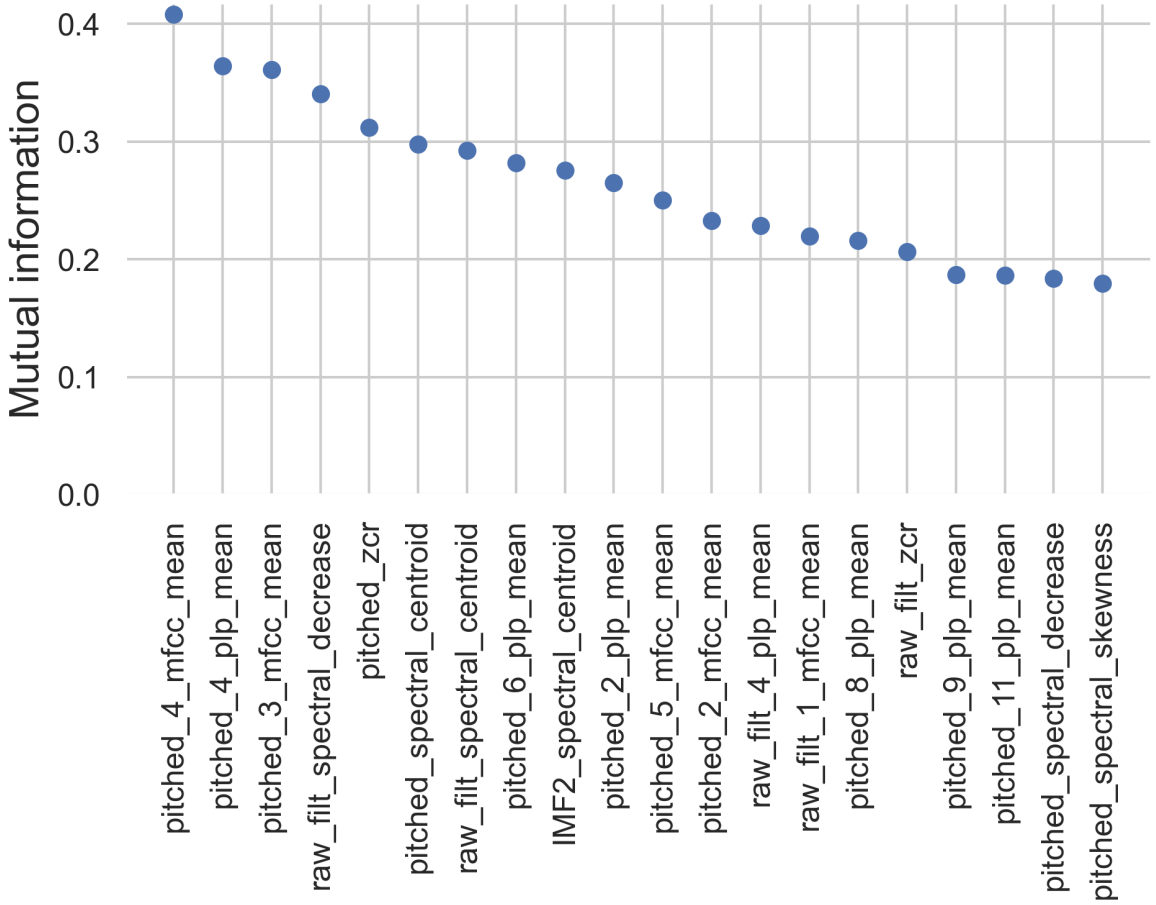

***Figure 4.9**. The top 20 features obtained by univariate ranking based on mutual information. Simulated dataset with -22 dB SNR used as an input.*

The case with the simulated dataset with -24.4 dB of SNR was the second of the group. Even more features from the pitch shifted signal subset were included in the best rankings, with 9 of them in the top 10 and 16 of them in the top 20. In other words, only 3 of the filtered signal-based and 1 IMF-based features made the list. The observed trend seemed to show that the more

corrupted S1 sounds are, the more they rely on the psychoacoustic features and less on the audio descriptors. As an additional remark, only the spectral centroid of the 2nd IMF was present in the best ranked features. The first five features in the rankings were: (1) mean of the 4th MFCC of the shifted audio signal, (2) mean of the 4th PLP coefficient of the shifted audio signal, (3) mean of the 3rd MFCC of the shifted audio signal, (4) spectral decrease of the filtered audio signal, and (5) mean of the 6th PLP coefficient of the shifted audio signal. This is shown in Figure 4.10.

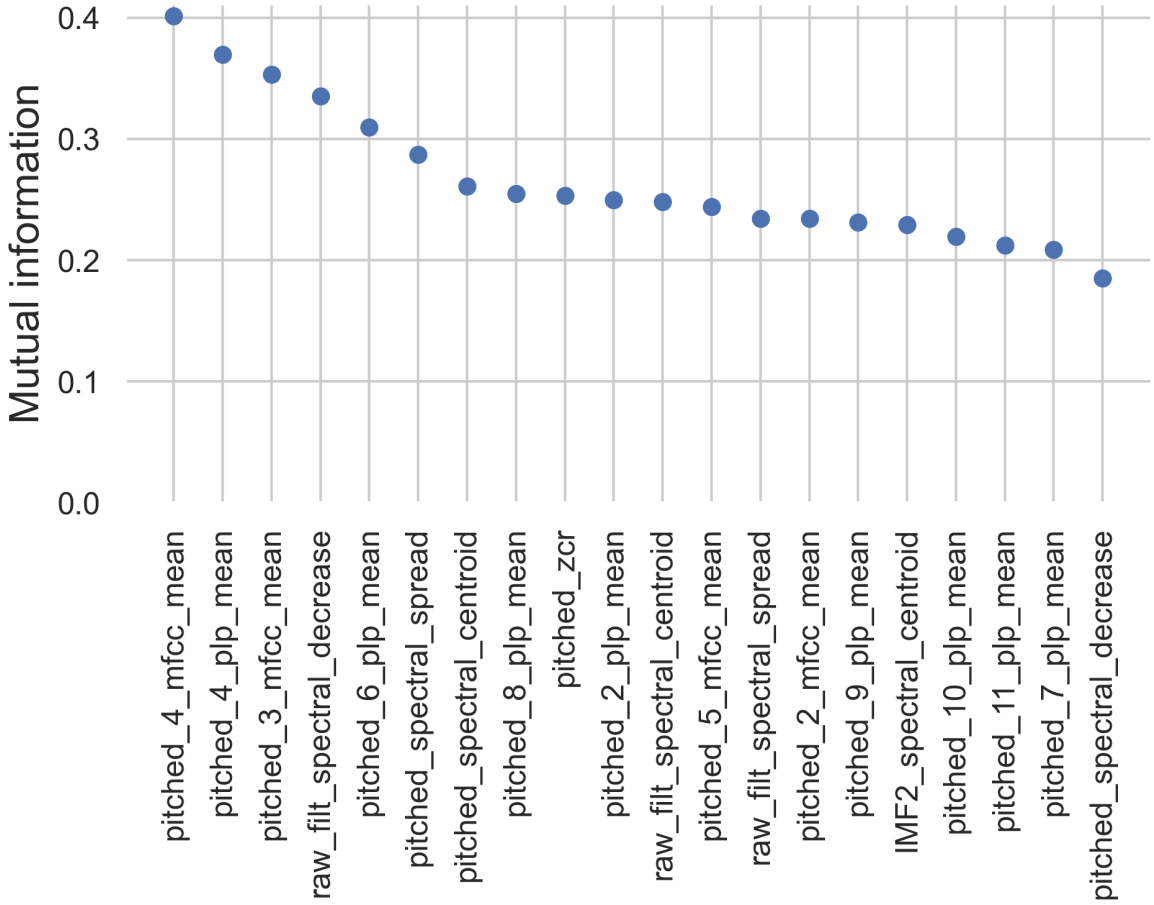

***Figure 4.10**. The top 20 features obtained by univariate ranking based on mutual information. Simulated dataset with -24.4 dB SNR used as an input.*

Regarding the simulated case with -26.7 dB of SNR (noise-wise, the most "severe" situation available), the trend seemed to closely follow the latter 2 cases: 8 out of 10 top features and 16 out of top 20 features were taken from the subset with pitch shifted input. Besides one place in the rankings taken by a zero-crossing rate feature, all others were based on psychoacoustics and spectral characteristics. The first five features with the highest MI in decreasing order were: (1) mean of the 4th MFCC of the shifted audio signal, (2) mean of the 4th PLP coefficient of the

shifted audio signal, (3) spectral decrease of the filtered audio signal, (4) mean of the 3rd MFCC of the shifted audio signal, and (5) mean of the 6th PLP coefficient of the shifted audio signal. Figure 4.11 displays the MI rankings for this final case.

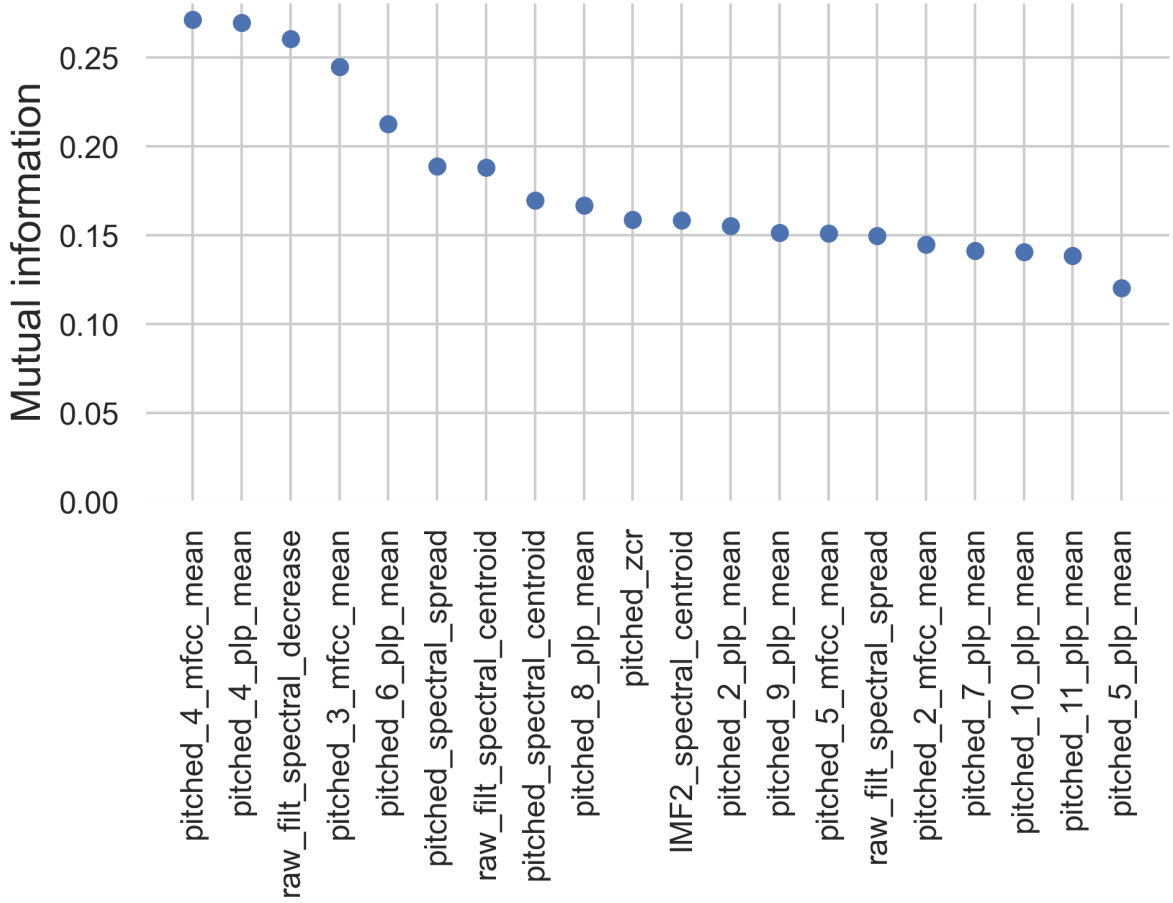

***Figure 4.11**. The top 20 features obtained by univariate ranking based on mutual information. Simulated dataset with -26.7 dB SNR used as an input.*

4.3.3 ANOVA

One-way ANOVA is utilized to check how much the means of specific features and target labels differ from one another. More specifically, it analyzes whether there are statistically relevant differences between means of the feature values for data that contains S1 sound labels and data that doesn't.

The null hypothesis [158] in ANOVA always states that there are no differences in means between populations, e.g. that the mean of the distribution for spectral centroid is the same for

target labels equal to 1 and target labels equal to 0. In this case, p-value describes the "acceptability" level of the null-hypothesis, i.e. p-value will be closer to 1 if the means of the aforementioned distributions are very similar and might drop very close to 0 if the means are very different. The latter situation implies a strong statistical significance of the specific feature in the context of target separability.

The negative natural logarithm was used in Scenario A to rank the features: the lower the p-value, the stronger the impact it has on the quality of classification. However, in the case of Scenario B, the p-values have gotten to incredibly small values such as 10^{-308}, with some of them becoming undetectable within the resolution of the double-precision floating-point format. This is a fairly regular case in ANOVA testing: the null hypothesis is rejected if p-value goes below a certain value. As it was shown that some of the features (mostly perception-based from the pitch shifted subset) cannot be ranked in the strictest sense, the methodology was redefined to output the number of features exhibiting $p < 0.01$ for each of the feature subsets.

4.3.3.1 Custom datasets

Figure 4.12 shows the percentage of features from specific subsets of custom datasets that were considered statistically relevant by the ANOVA method. For the custom dataset with 200 ms window size, the results revealed that 68 out of 79 (86%) features from the filtered subset were rendered as having impact on the separability of the dataset, with 62 out of 79 (78%) from the shifted signal and 41 out of 54 (76%) of the IMF-based features. The situation was somewhat different for the 150 ms window custom dataset, where 65 out of 79 (82%), 63 out of 79 (80%) and 44 out of 54 (81%) features from the filtered, pitch shifted, and EMD-processed subsets respectively had an impact on the target label. Finally, in the case of the custom dataset with 100 ms window, 70 from the total of 79 (89%) were taken from the bandpass-based features, 66 from the total of 79 (84%) from the shifted signals and 41 from the total of 54 (76%) from the IMF-based features. It was shown that more than 75% features from each subset are relevant for the FPCG classification task, implying good choice of the preprocessing methods and an adequate feature engineering process.

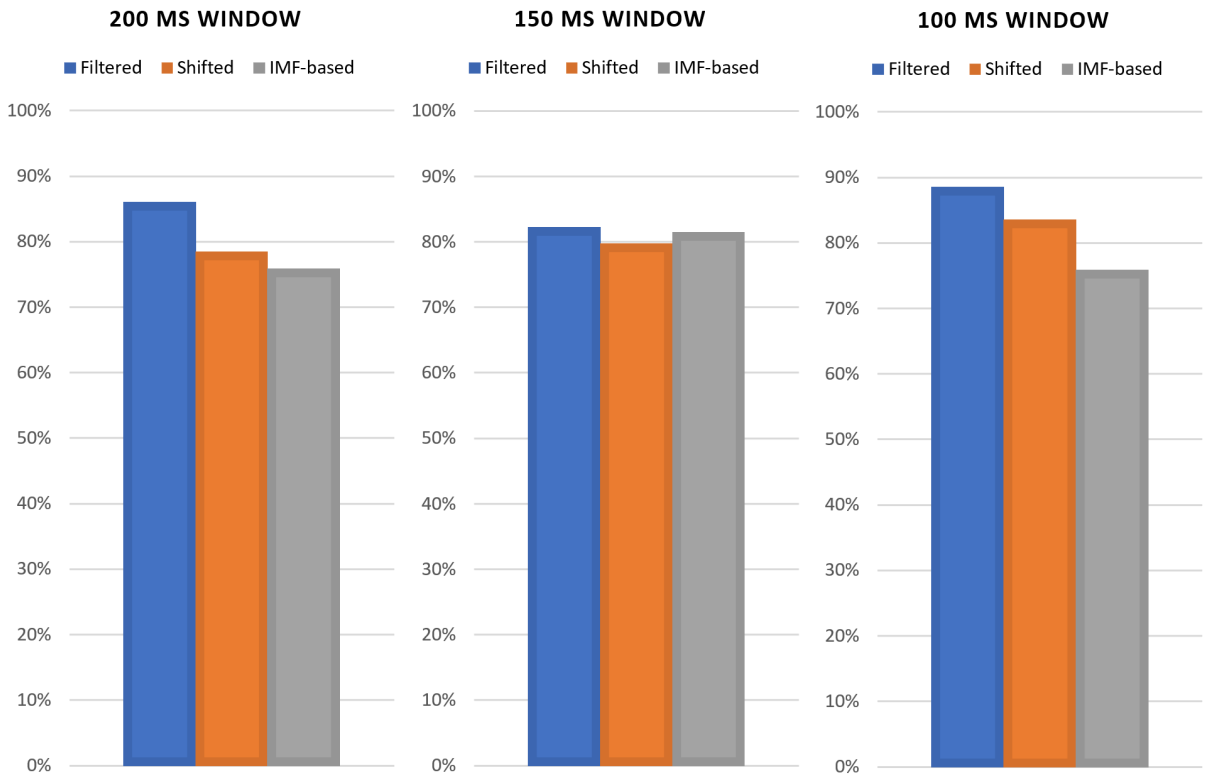

***Figure 4.12**. Percentages of statistically relevant features outputted by one-way ANOVA for custom datasets.*

4.3.3.2 Simulated datasets

Similar to custom datasets, simulated datasets were also used as an input to the ANOVA method. The results, shown in Figure 4.13, demonstrate similar behavior to the latter cases, with lower significance of the IMF-based features. This can be explained by the difficulty exhibited by the EMD sifting process in converging to relevant information in the case of artificially corrupted simulated datasets. Nevertheless, 50% or more IMF-based features were considered statistically significant even with suboptimal convergence. For the case of simulated dataset with -22 dB of SNR, 67 out of 79 features (85%) from the filtered subset, 67 out of 79 (85%) from the shifted subset and 27 out of 54 (50%) from the EMD subset were found relevant. Regarding the dataset with -24.4 dB of SNR, 64 out of 79 (81%), 66 out of 79 (84%) and 30 out of 54 (56%) were selected from the filtered, shifted and IMF groups respectively, while the final dataset with SNR of -26.7 dB demonstrated 60 relevant features out of 79 (76%) from the filtered signal version, 61 out of 79 (77%) from the pitch shifted signal and 28 out of 54 (52%) from the IMFs.

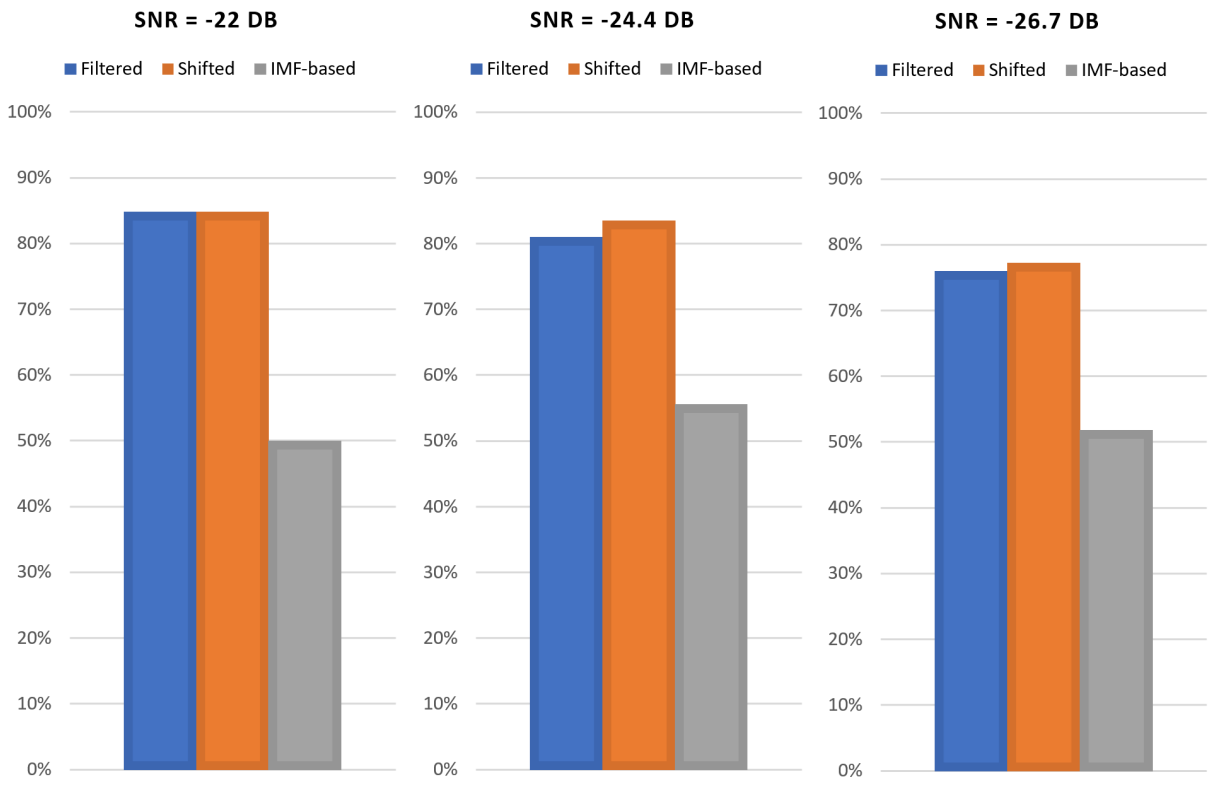

***Figure 4.13**. Percentages of statistically relevant features outputted by one-way ANOVA for simulated datasets.*

4.3.4 Embedded approach - Random Forest

Embedded methods evaluate feature importance values during the training process, combining the qualities, both as an innate or extended functionality [159]. Random forest algorithm includes feature ranking and selection as a built-in capability, averaging and assessing the decrease of prediction uncertainty after splitting the data on a certain feature. 100 decision trees were trained for every input dataset.

4.3.4.1 Custom datasets

First of all, taking results from the custom dataset with a 200 ms window started to show similar outcomes for feature rankings as in the previous univariate methods of Mutual Information and one-way ANOVA, especially in the occurrence frequency of psychoacoustic features. Furthermore, there was an increased prominence of IMF-based feature positioning, albeit after the first 8 places. The first five most relevant features in decreasing order of importance were: (1) standard deviation of the 4th PLP coefficients of the filtered audio signal, (2) standard deviation of the 3rd PLP coefficients of the shifted signal, (3) standard deviation of the 2nd PLP

coefficients of the shifted signal, (4) maximum of the filtered signal, and (5) standard deviation of the 1st MFCC of the shifted audio signal. Figure 4.14 shows feature importance values obtained using a random forest.

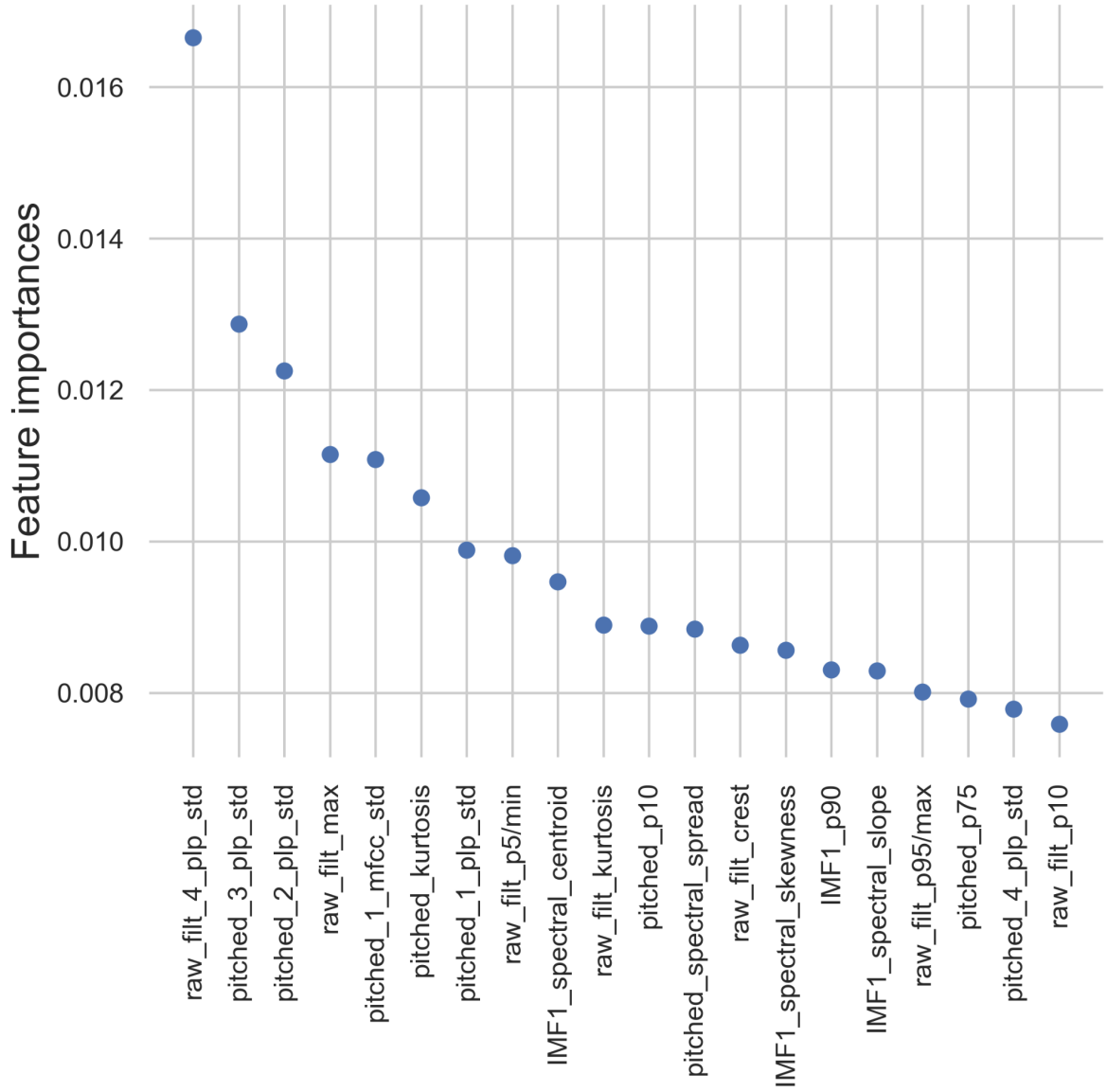

***Figure 4.14**. The top 20 features inherently obtained by the learning algorithm of the random forest classifier. Custom dataset with 200 ms window used as an input.*

The situation in the case of 150 ms window custom dataset was quite similar to the latter one, however with more significance added to the 1st IMF. As an interesting note, it could be observed that spectral features from the IMFs take higher rankings than the ones from the filtered signal, suggesting new information was gained from the sifting process. The first five places in the feature importance list were: (1) standard deviation of the 4th PLP coefficients of

the filtered audio signal, (2) standard deviation of the 3rd PLP coefficients of the shifted signal, (3) standard deviation of the 2nd PLP coefficients of the shifted signal, (4) spectral centroid of the 1st IMF, and (5) kurtosis of the filtered audio signal. This is shown in Figure 4.15.

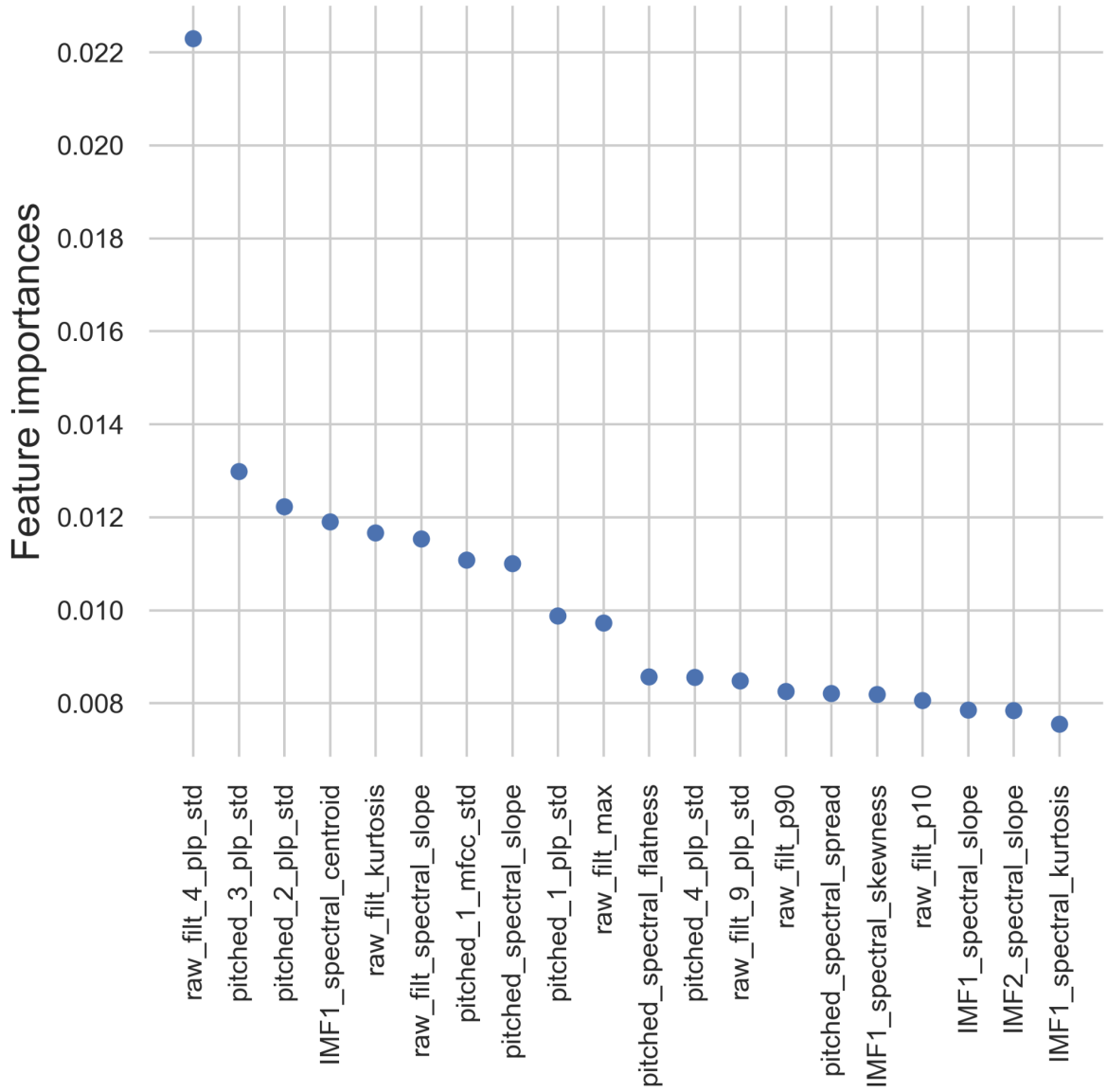

***Figure 4.15**. The top 20 features inherently obtained by the learning algorithm of the random forest classifier. Custom dataset with 150 ms window used as an input.*

For the case of the smallest window (100 ms) used in custom dataset segmentation, it was implied that IMFs have gone up in rankings even further compared to the latter cases. Besides that, the number of features from the pitch shifted domain in the first ranks has decreased slightly, and the first feature in rankings (PLP from the filtered signal) has been confirmed to exhibit most significance in all 3 custom cases. This might suggest minor dependence on the

effectiveness of psychoacoustics with decreased window sizes. However, the rankings of the aforementioned were still very high on average. The first five places in the feature importance list were: (1) standard deviation of the 4th PLP coefficients of the filtered audio signal, (2) standard deviation of the 3rd PLP coefficients of the shifted signal, (3) standard deviation of the 2nd PLP coefficients of the shifted signal, (4) spectral centroid of the 1st IMF, and (5) kurtosis of the filtered audio signal. This is depicted in Figure 4.16.

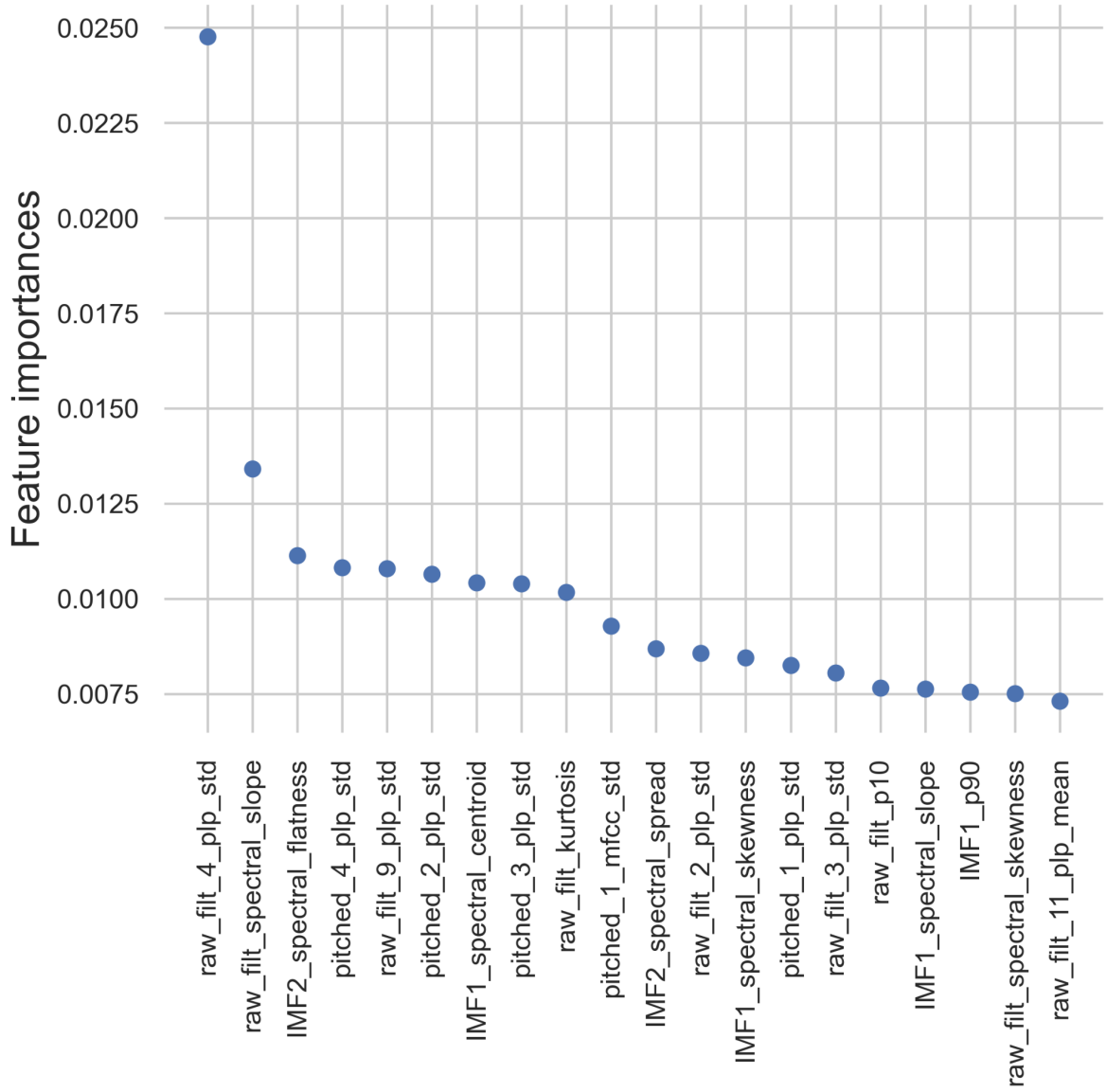

***Figure 4.16**: The top 20 features inherently obtained by the learning algorithm of the random forest classifier. Custom dataset with 100 ms window used as an input.*

4.3.4.2 Simulated datasets

Results on the simulated dataset with -22 dB of SNR showed a dramatic increase in the positioning of perceptually motivated parameters: 7 out of top 10 and 13 out of top 20 are gained from the PLP coefficients and MFCCs. Furthermore, the number of psychoacoustics-based features taken from the pitched audio signal severely outmatched the filtered version, with all 13 of the bio-inspired descriptors in the top 20 being from the shifted representation. The first five places in the feature importance list were: (1) mean of the 4th MFCC of the shifted audio signal, (2) mean of the 4th PLP coefficient of the shifted signal, (3) mean of the 3rd MFCC of the shifted signal, (4) spectral centroid of the 2nd IMF, and (5) spectral centroid of the pitch shifted audio signal. The top 20 ranking can be found in Figure 4.17.

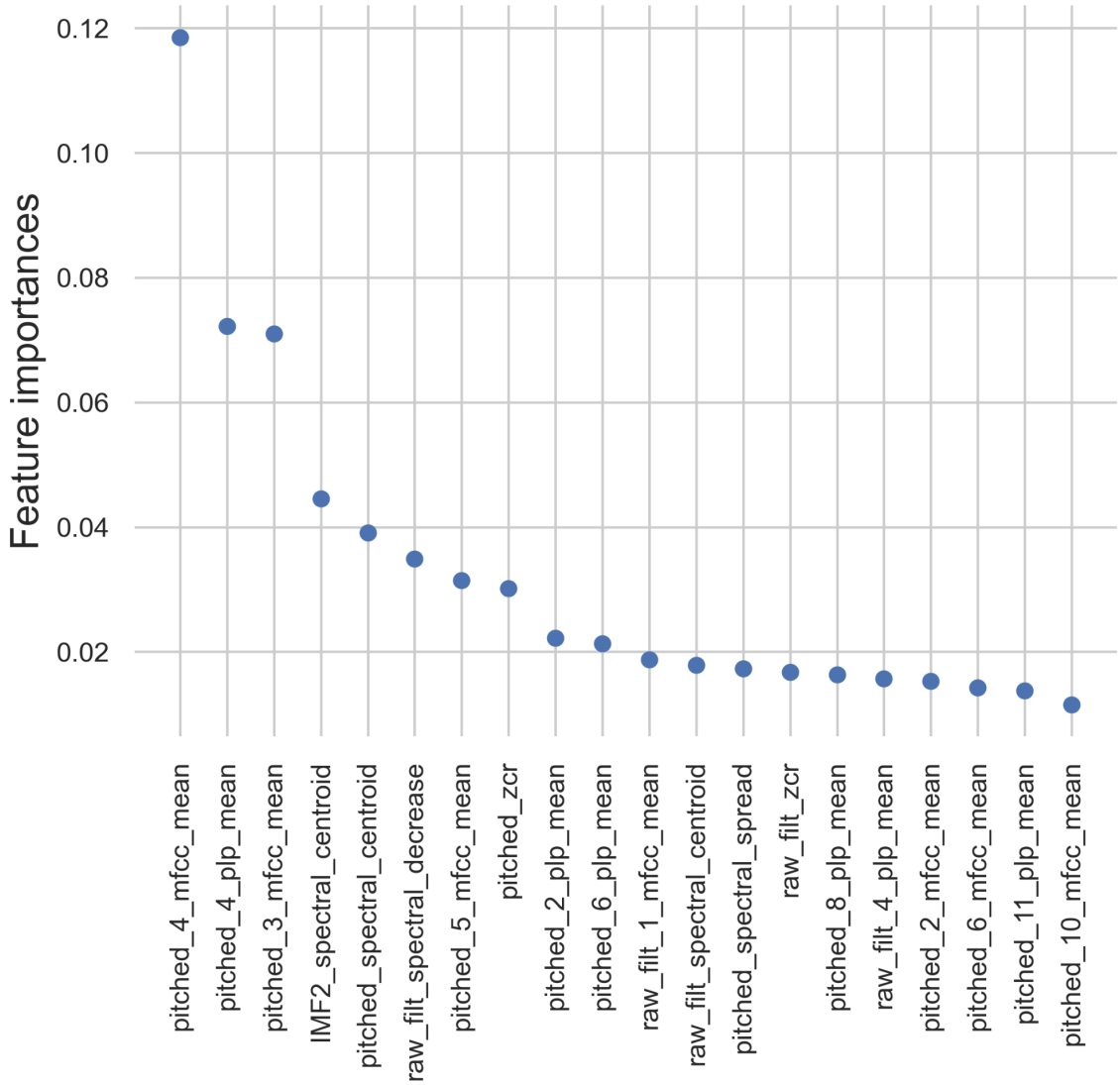

Figure 4.17*. The top 20 features inherently obtained by the learning algorithm of the random forest classifier. Simulated dataset with -22 dB SNR used as an input.*

A similar situation could be found in the rankings for simulated dataset with -24.4 dB of SNR. Combining pitched shifted signal with psychoacoustic feature extraction exhibited much higher rankings than other approaches, insinuating very high relevance for the classification task. The first five places in the feature importance list were: (1) mean of the 4th MFCC of the shifted audio signal, (2) mean of the 3rd MFCC of the shifted signal, (3) mean of the 4th PLP coefficient of the shifted signal, (4) spectral spread of the pitch shifted audio signal, and (5) spectral decrease of the filtered audio signal. The described results are given in Figure 4.18.

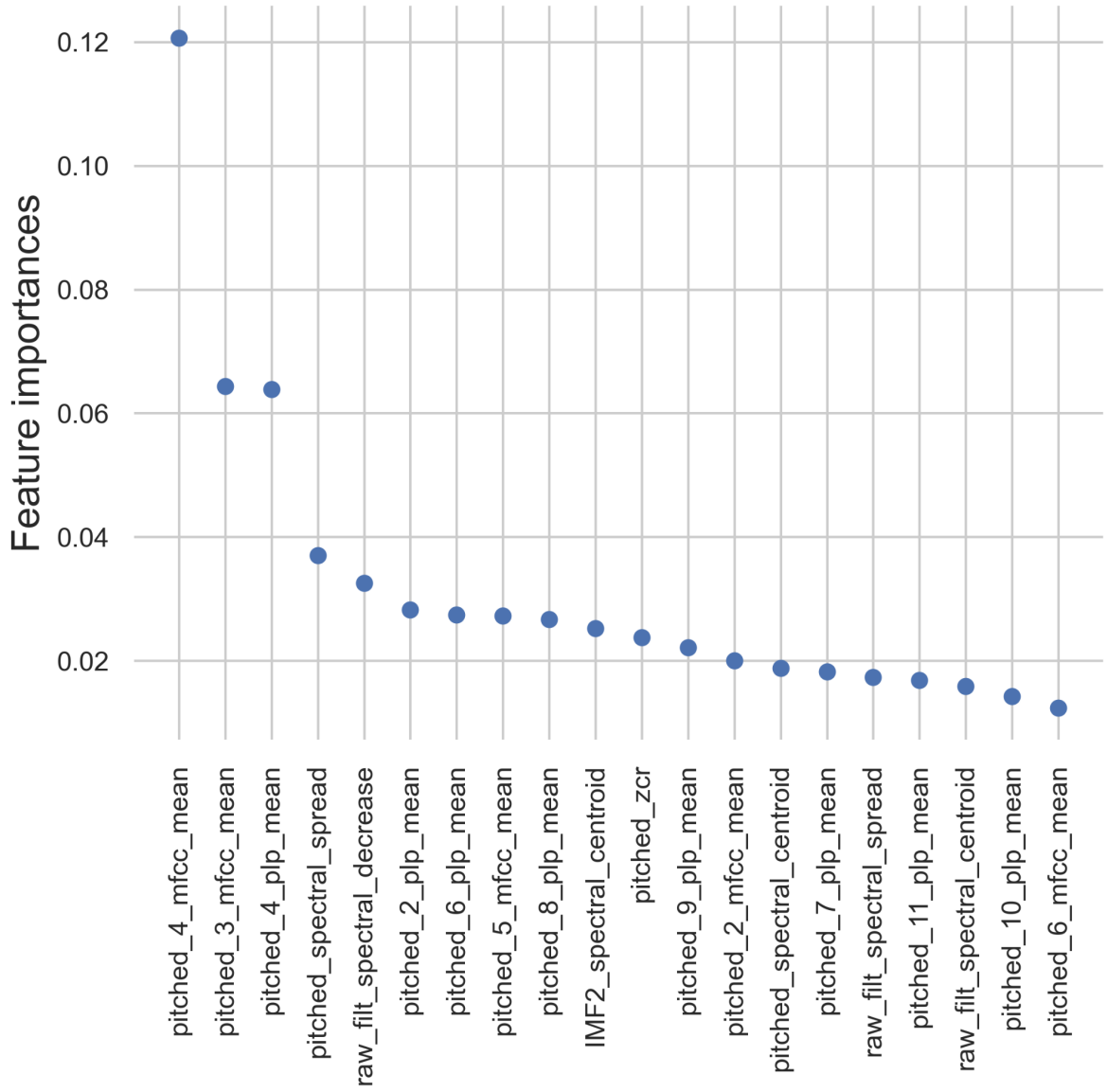

Figure 4.18. *The top 20 features inherently obtained by the learning algorithm of the random forest classifier. Simulated dataset with -24.4 dB SNR used as an input.*

In the last example, random forest was utilized to rank the features from the simulated dataset with -26.7 dB of SNR. Results were quite consistent to the previous 2 cases. As exhibited in the results with univariate feature selection methods, it was once again apparent that the IMFs underperformed in the simulated cases compared to the custom recorded ones. This is further explained in the Discussion section. Figure 4.19 displays the ranking results, with the top five ranked feature being: (1) mean of the 4th MFCC of the shifted audio signal, (2) mean of the 4th PLP coefficient of the shifted signal, (3) spectral decrease of the filtered audio signal, (4) mean of the 3rd MFCC of the shifted signal, and (5) spectral spread of the pitch shifted audio signal.

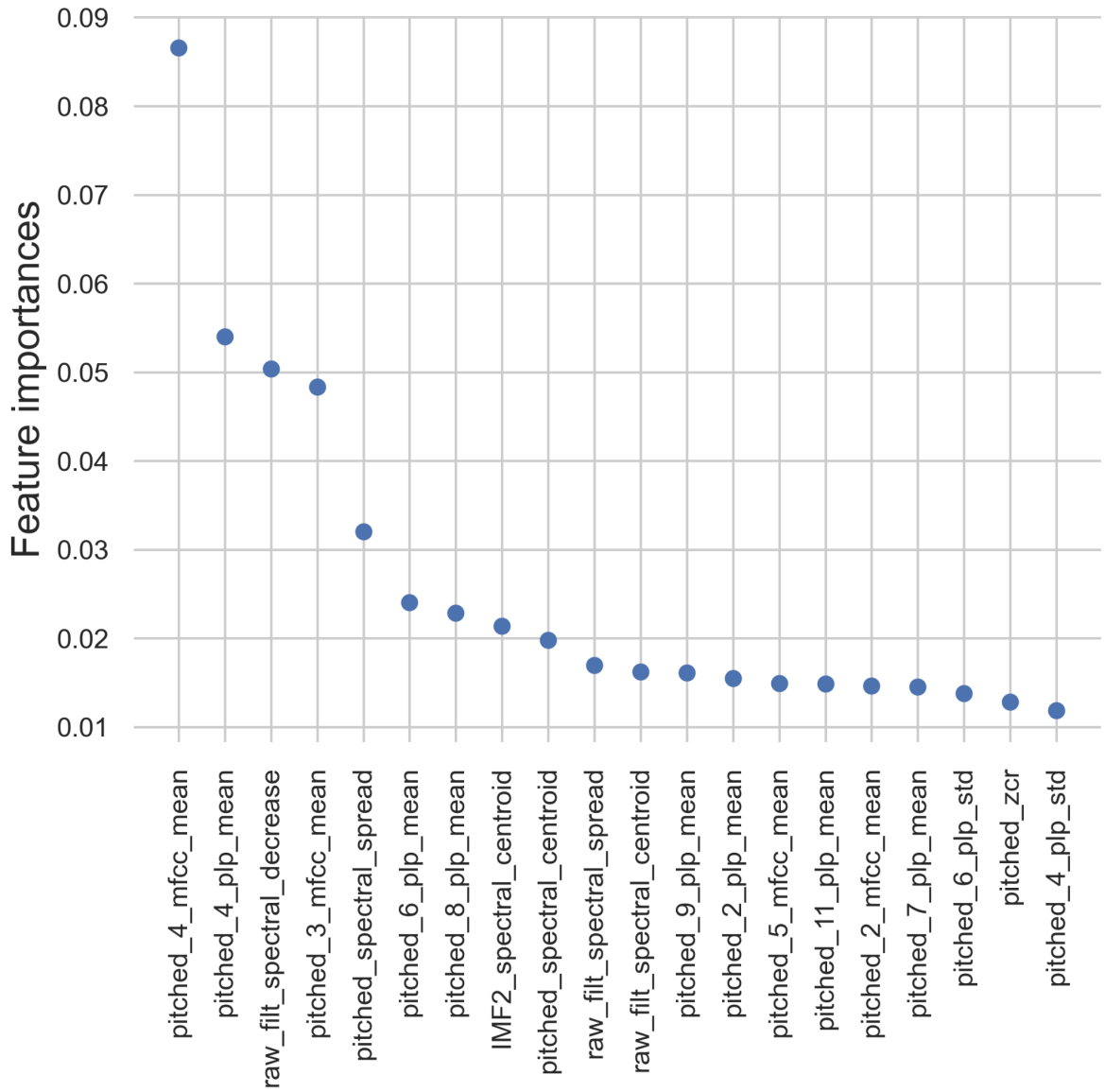

Figure 4.19. *The top 20 features inherently obtained by the learning algorithm of the random forest classifier. Simulated dataset with -26.7 dB SNR used as an input.*

4.3.5 Embedded approach - SVM

Support vector machine classifiers can also be used as an embedded method in feature ranking and selection [160]. As their training principles are different from the random forest ones, this research has employed a linear support vector machine for the feature selection process.

4.3.5.1 Custom datasets

As usual, the first case consisted of the custom dataset with 200 ms window size. The inclusion of this embedded method in the ranking process showed some expected, but also some new trends: even though the combination of pitch shifting, and psychoacoustics seemed to yield most information in the training process, it would appear that higher order PLP coefficients and MFCCs were more present in the top rankings. The top five ranked features were: (1) mean of the 9th PLP of the shifted audio signal, (2) mean of the 1st PLP coefficient of the shifted signal, (3) standard deviation of the 3rd PLP coefficient of the filtered signal, (4) mean of the 4th PLP coefficient of the shifted audio signal, and (5) standard deviation of the 5th PLP coefficient of the filtered audio signal. Figure 4.20 showcases the results in more detail.

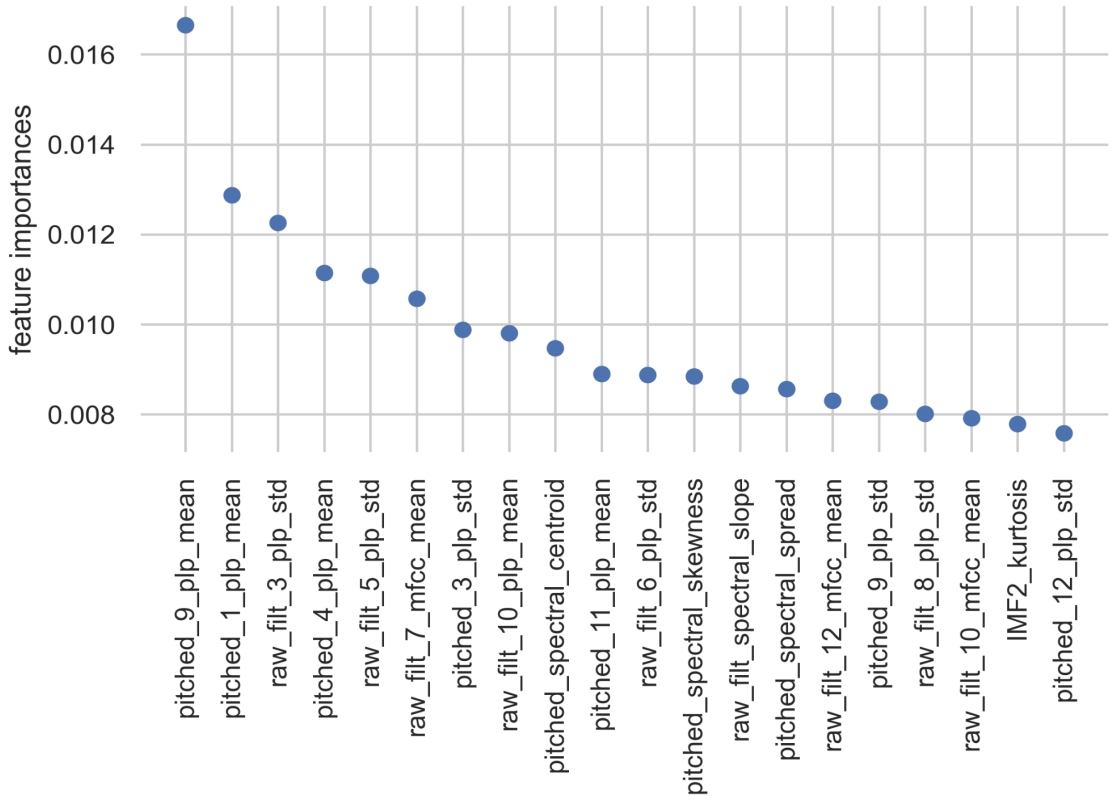

***Figure 4.20**. The top 20 features inherently obtained by the learning algorithm of the support vector machine classifier. Custom dataset with 200 ms window used as an input.*

Feature importance values for the custom dataset with 150 ms window showcased similar results as before. All 10 of the top 20 features are from the psychoacoustics domain, with 6 of them being from the pitch shifted signal. The list of five features with the highest importance values is given here: (1) mean of the 9th PLP of the shifted audio signal, (2) mean of the 1st PLP coefficient of the shifted signal, (3) mean of the 4th PLP coefficient of the shifted signal, (4) standard deviation of the 6th PLP coefficient of the filtered audio signal, and (5) standard deviation of the 3rd PLP coefficient of the filtered audio signal. This is depicted in Figure 4.21.

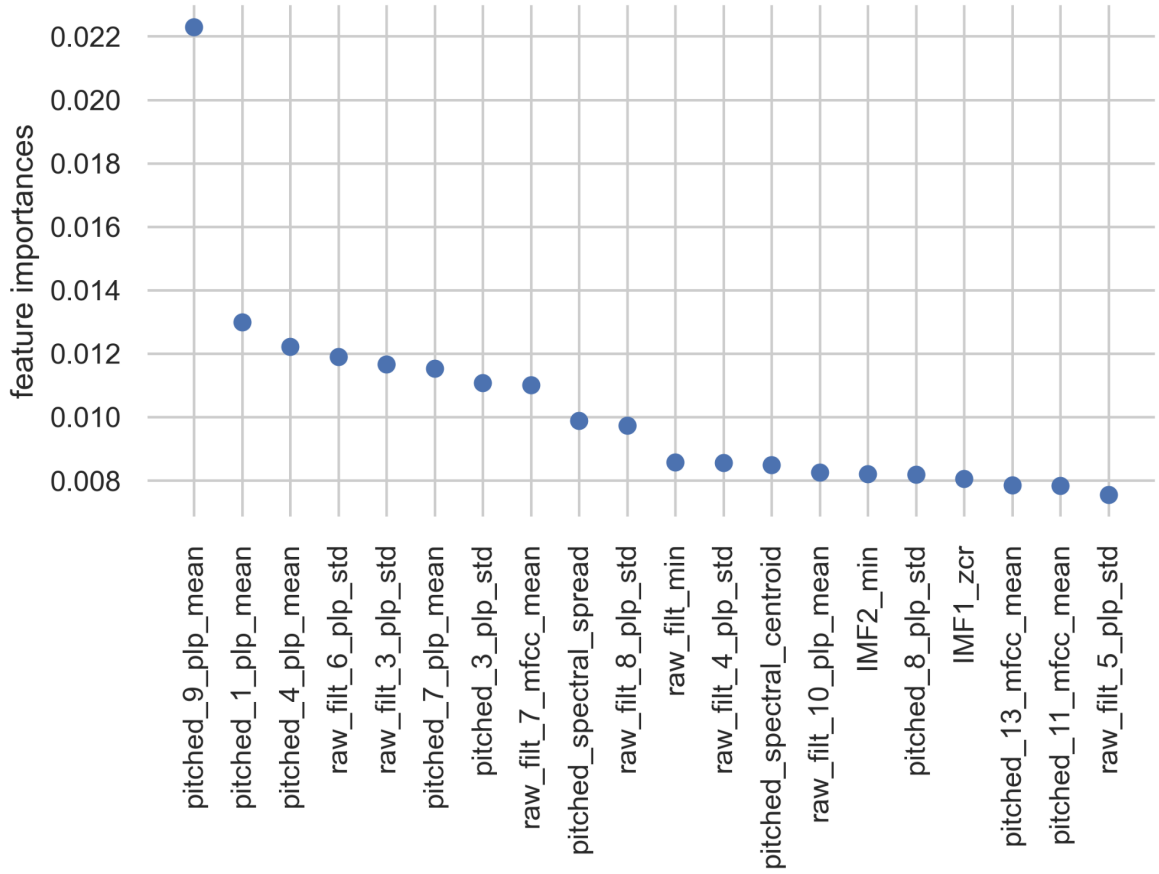

***Figure 4.21**. The top 20 features inherently obtained by the learning algorithm of the support vector machine classifier. Custom dataset with 150 ms window used as an input.*

The situation only changed by a minor degree for the case of 100 ms window custom dataset. One notable observation was regarding the number of the features from the shifted signals in the top 20 ranking, going down from 10 in the previous 2 cases to 8 in this one. However, since the rankings of the most prominent descriptors remained quite similar, it can be noted that the window size only affects the impact of pitch shifted signal version and the corresponding feature extraction in a very minor way. Top five features are given here: (1) mean of the 4th

PLP of the shifted audio signal, (2) mean of the 7th PLP coefficient of the shifted signal, (3) mean of the 1st PLP coefficient of the shifted signal, (4) standard deviation of the 6th PLP coefficient of the filtered audio signal, and (5) mean of the 9th PLP coefficient of the shifted audio signal. Figure 4.22 displays the results in more detail.

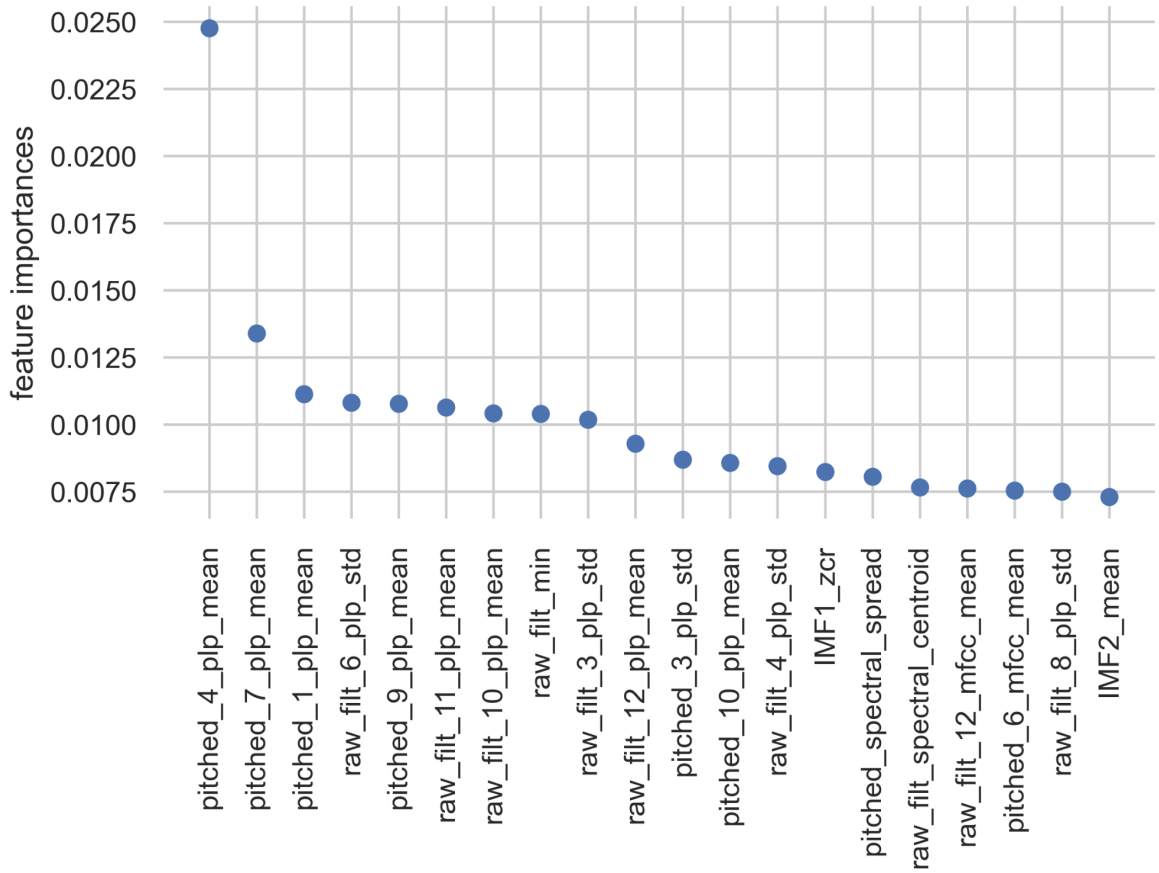

***Figure 4.22**. The top 20 features inherently obtained by the learning algorithm of the support vector machine classifier. Custom dataset with 100 ms window used as an input.*

4.3.5.2 Simulated datasets

Results with the simulated datasets, first of all the one with the SNR = -22 dB, exhibited the same trend compared to the custom datasets: PLP coefficients and MFCCs from the pitch shifted signals were ranked highest, with the smaller number of the aforementioned taken from the filtered unshifted signal also being present in the top 20. In other words, out of 18 features calculated through psychoacoustic modeling, 12 were gained from the pitch shifted representation. The top five features were: (1) mean of the 3rd MFCC of the shifted audio signal, (2) mean of the 7th PLP coefficient of the shifted signal, (3) standard deviation of the 9th PLP coefficient of the shifted signal, (4) standard deviation of the 7th PLP coefficient of

the filtered audio signal, and (5) standard deviation of the 3rd PLP coefficient of the shifted audio signal. The described results are displayed in Figure 4.23.

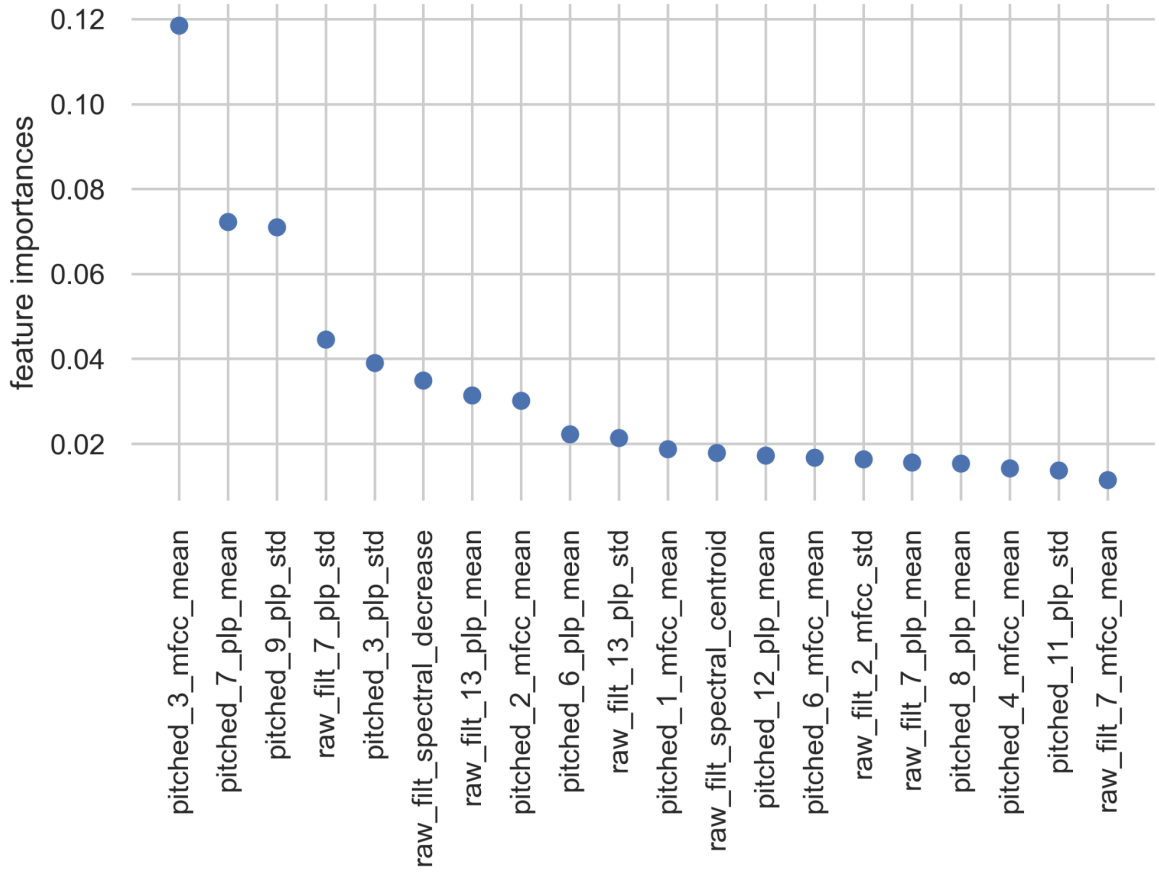

Figure 4.23. *The top 20 features inherently obtained by the learning algorithm of the support vector machine classifier. Simulated dataset with -22 dB SNR used as an input.*

Using the second simulated dataset (SNR of -24.4 dB) showed that 17 out of the best 20 features come from the psychoacoustics domain, while 13 of the top 20 came from the pitch shifted version of the input signal. The top five features ranked by feature importance are given here: (1) standard deviation of the 9th PLP coefficient of the shifted audio signal, (2) mean of the 7th PLP coefficient of the shifted signal, (3) spectral decrease of the filtered audio signal, (4) standard deviation of the 3rd PLP coefficient of the filtered audio signal, and (5) mean of the 6th PLP coefficient of the shifted audio signal. All of the aforementioned is taken from Figure 4.24 found hereafter.

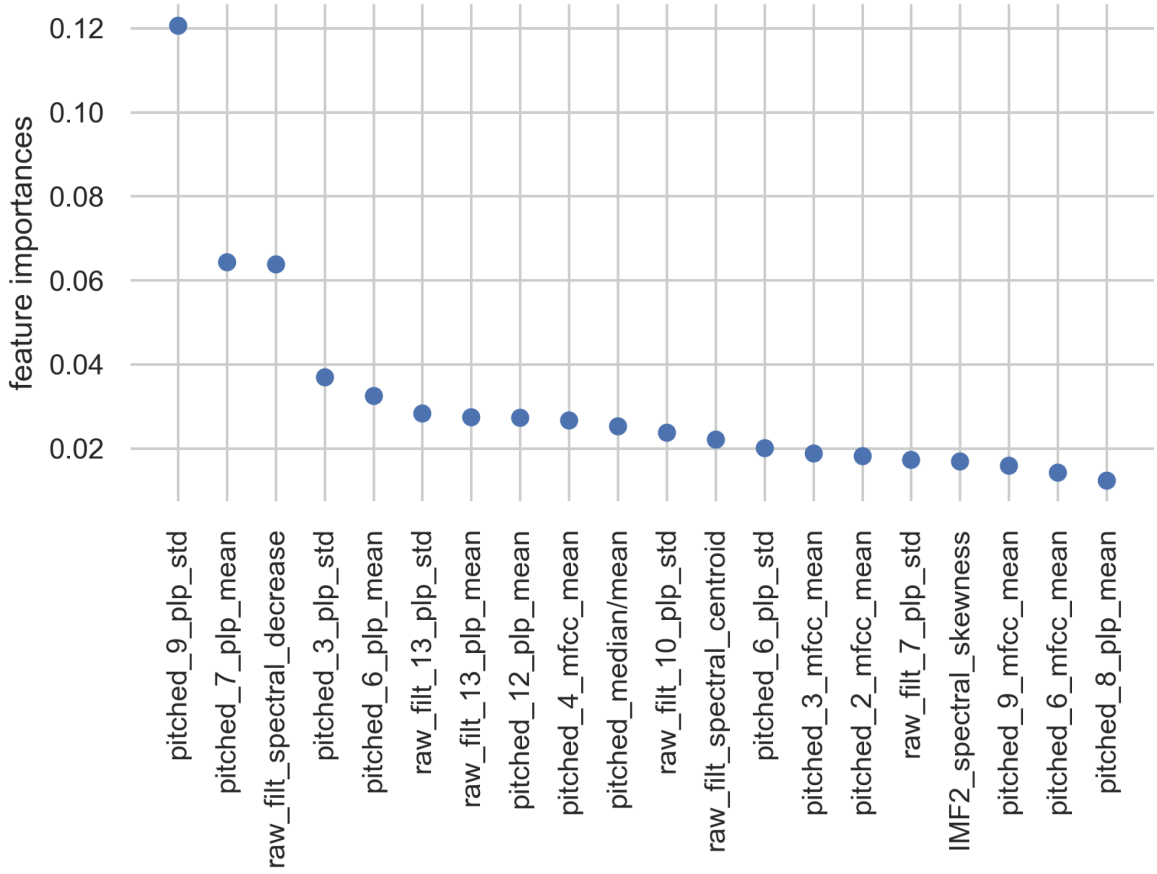

Figure 4.24*. The top 20 features inherently obtained by the learning algorithm of the support vector machine classifier. Simulated dataset with -24.4 dB SNR used as an input.*

The final case of the analysis included utilization of the simulated dataset with -26.7 dB of SNR. The results demonstrated in Figure 4.25 indicated repeatability of the subsection of the most important features: even though the specific order of the feature might change, PLP and MFCC-based descriptors outperform others, especially if employed on a frequency shifted signal. The top five features for this case were: (1) standard deviation of the 9th PLP coefficient of the shifted audio signal, (2) spectral decrease of the filtered audio signal, (3) mean of the 7th PLP coefficient of the shifted signal, (4) mean of the 13th PLP coefficient of the filtered signal, and (5) mean of the 6th PLP coefficient of the shifted audio signal.

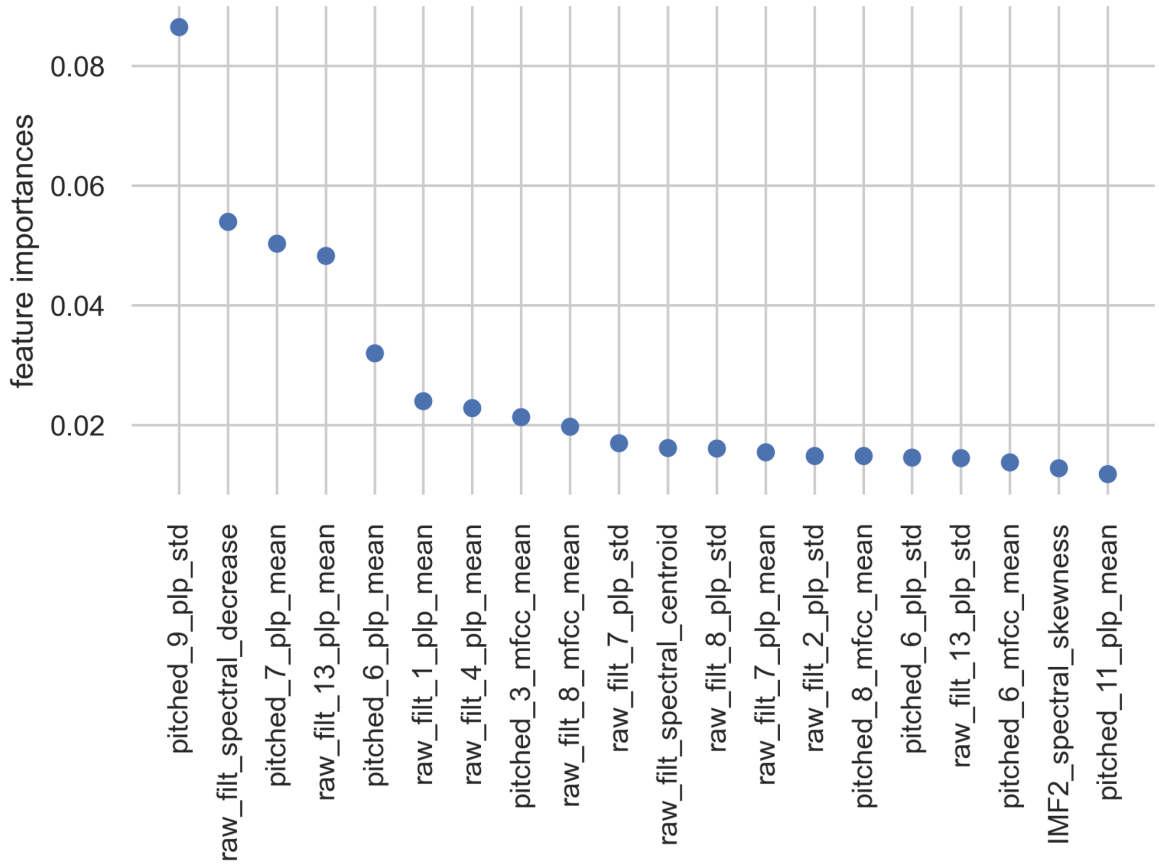

***Figure 4.25**. The top 20 features inherently obtained by the learning algorithm of the support vector machine classifier. Simulated dataset with -26.7 dB SNR used as an input.*

4.3.6 Recursive feature elimination

Recursive feature elimination was employed similarly to the approach taken in Scenario A. This time, the random forest ensemble was trained with the full set of features containing the filtered, pitch shifted and IMF-based subsets of descriptors, with the same approach utilized once again on separate subsets as well. 5-fold cross-validation was used as a way to average the accuracy results by choosing different data points for training and validation at each iteration. Cases with custom extracted datasets that include 150 and 100 ms window lengths were omitted from the analysis as they include a rather imbalanced distribution of positive and negative data points. In other words, these cases were used to assess the impact of window length on specific features, while their respective accuracy scores would contain less insight than for the tests with balanced datasets. More concretely, as the target variable gets much more observations in one class than the other, accuracy score can become meaningless [161]: for example, a model with custom dataset of 100 ms window size that contained 17:83 positive to negative label ratio

could have gotten to a fairly high accuracy of 83% by just labeling every data point as a negative observation. Balancing the dataset artificially was always a possible option, however avoided in this research due to the availability of "naturally" balanced datasets obtained through recorded and simulated data.

4.3.6.1 Custom dataset

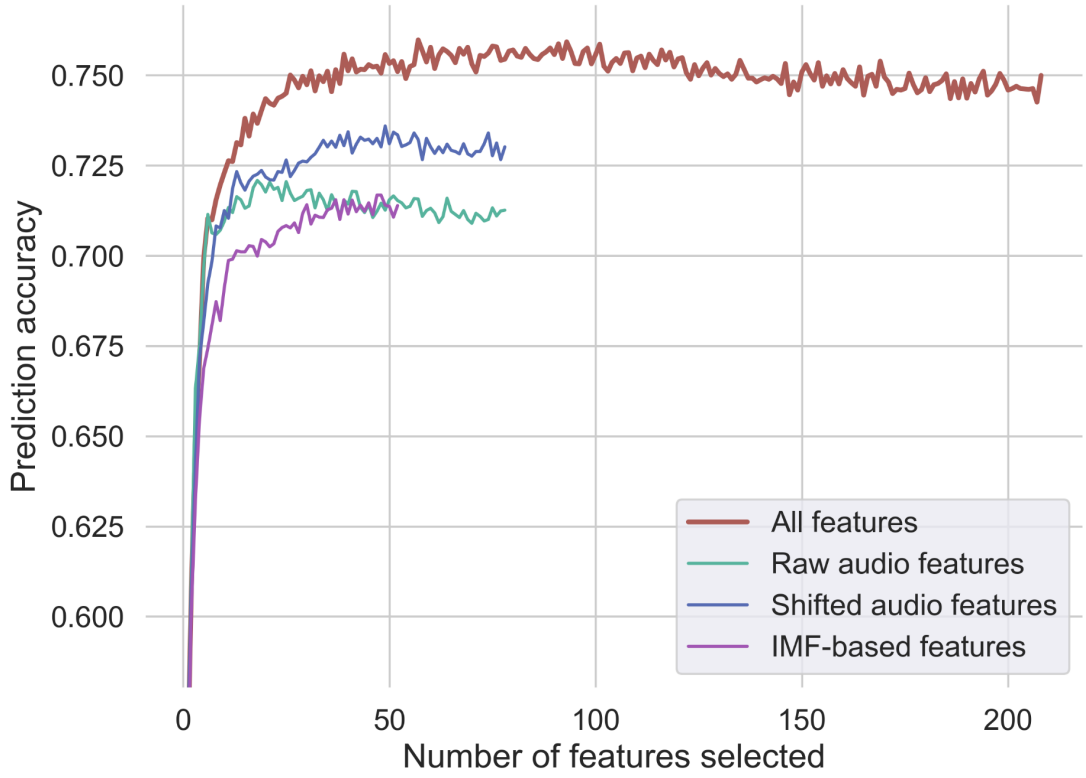

***Figure 4.26**. Prediction accuracies of random forest models trained with different feature subsets selected through recursive feature elimination. Custom dataset with 200 ms window used as an input.*

Recursive feature elimination returned the selection of 57 features for a full dataset of 212 features incorporated into a custom dataset with a 200 ms window. This included 14 features from filtered audio, 24 from pitch shifted and 19 from the IMF-based signals. Utilized on each of the 3 feature subsets, the method selected 18 out of 79 features from the filtered signal, 49 out of 79 features from the shifted audio and 48 out of 54 IMF-based features. The latter suggested a much "flatter" structure of feature importance values in the case of IMF-based descriptors compared to the ones taken from filtered audio. As the sifting process of EMD can converge to different states in any of the observations, making it more unstable than the statically filtered version, different IMF-based features might have contained variable levels of

relevant information through the entire dataset. Still, signal characteristics found in the noisy FPCG signal could definitely be used for the improvement of the model performance.

Training the random forest ensemble with varied number of features from the selected sets showcased the following results for the maximum achieved accuracy: 75.97% for the entire dataset, 72.08% for features from filtered audio, 73.59% for features from pitch shifted audio and 71.67% for IMF-based features. This exhibited an increase in accuracy of 3.89% compared to using only bandpass filtering as a preprocessing method. The overall results can be found in Figure 4.26.

4.3.6.2 Simulated datasets

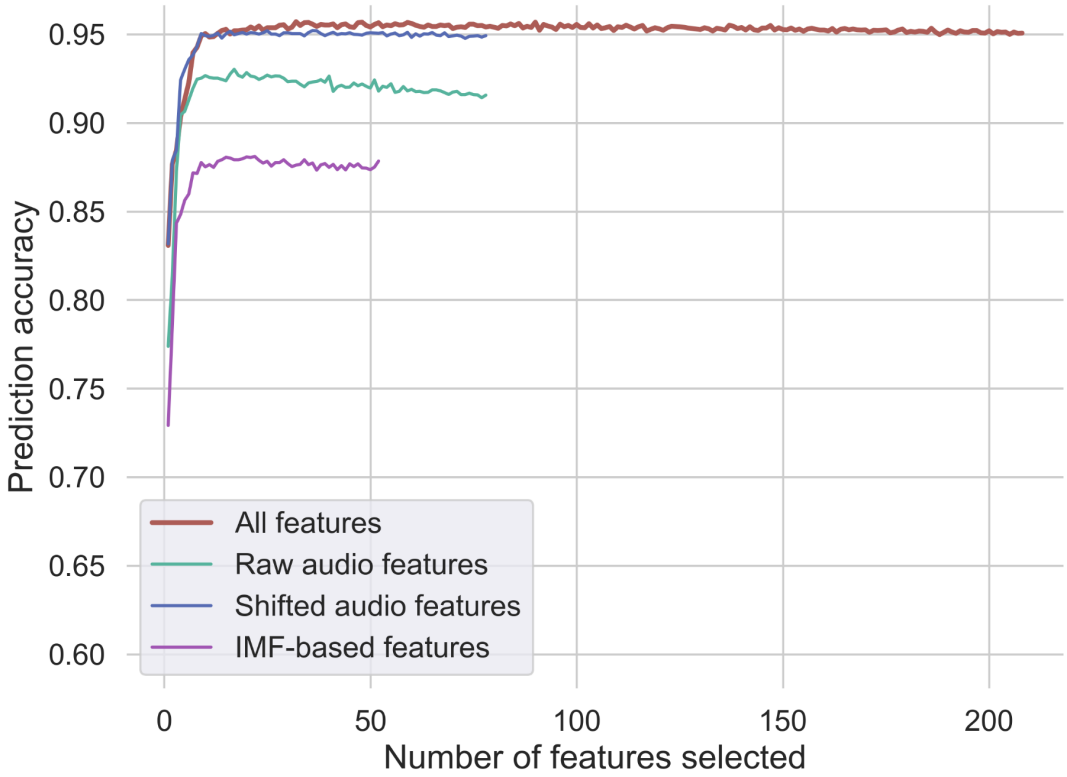

***Figure 4.27**. Prediction accuracies of random forest models trained with different feature subsets selected through recursive feature elimination. Simulated dataset with -22 dB SNR used as an input.*

Running the recursive feature elimination on a simulated dataset with -22 dB SNR showed the selection of only 32 features from the entire feature set, containing 5 from the filtered audio subset, 24 from the pitch shifted audio, and 4 from the IMFs. If each subset was observed

separately and RFE applied in this manner, 17 out of 79 features were selected from the filtered signal, 36 out of 79 from the pitch shifted version and 22 out of 54 from the IMFs.

The results depicted in Figure 4.27 show that the maximum accuracy of a random forest ensemble for the entirety of select features was 95.71%, with 93.02%, 95.21% and 88.05% for the subsets from filtered, shifted and IMF-based signals, respectively. This showed a boost in accuracy of 2.69% and a 38.5% reduction in error rate between the full selected set and the selected set from the filtered signal version.

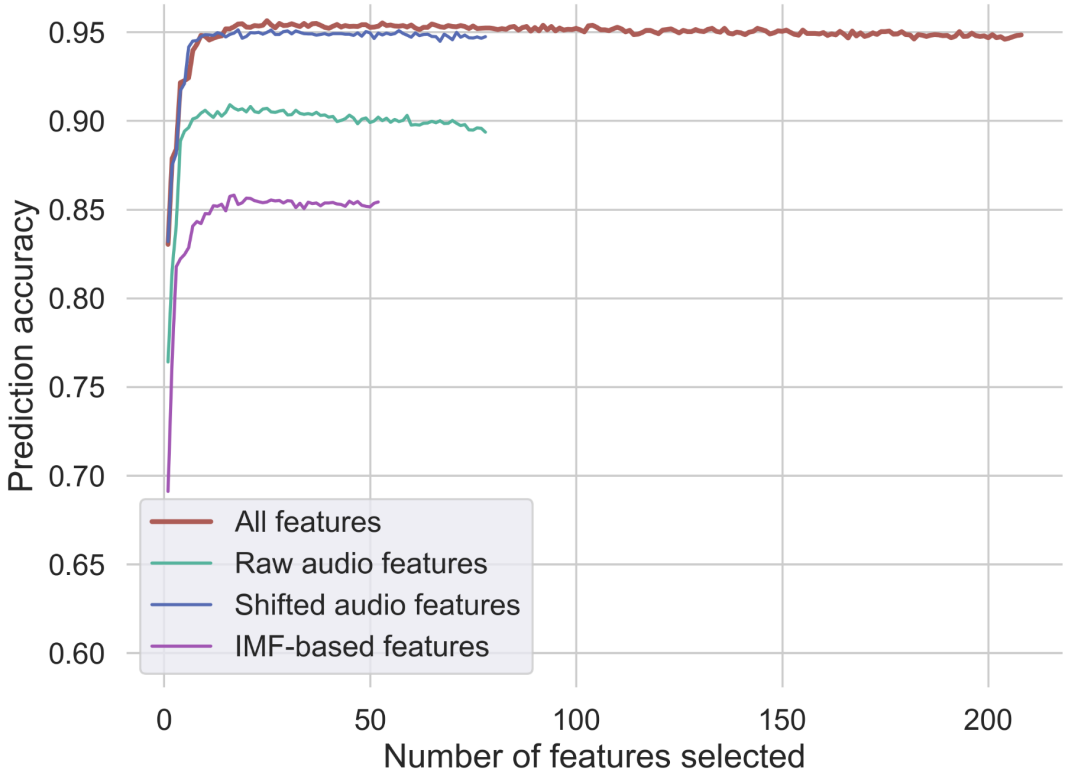

***Figure 4.28**. Prediction accuracies of random forest models trained with different feature subsets selected through recursive feature elimination. Simulated dataset with -24.4 dB SNR used as an input.*

Having a noisier simulated dataset with -24.4 dB of SNR yielded a selected set of only 25 features, with 4 from the filtered signal, 19 from the shifted version and only 2 from the IMFs. Observing the feature subsets only, RFE methods returned 16 out of 79 features from the filtered audio, 18 out of 79 from the shifted one and 17 out of 54 from the IMFs.

The maximum accuracies from the random forest ensembles were: 95.63% for the overall selected set, 90.89% for the statically filtered, 95.12% for the pitch shifted and 85.8% for the

IMF-based subsets. The increase in accuracy was 4.74%, decreasing the error rate by 52% if descriptors from the pitch shifted and IMF signals were added to the original filtered audio features. Figure 4.28 shows the evolution of prediction accuracy when more features from the selected subsets are added as an input to the model.

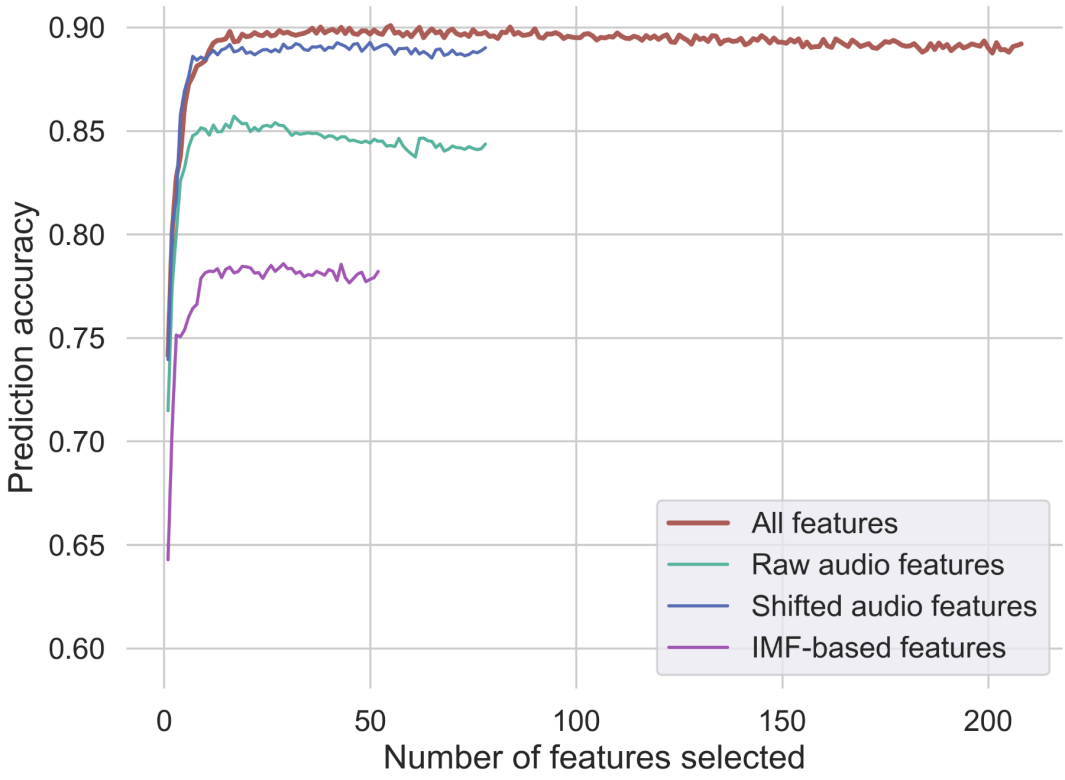

***Figure 4.29**. Prediction accuracies of random forest models trained with different feature subsets selected through recursive feature elimination. Simulated dataset with -26.7 dB SNR used as an input.*

Taking the noisiest available simulated dataset as an input (SNR = -26.7 dB), RFE found 55 features from the entire feature set, including 14 from the filtered signal, 35 from the shifted signal and 6 from the IMFs. Using the feature subsets in standalone yielded 17 out of 79 filter-based features, 50 out of 79 PS-based and 29 out of 54 IMF-based characteristics.

The accuracy scores (displayed in Figure 4.29) showed a maximum of 90.08% for the entire selected set, 85.71% for the filtered subset, 89.27% for the pitch shifted subset and 78.58% for the EMD subset. Prediction accuracy was increased by 4.37%, reducing the error by almost 31%.

4.3.7 Results with the chosen classifiers

Besides the demonstrated results of feature selection and ranking, the general predictive capability of the complete feature subsets and their corresponding combinations was evaluated through the 10-fold cross-validation employed on 2 different machine learning models. The chosen classifiers were the cubic SVM and bagged trees.

An SVM model with the nonlinear kernel was considered an appropriate option for assessing the overall validation accuracy outside the scope of feature selection, as the data itself has exhibited a highly complex separability [162]. The latter can be seen through the selection of a relatively large number of parameters, while the feature importance values were shown to include a quite gradual decline with fairly low values. To rephrase it, the FPCG signal classification requires a lot of parameters in order to maximize the quality of prediction, as the acoustic environment was heavily imbued with noise and relevant signal variations, both temporally and spectrally. Regarding the choice of the cubic kernel, it was taken as the common higher order polynomial kernel used in binary classification problems, while taking an even higher degree was discouraged due to a stronger possibility of overfitting the model [163].
The bagged trees classifier is an ensemble of n classification tree models (in this research, n was chosen to be 100) trained on different subsets of observations which are then included in the voting process. This approach can decrease the assessed error metric (e.g. mean square error) of the classifier and improve performance [164].

Besides the regularly used accuracy as a metric describing the classification quality, precision and recall [165] are also introduced in these results. In the context of the specific classification problem described in this research, precision is the probability of the positive identification actually being an S1 sound (true S1 sounds detected vs. all observations detected as S1 sounds), while recall describes the probability that the S1 sound is properly identified (true S1 sounds detected vs. all S1 sounds that should have been detected).

Table 4.2 exhibits the results for 4 balanced datasets (custom with 200 ms window and all simulated ones) taken on 7 different feature subset combinations: each of the 3 subsets (originating from filtered, pitch shifted and IMF signal representations), combinations of 2 subsets and a combined set of all features. “C” and “S” represent the custom and simulated datasets, respectively.

***Table 4.2**. Results for accuracy, precision and recall aggregated through 10-fold cross-validation. One custom dataset and 3 simulated ones were used as an input to the cubic SVM and bagged trees models. In the final 3 columns, F represents „Filtered“, S represents „Shifted“ and I represents „IMF-based“.*

Classifier	Case	All	Filtered	Shifted	IMF	F+S	F+I	S+I
Accuracy								
Cubic SVM	**C 200 ms**	76.16%	69.24%	72.74%	67.83%	73.73%	72.42%	75.48%
	S 22 dB	97.65%	94.79%	96.58%	87.31%	97.15%	95.22%	97.17%
	S 24 dB	97.42%	92.69%	96.49%	84.56%	97.07%	94.38%	96.95%
	S 26 dB	92.76%	87.62%	91.09%	77.05%	92.30%	88.44%	91.35%
Bagged trees	**C 200 ms**	74.60%	70.09%	72.05%	70.40%	72.34%	72.16%	72.90%
	S 22 dB	95.89%	91.40%	94.76%	86.87%	95.01%	92.43%	94.79%
	S 24 dB	94.36%	89.25%	94.86%	84.42%	94.35%	92.16%	94.61%
	S 26 dB	88.55%	83.99%	88.87%	77.18%	88.42%	84.78%	88.77%
Precision								
Cubic SVM	**C 200 ms**	76.70%	66.60%	70.32%	64.33%	71.89%	69.85%	73.69%
	S 22 dB	97.85%	95.32%	96.89%	87.82%	97.40%	95.40%	97.37%
	S 24 dB	97.51%	93.07%	96.76%	85.44%	97.27%	94.72%	97.11%
	S 26 dB	92.81%	88.73%	91.78%	78.73%	92.59%	88.72%	91.93%
Bagged trees	**C 200 ms**	75.47%	70.84%	72.98%	70.30%	73.95%	72.30%	73.65%
	S 22 dB	95.95%	91.80%	94.98%	89.52%	95.29%	93.30%	95.20%
	S 24 dB	94.63%	90.03%	94.84%	87.90%	94.60%	92.91%	94.72%
	S 26 dB	89.05%	85.51%	89.24%	80.61%	89.05%	86.53%	89.07%
Recall								
Cubic SVM	**C 200 ms**	65.47%	57.95%	63.47%	57.51%	64.51%	63.76%	67.35%
	S 22 dB	97.75%	94.91%	96.70%	88.50%	97.27%	95.66%	97.32%
	S 24 dB	97.67%	93.25%	96.68%	85.69%	97.24%	94.75%	97.19%
	S 26 dB	93.70%	88.00%	91.51%	78.14%	93.03%	89.76%	92.12%
Bagged trees	**C 200 ms**	61.17%	52.46%	56.19%	54.69%	55.68%	57.78%	58.15%
	S 22 dB	96.38%	92.12%	95.23%	85.40%	95.36%	92.47%	95.04%
	S 24 dB	94.83%	89.82%	95.58%	82.13%	94.83%	92.36%	95.20%
	S 26 dB	89.58%	84.30%	90.01%	75.40%	89.31%	84.67%	90.01%

Chapter 5

Discussion

The bio-inspired (perceptual) approach to extracting different features of sound signals, as well as using a data-driven method for sifting a number of representations of signal components, is fundamentally different from applying clear mathematical principles in order to capture a signal characteristic. The theoretical background in utilization of such procedures might be hard (or impossible) to provide, however their benefits seem to be substantial and make an excellent basis for discussion. This is especially apparent in the case of classification of extremely unclear, noisy and volatile presence of FPCG sounds in a recorded stream. Numerous sources of noise, such as mother's heart sound and fetal movements, are representing only some of the challenges that have to be prevailed by signal processing and feature extraction mechanisms employed for the task. Beyond the absolute levels of noise induced in the recording, the amplitude of the relevant FPCG sounds can vary drastically, depending on the position of the fetus, vicinity to the acoustic transducer and the acoustic impedance setup present in a particular moment.

The presented results have, as the most apparent point, shown clear increase in model metrics and separability if sets of features inspired by iterative and perceptually pertinent methods were put in conjunction with objective, mathematically clear signal descriptors from the temporal and spectral domains. In addition, the results have also displayed very interesting feature ranking scores.

5.1 EMD insights

5.1.1 Scenario A

The extracted dataset was gained through the usage of 7 raw data recordings, giving 7604 data points. The preprocessing and feature extraction steps resulted with three groups of features extracted from: 1) raw audio signals, 2) audio signals filtered with a bandpass filter with 50 and 150 Hz cut-off frequencies, 3) IMF signals obtained from audio signals filtered with a high-pass filter (cut-off frequency of 50 Hz). A valid question that may arise is why the same type

of preprocessing filter was not applied on the second signal used as a basis for calculating IMFs. Even though the majority of the FPCG energy was contained in the spectral band between 50 and 150 Hz, applying a bandpass filter here might have hindered the inherent ability of the EMD to converge to a relevant signal component. In other words, applying the sifting process for the extraction of useful harmonic information from the noisy spectrum is one of the core functionalities and applications of EMD. In order to achieve this and following the analysis of the FPCG signal noise spectrum in the custom raw dataset, EMD was used on signals that had noise components up to 1000 Hz in spectrum (Nyquist frequency for a 2 kHz sampling rate). Regarding the introduction of unfiltered signal representation in addition to the band-pass version, the motivation was to enrich the feature space so that the algorithms for feature ranking and selection would have a wider basis for finding the most useful and relevant features.

The 1st IMF has been shown to be highly relevant in the classification process and the accompanying machine learning procedures. First, a consistent result of the applied feature ranking and selection methods was the presence of spectral features of the 1st IMF as the most important IMF-based features. For example, the list of 48 features selected using the RFE method contained 8 of 9 spectral features of the 1st IMF, while 7 out of 10 best ranking features from ANOVA were taken from the 1st IMF as well. Even though these features are moderately correlated with the statistical features of the filtered audio signal, it seems that the IMF manages to capture the quality of intrinsic oscillation that is highly relevant for the fetal heartbeat classification task both when features are considered individually and when they are observed in the presence of other available features.

In the case of the 2nd and the 3rd IMFs, convergence to lower frequencies compared to the 1st IMF might explain their lower ranking and relevance. As the iterative procedure of EMD originally uses the higher frequency information and then moves to the lower parts of the spectrum in the subsequent runs, the results indicated the divergence from relevant FPCG information in the cases of higher IMFs. These signals might have contained information with less predictive power and higher levels of inhibiting noises present in the lower frequencies, such as maternal heart sounds and gastrointestinal activity.

Regarding features based on filtered audio signals, they were shown to be more relevant for the classification task than features based on raw audio. The results from the feature ranking methods demonstrated superior predictive power of the filtered audio signals compared to the

unfiltered raw audio version, suggesting that the noise contained in the entire used bandwidth (0-1000 Hz) decreased feature relevance and thus the quality of the trained model. However, utilization of EMD has also indicated the presence of useful information hidden in noise above the cut-off frequency of 150 Hz used in the signal representation preprocessed through the bandpass filter.

5.1.2 Scenario B

IMF-based features were also assessed in a more extensive Scenario B, containing 6 different extracted datasets taken from the custom recorded and the simulated data. As the first step, all 8 recordings from the raw dataset were employed this time, with signal window length for segmentation being changed in 3 test cases: 200 ms, 150 ms and 100 ms. This was done mostly to validate that the choice of the window size did not favor one feature set more than the other.

The introduction of the PS-processed signal representation, as well as perception-based feature extraction mechanism to the analysis, have given an additional dimension in comparing different feature sets and their impact on the classification process. In the case of extracted datasets from the recorded data, IMF-based features were present in most of the "Top n" lists in feature ranking methods, which is especially interesting given the fact that EMD-processed signals haven't employed any bio-inspired feature engineering. In other words, some features taken from the IMFs were deemed important enough to be placed high in the rankings, despite the strong "competition" provided by audio and psychoacoustic features taken from the filtered and pitch shifted signals. RFE method has also revealed intriguing results in selecting a high number of features taken from the IMFs for the "best" subset used for optimized model training and classification procedures. For example, in the case of the 200 ms window length, 20 out of 57 overall features were selected from the IMFs, while only 11 of the features were taken from the filtered version (both audio and psychoacoustic features). The ranking results have manifested similar performance in all datasets extracted from recorded data, with the following IMF-based features showing up in multiple places across all 3 dataset cases: spectral centroid, the 95th percentile divided by maximum, spectral kurtosis, spectral skewness, spectral slope, the 5th percentile divided by minimum, and the 90th percentile from the 1st IMF; the 95th percentile divided by maximum, the 5th percentile divided by minimum, kurtosis and minimum from the 2nd IMF. To sum it up, this strongly implied that the 1st and 2nd IMF shapes in the

temporal domain, along with the spectral content of the 1st IMF, were important for the task of FPCG classification in the case of custom recorded data.

The situation was more complicated in the case of datasets calculated from the simulated data. First, the modality of the simulation process (done by authors in [27]) and the labeling procedure introduced for the purposes of this work have removed any uncertainties originating from the labeling noise and severe variations of FPCG signal levels. Namely, the challenges and obstacles that were very apparent in the custom data did not exist in the simulated data, making the training process more straightforward and with much higher accuracy. Secondly, the noise distribution of the simulated data is a great deal "flatter" than in the case of custom data, since different noise components are being added through the entire spectrum, including the ambient noise originating from the environment. Rudimentary tests that consisted of visually inspecting the waveform have implied clearest IMF peaks if 250 Hz was chosen as the high cut-off frequency for the preprocessing bandpass filter, which is substantially lower than the chosen cut-off of 1000 Hz in the case of recorded data. This difference can be explained with the data-dependent sifting process of EMD being dependent on the levels of noise in the system, meaning the noise distribution is a big factor in choosing the best parameters for preprocessing. Some methods originating from the EMD, such as EEMD, include the addition of white noise to the original signal to ensure better convergence, but the amount of added noise then becomes another parameter that needs tuning.

IMF-based features were thus considerably less present in the feature rankings for datasets extracted from simulated data. Besides the (potential) imperfect convergence towards the optimal set of relevant FPCG signal components, it would seem that the non-iterative procedures such as IIR filtering and pitch shifting have extracted sufficiently important information from the raw signal beyond the high levels of noise artificially superimposed on the FPCG sounds. Apart from that, results on very low levels and amplitude variations of fetal heart sounds in the recorded data compared to the cleaner situation of simulated data suggest the ability of EMD to converge towards meaningful data in "difficult" scenarios. To put it another way, IMFs may yield additional insight into the signal characteristics, especially if the original data is corrupted by a multitude of factors, such as very low SNR, varied signal amplitude levels, dubious signal presence and spectral signature changes.

5.2 Remarks on pitch shifting and psychoacoustics

5.2.1 Pitch shifting

The main motivation for the utilization of pitch shifting as a preprocessing method was to exploit the non-linear placement of the critical bands, i.e. to transpose the frequency content to higher values where human hearing has higher sensitivity. The results for all six dataset cases in Scenario B showed superior performance of PS-processed features compared to both the features taken from the filtered audio and IMF-based descriptors, which could be seen in the vast majority of feature ranking and RFE methods. As another intriguing outcome, besides the psychoacoustic features, some audio features based on the pitch shifted signal were also ranked highly in the results, mainly including the characteristics based on frequency content such as spectral centroid, spectral kurtosis and spectral spread.

The impact of pitch shifting on audio features can be explained by the following: the instantaneous frequency calculation for any given subwindow is imperfect and may change the phase of the particular frequency component, thus changing the temporal and spectral content of the resulting signal through constructive and destructive interference of the signals in adjacent subwindows. This might extract some additional relevant information from the audio. Furthermore, the “lossy” process of phase vocoding that includes FFT and its inverse can warp and reshape the spectral content of the shifted signal compared to the original one (besides the simple frequency transposition).

The taxonomy is not clear on the topic of subjectivity of audio features taken from the pitch shifted signal. On one side, the modality of the audio features implies objective characterization of the descriptors, while the frequency shifting procedure has made the input signal more appropriate for human hearing sensitivity. In any case, good ranking of both audio and psychoacoustic features from pitch shifted signals is highly encouraging for future research.

5.2.2 Psychoacoustics

Employment of bio-inspired features for the classification of FPCG signals has proven to be a critical point of this research. PLP coefficients and MFCCs computed on the bandpass filtered and pitch shifted audio representations were regularly ranked as features with the most significance, especially in the case of the spectrally transposed signal. Even further, it has been

shown that the utilization of psychoacoustics with pitch shifting as preprocessing increased accuracy of the model by a noticeable degree for all 6 dataset cases.

As a first interesting point, it must be noted that psychoacoustic modelling done on the pitch shifted signal has demonstrated superior performance compared to the one processed on the filtered audio. As an example, ranking results for mutual information in the case of extracted custom dataset with 200 ms window length have shown that 13 out of 20 best ranked features come from the PS + psychoacoustics family of features, while there are no features based on psychoacoustics and raw filtered signals. The situation was almost identical in the case of simulated datasets. Secondly, RFE analysis done on all features for the custom datasets has shown that more than two thirds of psychoacoustics-based features were chosen from the PS-processed inputs: 12 out of 15 for the extracted custom dataset on 200 ms window and 28 out of 37 extracted from the simulated data with SNR of -26.7 dB.

Psychoacoustic features also outperform audio features taken both from the filtered audio and the IMFs. As an example, embedded ranking through a support vector machine classifier done on custom raw data with 150 ms window size demonstrated that all of the first 10 places have been occupied by the MFCC and PLP features. There are several aspects that influence the aforementioned results:

1. Audio descriptors based on psychoacoustics have been shown to exhibit near human-like performance in classification, such as in the case of auditory scene recognition [166]. Modelling of the auditory system was feasible in the case of FPCG, especially due to their potential of being perceived even by untrained personnel, as demonstrated in [91].

2. Shifting the audio to higher frequencies could have increased the excitation of a multitude of critical bands, both in the case of MFCCs (mel bands) and PLP coefficients (Bark bands). As pitch shifted representations have outperformed the non-pitch shifted counterparts, it is reasonable to imply that the widening of the spectrum and the shift towards biologically more sensitive frequencies have significantly influenced this.

3. Nonlinearities induced in the processing stage (logarithmic operator in MFCCs and cubic root in PLP coefficients) have increased the noise robustness of specific features, as described in [167].

4. Spectral shaping utilized in the process has helped to close the gap between objective representations and human hearing. This is especially apparent in the case of PLP, where equal loudness contours have been calculated and applied to the input.

Taking the discussion further in, it was clear from the results that the features based on PLP were ranked somewhat higher than the MFCCs. Besides the already mentioned equal loudness contouring introduced in PLP coefficients (compared to preemphasis and liftering found in MFCCs), the second possible reason for better feature ranking is the cubic-root conversion of amplitude found in PLP generation, showing superior noise robustness to the logarithmic compression used by the MFCC calculation process [168]. Secondly, the autoregressive process for PLP calculation smooths out some of the details from the audible spectrum, making it potentially more robust than the liftering postprocessing found in MFCCs.

5.3 Impact of feature subsets on classification quality

The overall results indicated that both empirical mode decomposition and pitch shifting, as well as psychoacoustics modelling of features, have a positive impact on the quality of the FPCG signal classification model.

Focusing solely on Scenario A, which contained audio and IMF-based features, the results indicated that EMD seems to be an adequately robust and appropriate method to extract the information from the higher-frequency range. Features processed with EMD thereby exhibit their significant contribution to the classification performance, confirmed by training and cross-validating multiple classifiers. The improvements of accuracy as a result of using selected IMF and audio features instead of all audio features ranged from 3.75% for linear support vector machine and up to 10.28% for multilayer perceptron. The combination of different feature subsets leads to the construction of a smaller, however more efficient and relevant set of characteristics.

Scenario B introduced a more complex collection of tests, with 6 different dataset cases and more preprocessing and feature extraction steps undertaken in the analysis. In recursive feature elimination, all test cases exhibited superior performance of the selected subset of features, containing all 3 groups of different processing steps (static bandpass filter, EMD and pitch shifting) and 3 groups of features (statistical, spectral and perceptual). Compared to using only

features on static bandpass filtering (taken as a baseline), the accuracy and robustness of the model increased for each one of the input datasets.

Additionally, two classifiers (cubic SVM and bagged trees) were chosen to assess the performance of the overall feature set and the combinations of specific subsets, in order to evaluate the general predictive power of the introduced preprocessing and feature extraction methods. Cross-validation with 10 folds was used as an aggregating mechanism for the results, which rated the prediction accuracy, precision and recall. Starting with accuracy, cubic SVM showed an increase from 69.24% for filtered features only to 76.16% if all features were used, yielding a difference of almost 7% if custom dataset was taken as an input. The results on the same classifier demonstrated that the combination of features based on pitch shifted signals and IMFs outperformed all other combinations of 2 subsets. The results were similar for the bagged trees classifier, exhibiting an improvement of almost 4% if pitch shifted and EMD subsets of features were added to the ones originating from the filtered signal. Regarding simulated datasets, the accuracies were shown to be slightly higher than 90% if all features were used, with very strong reductions in error rates compared to using only bandpass filter-based features. For example, in the case of the simulated dataset with -24.4 dB SNR, adding descriptors from the proposed methods to features based only on static filtered signal version increased the accuracy from 92.69% to 97.42%, reducing the misclassification rate almost 3 times. Precision was also increased by a high degree for all 4 introduced datasets, especially in the case of a custom dataset with 200 ms window, where it increased from 66.60% to 76.70% if EMD and PS-based features were added to the ones based on frequency filtering. The same was the case for recall as well, increasing with the introduction of proposed features in all 4 scenarios.

Using a separate test set for the validation of classification accuracy on data coming from "unseen" sources (as expected in production-ready models) was contemplated, but subsequently discarded from the research. The impact of various methods for improving the quality of machine learning models for FPCG classification was demonstrated on two fundamentally different datasets that yielded very different model accuracies through cross-validation (roughly 75% for the raw dataset and approximately 90-95% for the simulated datasets), meaning the predicting potential of the trained model is extremely correlated with the quality of the labeled data. Additionally, it would be critical to tune the hyperparameters for the best performing classifier trained with data coming from the optimal distribution expected in

real-world scenarios, making this a challenge best addressed in a professional and potentially non-academic environment.

5.4 The proposed algorithm

This research has developed and analyzed an algorithm for the detection of S1 sounds taken from the stream of FPCG signals that can be used in real time. On top of the proposed approaches for preprocessing and feature extraction, a list of feature selection and ranking methods has been introduced as means of increasing the accuracy and robustness of the classification process within the machine learning principles. The showcased classifiers were not tuned in terms of hyperparameters, as the most appropriate choice of the model heavily depends on the distribution of data used as an input for the training process.

The components of the algorithm (block diagram given in Figure 5.1) can be found in different section of this thesis:

- Signal preprocessing
    - Empirical mode decomposition (Methodology section, subsection Empirical mode decomposition)
    - Pitch shifting (Methodology section, subsection Psychoacoustics)
- Data preparation (Methodology section, subsection Data preparation)
- Feature extraction
    - Psychoacoustics (Scenario B section, subsection Feature extraction)
    - “Conventional” audio features (Scenario A section, subsection Feature extraction)
- Feature selection and ranking (Scenario A and Scenario B sections, subsection Results)
- Model training and validation (Scenario A and Scenario B sections, subsection Results).

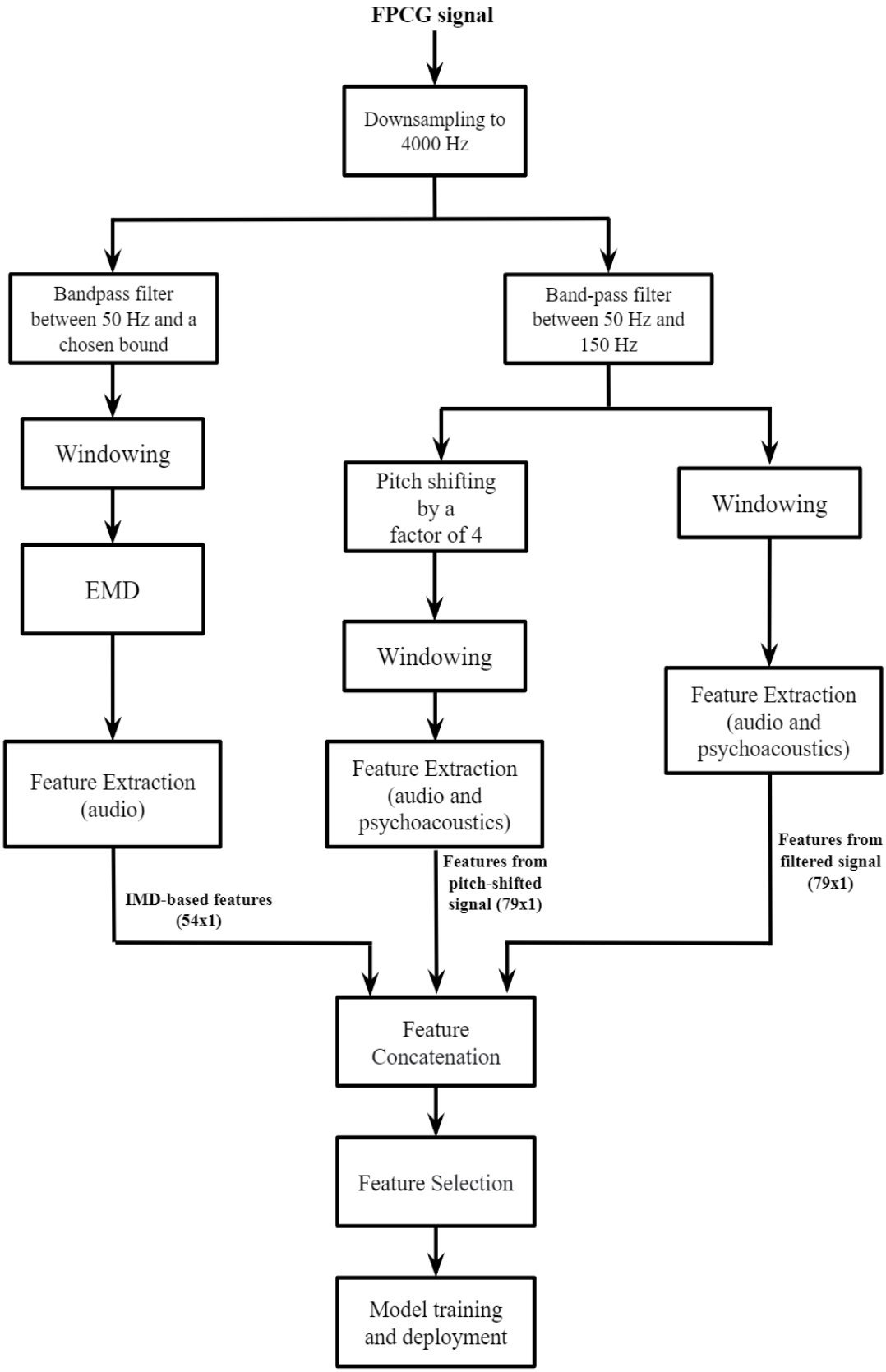

***Figure 5.1**. Block diagram of the overall proposed algorithm for FPCG signal classification.*

Chapter 6

Conclusion

Fetal phonocardiography is the oldest method for assessing the welfare of the fetus, inherently suffering from a number of issues, such as poor signal-to-noise ratio and strong dependency on probe positioning. However, the method itself works with just one passive sensor (microphone), giving it a very strong potential to be used as an easily accessible system in prenatal monitoring. This research has aimed to demonstrate the impact of empirical mode decomposition, pitch shifting and psychoacoustic-based feature extraction on the classification of fetal heartbeat sounds. Furthermore, the utilization of machine learning principles for automatic detection seemed to be a good direction for validating important parameters of fetal health.

The results have demonstrated a very positive impact of all three methods on the final classification metrics. In Scenario A, IMF-based features were shown to improve the fetal heartbeat detection accuracy when added to the set of conventional audio features. Four different classifiers were chosen for assessing the detection accuracy for a rather noisy dataset (both in raw data and labels), with models trained on a selected subset of features containing both audio and IMF-based features outperforming the models having only audio features as input data. The detection accuracies improved by 4.6%, 3.92%, 3.75%, and 10.28%, taking random forest ensemble, logistic regression model, linear support vector machine, and a multi-layer perceptron respectively as the chosen classifiers. In Scenario B, features calculated through EMD have shown high ranking in the cases with raw custom data compared to the ones gained through filtered audio, albeit lower than the newly introduced pitch shifted signal representation and the corresponding features. Interestingly enough, the ranking and relevance of IMF-based features started to drop with the introduction of simulated data as the input, where the overall model accuracy levels were above 90%. The reasoning for this is the following: very low SNR levels (especially for the real-world, recorded data), drastic amplitude variations for FPCG signal (i.e. fetus repositions itself) and spectral content changes (originating from the probe positioning and acoustic impedance shifts) are well suited for an iterative, data-driven method to “draw out” the relevant signal characteristics. In the case of mostly stable signal levels (even with very low SNR) as in the simulated data, it would seem that the method cannot

contribute to the predicting power of the overall model as much as in the more volatile situations. Furthermore, the quality of the sifting process convergence depends on the noise content found in frequencies above the relevant spectrum. As explained in the Discussion section, varied cut-off frequencies were required for different datasets for EMD to converge to the relevant content. This can be mitigated by various EMD extensions, however the stability of the process driven by its own input data might be challenging.

The situation is somewhat clearer for the pitch shifted preprocessing and perceptual feature extraction. Regarding the former, originally introduced as a preprocessing step for psychoacoustic modelling, it has been suggested that the ranking of some statistical and spectral features employed on the shifted signal was also high, giving insight on the capacity of phase vocoding to reveal some more information in the original data. In order to check the impact of pitch shifting on psychoacoustics, the same perceptual features were also extracted from the filtered non-shifted signal. The results have shown superior ranking of the psychoacoustic features achieved through pitch shifted input, and not only to the psychoacoustics aimed towards non-shifted signals, but in general. This implies that the number of bio-inspired processing steps introduced to close the gap between automatic classification and human hearing are successfully utilized to increase the noise robustness and draw additional insight and dimension separability from the input data, especially if this data is pushed towards the more sensitive spectral range. PLP features have demonstrated somewhat better results than MFCC, which can be explained by several steps that PLP calculation utilizes in order to make itself more suitable for the task.

Regarding concrete accuracy gains in Scenario B and by having a larger group of filtered audio-based features as a baseline, the subsets of descriptors containing all 3 groups of different processing steps (static bandpass filter, EMD and pitch shifting) and 3 groups of features (statistical, spectral and perceptual) were used to achieve machine learning models that surpass the ones containing only the baseline characteristics. This was confirmed by training two robust models (cubic SVM and bagged trees) on all combinations of feature subsets. For the label-balanced case in the custom raw dataset (200 ms window length), the classification accuracy jumped from 69.2% to 76.2% with the combination of all feature subsets if the cubic SVM classifier is used. For the simulated raw dataset (SNR of -24.4 dB) the accuracy moved from 92.7% to 97.4%, reducing the misclassification rate nearly 3 times. The results for other dataset cases were fairly similar to the ones mentioned here. It was shown that very promising gains in

prediction accuracy, precision and recall can be achieved by adding audio and psychoacoustic features based on pitch shifting and EMD to the conventional set of audio features extracted from the filtered audio.

These encouraging findings heavily imply the strong impact of EMD, pitch shifting and psychoacoustics on the overall classification process. The algorithm that does segmentation, preprocessing and feature extraction for predicting the label of new data points in the FPCG signal stream can work in real time, making it feasible for implementation as a proof-of-concept FPCG signal classifier, given a large enough and properly labeled dataset that would conform to the real-world distribution. The feature ranking and selection routines seem critical for the entire prediction process, highlighting important characteristics and maximizing predictive power of the chosen subset of features. As shown in this research, features chosen for the final subset in both Scenarios were taken from every group, with the combination of pitch shifting and psychoacoustic modelling being the most impactful.

Bibliography

[1] Signorini, M.G., Pini, N., Malovini, A., Bellazzi, R., & Magenes, G. (2020). Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. *Computer methods and programs in biomedicine*, *185*, 105015.

[2] Di Maria, C., Liu, C., Zheng, D., Murray, A., & Langley, P. (2014). Extracting fetal heart beats from maternal abdominal recordings: selection of the optimal principal components. *Physiological measurement*, *35 8*, 1649-64.

[3] Kahankova, R., Martínek, R., Jaros, R., Behbehani, K., Matonia, A., Jezewski, M., & Behar, J.A. (2020). A Review of Signal Processing Techniques for Non-Invasive Fetal Electrocardiography. *IEEE Reviews in Biomedical Engineering*, *13*, 51-73.

[4] Abdulhay, E.W., Oweis, R.J., Alhaddad, A.M., Sublaban, F.N., Radwan, M.A., & Almasaeed, H.M. (2014). Review Article: Non-Invasive Fetal Heart Rate Monitoring Techniques. *Biomedical Science and Engineering, 2 3,* 53-67.

[5] Martínek, R., Barnova, K., Jaros, R., Kahankova, R., Kupka, T., Jezewski, M., Czabanski, R., Matonia, A., Jezewski, J., & Horoba, K. (2020). Passive Fetal Monitoring by Advanced Signal Processing Methods in Fetal Phonocardiography. *IEEE Access, 8*, 221942-221962.

[6] Scarpato, N., Pieroni, A., Nunzio, L.D., & Fallucchi, F. (2017). E-health-IoT Universe: A Review. *International Journal on Advanced Science, Engineering and Information Technology, 7*, 2328-2336.

[7] Mhajna, M., Schwartz, N., Levit-Rosen, L., Warsof, S.L., Lipschuetz, M., Jakobs, M., Rychik, J., Sohn, C., & Yagel, S. (2020). Wireless, remote solution for home fetal and maternal heart rate monitoring. *American journal of obstetrics & gynecology MFM, 2 2*, 100101.

[8] Lanssens, D., Vonck, S., Storms, V., Thijs, I.M., Grieten, L., & Gyselaers, W. (2018). The impact of a remote monitoring program on the prenatal follow-up of women with gestational hypertensive disorders. *European journal of obstetrics, gynecology, and reproductive biology, 223*, 72-78.

[9] Lanssens, D., Vandenberk, T., Smeets, C.J., De Cannière, H., Molenberghs, G., Van Moerbeke, A., van den Hoogen, A., Robijns, T., Vonck, S., Staelens, A., Storms, V., Thijs, I.M., Grieten, L., & Gyselaers, W. (2017). Remote Monitoring of Hypertension Diseases in Pregnancy: A Pilot Study. *JMIR mHealth and uHealth, 5 3*.

[10] van den Heuvel, J.F., Groenhof, T.K., Veerbeek, J.H., van Solinge, W.W., Lely, A.T., Franx, A., & Bekker, M.N. (2018). eHealth as the Next-Generation Perinatal Care: An Overview of the Literature. *Journal of Medical Internet Research, 20 6*.

[11] Angelov, G.V., Nikolakov, D.P., Ruskova, I.N., Gieva, E.E., & Spasova, M.L. (2019). Healthcare Sensing and Monitoring. *Enhanced Living Environments*.

[12] Banik, S., Melanthota, S.K., Arbaaz, Vaz, J.M., Kadambalithaya, V.M., Hussain, I., Dutta, S., & Mazumder, N. (2021). Recent trends in smartphone-based detection for biomedical applications: a review. *Analytical and Bioanalytical Chemistry, 413*, 2389 - 2406.

[13] Ceylan Koydemir, H., & Ozcan, A. (2018). Smartphones Democratize Advanced Biomedical Instruments and Foster Innovation. *Clinical Pharmacology & Therapeutics, 104 1,* 38-41.

[14] Vashist, S.K., Schneider, E.M., & Luong, J.H. (2014). Commercial Smartphone-Based Devices and Smart Applications for Personalized Healthcare Monitoring and Management. *Diagnostics, 4*, 104-128.

[15] Adam, J. (2012). The future of fetal monitoring. *Reviews in Obstetrics and Gynecology, 5 3-4*.

[16] Chakladar, A., & Adams, H. (2009). Dangers of listening to the fetal heart at home. *BMJ: British Medical Journal, 339*.

[17] Ang, E., Glunčić, V., Duque, A., Schafer, M.E., & Rakic, P. (2006). Prenatal exposure to ultrasound waves impacts neuronal migration in mice. *Proceedings of the National Academy of Sciences, 103*, 12903-12910.

[18] Schneider-Kolsky, M.E., Ayobi, Z., Lombardo, P., Brown, D., & Gibbs, M.E. (2009). Ultrasound exposure of the foetal chick brain: effects on learning and memory. *International Journal of Developmental Neuroscience, 27*, 677-683.

[19] Church, C.C., & Miller, M.W. (2007). Quantification of risk from fetal exposure to diagnostic ultrasound. *Progress in biophysics and molecular biology, 93 1-3*, 331-53.

[20] Romano, M., Cesarelli, M., D'Addio, G., Mazzoleni, M.C., Bifulco, P., Ferrara, N., & Rengo, F. (2010). Telemedicine Fetal Phonocardiography Surveillance: an Italian Satisfactory Experience. *Studies in health technology and informatics, 155*, 176-81.

[21] Moghavvemi, M., Tan, B.H., & Tan, S.Y. (2003). A non-invasive PC-based measurement of fetal phonocardiography. *Sensors and Actuators A-physical, 107*, 96-103.

[22] Lamonaca, F., Polimeni, G., Barbé, K., & Grimaldi, D. (2015). Health parameters monitoring by smartphone for quality of life improvement. *Measurement, 73*, 82-94.

[23] O'Dowd, M., & Philipp, E.E. (1994). The History of Obstetrics and Gynaecology. *CRC Press*.

[24] Lewis, D., & Downe, S. (2015). FIGO consensus guidelines on intrapartum fetal monitoring: Intermittent auscultation. *International Journal of Gynecology & Obstetrics, 131 1*, 9-12.

[25] Mdoe, P.F., Ersdal, H.L., Mduma, E., Moshiro, R., Kidanto, H., & Mbekenga, C. (2018). Midwives' perceptions on using a fetoscope and Doppler for fetal heart rate assessments during labor: a qualitative study in rural Tanzania. *BMC pregnancy and childbirth*, *18 1*, 103.

[26] Abbas, A.K., & Bassam, R. (2009). Phonocardiography Signal Processing. *Synthesis Lectures on Biomedical Engineering, 4*, 1-194.

[27] Cesarelli, M., Ruffo, M., Romano, M., & Bifulco, P. (2012). Simulation of foetal phonocardiographic recordings for testing of FHR extraction algorithms. *Computer methods and programs in biomedicine, 107 3*, 513-23 .

[28] Várady, P., Wildt, L., Benyó, Z., & Hein, A. (2003). An advanced method in fetal phonocardiography. *Computer methods and programs in biomedicine, 71 3*, 283-96 .

[29] Mittra, A.K., Choudhary, N.K., & Zadgaonkar, A. (2008). Development of an artificial womb for acoustical simulation of mother's abdomen. *International Journal of Biomedical Engineering and Technology, 1*, 315.

[30] Adithya, P.C., Sankar, R., Moreno, W.A., & Hart, S. (2017). Trends in fetal monitoring through phonocardiography: Challenges and future directions. *Biomedical Signal Processing and Control., 33*, 289-305.

[31] Kovács, F., Kersner, N., Kádár, K., & Hosszú, G. (2009). Computer method for perinatal screening of cardiac murmur using fetal phonocardiography. *Computers in biology and medicine, 39 12*, 1130-6 .

[32] Hornberger, L.K., & Sahn, D.J. (2007). Rhythm abnormalities of the fetus. *Heart, 93*, 1294-300.

[33] Leung, T.S., White, P.R., Collis, W.B., Brown, E., & Salmon, A.P. (2000). Classification of heart sounds using time-frequency method and artificial neural networks. *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2*, 988-991.

[34] Ţarălungă, D.D., Tautan, A., & Ungureanu, G.M. (2018). An Efficient Method for Fetal Heart Sounds Detection Based on Hilbert Transform. *2018 International Conference and Exposition on Electrical And Power Engineering (EPE)*, 916-9.

[35] Chourasia, V.S., & Tiwari, A.K. (2013). Design Methodology of a New Wavelet Basis Function for Fetal Phonocardiographic Signals. *The Scientific World Journal, 2013 5-6*.

[36] Koutsiana, E., Hadjileontiadis, L.J., Chouvarda, I., & Khandoker, A.H. (2017). Fetal Heart Sounds Detection Using Wavelet Transform and Fractal Dimension. *Frontiers in Bioengineering and Biotechnology, 5*.

[37] Tomassini, S., Strazza, A., Sbrollini, A., Marcantoni, I., Morettini, M., Fioretti, S., & Burattini, L. (2019). Wavelet filtering of fetal phonocardiography: A comparative analysis. *Mathematical biosciences and engineering: MBE, 16 5*, 6034-6046 .

[38] Akay, M., & Mulder, E.J. (1996). Examining fetal heart-rate variability using matching pursuits. *IEEE Engineering in Medicine and Biology Magazine, 15*, 64-67.

[39] Kovács, F., Horváth, C., Balogh, Á.T., & Hosszú, G. (2011). Extended Noninvasive Fetal Monitoring by Detailed Analysis of Data Measured With Phonocardiography. *IEEE Transactions on Biomedical Engineering, 58*, 64-70.

[40] Tang, H., Li, T., Qiu, T., & Park, Y. (2016). Fetal Heart Rate Monitoring from Phonocardiograph Signal Using Repetition Frequency of Heart Sounds. *Journal of Electrical and Computer Engineering, 2016*, 2404267:1-2404267:6.

[41] Khandoker, A.H., Ibrahim, E.A., Oshio, S., & Kimura, Y. (2018). Validation of beat by beat fetal heart signals acquired from four-channel fetal phonocardiogram with fetal electrocardiogram in healthy late pregnancy. *Scientific Reports, 8*.

[42] Ibrahim, E.A., Al Awar, S., Balayah, Z., Hadjileontiadis, L.J., & Khandoker, A.H. (2017). A Comparative Study on Fetal Heart Rates Estimated from Fetal Phonography and Cardiotocography. *Frontiers in Physiology, 8*.

[43] Chourasia, V.S., Tiwari, A.K., & Gangopadhyay, R. (2014). A novel approach for phonocardiographic signals processing to make possible fetal heart rate evaluations. *Digital Signal Processing, 30*, 165-183.

[44] Ruffo, M., Cesarelli, M., Romano, M., Bifulco, P., & Fratini, A. (2010). An algorithm for FHR estimation from foetal phonocardiographic signals. *Biomedical Signal Processing and Control, 5*, 131-141.

[45] Dutta, S., Singh, M., & Kumar, A. (2018). Classification of non-motor cognitive task in EEG based brain-computer interface using phase space features in multivariate empirical mode decomposition domain. *Biomedical Signal Processing and Control, 39*, 378-389.

[46] Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N., Tung, C.C., & Liu, H.H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 454*, 903-995.

[47] Karagiannis, A., Loizou, L., & Constantinou, P. (2008). Experimental respiratory signal analysis based on Empirical Mode Decomposition. *2008 First International Symposium on Applied Sciences on Biomedical and Communication Technologies*, 1-5.

[48] Charleston-Villalobos, S., González-Camarena, R., Chi-Lem, G., & Aljama-Corrales, T. (2007). Crackle sounds analysis by empirical mode decomposition. Nonlinear and nonstationary signal analysis for distinction of crackles in lung sounds. *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society, 26 1*, 40-7 .

[49] Sweeney-Reed, C.M., Nasuto, S.J., Vieira, M.F., & Andrade, A.D. (2018). Empirical Mode Decomposition and its Extensions Applied to EEG Analysis: A Review. *Advances in Data Science and Adaptive Analysis, 10*, 1840001:1-1840001:34.

[50] Papadaniil, C.D., & Hadjileontiadis, L.J. (2014). Efficient Heart Sound Segmentation and Extraction Using Ensemble Empirical Mode Decomposition and Kurtosis Features. *IEEE Journal of Biomedical and Health Informatics, 18*, 1138-1152.

[51] Bajelani, K., Navidbakhsh, M., Behnam, H., Doyle, J.D., & Hassani, K. (2013). Detection and identification of first and second heart sounds using empirical mode decomposition. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 227*, 976 - 987.

[52] Marzbanrad, F., Kimura, Y., Funamoto, K., Sugibayashi, R., Endo, M., Ito, T., Palaniswami, M.S., & Khandoker, A.H. (2014). Automated Estimation of Fetal Cardiac Timing Events From Doppler Ultrasound Signal Using Hybrid Models. *IEEE Journal of Biomedical and Health Informatics, 18*, 1169-1177.

[53] Zheng, W., Li, X., Jin, Z., Wei, X., & Liu, H. (2018). Foetal heart rate estimation by empirical mode decomposition and MUSIC spectrum. *Biomedical Signal Processing and Control, 42*, 287-296.

[54] Saleem, S., Naqvi, S.S., Manzoor, T., Saeed, A., Ur Rehman, N., & Mirza, J. (2019). A Strategy for Classification of "Vaginal vs. Cesarean Section" Delivery: Bivariate Empirical Mode Decomposition of Cardiotocographic Recordings. *Frontiers in Physiology, 10*.

[55] Ţarălungă, D.D., & Neagu, G.M. (2018). An Ensemble Empirical Mode Decomposition Based Method for Fetal Phonocardiogram Enhancement. *World Congress on Medical Physics and Biomedical Engineering 2018, IFMBE Proceedings*.

[56] Vican, I., Kreković, G., & Jambrošić, K. (2021). Can empirical mode decomposition improve heartbeat detection in fetal phonocardiography signals? *Computer methods and programs in biomedicine, 203*, 106038.

[57] Huang, N.E., & Attoh-Okine, N.O. (2005). The Hilbert-Huang Transform in Engineering. *CRC Press*.

[58] Taran, S., & Bajaj, V. (2019). Emotion recognition from single-channel EEG signals using a two-stage correlation and instantaneous frequency-based filtering method. *Computer methods and programs in biomedicine, 173*, 157-165.

[59] Rizi, F.Y. (2019). A Review of Notable Studies on Using Empirical Mode Decomposition for Biomedical Signal and Image Processing. *Signal Processing and Renewable Energy*, *2 3*, 89-113

[60] Tanaka, T., & Mandic, D.P. (2007). Complex Empirical Mode Decomposition. *IEEE Signal Processing Letters, 14*, 101-104.

[61] Wu, Z., & Huang, N.E. (2009). Ensemble Empirical Mode Decomposition: a Noise-Assisted Data Analysis Method. *Advances in Adaptive Data Analysis, 1*, 1-41.

[62] Yeh, J., Lin, T., Shieh, J.S., Chen, Y., Huang, N.E., Wu, Z., & Peng, C. (2008). Investigating complex patterns of blocked intestinal artery blood pressure signals by empirical mode decomposition and linguistic analysis. *Journal of Physics: Conference Series*, *61*.

[63] Torres, M.E., Colominas, M.A., Schlotthauer, G., & Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4144-4147.

[64] Colominas, M.A., Schlotthauer, G., & Torres, M.E. (2014). Improved complete ensemble EMD: A suitable tool for biomedical signal processing. *Biomedical Signal Processing and Control, 14*, 19-29.

[65] Ge, H., Chen, G., Yu, H., Chen, H., & An, F. (2018). Theoretical Analysis of Empirical Mode Decomposition. *Symmetry, 10*, 623.

[66] Ballou, G.M. (2015). Handbook for Sound Engineers (5th edition). *Focal Press*.

[67] Coro, G., Massoli, F.V., Origlia, A., & Cutugno, F. (2021). Psycho-acoustics inspired automatic speech recognition. *Computers and Electrical Engineering, 93*, 107238.

[68] Siegert, I., Lotz, A.F., Egorow, O., & Wendemuth, A. (2017). Improving Speech-Based Emotion Recognition by Using Psychoacoustic Modeling and Analysis-by-Synthesis. *International Conference on Speech and Computer* (*SPECOM 2017)*.

[69] Herre, J., & Dick, S. (2019). Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review. *Applied Sciences, 9 14*.

[70] Kane, P., & Andhare, A.B. (2016). Application of psychoacoustics for gear fault diagnosis using artificial neural network. *Journal of Low Frequency Noise, Vibration and Active Control, 35*, 207 - 220.

[71] Guan, S., & Brookens, T.J. (2021). The Use of Psychoacoustics in Marine Mammal Conservation in the United States: From Science to Management and Policy. *Journal of Marine Science and Engineering, 9*, 507.

[72] Miqueau, V., Parizet, E., & Germès, S. (2021). Psycho-acoustic evaluation of the automotive acoustic comfort using vibro-acoustic prediction methods. *INTER-NOISE and NOISE-CON Congress and Conference Proceedings (InterNoise21)*.

[73] Meintjes, A. (2020). Computer Assisted Cardiac Auscultation: Probabilistic Modelling and Psychoacoustic Feature Extraction for Heart Sound Descriptions (Doctoral Thesis). *Auckland University of Technology*.

[74] Patil, K.K., Nagbhushan B.S., & Kumar V. (2010). An efficient retrieval technique for heart sounds using psychoacoustic similarity. *International Journal of Engineering Science and Technology, 2 12*.

[75] Wisniewski, M., & Zielinski, T.P. (2011). Tonal Index in digital recognition of lung auscultation. *Signal Processing Algorithms, Architectures, Arrangements, and Applications (SPA 2011)*, 1-5.

[76] Plack, C.J. (2018). The Sense of Hearing (3rd edition). *Routledge*.

[77] Royalty-free image from *www.shutterstock.com*, by the user *boscorelli*.

[78] Batteau D.W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London. Series B, Biological sciences*, *168*, 158–180.

[79] Fastl, H., & Zwicker, E. (1990). Psychoacoustics: Facts and Models (2nd edition). *Springer*.

[80] Schnupp J., Nelken I., & King A.J. (2012). Auditory Neuroscience: Making Sense of Sound. *MIT Press*.

[81] Howard, D.M., & Angus, J.A. (2009). Acoustics and Psychoacoustics (4th edition). *Focal Press*.

[82] Stevens, S.S., Volkmann, J.E., & Newman, E.B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America, 8*, 185-190.

[83] Moore, B.C., & Glasberg, B.R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America, 74 3*, 750-3.

[84] Olson, H.F. (1972). The Measurement of Loudness. *Audio*, *56 2*, 18-22.

[85] Poulsen T. (1981). Loudness of tone pulses in a free field. *The Journal of the Acoustical Society of America, 69*, 1786-1790.

[86] Lyon, R.F. (2010). Machine Hearing: An Emerging Field [Exploratory DSP]. *IEEE Signal Processing Magazine, 27*, 131-139.

[87] Alías, F., Socoró, J.C., & Sevillano, X. (2016). A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. *Applied Sciences, 6*, 143.

[88] Oxenham, A.J. (2018). How We Hear: The Perception and Neural Coding of Sound. *Annual Review of Psychology, 69*, 27–50.

[89] Sun, M., Scheuer, M.L., & Sclabassi, R.J. (2000). Decomposition of biomedical signals for enhancement of their time-frequency distributions. *Journal of The Franklin Institute, 337*, 453-467.

[90] Härmä, A., Karjalainen, M., Savioja, L., Välimäki, V., Laine, U.K., & Huopaniemi, J. (2000). Frequency-warped signal processing for audio applications. *Journal of The Audio Engineering Society, 48*, 1011-1031.

[91] Chen, J., Phua, K., Song, Y., & Shue, L. (2006). A portable phonocardiographic fetal heart rate monitor. *2006 IEEE International Symposium on Circuits and Systems*.

[92] Royer, T. (2019). Pitch-shifting algorithm design and applications in music (Master Thesis). *KTH, School of Electrical Engineering and Computer Science*.

[93] Laroche, J., & Dolson, M. (1999). New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99)*, 91-94.

[94] Laroche, J., & Dolson, M. (1997). Phase-vocoder: about this phasiness business. *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*.

[95] Götzen, A.D., Bernardini, N., & Arfib, D. (2000). Traditional (?) implementations of a phase vocoder: the tricks of the trade. *Proceedings of the 3rd International Conference on Digital Audio Effects*.

[96] Grondin F. (2009). Guitar Pitch Shifter - Algorithm section. *www.guitarpitchshifter.com*.

[97] Russell, S.J., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th edition). *Pearson*.

[98] Kok J.N., Boers E.J.W., Kosters W.A., van der Putten P., & Poel M. (Artificial Intelligence: Definition, Trends, Techniques and Cases. *Encyclopedia of Life Support Systems (EOLSS)*.

[99] Copeland, B.J. (2021). "artificial intelligence". *Encyclopedia Britannica*, *https://www.britannica.com/technology/artificial-intelligence*.

[100] Jordan, M.I., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*, 255 - 260.

[101] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors (Basel, Switzerland)*, *18 8*, 2674.

[102] Shanthamallu, U.S., Spanias, A., Tepedelenlioğlu, C., & Stanley, M. (2017). A brief survey of machine learning methods and their sensor and IoT applications. *8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1-8.

[103] Papernot, N., Mcdaniel, P., Sinha, A., & Wellman, M.P. (2016). Towards the Science of Security and Privacy in Machine Learning. *ArXiv, abs/1611.03814*.

[104] Rajkomar, A., Dean, J., & Kohane, I.S. (2019). Machine Learning in Medicine. *The New England Journal of Medicine, 380*, 1347–1358.

[105] Padmanabhan, J., & Johnson, M. (2015). Machine Learning in Automatic Speech Recognition: A Survey. *IETE Technical Review, 32*, 240 - 251.

[106] Palaniappan, R., Sundaraj, K., & Ahamed, N.U. (2013). Machine learning in lung sound analysis: a systematic review. *Biocybernetics and Biomedical Engineering, 33*, 129-135.

[107] Gemein, L., Schirrmeister, R.T., Chrabaszcz, P., Wilson, D., Boedecker, J., Schulze-Bonhage, A., Hutter, F., & Ball, T. (2020). Machine-learning-based diagnostics of EEG pathology. *NeuroImage, 220*.

[108] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access, 7*, 81542-81554.

[109] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2020). A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. *IEEE Access, 8*, 20991-21002.

[110] Cömert, Z., & Kocamaz, A.F. (2017). Comparison of Machine Learning Techniques for Fetal Heart Rate Classification. *Acta Physica Polonica A, 132*, 451-454.

[111] Diker, A., Avci, E., Cömert, Z., Avci, D., Kacar, E., & Serhatlioglu, I. (2018). Classification of ECG signal by using machine learning methods. *26th Signal Processing and Communications Applications Conference (SIU)*, 1-4.

[112] Sahin, H., & Subasi, A. (2015). Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques. *Applied Soft Computing, 33*, 231-238.

[113] Mohri, M., Rostamizadeh, A., & Talwalkar, A.S. (2018). Foundations of Machine Learning (2nd edition). *Adaptive computation and machine learning*. *MIT Press*.

[114] Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science, 2*.

[115] Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2021). Special issue on feature engineering editorial. *Machine Learning*.

[116] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300*, 70-79.

[117] Haq, A.U., Zhang, D., Peng, H., & Rahman, S.U. (2019). Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection. *IEEE Access, 7*, 151482-151492.

[118] Vican, I., Kreković, G. & Jambrošić K. (2018). Relevance of Empirical Mode Decomposition for Fetal Heartbeat Detection on Smartphone Devices. *Proceedings of the 8th Congress of the Alps Adria Acoustics Association*.

[119] Han, Y., Kim, J., & Lee, K. (2017). Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*, 208-221.

[120] Herrera-Boyer, P., Peeters, G., & Dubnov, S. (2003). Automatic Classification of Musical Instrument Sounds. *Journal of New Music Research, 32*, 21 - 3.

[121] Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., & Fan, H. (2020). Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. *Neural networks: the official journal of the International Neural Network Society, 130*, 22-32 .

[122] Singh, M., & Cheema, A. (2013). Heart Sounds Classification using Feature Extraction of Phonocardiography Signal. *International Journal of Computer Applications, 77*, 13-17.

[123] Alnuaimi, S., Jimaa, S.A., & Khandoker, A.H. (2017). Fetal Cardiac Doppler Signal Processing Techniques: Challenges and Future Research Directions. *Frontiers in Bioengineering and Biotechnology, 5*.

[124] Posner, G.D., Oxorn, H., & Foote, W.R. (2013). Oxorn-Foote Human labor & birth (6th edition). *McGraw Hill*.

[125] Based on an image created by user *cookie_studio on www.freepik.com*.

[126] Delgado-Contreras, J.R., García-Vázquez, J., & Brena, R.F. (2014). Classification of environmental audio signals using statistical time and frequency features. *2014 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, 212-216.

[127] Ramachandran, K.M., Tsokos, R., & Tsokos, C.P. (2009). Mathematical Statistics with Applications. *Academic Press*.

[128] Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *Technical Report, IRCAM Paris*.

[129] Alías, F., Socoró, J.C., & Sevillano, X. (2016). A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. *Applied Sciences, 6*, 143.

[130] Rodgers, J., & Nicewander, A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *American Statistician, 42*, 59-66.

[131] Sánchez, R.R., Aguilar-Ruiz, J.S., Santos, J.C., & Díaz-Díaz, N. (2005). Analysis of Feature Rankings for Classification. *International Symposium on Intelligent Data Analysis (IDA)*.

[132] Kohavi, R., & John, G.H. (1997). Wrappers for Feature Subset Selection. *Artificial Intelligence, 97*, 273-324.

[133] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*, 16-28.

[134] Bekkerman, R., El-Yaniv, R., Tishby, N., & Winter, Y. (2003). Distributional Word Clusters vs. Words for Text Categorization. *Journal of Machine Learning Research, 3*, 1183-1208.

[135] Caruana, R., & Sa, V.R. (2003). Benefitting from the Variables that Variable Selection Discards. *Journal of Machine Learning Research, 3*, 1245-1264.

[136] Velusamy, D., & Ramasamy, K. (2021). Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. *Computer methods and programs in biomedicine, 198*, 105770 .

[137] Li, J., & Liu, H. (2017). Challenges of Feature Selection for Big Data Analytics. *IEEE Intelligent Systems, 32*, 9-15.

[138] Sawyer, S.F. (2009). Analysis of Variance: The Fundamental Concepts. *Journal of Manual & Manipulative Therapy, 17*, 27E - 38E.

[139] Genuer, R., Poggi, J., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*, 2225-2236.

[140] Guyon, I., Weston, J., Barnhill, S.D., & Vapnik, V.N. (2004). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning, 46*, 389-422.

[141] Tuv, E., Borisov, A., Runger, G.C., & Torkkola, K. (2009). Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. *Journal of Machine Learning Research, 10*, 1341-1366.

[142] Hughes, G.F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, 14*, 55-63.

[143] Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America, 33*.

[144] Smith, J.O., & Abel, J.S. (1995). The Bark bilinear transform. *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*, 202-205.

[145] Prusa, Z., & Holighaus, N. (2017). Phase vocoder done right. *25th European Signal Processing Conference (EUSIPCO)*, 976-980.

[146] Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. *International Society for Music Information Retrieval (ISMIR 2000)*.

[147] Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America, 87 4*, 1738-52 .

[148] Ellis D.P.W. (2005). PLP and RASTA (and MFCC, and inversion) in Matlab. *https://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat*

[149] Han, W., Chan, C., Choy, O.C., & Pun, K.P. (2006). An efficient MFCC extraction method in speech recognition. *2006 IEEE International Symposium on Circuits and Systems*.

[150] Jensen, J.H., Christensen, M.G., Murthi, M.N., & Jensen, S.H. (2006). Evaluation of MFCC estimation techniques for music similarity. *2006 14th European Signal Processing Conference*, 1-5.

[151] Krishna Kishore, K.V., & Krishna Satish, P. (2013). Emotion recognition in speech using MFCC and wavelet features. *3rd IEEE International Advance Computing Conference (IACC)*, 842-847.

[152] Elliott, D. (1988). Handbook of Digital Signal Processing: Engineering Applications. *Academic Press*.

[153] Paliwal, K.K. (1999). Decorrelated and liftered filter-bank energies for robust speech recognition. *European Conference on Speech Communication and Technology (EUROSPEECH'99)*.

[154] Qaisar, S.M., Hainmad, N., Khan, R., & Asfour, R. (2019). A Speech to Machine Interface Based on Perceptual Linear Prediction and Classification. *2019 Advances in Science and Engineering Technology International Conferences (ASET)*, 1-4.

[155] Xiao, D., Mo, F., Zhang, Y., Zhao, M., & Ma, L. (2018). An extended Levinson-Durbin algorithm and its application in mixed excitation linear prediction. *Heliyon, 4*.

[156] Brockwell, P.J., & Dahlhaus, R. (2004). Generalized Levinson-Durbin and Burg algorithms. *Journal of Econometrics, 118*, 129-149.

[157] Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64 3*, 153-181.

[158] Cuevas, A., Febrero-Bande, M., & Fraiman, R. (2004). An anova test for functional data. *Computational Statistics & Data Analysis, 47*, 111-122.

[159] Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200-1205.

[160] Sanz, H., Valim, C., Vegas, E., Oller, J.M., & Reverter, F. (2018). SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics, 19*.

[161] Li, J., Fong, S.J., Mohammed, S., & Fiaidhi, J. (2015). Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms. *The Journal of Supercomputing, 72*, 3708-3728.

[162] Soofi, A.A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic and Applied Sciences, 13*, 459-465.

[163] Ali, M.Z., Shabbir, M.N., Liang, X., Zhang, Y., & Hu, T. (2019). Machine Learning-Based Fault Diagnosis for Single- and Multi-Faults in Induction Motors Using Measured Stator Currents and Vibration Signals. *IEEE Transactions on Industry Applications, 55*, 2378-2391.

[164] Bauer, E., & Kohavi, R. (2004). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning, 36*, 105-139.

[165] Goutte, C., & Gaussier, É. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *European Conference of Information Retrieval (ECIR 2005)*.

[166] Peltonen, V.T., Tuomi, J.T., Klapuri, A., Huopaniemi, J., & Sorsa, T. (2002). Computational auditory scene recognition. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2*, II-1941-II-1944.

[167] Mesgarani, N., David, S.V., Fritz, J.B., & Shamma, S.A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences, 111*, 6792 - 6797.

[168] Zhao, X., & Wang, D. (2013). Analyzing noise robustness of MFCC and GFCC features in speaker identification. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7204-7208.

Biography

Ivan Vican was born on the 18th of September 1990 in Imotski, Croatia. He got his bachelor's degree from the University of Zagreb, Faculty of Electrical Engineering and Computing in 2012 and a master's degree from the same faculty in 2014. After several years of working as a Research Associate, Acoustics Engineer, and Signal Processing Engineer, he made a career pivot to Data Science, specialized in audio and biomedical signals. For the last 4 years, he has been functioning as a freelance Data Science and Algorithms expert for dozens of companies and startups, mostly situated in the United States and Western Europe. His domains of expertise are the following: audio & biomedical signal processing, machine learning, deep learning, audio codecs, noise cancellation algorithms, time series synchronization and more. Ivan Vican has published 4 conference papers, 2 journal papers and has submitted 3 patent applications.

List of publications

Patent applications & grants

[1] Brickner, S., Williams, M.T., Viss, P., & Vican, I. (2022). Systems and Methods for Providing Survey Data. *Patent number: 11277663*

[2] Abad, A., Chinimilli, P.T., & Vican, I. (2019). Systems and Method for Angle Calculations for a Plurality of Inetrial Measurement Units. *Publication number: 20190293404*

[3] Jambrošić, K., Vican, I., & Domitrović, H. (2018). Resonator absorber with Adjustable Acoustic Characteristics (grant). *Patent number: 10032444*

Journal papers

[1] Vican, I., Kreković, G., & Jambrošić, K. (2021). Can empirical mode decomposition improve heartbeat detection in fetal phonocardiography signals? *Computer methods and programs in biomedicine, 203*, 106038.

[2] Vican, I., Jambrošić, K., & Domitrović, H. (2015). Improvement of acoustic resistance equations in perforated plate absorbers with thin porous layers. *Noise control engineering journal*, *63 5,* 415-423.

Conference papers

[1] Vican, I., Kreković, G. & Jambrošić K. (2018). Relevance of Empirical Mode Decomposition for Fetal Heartbeat Detection on Smartphone Devices. *Proceedings of the 8th Congress of the Alps Adria Acoustics Association*.

[2] Kreković, G., & Vican, I. (2017). Towards a Parallel Computing Framework for Direct Sonification of Multivariate Chronological Data. *Proceedings of the 12th International*

Audio Mostly Conference on Augmented and Participatory Sound and Music Experience (AM'17).

[3] Vican, I., Budimir, M., Nageswaran, C., et al. (2015). High Temperature Pipe Structural Health Monitoring System Utilising Phased Array Probes On TOFD Configuration. *24th International Conference Nuclear Energy for New Europe (NENE 2015)*.

[4] Vican, I., Jambrošić, K., & Horvat, M. (2014). Comparison of Acoustic Resistance of a Perforated Plate Absorbers with a Tightly and Loosely Placed Thin Porous Layer. *Proceedings of the 6th Congress of the Alps Adria Acoustics Association*.

Biografija

Ivan Vican rođen je 18.9.1990. u Imotskom. 2012. godine je stekao zvanje prvostupnika na Sveučilištu u Zagrebu, Fakultetu Elektrotehnike i Računarstva. Dvije godine kasnije stekao je zvanje magistra struke na istom fakultetu. Nakon više godina rada kao zavodski suradnik, inženjer za akustiku i inženjer za obradu signala, odlučio se za novu karijeru u podatkovnoj znanosti, dodatno specijaliziran za signale iz domene audia i biomedicine. Posljednje 4 godine djeluje kao slobodni stručnjak iz područja podatkovne znanosti i algoritama za desetke kompanija i startupova, prvenstveno smještenih u Sjedinjenim Američkim Državama i zapadnoj Europi. Njegove domene ekspertize su sljedeće: obrada audio i biomedicinskih signala, strojno učenje, duboko učenje, audio kodeci, algoritmi za poništavanje buke, sinkronizacija vremenskih nizova itd. Ivan Vican je objavio 4 rada na znanstvenim konferencijama i skupovima, 2 članka u znanstvenim časopisima te je podnio 3 patentne prijave.