KAUNAS UNIVERSITY OF TECHNOLOGY

JONAS MINELGA

COMPUTATIONAL INTELLIGENCE METHODS FOR NON-INVASIVE LARYNX PATHOLOGY SCREENING

Doctoral Dissertation
Technology sciences, informatics engineering (07T)

2018, Kaunas

This doctoral dissertation was prepared at Kaunas University of Technology, Faculty of Electrical and Electronics Engineering, Department of Electric Power Systems during the period of 2012 – 2017.

Scientific Supervisor:
Prof. Dr. Habil. Antanas VERIKAS (Kaunas University of Technology, technology sciences, Informatics engineering, 07T).

Doctoral dissertation has been published in:
http://ktu.edu

Editor:
Matas Petronis (Kaunas University of Technology, Centre of Foreign Languages)

ISBN 978-609-02-1527-2

The bibliographic information about the publication is available in the National Bibliographic Data Bank (NBDB) of the Martynas Mažvydas National Library of Lithuania.

KAUNO TECHNOLOGIJOS UNIVERSITETAS

JONAS MINELGA

SKAITINIO INTELEKTO METODAI NEINVAZINEI GERKLŲ LIGŲ DIAGNOSTIKAI

Daktaro disertacija
Technologijos mokslai, informatikos inžinerija (07T)

2018, Kaunas

Disertacija rengta 2012 – 2017 metais Kauno technologijos universiteto Elektros ir elektronikos fakultete Elektros energetikos sistemų katedroje.

Mokslinis vadovas:
Prof. Dr. Habil. Antanas VERIKAS (Kauno Technologijos Universitetas, technologijos mokslai, informatikos inžinerija, 07T).

Interneto svetainės, kurioje skelbiama disertacija, adresas:
http://ktu.edu

Redagavo:
Matas Petronis (Kauno technologijos universitetas, Užsienio kalbų centras)

ISBN 978-609-02-1527-2

Leidinio bibliografinė informacija pateikiama Lietuvos nacionalinės Martyno Mažvydo bibliotekos Nacionalinės bibliografijos duomenų banke (NBDB).

Abstract

Laryngeal disorders affect roughly 5-6% of the general human population and larynx-related cancer alone causes around 200,000 annual deaths worldwide. This being one of only a few areas where annual deaths are increasing, requires effort in targeting easy, effective and accessible preventive laryngeal health care. This research analyzes computational intelligence techniques for larynx pathology detection, using non-invasive measurements, such as human voice recordings and answers to specific questionnaires. The intention was to develop a technique for voice and query data analysis, which would be capable of detecting voice pathology and providing support in screening for laryngeal disorder.

This study is performed by using a subject's voice recordings and answers to specific questionnaire, obtained by otolaryngology specialists. The collected data is gathered into 3 databases. Voice database contains sustained phonation (/a/ as in word "large") recordings of 273 subjects (163 healthy and 110 pathological voices) varying in sex and age (from 19 to 85 years old), where each recording is labeled by a clinical diagnosis, obtained from clinical voice specialists. There are up to three recordings done for each patient, where part of them is also recorded with contact microphone. Query database contains data from 596 subjects (327 healthy and 269 pathological) also varying in sex and age. All subjects from the voice database are present in the query database. Patient classification discerns patients into *healthy* and *pathological* classes.

The methodological approach towards the analysis of voice and query data consisted of several steps. Firstly, characterization of audio recordings was obtained by using various techniques to extract diverse features from voice recordings. 14 different feature sets of varying size were extracted from each recording, resulting in 927 features per recording. Secondly, 6 audio parameters (extracted using "Dr. Speech" software) were provided by otolaryngology specialists, together with query data (answers to 25 questions). The three mentioned data sets were used for data classification either individually or in fusion. For data fusion task, new data dependent random forest-based way of available information from multiple data sets combination was introduced.

Random forest (RF) algorithm was used as base classifier for audio data classification and as a meta-learner in decision-level fusion cases. Both variations were used for classification of audio recordings from two types of microphones, which allowed assessment if contact microphones can provide useful information for classification accuracy improvement compared to acoustic ones or used together. Data dependent random forest-based data combination and classification technique was proposed and applied for voice data classification. In this work, affinity analysis of the query data was used to extract rules for each class (*healthy* and *pathological*), which allows identification if patient has larynx

pathology. Classification of query data was performed by extracted association rules and Decision tree (DT) algorithm.

Two distinct techniques were used to represent voice and query data visually for inspection and analysis. Voice data was mapped to 2D space using t-SNE algorithm, which also allows analysis of subject's similarity. Meanwhile, graphs of probability density functions (PDFs) were obtained for analysis of query data.

Our introduced data-dependent random forest-based technique of data combination and classification helped to achieve the highest classification accuracy of 86.37%, compared to the results achieved by using only one single data set. The used affinity analysis highlighted 17 important questions which allows reduction of the questionnaire. Our developed association rules technique for query data classification together with DT are completely transparent, which allows deeper exploration of decision-making process and is very useful for teaching/learning purposes, early preventive health care. Results of acoustic and contact microphone comparison revealed that the acoustic microphone is superior to the contact one. However, contact microphones may be more useful in the noisy environments, but additional research is required to determine the noise level when contact microphone becomes superior to the acoustic one.

Applied visual data analysis techniques is another useful contribution of this research, which allows detection of incorrectly labeled or more thorough examination requiring subjects. PDFs provide additional information (statistical) about a patient and serve as a learning material, as well as show which type of data is not present in the used data set and might affect classification accuracy. Accurate pathology detection was observed for unseen subjects with equal error rate (EER) of 11.11% by using association rules and EER of 10.26% by using the decision tree. When using decision-level fusion, an even lower EER of only 9.52% was achieved.

In conclusion, results of this research indicate that the developed techniques can be very useful for diagnostics, education and exploratory tasks in Otolaryngology departments.

Contents

List of Figures

List of Tables

List of abbreviations

AM	Acoustic microphone.
ANN	Artificial neural networks.
APQ	Amplitude perturbation quotient.
AUC	Area under the curve.
AVPD	Arabic Voice Pathology Database.
CART	Classification And Regression Tree.
C_{llr}	Cost of log-likelihood ratio - the comprehensive goodness-of-detection criterion.
CM	Contact microphone.
CPP	Cepstral Peak Prominence.
DA	Discriminant analysis.
DCT	Discrete cosine transform.
DET	Detection error trade-off.
DFT	Discrete Fourier transform.
DT	Decision Tree.
EER	Equal error rate.
F0	Fundamental frequency.
FATR	Frequency amplitude tremor.
FCM	Fuzzy c-means.
FDR	Fisher discrimination ratio.
FFT	Fast Fourier Transformation.
FKM	Fuzzy k-means.
FVQ	Fuzzy vector quantization.
GMM	Gaussian mixture models.
GNE	Glottal to noise excitation.
GUI	Graphical User Interface.
HMM	Hidden Markov model.
HNR	Harmonics to noise ratio.
ISO	International Standardization Organization.
k-NN	K-nearest neighbors algorithm.
kPCA	Kernel Principal Component Analysis.

LLD	Low-level descriptor.
LOGP	Logarithmic opinion pool.
LOP	Linear opinion pool.
LPC	Linear Prediction Coefficients.
LPCT	Linear prediction cosine transform.
LSP	Line spectral pairs.
LVQ	Learning vector quantization.
MEEI	Disordered Voice Database, MEEI, Voice and Speech Lab, Kay Elemetrics Corp..
MFCC	Mel-frequency cepstrum coefficient.
MFT	Maximum phonation time.
ML	Machine learning.
MLP	Multilayer perceptron.
MRBF	Median radial basis function.
NBC	Naive Bayes Classifier.
NLPCA	Auto-associative Neural Network.
NNE	Normalized noise energy.
OOB	Out-of-bag.
PCA	Principal Component Analysis.
PCM	Pulse-code modulation.
PDFs	Probability Density Functions.
PLP	Perceptual linear predictive.
PLPCC	Perceptual Linear Predictive Cepstral Coefficients.
PPQ	Pitch perturbation quotient.
PVSQ	Pediatric Voice Symptom Questionnaire.
R1	First Rahmonic.
RF	Random Forest.
ROC	Receiver operating characteristic.
ROCCH	ROC convex hull.
SNE	Stochastic neighbor embedding.
SNR	Signal to noise ratio.
SOMI	Self-organizing map imputation.
SPI	Soft phonation index.
SVD	Singular Value Decomposition.
SVM	Support vector machine.

t-SNE	T-distributed stochastic neighbor embedding.
TVQ	Thyroidectomy-Related Voice Questionnaire.
UI	User interface.
VFHQ	Vocal Fatigue Handicap Questionnaire.
VQ	Vector quantization.
V-RQOL	Voice-related Quality of Life.
VTDS	Vocal tract discomfort scale.
VTI	Voice turbulence index.
WAV	Waveform Audio File Format.
ZCR	Zero-crossing rate.

1. INTRODUCTION

1.1. Research Area

In this research pathological voice detection was analyzed in a context of computational intelligence methods. Non-invasive data, such as voice recording and questionnaire were used. New methods for voice and query data classification were proposed. Methods for visual representation of data were also analyzed to provide additional information for deeper analysis and exploration.

The human voice is produced by a glottal excitation that is filtered by a vocal tract, controlled by our brain and muscle movements, as well as hearing sensory system [58]. If any of these fail to do their part, voice signal becomes distorted. In most cases voice changes are made by vocal folds, which can lose elasticity, gain weight, or lose the ability to close properly depending on the disease. These changes in voice signal allow to differentiate between healthy and pathological voice.

The newer version of same databases, used in [157, 99], was used here. Data was extended by adding additional feature sets of new observations. The decision level fusion of voice and query data classification was analyzed for classification accuracy improvement.

The aim of this study was to research usage of non-invasive modalities, such as voice recording or query data, to discern pathological voice and to propose new techniques for classification accuracy improvement. A new association rules-based method was introduced and applied for query data analysis. To improve voice data classification accuracy, a novel technique for building data dependent random forest was proposed. Both data modalities were used separately by applying different methods and results were combined by meta-learner. The main areas of this study are data mining and machine learning techniques, however visual data representation was also analyzed as a tool for better data exploration. As an additional outcome of this research, a computer program was created for the use by otolaryngologists.

1.2. Problem Relevance

Laryngeal disorders affect around 5% [16] or 6.6% according to [59] of the general population. Larynx related cancer alone causes roughly 200,000 annual deaths worldwide, and this number continues to increase, while deaths from other types of cancer are in decreasing tendencies. As we can see, preventive laryngeal health-care technologies are required.

Laryngeal pathology detection is a rather complex task, requiring multiple types of data analysis. Patient complaints are usually collected as answers to questionnaires, old medical history is provided in a form of text and instrumental analysis tools usually provide images and/or voice recordings.

As proved in many studies, non-invasive data, such as answers to questionnaires and voice recordings, can be used for automatic analysis, which allows early pathology detection, as well as deeper voice quality assessment. This can be used as a diagnostic tool and as a preventive health-care measure. According to [7, 106, 102, 47, 9, 130, 128, 52, 157], audio analysis is increasingly being applied in this field of research. However, there are very few attempts to use query data for this task [159, 156, 155, 13, 158, 151].

Despite growing popularity of non-invasive voice pathology diagnostic techniques, they are still not applicable for use as a completely trusted tool. Many researchers have achieved a very low rate of error, but in most cases only relatively small training and testing databases were used. This indicates the need to investigate new methods for voice pathology detection and to improve the accuracy of currently used techniques.

1.3. Research Object

The object of this research is voice pathology detection by using voice and query data analysis in separate and fusion matter. Non-invasiveness is the main advantage of this data.

1.4. Objectives and Tasks

The main objective of this research is to improve voice pathology classification accuracy by developing new classification technique, which would be capable of classifying data from multiple sources. Second objective is to improve classification accuracy by the use of voice and questionnaire data decision-level fusion.

Tasks formulated for this study:

1. Analyze state-of-the-art work of laryngeal pathology detection identifying the drawbacks of techniques used.

2. Review classification techniques using voice and query data fusion and identify their limitations.

3. Propose a new technique for combination and classification of data available in different feature sets.

4. Develop a transparent technique for voice and query data analysis.

5. Experimentally validate the proposed techniques with out of bag (OOB) data.

6. Develop a non-invasive decision support system for voice pathology detection, providing multiple approaches for graphical patient data analysis and employing several data sources.

1.5. Research Methods

Many machine learning methods are proved to be useful for successful voice pathology detection. Studies are done using GMM, HMM, k-NN, LVQ, MLP, SVM, Random Forest, Decision Tree, discriminant analysis and other methods. In some cases, a combination of multiple algorithms or additional data (or both - multiple algorithms and multiple data sets), such as questionnaire answers, are used to improve accuracy. No matter which algorithm is selected, non-invasive laryngeal pathology detection from human voice and questionnaire data is a complex task, which can be divided into several steps:

1. Collection of questionnaire data from patients.
2. Recording of patient voice signal.
3. Extraction of audio features from voice recordings.
4. Collected data analysis to determine healthy or pathological class.

Random Forest was selected as the main meta-learner classifier in this work. To improve classification accuracy, voice and query data decision-level fusion was used. The new proposed technique for building data dependent meta-learner was applied. As a base classification algorithm, an association rules algorithm was introduced, which was used only for query data classification. The decision tree algorithm was used for additional classification, using only 6 basic (well-known for doctors) audio parameters and GFI parameter from query data. This allows to provide the user with visual sample of decision tree, which makes it easier to interpret the results. To classify voice data, the Random Forest algorithm was used. A new technique was proposed to build a data-dependent Random Forest which combines data from multiple data sets. For data visualization purposes, t-SNE algorithm was used to map data to two-dimensional space.

1.6. Scientific Novelty

The main scientific contributions of this research are these:

- The proposed novel, data dependent random forest based, technique for combination and classification of data in multiple data sets. This new method classifies multiple voice parameter data sets with the Random forest algorithm and constructs another RF from previously used RFs to provide the final result.
- The developed novel association rules-based algorithm for query data analysis. This technique relies on association rules, extracted from the questionnaire data by using affinity analysis. It is a completely transparent technique which is very beneficial in preventive health care.

- Visual data representation for easier patient comparison and better understanding of decision-making process. The t-SNE algorithm, graphs of probability density functions, and visual representation of decision tree were employed for data visualization. Proximity matrix, mapped to 2D space allows visual comparison of patients, PDFs represents the distribution of query data and decision tree provides view of algorithm logic.

1.7. Practical Significance

Successful laryngeal pathology detection can sometimes be very difficult even for a well-trained physician. Automated algorithms and tools, such as used and developed in this work, can significantly improve doctors work, by providing additional easily accessible diagnostic information. This can be used as a single tool, to ease the patient diagnostic process, or as a tool to get additional information about the patient: additional class from association rules, visual graphs of the decision tree, visual views of patient position among others (t-SNE) and patient information comparison to others. In addition to all this, our developed software could be modified and provided for public access as a web page or mobile phone application, which would enable anyone to check their larynx health in just a few simple steps.

1.8. Defended Statements

1. Human voice signals and answers to specific questions contain information about healthiness of larynx, which can be used for non-invasive pathology detection.

2. Computational intelligence methods are capable of distinguishing between healthy and pathological voice signals with accuracy of more than 90%.

3. The proposed data dependent Random Forest based technique for data combination from multiple data sets and classification outperforms data-fusion and decision-fusion techniques for the data used in the experiments with the maximum achieved accuracy of 86.37%.

4. A pathological larynx can be successfully detected using only query data, which represents patient voice function and quality evaluation.

5. Voice and query data decision-level fusion analysis can separate a healthy larynx from pathological one with accuracy of 90.48%, and outperforms classification when data is used separately.

1.9. Approval of Research Results

Results of this research were presented in:

1. Comparing throat and acoustic microphones for laryngeal pathology detection from human voice, 9^{th} *International Conference on Electrical and Control Technologies ECT-2014*, 2014 May 8-9, Kaunas, Lithuania.

2. Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders, *2014 IEEE Symposium Series on Computational Intelligence: IEEE Symposium on Computational Intelligence in Healthcare and e-health*, 2014 December 9-12, Orlando, Florida, USA.

3. Towards Voice and Query Data-based Non-invasive Screening for Laryngeal Disorders, *14th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED '15)*, 2015 January 10-12, Tenerife, Canary Islands, Spain.

1.10. Publications of the Results

The main results of this research are published in:

1. A. Verikas, A. Gelzinis, E. Vaiciukynas, M. Bacauskiene, J. Minelga, M. Hallander, V. Uloza, E. Padervinskis, Data dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: acoustic versus contact microphone, *Medical Engineering & Physics*, 37, 2015, 210-218.

2. E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, J. Minelga, M. Hallander, E. Padervinskis, V. Uloza, Fusing voice and query data for non-invasive detection of laryngeal disorders, *Expert Systems with Applications*, 42(22), 2015, doi:10.1016/j.eswa.2015.07.001.

3. J. Minelga, A. Verikas, E. Vaiciukynas, A. Gelzinis, M. Bacauskiene, A transparent decision support tool in screening for laryngeal disorders using voice and query data, *Applied Sciences*, 2017, doi:10.3390/app7101096.

1.11. Structure of the Dissertation

This study begins with the review of related work in Chapter 2. Many studies from the same field and methods used are reviewed. Chapter 3 contains a description of data used in this study. Voice feature extraction from audio files and query data collection is described in Chapter 4. We used t-SNE dimensionality reduction algorithm to map audio and query data to 2D space and provide visualizations.

Chapter 4 is dedicated to describing voice features and data analysis methods used in this study:

- Section 4.1 describes voice parameters (features) extracted from voice signal and used for classification.

- Sections 4.2, 4.3, 4.4 and 4.5 explain classification algorithms used in this study. Our proposed data dependent Random Forest and Association rules as well as Random Forest and Decision Tree algorithms are explained in detail, providing main features and steps of development.

- Section 4.7 covers methods of missing data control and management.
- In Section 4.11 data visualization techniques used in this study are described and sample images are provided. Created computer program is analyzed in detail while providing examples and deeper explanations of visualization by t-SNE, Probability density functions and Decision Tree.

Experimental results of this study are provided in Chapter 5. Pathology detection is evaluated in Section 5.3, while other sections cover evaluation and results of comparison of acoustic and contact microphones, association rules, visualization of data and decisions, computer program's ease of use. Discussion of the research results and conclusions are provided in Chapter 6 and Chapter 7 respectively.

2. RELATED WORK

Acoustic and questionnaire data analysis proved to be an excellent way to assess voice quality and detect laryngeal pathology. Query data-based detection consistently outperforms voice data-based detection and according to [151], fusion of these modalities provides even better performance. For voice data-based classification accuracy improvement, multiple ways of voice signal recording are used, but as mentioned in [157], in a controlled noise environment acoustic microphone outperforms a contact one. Several studies have been carried out where automatic recognition of vocal fold pathology was performed using acoustic and questionnaire data. These studies can be separated into 3 groups:

1. Studies which analyze vocal audio data and pick out most important audio parameters.

2. Studies which analyze questionnaire data and select most important questions.

3. Studies which construct the best performing classifier for voice pathology detection using audio data and/or questionnaire data.

The scope of this study was to analyze all previously mentioned groups and build a classification method with improved accuracy. As this is a continues work based on [152], the goal of this research was to improve classification algorithms analyzed in [152] and propose new techniques to achieve higher voice pathology detection accuracy. As an improvement for techniques used in [152], data dependent random forest technique, newly extracted association rules classification algorithm and fusion of voice and query data in classification was proposed. A computer program was developed as an additional product of this research for the use by otolaryngologists. The program provides user not only with the classified patient class label, but also with classification certainty, 2D data map (created using *t*-SNE algorithm) and visual interpretation of decision tree (DT). These features allow a more thorough investigation of a patient and ensure maximum transparency of the whole classification process.

2.1. Voice function and quality assessment

Successful vocal assessment process should include such techniques as visual examination, aerodynamic measures, acoustic voice analysis and patient self-assessment [85]. There are many techniques able to provide relevant information about the voice disorder, however, current life conditions do not allow the application of these techniques to everyone, and only the ones requiring least time and effort are used for the initial patient analysis. In medicine, invasive and non-invasive techniques are used for voice diagnostics, where indirect laryngoscopy and video laryngostroboscopy are the most important ones [159]. Usually, invasive techniques are used, but that may be very expensive and difficult to use in home-care setting [151], as well as the fact that it can result in unnecessary patient discomfort [162, 53]. Nevertheless, in most cases all available information can be collected, and a physiological state can be successfully assessed using non-invasive techniques. Today, automatic non-invasive voice analysis is increasingly used as an objective method for laryngeal pathologies detection [130, 5, 74, 61, 108, 157, 44, 164, 66, 32, 93, 167, 162]. As mentioned in [102, 149] even voice signals, transmitted through telephone lines, can be successfully used for this kind of analysis.

This kind of voice analysis requires signal processing algorithms to be used for feature extraction from glottal signals (a voice signal, which is obtained between vocal folds and vocal tract to avoid vocal fold structure changes depending on age) [95]. When the features are extracted, classification methods can be applied, to distinguish normophonic and dysphonic voices. Classification success is greatly dependent on features extracted from voice, because some pathological voice changes might be "visible" only in specific features.

In clinical practice of voice pathology detection, sustained phonation is used for patients' voice quality assessment, because they circumvent linguistic artefacts [121, 146], are time-effective and reduce variance in sustained vowels [167, 163]. According to [90, 94, 53, 121] most important parameters include fundamental frequency, jitter, shimmer, amplitude perturbation quotient (APQ), pitch perturbation quotient (PPQ), harmonics to noise ratio (HNR), normalized noise energy (NNE), voice turbulence index (VTI), soft phonation index (SPI), frequency amplitude tremor (FATR), glottal to noise excitation (GNE) and Mel-frequency cepstrum coefficients (MFCC). Fundamental frequency, jitter and shimmer are the easiest to extract parameters, so they are used very often. Fundamental frequency is the lowest frequency of the sinusoidal waves and it is used to measure the pitch of a voice signal [148]. Jitter and shimmer, accordingly, represent the variation in frequency and amplitude of voice fluctuations. According to [148] findings, if shimmer value is too high, patient might have a speech disorder. In some studies, together with previously mentioned parameters, Cepstral Peak Prominence (CPP) is used as well. It is claimed that CPP is the most powerful predictor of perceived hoarseness [10]. Also, according to [166] this parameter correlates well with breathy speech. Some researchers are using voice analysis tools or libraries providing sets of audio features like openSMILE or MPEG-7.

Researchers try to use different voice parameters to find a better way for pathological voice detection. For example, [107] use only first two formants from two Arabic vowels. Classification is done using vector quantization (VQ) and artificial neural networks (ANN) separately, in order to compare them. The achieved 78.72% best accuracy looks not as promising as the 92.86% achieved by [130]. Another high accuracy was achieved by [121], where it varies from 87%

to 100% depending on the disease and sex of the patient. Authors applied such techniques as principal component analysis (PCA), kernel principal component analysis (kPCA) and auto-associative neural network (NLPCA). These results look very impressive, especially taking into account that the database of more than 2000 patients (212 pathological) were used. However, less than 11% was pathological, therefore it might be possible that these high-accuracy results were achieved due to data set imperfection. Highly satisfactory results were achieved by [109], where authors managed to reach accuracy of 99.994%. Support vector machine (SVM) and MPEG-7 audio low level features after reduction by Fisher discrimination ratio (FDR) were used, however the database contained audio recordings of only 226 patients.

Very often, random forest (RF) is used a base classifier for voice pathology detection and provides very high accuracy. As [62] shows, by using RF it is possible to achieve classification accuracy as high as 100%. Publicly available Saarbruecken Voice Database (SVD) was used for these experiments and only 28 audio features were extracted. [111] used database of 3126 audio recordings for sounds of heart classification. These recordings are not voice sounds, however the audio features are very similar. RF application in this research provided best achieved accuracy of 92%.

In some studies, authors try to use different techniques for accuracy improvement, like data set optimization (reduction) or utilization of less often used classification algorithms. For example [120] reduced their voice parameters vector to only 17 features and use confusion matrix for classification. Their best achieved classification accuracy was 83.7%. [29] achieved 100% classification accuracy while discriminating between healthy and pathological classes and 87% discrimination accuracy between nodules and Reinke's edema. Such high classification accuracy was probably achieved due to extremely small database, consisting of only 47 patients. In [6], authors apply Auto-Correlation and Entropy features for voice pathology detection. They employ 3 different publicly available voice databases: Disordered Voice Database (MEEI), Saarbrucken Voice Database (SVD), Arabic Voice Pathology Database (AVPD). Achieved accuracies of 99.69%, 92.79%, and 99.79% respectively, looks very impressive. The same three previously mentioned databases were also used by [8]. Fisher discrimination ratio (FDR) was applied for parameters vector reduction and SVM was used for classification. Best obtained accuracies were 99.68%, 88.21% and 72.53% for SVD, MEEI and AVPD databases respectively. Others experiment using different types of recording equipment, most commonly - contact microphone. However, as noted in [157], the contact microphone does not bring any additional information useful for the classification of voice signals, and is outperformed by an acoustic microphone.

Questionnaire data collected by otolaryngologists is another source of data that can be used for voice pathology detection. As noted in [151], responses to some specific questions may contain information, which is not available in acoustic analysis. There are many questionnaires dedicated to voice quality assessment, however they are rarely used for preventive larynx health care. [68] analyses Pediatric Voice Symptom Questionnaire (PVSQ), [122] developed Vocal Fatigue Handicap Questionnaire (VFHQ), and they prove both questionnaires to be useful voice assessment tools, nevertheless, both these questionnaires are used only for voice quality evaluation. [151] shows, that query data based laryngeal pathology detection perpetually outperforms acoustic analysis and the fusion of both data sets improves the performance even more. Query data can also be described as self-assessment data, which captures patients' perception of their own voice problems [85]. This also includes information about the impact of voice problems on the patients' daily lives, and provides additional information regarding other evaluation methods [75].

Some researchers are creating and using new voice evaluation techniques, such as vocal tract discomfort scale (VTDS) in [84]. Likert scale of seven points is used in VTDS to measure patient vocal discomfort using eight sensory symptoms: burning, tightness, dryness, aching, tickling, soreness, irritation and lump in the throat [84]. Each of these items has to be given a value from 0 to 6 and has a cut-off value, which indicates a presence of the symptom. The use of a VTDS scale allows easy development of the questionnaire and data validation process, because each required value can be a numerical answer to the question. Researches like [86] and [127] adapt the VTDS scale for German and Italian speaking patients respectively. Their results show that this scale is reliable, consistent, and has high clinical validity.

Another questionnaire used in some of the studies is developed by Hogikyan and Sethurman, and is called Voice-related Quality of Life (V-RQOL) [79]. This questionnaire contains 10 questions addressing the impact of voice on daily activities [30]. Each answer is a numerical value ranging from 0 to 100, where lower value indicates higher impact. Findings of [103] gives grounds for application of the V-RQOL as a reliable tool for screening occupational voice disorders.

Researchers use different techniques for query data classification. [147] generates association rules by modified apriori algorithm for medical data classification. Despite the initial statement that association rules mining is useless in a domain like medicine, [147] manages to achieve best average classification accuracy of 96.48%. [1] took a completely different approach and used RF for query data classification. Depending on the data aggregation technique, the average achieved accuracy varies around 80%.

As we can see from the results of other researchers, Random Forest and Support Vector Machine algorithms can be distinguished as the best performing classification algorithms for audio features classification. Such results as 99.994% by [109], 100% by [62], 92% by [111] or 86.62% by [157] indicates high potential of these algorithms, which is why RF or SVM should be highly considered for classification task in voice pathology detection. As mentioned in [18, 17, 39, 48], RF shows high performance and accuracy when dealing with high dimensional data, which makes it suitable for classification of data used in this research. Query data classification results show, that high classification accuracy can be achieved using only questionnaire data. However, as [54, 159, 151] shows, a combination and classification of query and voice data provide even better results, which indicates the high potential of this technique to achieve higher accuracy than other existing techniques.

There are many commercial voice analysis tools that are being used by laryngologists, such as LingWaves, Computerized Speech Lab and Dr. Speech. Even though Dr. Speech is very popular due to its low price and good documentation, it is not specifically clear how to use analysis results available from the software [134]. As far as we know, there is no existing tool which would be capable of analyzing both voice and query data in laryngeal pathology detection.

2.2. Acoustic and contact microphones

The human voice can be captured using different types of microphones. Acoustic microphones (AM) are the most commonly used, and capture the voice in the same manner as the human ear. Contact microphones (CM) are able to capture the voice, because vibrations of vocal folds are transmitted through the vocal tract and reaches the surface of the skin [12, 110]. Schematics of how each type of microphone is used are shown in Figure 2.1. As mentioned in [138], the microphone used for human voice recordings should capture frequencies from the lowest possible up to the highest perceivable by the human ear – 16000-20000 Hz. Multiple researchers show that the contact microphone is useful for extraction of voice fundamental frequency [66], voiced speech sound pressure levels estimation [139], recording subglottal pressure waves [116], neck surface vibrations mapping during vocalized speech [114] and detecting glottal vibrations [139].

Background noise highly affects the validity and reliability of acoustic measurements [38, 37]. As it is shown in multiple studies, the CM is less sensitive to background noise because of its vicinity to the voice source [110, 139, 160, 142]. Audio recordings made with contact microphone have different frequency contents, which indicates that some parts of information might be lost. Multi-location contact microphone usage is suggested by [110], to minimize information loss, because some frequencies can be recorded better in other locations than throat. It is also suggested by [38], that for valid reproduction of results in audio analysis, acoustic environment should have a signal-to-noise ration of at least 30 dB. This requirement can be easily fulfilled by performing all recordings in a sound-proof booth. However, when recording has to be done in an ordinary environment, this solution is not feasible.

Several studies proved that in cases where background noise is non-stationary, contact microphones can significantly improve classification accuracy [113, 35, 34]. A combination of recordings done by both types of microphones helped [41] to achieve speech recognition accuracy of 80%. An increase in performance while combining features of one type extracted from both types of recordings

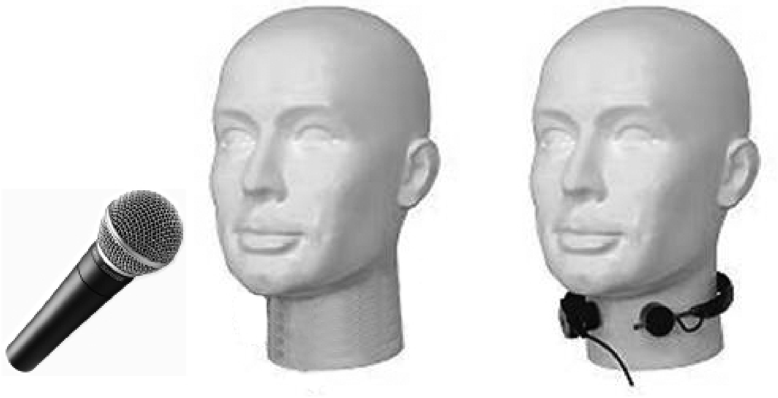


Figure 2.1 Acoustic (left) and contact (right) microphone usage schematics

was mentioned by [104]. Significant improvement of throat-only speech recognition by using a new framework introduced by [45], which learns joint sub-phone patterns of contact and acoustic microphone recordings using a parallel branch HMM structure.

2.3. Audio feature set extraction

Effective voice pathology analysis requires feature extraction in order to separate healthy and pathological voices. Some researchers use features from multiple data sources, such as audio recordings or questionnaires, while others use different representations of voice signal and features extracted from them. To obtain standard features or feature sets from voice signal, many available tools can be used, such as Dr. Speech, LibXtract, YAAFE, jAudio, Librosa, Marsyas, Aubio, Essentia, Meyda or MIRtoolbox. These are the most commonly used tools and complete list contains a lot more [101, 36, 20, 46]. As mentioned in [101], these tools usually come in one of the following formats:

- Stand-alone application.
- Plug-in for a host application.
- Software library.

Many such tools are developed every year, where some of them are completely new and others are extended or improved versions of older ones. Each tool offers a various functionality requiring different levels of computer usage skills. However, standard tools and features often provide unsatisfactory (not high enough accuracy) results, so researchers are developing new calculated features and feature extraction techniques.

Over the past years, certain measures of voice signal have been introduced, such as fundamental frequency (f_0), pitch perturbation (jitter), amplitude perturbation (shimmer), harmonic to noise ratio (HNR), normalized noise energy (NNE), signal to noise ratio (SNR) and mel-frequency cepstral coefficients (MFCC) [53]. MFCC coefficients are derived from the type of cepstral representation of the audio recording, and in most of the studies they are used in combination with other features [11, 161, 58]. Jitter and Shimmer are probably the most popular, so they are used in many studies and voice analysis tools, which explains otolaryngologists' familiarity with them.

For pathology detection improvement HNR and NNE features are used by [117, 90, 166, 125, 71]. Harmonic to Noise Ratio shows relative level of spectral noise in voice recording, which has lower value in pathological voices [71]. According to [166, 125], high level HNR indicates less noise. NNE together with HNR is widely used to evaluate voice quality [117]. As a pathological voice loses its quality, previous features can improve pathology detection. In some studies, first Rahmonic (R1) feature is used, which is proportional to the geometric mean of harmonic-to-noise ratio [10]. This means that R1 describes the voice quality globally.

2.4. Query data feature extraction

Multiple types of questions can be provided in questionnaires requiring different types of answers. Depending on the task, which questions are used for, answers can be numerical values, text or checkboxes. Numbers can be selected from a provided range (for example when patient is asked to evaluate its voice quality from 0 (being worst) to 10 (being healthy)), or written freely (such as age). Textual answers allow us to provide more detailed responses, but are more complicated to analyze. Today, most of textual analysis methods are still manual, which is expensive and time-consuming [115, 80]. Researchers are developing various automatic text analysis techniques, where machine learning approach looks the most promising [100, 3]. Checkboxes may allow selecting multiple answers for one question, which also provides more detailed response, but analysis requires more complicated techniques.

Questions used in this research are provided in Table 3.2. As it can be seen, most of the answers are single numerical values, and two questions require textual answers of predefined words (*Yes*, *No*, *Man*, *Woman*). Answers, expressed as numbers, require no additional modifications and are left as they are. Meanwhile, an automatic technique is used to transform textual responses to numbers. Answers of type *Yes/No* are expressed as 1 for *Yes* and 0 for *No*. Answers of type *Man/Woman* are expressed as 0 for *Man* and 1 for *Woman*. After editing required answers, we get single numeric data vector for each questionnaire, which can be used in classification algorithm.

2.5. Query Data validation

Before taking any further action and trying to use questionnaire data for classification, the consistency and completeness of it has to be ensured. In some cases, certain parts of data might be missing, therefore policy for handling such situations must be prepared and applied.

According to [83, 165, 112, 64, 50, 140] in some cases it is possible to use various techniques to fill missing data, such as expectation-maximization, multiple imputation, K-nearest neighbors (k-NN), Fuzzy c-means (FCM), Self-organizing map imputation (SOMI), Random Forest (RF), association rules. Each time when missing data are imputed, performance of imputation algorithm must be evaluated, because missing data imputation has to improve classification accuracy [64]. However, this technique is not suitable when the amount of missing data is excessive and values are not numerical. It must be taken into consideration that the estimation of missing values is quite unreliable and different imputations might lead to completely different classification results [50].

If a large part of respondents skipped the same questions, it would be more appropriate to remove those questions and analyze the data set without them. When there are only few answer collections where some responses are missing, usually it is the best solution to remove those collections from data set [77]. As noted in [50], this solution is acceptable only when the amount of missing data is relatively small (e.g. less than 5% of whole data set).

As [77] suggests, in certain occasions we can use all the questionnaires, even if some of them are incomplete. In this case we would have different sample sizes for each question. Unfortunately, this scenario is not suitable for correlation or regression studies, that is why in this work the most suitable approach was taken and incomplete questionnaires were removed.

2.6. Data and decision level fusion

Data fusion is used to achieve a complete data set from different sources which do not contain the same data [168]. Often it is possible to collect different types of data for single classification task, and in order use it all in computational intelligence methods, the data has to be joined into a single data set. It is also known that integration and analysis of data from multiple sources can be used to develop insights that are more detailed and more accurate than those resulting from single data source [136]. There are two most commonly used data fusion techniques: data-level fusion and decision-level fusion.

Data level fusion, often called feature level fusion, can be described as basic integration of two data sources [60]. Both heterogeneous sets of data must contain the same number of subjects, about which the data is collected. Then all the data is copied to a single collection, while joining by the identifier of the subject. When one data set contains less subjects than the other, some kind of data imputation technique has to be applied in order to resolve missing data problem.

Decision level fusion can be defined as the process of fusing information from individual data sources after each data source has undergone a preliminary classification [123]. According to [144, 123], it can improve recognition performance compared to simple individual classifiers. However, selection of classification algorithms has to be done carefully, because it is possible to run into a situation where selected classifiers are so bad that the combined result is worse than that of some of the individual classifiers [123]. Decision-level fusion is used when multiple classification algorithms can be applied or data level fusion does not provide enough accuracy. It also allows using multiple types of data sources, even when some of them cannot be fused into a single one.

Many decision-level fusion techniques are used by different researchers. Often results from base classifiers are combined by using linear opinion pool (LOP), the logarithmic opinion pool (LOGP), fuzzy k-means (FKM), fuzzy vector quantization (FVQ), median radial basis function (MRBF) network. However, usually the most simple technique is used where the final class is obtained from meta-learner classifier. According to [144], three types of decision level fusion can be distinguished:

1. *Abstract Level Fusion.* The most simple decision level fusion, where final decision is achieved by combining outputs of base classification algorithms to a test sample. Most commonly, the majority voting method is applied here, but often an obtained test sample is used in another classifier (meta-learner).

2. *Rank Level Fusion.* Outputs of each used classifier are sorted in decreasing order of confidence so that each class has its own rank. Fusion is performed by summing up ranks of each class and final decision is given by choosing class of the highest rank.

3. *Score (Measurement) Level Fusion.* When using this technique, fusion rules on the data vectors are derived to represent the distance between the test and training subjects. Each classifier's output is represented by scores or measurements. Fusion is achieved by combining vectors of scores and decision is given by the class that has the smallest value.

2.7. Visual voice and query data analysis

In some cases, results provided by classification algorithms is not accurate enough and deeper data analysis is required. Data visualization is an important part of exploratory data analysis and helps to better understand the data which is used [43]. When high dimensional data is used, modern visualization techniques can be applied for data visualization in lower, but still meaningful dimensions [87]. In other situations, statistical data distribution visualization might provide additional insights.

In this study, voice data is in high dimensionality, therefore a visualization algorithm with dimensionality reduction is required. High-dimensional data visualization is still an important active research field [70, 55, 141] providing several algorithms, such as SP matrices, parallel coordinate plots, Isomap

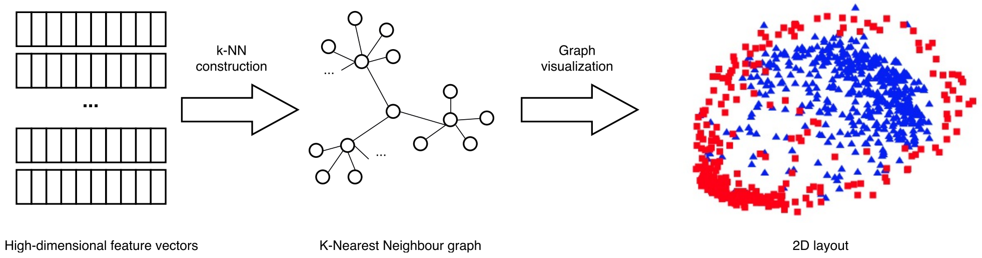


Figure 2.2 Typical workflow of data projection to 2D space by first constructing K-nearest neighbor graph

and t-distributed stochastic neighbor embedding. Most commonly used algorithms (including *t*-SNE) first construct similarity structure which then is used to project data to 2D space. An example of such workflow is provided in Figure 2.2.

t-SNE algorithm was used for voice data visualization in the present study. According to [154], this algorithm has shown great performance working both with real world and with artificial data. This algorithm defines data similarities in terms of conditional probabilities in the high-dimensional data space and their low dimensional projection [76]. Another useful feature of this algorithm is the ability to embed new observation onto a beforehand generated map. Voice data, mapped to 2D space, provides useful information allowing to spot misclassified, incorrectly labelled data or data mapped very closely but originating from different sources [99]. Misclassified or incorrectly labelled data can be spotted, when observation of one class appears (in the 2D data map) inside or close to the group of observations from another class.

For query data visual analysis, Probability Density Functions (PDFs) were used. PDFs have a wide range of applications, including anomaly detection, two-sample comparison, binary classification and clustering [26]. In our study, PDFs are used to create and visualize distributions of query and several audio data parameters (in a statistical sense). Provided graphs give better understanding about the position of analyzed observation - closer to healthy or pathological group. As recommended in [133], probability density functions are calculated using Epanechnikov kernel smoothing.

2.8. Computer software usability evaluation

Computer programs are a very important part of the medical diagnostics process. First, health care information systems were introduced in early 1970's [92] and are increasingly used since. Therefore, continuous assessment and improvement of such computer programs is required. As mentioned in [129], such evaluation requires analysis of a user's understanding of the software, because user satisfaction guarantees successful implementation.

Because of the area where medical computer software is used, it requires

Table 2.1 Usability definitions in various Standards

Standard	Usability definition
ISO/IEC9126-1, 2000	The capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions.
ISO9241-11, 1998	The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.
IEEE Std.610.12-1990	The ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component.

great usability and high quality [118]. As mentioned in [118], usability (often called "ease of use") can be applied to any object that is used for some tasks. It also can be explained as how easily, effectively, efficiently and satisfactorily computer program allows to achieve your task. Computer program usability is defined by International Standardization Organization (ISO) as "the extent to which a product can be used by specified users to achieve specified goals with efficiency, effectiveness, and satisfaction in a specified context of use" [69]. Other definitions from various standards are provided in Table 2.1.

According to [96], computer program quality can be defined in two ways:

1. The degree to which the whole system, it's component or a single process meets the specified requirements.

2. The degree to which the whole system, it's component or a single process meets the needs or expectations of a user.

The second definition indicates that system quality and usability are similar. It is still a big debate in which one affects the other, and opinions depend on the research domain. However, it is proved that system usability and quality are related to each other [118]. There are many techniques developed for successful software quality and usability evaluation, but the latter one usually is considered more important. It reveals user satisfaction, which highly depends on program quality (higher satisfaction - better quality), so at the same time it provides evaluation results for both - quality and usability. ISO 9241 standard is widely analyzed in many studies and is the most commonly used technique for computer software usability evaluation. In this research we are using ISO 9241-11, which is better known as "ISO 9241-11:1998 Guidance on usability".

3. VOICE & QUERY DATA

Non-invasive data, such as voice recordings and questionnaire answers, were used in this study. Data was provided by the Department of Otolaryngology from Lithuanian University of Health Sciences. All data was divided into 3 different data sets - two for audio data, and one for query data. In cases where data of different types was used, validation techniques were applied, which are more thoroughly described in section 4.7.

3.1. Voice data

Audio recordings of sustained phonation of the vowel /a/ (as in English word "large") were used for representation of subject's voice information. The decision to use steady-state phonation was made because it is simple, time effective, reduces variance in sustained vowels, and enables reliable computation of acoustic features [167, 163]. Moreover, [93] mentions that sustained vowels are not influenced by speech rate and stress, usually does not contain voiceless phonemes, fast voices onsets and terminations, prosodic fluctuations in pitch and amplitude. All voice signals were digitally recorded in a sound-proof booth using an acoustic microphone AKG Perception 220 (AKG Acoustics, Vienna, Austria) with frequency range from 20 Hz to 20 kHz. Microphone was placed 10 cm away in front of the subjects mouth, keeping about 90° microphone-to-mouth angle and all subjects were seated (Figure 3.1). WAV audio format was used (mono PCM, 16 bit samples at 44 kHz rate). For each subject there were 1-3 recordings done (0.5 to 3 s length).

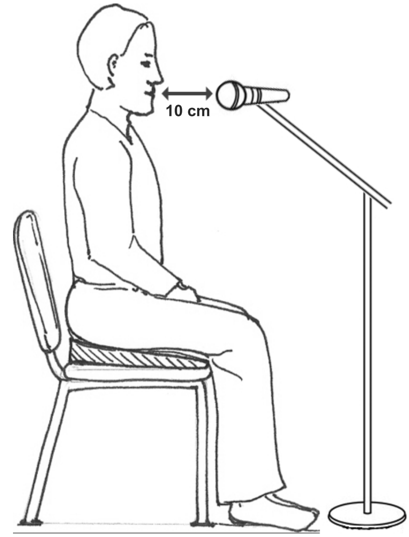


Figure 3.1 Set up scheme for patient voice recording

The database contains 273 mixed gender and age (19 to 85) subjects (distribution visible in Figure 3.2) as in [157] (163 normal and 110 pathological voices). Healthy voice observations were collected from healthy volunteers, who considered their voices as normal, had no history of laryngeal diseases and had no complaints about their voice. These subjects were also checked by otolaryngologists, using laryngostroboscopy method, to ensure that there were no pathological alterations in the larynx. Likewise, the recordings of these individuals were evaluated as normal by clinical voice specialists.

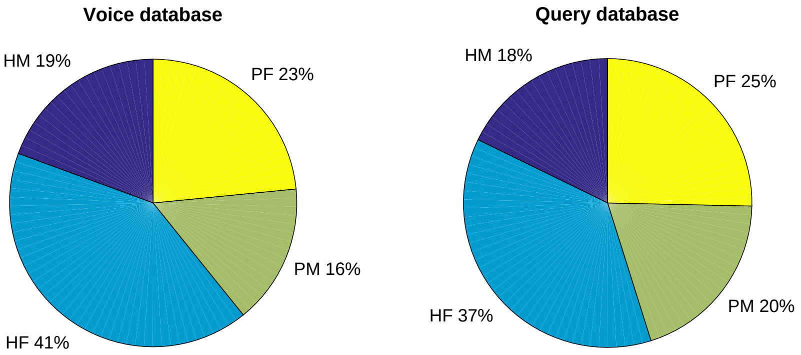


Figure 3.2 Distribution of healthy and pathological males and females in voice and query data sets. HM - Healthy Males, HF - Healthy Females, PM - Pathological Males, PF - Pathological Females.

The pathological voice group included patients with mass lesions of vocal folds (nodules, polyps, and cysts) and diffuse lesions of vocal folds (papillomata, hyperplastic laryngitis with keratosis, and carcinoma). Information from laryngostroboscopy and direct microlaryngoscopy was used to visually evaluate the severity level of the pathology (from 1 to 3) and provide initial diagnosis. Histological examination of laryngeal samples taken during endolaryngeal microsurgical intervention was used for final diagnosis confirmation.

For some experiments in this research we used different modifications of audio recordings. While analyzing our suggested association rules algorithm and comparing query data classification accuracy with voice data (or fusion data) classification accuracy, we used either raw (V_0) audio signals for feature extraction or signals after pre-processing (V_1, V_2). We used longest continuous sequence of frames detection and only active frames were retained for further analysis. Two types of voice activity detection were used:

1. Simple voice activity detection used in [52]. It works by filtering absolute value of standardized signal by convolution with Gaussian window with the size of 750 frames. Standardized signal is divided by 4, and only resulting values above 0.04 of such filtering are used as indicators of voice activity.

2. Statistical model-based [135] activity detection available by `vadsohn` in

Table 3.1 Settings for `vadsohn` function

Parameter	Value
qq.pr	0.5
qq.ts	0.4*length(s)/fs
qq.tn	0.1*length(s)/fs
qq.ti	0.1
qq.tj	0.1
qq.gx	50
qq.gz	0.001
qq.ne	1

Voicebox [21] toolkit. We also changed some default settings from their defaults (see Table 3.1).

The goal of this study was to classify subjects' voices as *Healthy* or *Pathological*. That way, the severity and type of the disease was ignored, and in the final database all pathological patients were marked the same. Subjects' distribution by their class is shown in Figure 3.3.

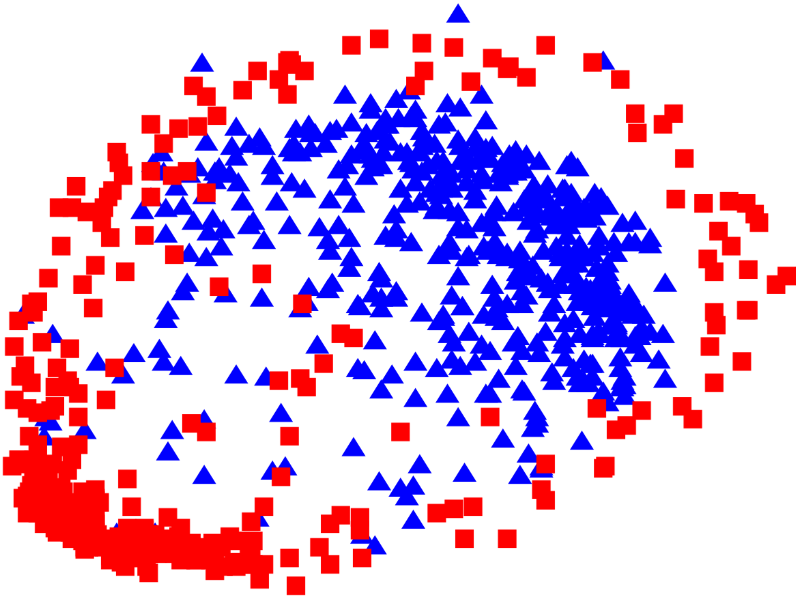


Figure 3.3 Visualization of audio database. Dimensionality reduction done by *t*-SNE.

3.2. Query data

Query data was collected during initial patient examination by otolaryngology specialists. Steps of the collection process are represented in Figure 3.4. A paper version of questionnaire was filled out by the doctor, or a patient was asked to fill it out on his/her own, and later the data was transferred onto computer. Afterwards, required data validation techniques were applied, to change textual answers to numerical values (for example answer "*Yes*" was changed to "1" and answer "*No*" was changed to "0") and remove the questionnaires with missing data.

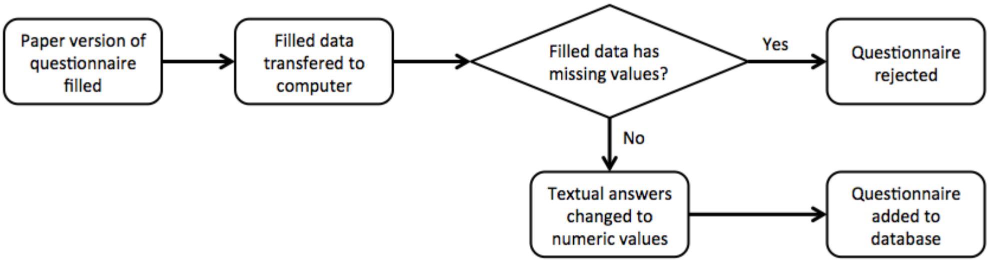

Paper version of questionnaire filled
Filled data transfered to computer
Filled data has missing values?
Yes
Questionnaire rejected
No
Textual answers changed to numeric values
Questionnaire added to database

Figure 3.4 Query data collection process

For questionnaire data sets, the best performing questions were selected by [13]. Database of 596 mixed gender subjects (106 healthy men, 221 healthy women, 118 pathological men and 151 pathological women) was used. Distribution can be seen in Figure 3.2. Questions used to collect data from patients are provided in Table 3.2. Visual patient distribution by their class is provided in Figure 3.5.

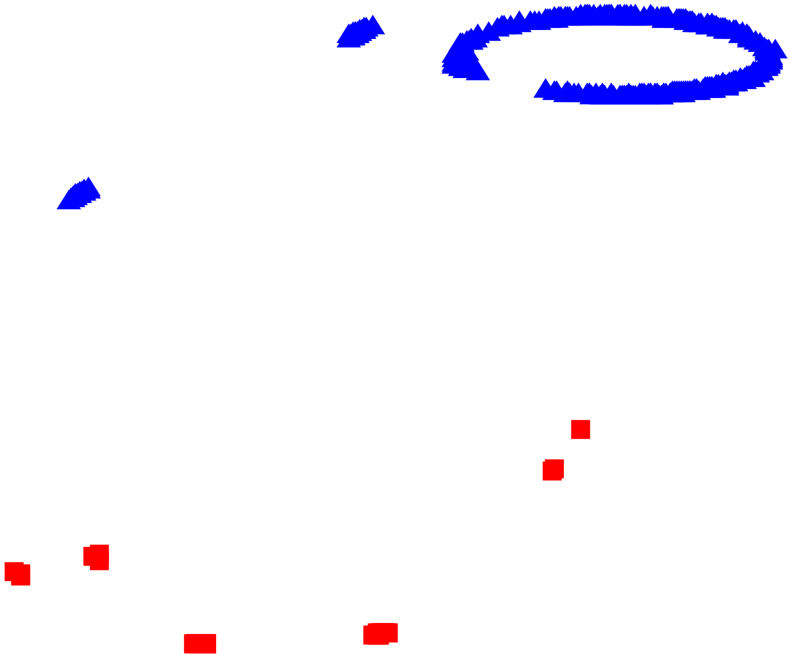


Figure 3.5 Visualization of the query database. Dimensionality reduction done by *t*-SNE.

Quantization was used in affinity analysis rule extraction, where responses to questions with a wider scale (having more than seven unique values) were re-coded into quartiles. This was performed because data used in affinity analysis have to be categorical, which means that there should be as few unique values as possible. Zero values were always put into a separate category and not considered for re-coding. Initial analysis revealed that many subjects responded to questions # 22-23 with answer 0 and in those questionnaires most of the other answers coincided. This allowed a reduction of the final number of extracted rules and make them more general, by replacing responses to questions # 22-25 with a single response G_0, indicating that at least one of those questions was answered by *no problem* answer:

$$G_0 = (G_1 = 0 \vee G_2 = 0 \vee G_3 = 0 \vee G_4 = 0) \tag{3.1}$$

Also, while performing affinity analysis and extracting association rules, we used two data sets of query data: short version Q_9 (questions # 1-9) and long version Q_{26} (questions # 1-26).

Table 3.2 Questionnaire items. SSA stands for subjective self-assessment.

#	Question content	Units (or scale) of measurement
1.	subject's gender	{Man, Women}
2.	subject's age	discrete number
3.	average duration of intensive speech use	hours / day
4.	average duration of intensive speech use	days / week
5.	smoking	{Yes, No}
6.	smoking intensity	cigarettes / day
7.	smoking history	years
8.	maximum phonation time	seconds
9.	SSA of voice function quality	visual analogue scale from 0 to 100
10.	SSA of voice hoarseness	from 0 (*no*) to 100 (*severe hoarseness*)
11.	voice handicap progressing	grade from 1 to 4
12.	SSA of daily experienced stress level	from 0 (*no*) to 100 (*very high stress*)
13.	frequency of singing	grade from 1 to 5
14.	frequency of talking/singing in a smoke-filled room	grade from 1 to 5
15.	SSA of experienced discomfort due to voice disorder	from 0 (*no*) to 100 (*high discomfort*)
16.	SSA of "too weak voice"	from 0 (*no*) to 100 (*very clear*)
17.	SSA of repetitive "loss of voice"	from 0 (*no*) to 100 (*very clear*)
18.	SSA of reduced voice	from 0 (*no*) to 100 (*very distinct*)
19.	SSA of reduced ability to sing	from 0 (*no*) to 100 (*very distinct*)
20.	frequency of voice cracks or aberrant voice	from 0 (*no*) to 100 (*very often*)
21.	level of vocal usage	level from 1 to 4
22.	speaking took extra effort (G_1)	from 0 (*no*) to 5 (*severe problem*)
23.	throat discomfort or pain after voice usage (G_2)	from 0 (*no*) to 5 (*severe problem*)
24.	voice weakens while talking, vocal fatigue (G_3)	from 0 (*no*) to 5 (*severe problem*)
25.	voice cracks or sounds different (G_4)	from 0 (*no*) to 5 (*severe problem*)
26.	glottal function index [14, 124] ($GFI=G_1+G_2+G_3+G_4$)	grade from 0 to 20

4. METHODOLOGY

In this work, multiple classification techniques were used for voice and query data classification in order to determine if patient is *healthy* or *pathological*. Our learning data sets were made of n observations, where each of it contains a vector of features (also called as predictors) and corresponding class label. We used 3 different data sets of voice features, extracted from audio recordings. Features from our suggested set are described in the following sections of this chapter.

Since our created tools are going to be used by otolaryngologists in their clinical practice, there was a task set that algorithms used for data classification and exploration should lend themselves for providing insights into automatic decisions and numerical data analysis - be transparent. To fulfill this requirement, Decision tree and Association rules algorithms were used. Both of them are transparent, based on parameters routinely used by otolaryngologists and can be very beneficial for perception of various conditions and treatment planning. The unsupervised Apriori algorithm [4] was used to mine association rules, that would reveal the most pronounced co-occurrences of responses. As a result, simple confidence-based laryngeal disorder detection algorithm was derived from association rules.

We used random forest as a basic model to detect laryngeal pathology and demonstrate how information available from this "black box" type model can be used to visually explore data and decisions. In our proposed data dependent random forest-based technique we used RF to fuse information available from several non-invasive data sets. In cases where data-level fusion was used, we performed an initial data analysis, which left data sets only with full data vectors (fused data set containing only subjects existing in all fused data sets).

Acoustic and contact microphones were compared for voice recording in order to determine if classification accuracy of methods, used in this study, can be improved by using different kind of microphone or features from two recordings, one with acoustic and another with contact microphone. For this task, Random Forest classification algorithm was used as a basic model. To test classification accuracy improvement using data from several microphones, different data fusion techniques were analyzed.

For data visualization purposes t-SNE (Stochastic Neighbor Embedding) algorithm was used to map the proximity matrix from RF and query data on to 2D space. The user is also provided with decision tree graph and probability density functions for a more detailed visual data and decision analysis. Moreover, visual data comparison of basic patient audio parameters and questionnaire answers are available in separate windows as a feature of our created tool.

Developed computer program usability evaluation was performed according to ISO-9421 standard, which revealed whether the program is ready for usage by specialists in their everyday work.

4.1. Voice features

For audio feature extraction from audio recordings, acoustic analysis and signal analysis techniques were applied. Features were extracted for multiple short segments of audio signal, which was divided into windows (also called frames) by applying short-term parametrization. This technique provides a set (or so-called vector) of values for each feature (for example 100 Cepstral energy values, when signal is divided into 100 segments). Depending on a task, obtained features can be used as they are or post-processed by applying compression or transformation techniques.

Table 4.1 Extracted audio feature sets used in this study (927 features in total)

#	**Type of features**	**Size**
1.	Pitch and amplitude perturbation measures	24
2.	Frequency (0-5000 Hz)	100
3.	Mel-frequency bands	35
4.	Cepstral energy	100
5.	Mel-frequency cepstral coefficients	35
6.	Autocorrelation	80
7.	Harmonics to noise ratio in spectral domain	11
8.	Harmonics to noise ratio in cepstral domain	11
9.	Linear predictive coefficients	77
10.	Linear predictive cosine transform coefficients	77
11.	Shape of signal envelope	128
12.	Levinson-Durbin reflection coefficients	24
13.	Vocal tract area irregularity	71
14.	Perceptual linear predictive cepstral coefficients	154

Features in the first data set were extracted from audio recordings during this research. To perform this task, Matlab software was used. For each subject, there are up to 3 separate audio recordings, and for that reason data set contains more feature vectors than there were subjects. Another audio feature set contains audio parameters from same patients from whom the query data was obtained, and were extracted in the Department of Otolaryngology from Lithuanian University of Health Sciences. The last data set contains features extracted from the same recordings as in the previous data sets, but different feature extraction techniques were used.

The first feature set is the most versatile characterization of the recorded voice samples. It consists of 14 diverse feature subsets and contains 927 features in total. This set of features was selected as by [52, 157] it was shown that the use of it allows achieving significantly high accuracy. A full list of the feature subsets and number of features in each of them is provided in Table 4.1. To

Table 4.2 Extracted audio features, which otolaryngologists are familiar with

#	**Feature**
1.	Fundamental frequency (F0)
2.	Jitter
3.	Shimmer
4.	Normalized noise energy (NNE)
5.	Harmonics to noise ratio (HNR)
6.	Signal to noise ratio (SNR)

extract audio features from recordings, various signal analysis techniques were used. Feature subsets from this set are described in the following sections, while a more detailed analysis can be found in [52, 157]. For the increase of feature diversity, the last 3 feature subsets from the list were added to the previously used ones in [52]. All feature sets in this data set were pre-processed before classification by centering and scaling to have zero mean and unit variance. This set of features was used for the experiments of RF algorithm.

The second feature set is a lot smaller and contains only 6 audio parameters. The list of these parameters can be found in Table 4.2. These features were extracted by otolaryngology specialists during patient examination and voice recording. "Dr. Speech" (http://www.doctorspeech.com) software, created by Tiger DRS Inc was used. This computer program offers a set of 23 features, such as HNR, pitch and amplitude perturbation measures, jitter and shimmer. Audio parameters from this data set were selected, as they have a great importance in voice pathology detection, and as it is mentioned in [99], they are widely adopted by otolaryngologists all over the world. Moreover, the use of these parameters with transparent techniques, such as Decision Tree or Association rules, is of a great value for the end user, because it allows exploration of data and automatic decisions.

Table 4.3 Overview of the `emobase.conf` file, used to extract features from audio recordings

Low-level descriptors	Statistical functionals
Intensity, loudness, 12 Mel-frequency cepstral coefficients (MFCCs), pitch, pitch envelope, probability of voicing, 8 frequencies of line spectral pairs (LSP), zero-crossing rate (ZCR)	Min (or max) value and its respective relative position within a signal, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1–3, and 3 inter-quartile ranges

The third feature set contains features computed by openSMILE [46]. A base feature set of this tool is designed for emotion recognition, however, it was successfully used in many studies like [46, 27, 51, 132] for various kinds of tasks, achieving accuracy of at least 70%. Such results were a reason to make use of openSMILE in this work. Features, extracted with openSMILE, were used while comparing audio parameters classification to query data classification using extracted rules.

This feature set contains 26 low-level descriptors (LLDs) and the 1^{st} derivative (delta or velocity) of each LLD. All features are specified and can be configured in the `emobase.conf` configuration file. A collection of statistical functionals is applied to summarize various aspects of the framebased data distribution for each LLD ant its delta. More details are provided in Table 4.3. Data set contains 988 features in total - (26 LLDs + 26 deltas) x 19 functionals. Following audio signal processing-related settings are set as default in `emobase.conf` file:

- Pitch and pitch envelopes are estimated using pre-emphasis (of 0.97) and overlapping Hamming windows. Default values for overlapping step and Hamming window are 10 ms and 25 ms respectively.
- Other LLDs are obtained without pre-emphasis and signal is windowed into overlapping Hamming windows. Overlapping step duration is 10 ms, while Hamming window duration is 40 ms.

A simple moving average filter with a window size of 3 is used to smooth all extracted LLDs. Only then they are compressed by the statistical functionals.

4.1.1. Pitch and amplitude perturbation measures

In this study, we used pitch and amplitude perturbation measures suggested in [102] and analyzed in [52], which contain 24 features. Audio signal was segmented into overlapping windows (segments) to calculate measures of both pitch and amplitude. Perturbation values were estimated for each window, where the length of the window was estimated as $3/F_{inf}$, where F_{inf} is the minimum allowed fundamental frequency (pitch) $F0$ [89]. We used $F_{inf} = 60Hz$, which resulted in windows of 50ms long. The same as in [102, 52], each window overlapped neighboring ones by 75%. The value of fundamental frequency F_i and corresponding amplitude measure A_i were calculated for each segment (window) of voice signal. For pitch detection, autocorrelation was used as in [102]. For our feature extraction, we used the continuous wavelet transform based pitch detection technique (same as in [52]) suggested by [89]. As mentioned in [150], such perturbation measures as jitter and shimmer are widely used by otolaryngologists which provide additional advantages.

4.1.1.1 Pitch perturbation features

Pitch and amplitude perturbation measure collection contains the following pitch perturbation features:

1. Mean frequency: $F_{av} = (1/n) \sum_{i=1}^{n} F_i$
2. Maximum frequency: $F_{max} = max_{i=1,...,n}(F_i)$
3. Minimum frequency: $F_{min} = min_{i=1,...,n}(F_i)$
4. Standard deviation of frequency:

$$F_{std} = (1/(n-1)) \sum_{i=1}^{n} (F_i - F_{av})^2 \tag{4.1}$$

5. Phonatory frequency range:

$$PFR = \frac{12}{log2} log \left(\frac{F_{max}}{F_{min}} \right) \tag{4.2}$$

6. Mean absolute jitter: $MAJ = (1/(n-1)) \sum_{i=1}^{n-1} |F_{i+1} - F_i|$
7. Jitter: $JITT = MAJ/F_{av}$
8. Pitch perturbation quotient smoothed over three windows:

$$PPQ_3 = \frac{(1/(n-2)) \sum_{i=2}^{n-1} \left| (1/3) \sum_{k=i-1}^{i+1} F_k - F_i \right|}{F_{av}} \times 100 \tag{4.3}$$

9. Pitch perturbation quotient smoothed over five windows:

$$PPQ_5 = \frac{(1/(n-4)) \sum_{i=3}^{n-2} \left| (1/5) \sum_{k=i-2}^{i+2} F_k - F_i \right|}{F_{av}} \times 100 \tag{4.4}$$

10. Pitch perturbation quotient smoother over 55 windows: $-PPQ_{55}$
11. Pitch perturbation factor:

$$PPF = \frac{N_{p \geq threshold}}{N_{voice}} \times 100 \tag{4.5}$$

where $N_{p \geq threshold}$ is the number of how many times the difference in fundamental period between the window i and window $i+1$ values is greater than 0.1 ms in magnitude. N_{voice} is the value, which describes how many times the pitch window i value differs from the pitch window $i+1$ value.

12. Directional perturbations factor:

$$DPF = \frac{N_{\delta\pm}}{N_{voice}} \times 100 \tag{4.6}$$

where $N_{\delta\pm}$ is the value, indicating how many times the pitch perturbation changes the algebraic sign across the windows.

4.1.1.2 Amplitude perturbation features

Pitch and amplitude perturbation measures collection contains the following amplitude perturbation features:

1. Mean amplitude: $A_{av} = (1/n)\sum_{i=1}^{n} A_i$
2. Maximum amplitude: $A_{max} = max_{i=1,\ldots,n}(A_i)$
3. Minimum amplitude: $A_{min} = min_{i=1,\ldots,n}(A_i)$
4. Standard deviation of amplitude: $A_{std} = (1/(n-1))\sum_{i=1}^{n}(A_i - A_{av})^2$
5. Mean absolute shimmer: $MAS = (1/(n-1))\sum_{i=1}^{n-1} |A_{i+1} - A_i|$
6. Shimmer %: $SHIM_p = MAS/A_{av}$
7. Shimmer in decibels:

$$SHIM_d = \frac{1}{n-1}\sum_{i=1}^{n-1} 20log\left(\frac{A_i}{A_{i+1}}\right) \tag{4.7}$$

8. Amplitude perturbation quotient smoothed over three windows:

$$APQ_3 = \frac{(1/(n-2))\sum_{i=2}^{n-1}\left|(1/3)\sum_{k=i-1}^{i+1} A_k - A_i\right|}{A_{av}} \times 100 \tag{4.8}$$

9. Amplitude perturbation quotient smoothed over five windows: $-APQ_5$
10. Amplitude perturbation quotient smoothed over 55 windows: $-APQ_{55}$
11. Amplitude perturbation factor:

$$APF = \frac{N_{p\geq threshold}}{N_{voice}} \times 100 \tag{4.9}$$

where $N_{p\geq threshold}$ represents the number of how many times the amplitude difference between the window i and window $i+1$ values is greater than 4% of the maximum amplitude.

12. Amplitude directional perturbation factor:

$$ADPF = \frac{N_{\delta\pm}}{N_{voice}} \times 100 \tag{4.10}$$

where $N_{\delta\pm}$ is the value, indicating how many times the amplitude perturbation changes the algebraic sign across the windows.

4.1.2. Frequency features

Frequency features were calculated by dividing the frequency range from 10 to 5000 Hz into $n = 100$ non-overlapping frequency bands of equal width. The total spectral energy of the i^{th} band is used to calculate the i^{th} frequency feature x_{2i}:

$$x_{2i} = sum_{\omega_k \in BAND_i} F(\omega_k),\ i = 1, ..., n \tag{4.11}$$

where $BAND_i$ is the i^{th} frequency band, ω_k addresses distinct frequencies, and the Fourier spectrum F is given by the short-time FFT

$$F = FFT\|(v \cdot h)\| \tag{4.12}$$

where v is the voice signal, $\| \cdot \|$ stands for the norm operator, and h is the Hanning window.

4.1.3. Mel-frequency features

Spectral energy of the i^{th} mel-window is used to calculate the i^{th} mel-frequency feature x_{3i} [57]:

$$x_{3i} = W_i F,\ 1 \leq i \leq M \tag{4.13}$$

where M is the number of the mel-windows in the mel-scale and W_i is the triangular weighting function associated with the i^{th} mel-window. The center frequencies of the weighting function were set to the mel-scale frequencies, so the weighting function vanished. Different number of mel-windows M is used by different authors. For example, in [57] this number was varying from 20 to 24. In this work the number of windows was $M = 32$ as in the experiments carried out by [52] it provided the best performance.

4.1.4. Cepstral energy features

For cepstral feature extraction, time (quefrency) range of 0.2 to 25 ms (40-5000 Hz) is divided into $n = 100$ non-overlapping bands of equal width. The ith feature x_{4i} is calculated by the total cepstral energy in the i^{th} band:

$$x_{4i} = \sum_{\tau_k \in BAND_i} Q(\tau_k) \tag{4.14}$$

where τ_k addresses distinct quefrencies and Q is given by

$$Q = \|IFFT(log(F))\| \tag{4.15}$$

where $IFFT$ stands for the inverse fast Fourier transform.

4.1.5. Mel-Frequency Cepstral Coefficients

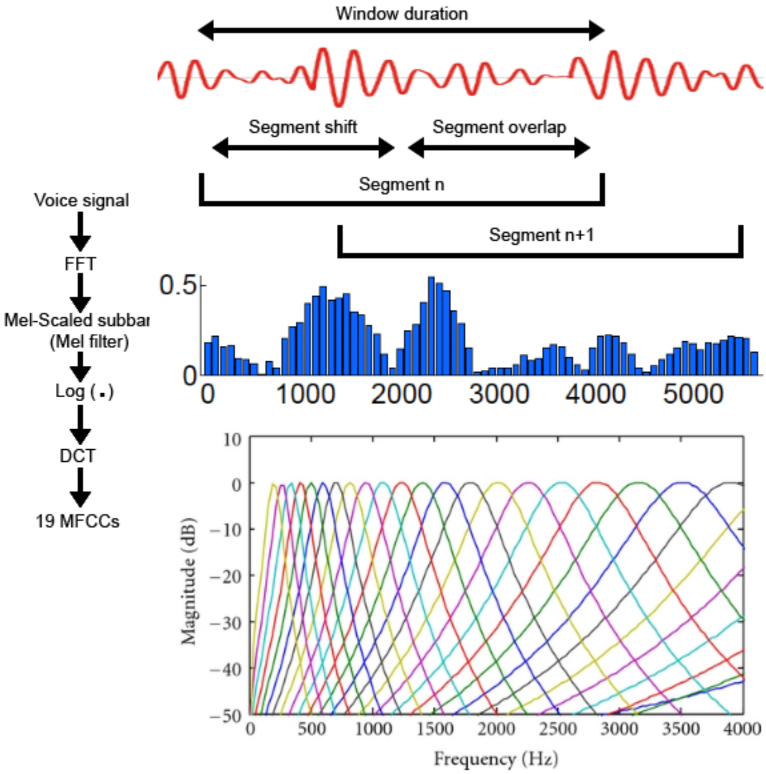


Figure 4.1 Mel-Frequency Cepstral Coefficients extraction from a voice signal

Mel-Frequency Cepstral Coefficients (MFCC) are widely used by researchers for audio signal characterization. MFCC can be estimated from short-term segments of input signal by applying parametric approach derived from Linear Prediction Coefficients (LPC) or non-parametric Fast Fourier Transformation (FFT) [161]. MFCC characterizes the energy distribution of a signal in the frequency domain. When voice signal contains additional noise, MFCC values are not very robust, so it is a common practice to normalize values in voice analysis systems to lessen the influence of the noise.

In the beginning audio signal is divided into segments (windows) in the time domain. Then, the amount of energy from each particular frequency range is obtained by converting windowed signal into the frequency domain by using FFT. For data amount reduction, triangular Mel-frequency filters are applied by summing filtered FFT bin values, and then Mel filter bank outputs are obtained. Mel-scaling is performed to get higher resolution at low frequencies and lower resolution at high frequencies.

The final step of MFCC extraction is application of discrete cosine transform (DCT) to logarithm or Mel filter bank outputs. DCT represents the signal by constant component and components of successively increasing frequency, respectively called first basis function and remaining basis function. Compacted MFCC vector of the relative frame is represented by the first components of DCT. The very first coefficient of MFCC vector (also known as 0^{th} or constant component, reflecting fundamental frequency), often is excluded.

Feature extraction from windowed signal results in great amount of data. The number of data depends on such parameters as signal length, window size and overlap settings. The basic steps of MFCC extraction are provided in Figure 4.1

4.1.6. Autocorrelation features

The process of autocorrelation features extraction is visually presented in Figure 4.2. The calculations which have to be performed to extract autocorrelation features are these:

1. Autocorrelation coefficients r_τ are calculated from the voice signal v in the interval $[0 \quad t]$:

$$r_\tau = \sum_{n=0}^{N-\tau-1} v_n v_{n+\tau},\ 0 \leq \tau \leq tF_s \tag{4.16}$$

   where N is the length of the vector v, F_s is the sampling rate of the voice recording ($44100 s^{-1}$), and the time span t is chosen larger than one period.

2. The average fundamental period t_{av} is estimated, which, together with half of the period, is marked in the Figure 4.2 by the thin line and ∇ respectively.

3. Coefficients with the lag less than $t_{av}/2$ are selected and the number of these coefficients equals to $t_{av}/2F_s$. In Figure 4.2 these coefficients are shown as dots in the bottom picture.

4. Finally, selected autocorrelation coefficients are processed by smoothing them with cubic spline smoothing algorithm [31] and sampling them over a linear mesh in the interval $[0 \quad t_{av}/2]$. Resulting values are denoted as $\hat{r}_{\tau_i}$ and considered as the estimation of the autocorrelation coefficient at the lad τ_i, which are shown as circles in Figure 4.2.

The autocorrelation features x_{6i} are then given by $\hat{r}_{\tau_i}$ values:

$$x_{6i} = \hat{r}_{\tau_i} \tag{4.17}$$

where $\tau_i,\ i = 1, ..., M$ are linearly spaced in the interval $[0 \quad t_{av}/2]$. In this study value of $M = 80$ autocorrelation features was used.

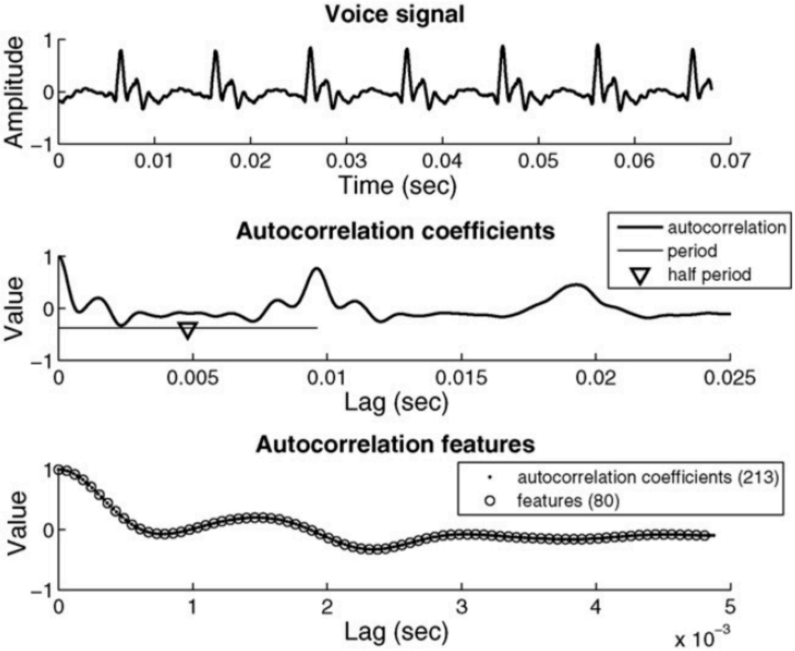


Figure 4.2 Autocorrelation features extraction process

4.1.7. Harmonics to noise ratio in spectral domain

Harmonics to noise ratio (HNR) is a widely used measure of voice quality [33, 78]. As is noted in [131], it is also very often used by otolaryngologists and speech pathologists for voice function evaluation. Because of its popularity, this measure is used in such commercial voice assessment software solutions as "Dr. Speech".

HNR is given from the energy of harmonics related to the noise energy in the voice. It can be estimated from the speech spectrum or from the speech cepstrum. In the first case, the harmonics energy (spectral domain) has to be filtered and removed, while in the second case the rahmonic energy (cepstral domain) has to be filtered and removed. The Fourier transform applied to the resulting filtered cepstrum or the filtered log spectrum in the spectral domain case provides the noise spectrum N. Then N is subtracted from the energy of the original log spectrum O to estimate HNR. Usually, HNR is calculated in different frequency bands, where the ones used in this study are provided in Table 4.4. We used the same frequency bands as in [102].

The feature x_{7i} (HNR for the ith frequency band) is calculated by

$$x_{7i} = mean(O_{f \in BAND_i}) - mean(N_{f \in BAND_i}) \tag{4.18}$$

Table 4.4 Frequency bands used to calculate HNR

Band number	Frequency Band (Hz)
1	0-500
2	0-1000
3	0-2000
4	0-3000
5	0-4000
6	0-5000
7	500-1000
8	1000-2000
9	2000-3000
10	3000-4000
11	4000-5000

where f addresses frequencies.

4.1.8. Harmonics to noise ratio in cepstral domain

HNR in cepstral domain components fills feature vector x_8. These values are computed in different quefrency bands in the cepstral domain. Process of HNR values computation is described in the previous section. Feature sets x_7 and x_8 form two redundant feature sets, the same as x_3 and x_5. In this study both HNR feature sets were used, to determine how well the classifiers are able to learn the decision boundaries defined by each of them.

4.1.9. Linear prediction coefficients

Linear prediction features were obtained from parameters of forward linear predictor determined by minimizing the squared prediction error. As mentioned in [52], the p^{th}-order linear predictor predicts the current value of the real-valued time series v based on past samples:

$$\hat{v}_n = -a_1 v_{n-1} - a_2 v_{n-2} - ... - a_p v_{n-p} \tag{4.19}$$

where $\hat{v}_n$ is the predicted value, a_i is the ith predictor parameter, and p is the predictor order. The feature vector x_9 is given by parameters of the predictor:

$$x_9 = \{a_1, a_2, ..., a_p\} \tag{4.20}$$

According to [25, 91, 88], to model male and female voices, different values of p may be required. In this study we used the same values as suggested in [25] - $p = 33$ for female voice and $p = 44$ for male voice.

4.1.10. Linear prediction cosine transform (LPCT) coefficients

The discrete cosine transform (DCT) is closely related to the discrete Fourier transform [52]. It is often possible to reconstruct a signal from only a few DCT coefficients with a very high accuracy. To obtain linear prediction cosine transform feature set x_{10i}, features are computed by the coefficients of the DCT applied to the linear prediction coefficients:

$$x_{10i} = \sum_{n=1}^{p} \left(x_{9n} cos \frac{\pi(2n-1)(i-1)}{2p} \right), \; i = 1, ..., p \tag{4.21}$$

Comparison is the purpose of using both x_9 and x_{10} sets of features, as in the case of $x_3 - x_5$ and $x_7 - x_8$.

4.1.11. Reflection coefficients and vocal tract area irregularity features

Human vocal tract can be modelled by using M tubes. This way, feature computation is based on M^{th} order linear prediction filter. M^{th} order prediction error E^m and area of m^{th} tube A_m are computed for each frame of voice recording. Later on, the Levinson-Durbin recursion algorithm is used:

$$Am = A_{m+1} \frac{1 + k_m}{1 - k_m}, \; m = M, ..., 2, 1 \tag{4.22}$$

where $A_{M+1} = 1$ and k_m is Levinson-Durbin reflection coefficient. The 12th feature subset from our collection contains only Levinson-Durbin reflection coefficients. Values for Vocal tract area irregularity subset are obtained by calculating the mean area $\overline{A}_m$, the variance of tube area S_m and the variance of ratio $S_m r$ for each tube. As it is mentioned in [105], to calculate these measures, tube area values $A_m k$ are computed for different frames k of a voice recording:

$$\overline{A}_m = \frac{1}{K} \sum_{k=1}^{K} A_{mk}, \; m = 1, ..., M \tag{4.23}$$

$$S_m = \frac{1}{K-1} \sum_{k=1}^{K} (A_{mk} - \overline{A}_m)^2, \; m = 1, ..., M \tag{4.24}$$

$$S_{mr} = \frac{1}{K-1} \sum_{k=1}^{K} \left(\frac{A_{mk-1}}{A_{mk}} - \frac{\overline{A}_{m-1}}{A_m} \right)^2, \; m = 2, ..., M \tag{4.25}$$

where K is the number of frames in a voice recording. If we assume that $M = 24$ tubes, then in total we have $24 + 24 + 23 = 71$ features in 13th feature subset.

4.1.12. Perceptual linear predictive cepstral coefficients

The perceptual linear predictive (PLP) analysis combines spectral analysis with linear prediction analysis [63]. However, in our study we calculate PLP cepstral coefficients (PLPCC). This allows to take psychophysical aspects of human hearing into consideration, according to the following steps:

1. *Signal windowing.* To split audio recording into separate parts, 10ms Hamming window was used with 5ms overlap. As 44kHz audio sampling rate was used, 10ms frame splits to 440 samples.

2. *Power spectrum.* Fast Fourier Transform (FFT) using 512 bins is applied to each obtained frame for transformation to frequency domain. After that, power spectrum is computed.

3. *Bark spectrum (Eq.(3) in [63]).* As it is explained in [63], the power spectrum is warped into the Bark scale using this equation:

$$F_{Bark} = 6 \times sinh^{-1}\left(\frac{F_{Hz}}{600}\right) = 6 \times ln\left(\frac{F_{Hz}}{600} + \sqrt{\left(\frac{F_{Hz}}{600}\right)^2 + 1}\right) \quad (4.26)$$

   where F_{Bark} is frequency in Barks, F_{Hz} is frequency in Hertz, $sinh^{-1}$ denotes the inverse of hyperbolic sine, and ln stands for natural logarithm.

4. *Critical-band spectral resolution.* Energy in FFT bins is collected through 1 - *Bark* wide overlapping triangle filter-banks. It is also equally spaced by a 1 - *Bark* interval. Simulated critical-band masking curve (Eq.(4) in [63]) from [63] was used to convolve each triangle of the filter-bank.

5. *Loudness equalization (Eq.(7) in [63]).* This equal loudness pre-emphasis approximates the non-equal sensitivity of human ear at different frequencies at about 40 dB level.

6. *Intensity-loudness conversion (Eq.(8) in [63]).* Amplitude compression was used as another modification for human hearing approximation. The resulting spectrum could be regarded as perceived loudness, which is measured in Son units.

7. *Autoregressive modelling.* The perceived loudness spectrum is approximated by the spectrum of all-pole spectral modelling: the inverse DFT is applied to he perceived loudness spectrum; to solve the Yule-Walker equations for the autoregressive coefficients of the M^{th}-order all-pole model, first $M + 1$ values are used; the coefficients are hereafter transformed to cepstral coefficients. In this study $M = 14$.

The extraction of PLPCC was done using Matlab code from [42]. We used audio recordings of varying length, which resulted in different number of short-time frames, where each provided 14 PLPCCs. To summarize distribution of cepstral coefficients, descriptive statistics were used: (1) min, (2) max, (3) range $(min - max)$, (4) mean, (5) median, (6) standard deviation, (7) lower quartile (Q_{lower}), (8) upper quartile (Q_{upper}), (9) inter-quartile range $(Q_{upper} - Q_{lower})$, (10) skewness, (11) kurtosis. The overall size of PLPCC feature subset resulted in 14 x 11 = 154 features.

4.1.13. Dr. Speech

In this study 6 audio parameters were used, which were obtained by otolaryngologists using voice assessment software "Dr. Speech". Otolaryngology specialists are very familiar with this software, because they are often using it in their work. This software is based on a set of 23 features of various types. We adopted several audio parameters from this feature set, such as jitter, shimmer, HNR, NNE, F0, SNR, and included them into our feature database.

4.2. Decision tree classifier

Decision tree (DT) is a classifier, represented as a tree-like hierarchical structure, containing internal nodes, leaf nodes and branches [24]. DT structure begins with root node and two sub trees - left and right. All nodes, except leaf nodes, have a left and right sub-tree. Each leaf node represents a class label which serves as a classification result. Example of very basic decision tree is provided in Figure 4.3. This DT structure is created during the learning phase. Once the tree is created, data is taken randomly from data set and classification accuracy is tested [81].

There are multiple methods for creating decision tree, but in this study standard CART [19] algorithm was used. The following steps were performed using this method:

1. In the first step whole training data set was used. All conceivable binary splits were examined on every predictor.

2. The split with the best optimization criterion was selected in the second step.

3. Selected split was applied.

4. All previous steps were repeated recursively for both child nodes.

Recursive split continues until concerned node contains observations of only one class. Other stop rules are: the node has fewer observations than it is allowed, imposed split produces node with fewer observations than allowed, or maximum allowed number of splits is achieved [19]. When the tree is built, it is pruned to check for over fitting and noise.

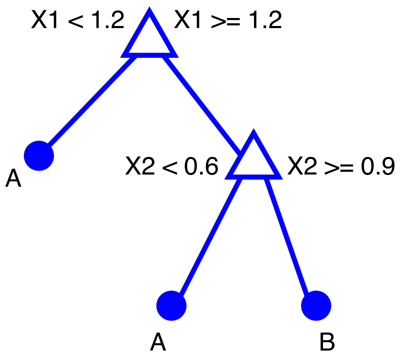


Figure 4.3 Structure of a very basic decision tree. Triangles represent root and inner nodes, while filled circles - leaves.

The main advantage of the decision tree is that it is easy to construct and is readily interpretable (human readable) [24]. It is a "white box" algorithm, which can be analyzed in visual form and work as a decision support system. As mentioned in [81], DT is also robust to outliers and missing data values. The fact, that DT is susceptible to noisy data and only single output is allowed, is often presented as a disadvantage of DT [24].

Many researchers successfully used decision tree classifiers for the diagnosis of various diseases, such as breast or ovarian cancer, heart sound diagnosis and so on [81]. In this research, a decision tree is used for voice pathology detection. It is constructed using six audio parameters (F0, Jitter, Shimmer, NNE, HNR, SNR) and most descriptive parameter from query data - GFI.

4.3. Association Rules

Association rules was proposed as a technique for query data classification. The idea of this algorithm was to classify query data according to the rules, extracted from the questionnaire data used in this research. This approach would allow having a transparent technique for voice pathology detection, requiring no special hardware equipment. As an additional benefit, association rules extraction would provide useful information indicating most important questions and reducing the size of questionnaire.

To extract rules from our training data set, affinity analysis was used. Affinity analysis, as described in [151], determines the relationship of observations and features in a dataset. Set of association rules, with the type of `if <antecedent> then <consequent>`, is a result of such an analysis. In this set, *antecedent* (A) is a specific response or set of co-occurring responses regarding questionnaire statements, while *consequent* (C) is the subject's diagnosis. The following measurements were calculated to evaluate importance and usefulness of each rule:

$$support(A \to C) = P(A \wedge C) \tag{4.27}$$

$$confidence(A \to C) = \frac{P(A \wedge C)}{P(A)} \tag{4.28}$$

$$lift(A \to C) = \frac{confidence(A \to C)}{P(C)} \tag{4.29}$$

where $P(condition)$ is the probability of condition (fraction of data having the condition), $\wedge$ is logical AND, *support* is the popularity of the rule (fraction of data containing both A and C), *confidence* describes the strength or purity of the rule (how often having A leads to C), *lift* is a measure of surprise (the increased likelihood of C being found in combination with A).

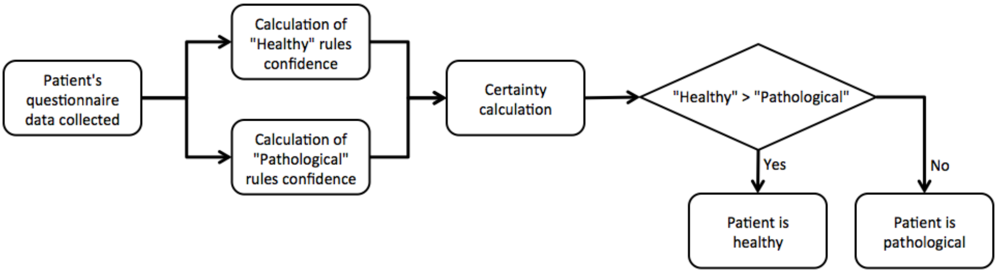

Patient's questionnaire data collected
Calculation of "Healthy" rules confidence
Calculation of "Pathological" rules confidence
Certainty calculation
"Healthy" > "Pathological"
Yes
No
Patient is healthy
Patient is pathological

Figure 4.4 Process diagram of classification by association rules

To apply extracted rule set for classification in this work we were using simple approach, derived from prediction by weighted majority [119]. As visible in Figure 4.4, to obtain prediction and certainty, the following 3 steps were used:

1. Confidences of 'healthy' rules, triggered by subject's responses, were summed.
2. Confidences of 'pathological' rules, triggered by subject's responses, were summed.
3. The difference between summed confidences was divided by the total triggered confidence to obtain a rough estimate of certainty.

Certainty was calculated as follows:

$$Certainty = 100 \cdot \frac{\sum_{i=1}^{J} P_i - \sum_{i=1}^{K} H_i}{\sum_{i=1}^{J} P_i + \sum_{i=1}^{K} H_i} \tag{4.30}$$

where J is the number of 'pathological' rules triggered, K is the number of 'healthy' rules triggered, P is the confidence of the triggered 'pathological' rule, and H is the confidence of the triggered 'healthy' rule. The sign of the resulting certainty value helps to determine the diagnosis: a positive value means pathological and a negative value means a healthy case.

The generated rules were filtered to leave only those having diagnosis in the consequent and being relatively significant: *support* > 0.28 and $confidence >$

Table 4.5 Rule coverage for mined antecedent items. Mapping of item's short name to the question number is as in Table 3.2. Absolute frequency, as the number of rules the specific response participates in, is in "Healthy" and "Pathological" columns.

Responses	Healthy	Pathological
G_0=0	11	0
H=2	11	0
C=0	6	1
Y=0	6	0
U=7	3	2
MFT=[2,12]	0	4
L=3	4	0
VAS=(63,100]	0	2
G_0=0	0	1
X=(73,100]	0	1
S=(75.8,100]	0	1
R=(60,100]	0	1
W=(60,100]	0	1
D=(65,100]	0	1
Gender=F	1	0

0.9 for a *healthy* subject and *support* > 0.16 and $confidence$ > 0.9 for a *pathological* subject. Gravity of each specific response, or absolute frequency of each retained antecedent component belonging to the healthy or pathological rules, are provided in Table 4.5. Final lists of generated rules and their parameters are provided in Table 5.6 (rules for *healthy* subjects) and Table 5.7 (rules for *pathological* subjects). Answers to questions # 22-25 (G_1 - G_4) were replaced by single variable G_0 (Formula 4.31), which indicates that patient answered at least to one of those questions by answer "no problem". As it was presented in [151], 'Healthy rules' turned out to be more complex, but redundant, where the most persistent components had the highest gravity. 'Pathological rules' were found to be shorter and their components slightly more diverse.

$$G_0 = (G_1 = 0 \vee G_2 = 0 \vee G_3 = 0 \vee G_4 = 0) \tag{4.31}$$

4.4. Random Forest Classifier

A random forest (RF) classifier is a committee of decision trees [18], where majority voting is used for tree aggregation, in order to solve classification tasks, see Fig. 4.5. As mentioned in [82, 40], RF is fast to train and to evaluate, it is parallelisable, robust to noise and provides good performance for high-dimensional data. RF is also known to be robust against over-fitting, and generalization error converges to a limit as the number of trees increases [18]. According to [126], generalization error is determined by the correlation of individual trees and average strength of them. Out-of-bag (OOB) validation is used to evaluate generalization performance of random forest, where only trees that did not use subject data for their construction, vote. Low bias and low correlation are essential for accuracy, which are achieved by growing trees to the maximum depth and applying randomization:

- Each separate tree of RF is grown on a bootstrap sample of the training set.
- In tree growing process at each node only n variables are randomly selected out of N available and only one variable, providing the best split, is used out of n selected.

While constructing RF, n is the only parameter, which has to be selected experimentally. If a single tree is grown using only part of the whole data set (bootstrap data), not used data (out-of-bag data) can be used for testing purposes of that tree. OOB also can be used for variable importance estimation, which is useful when feature selection is applied for the data set.

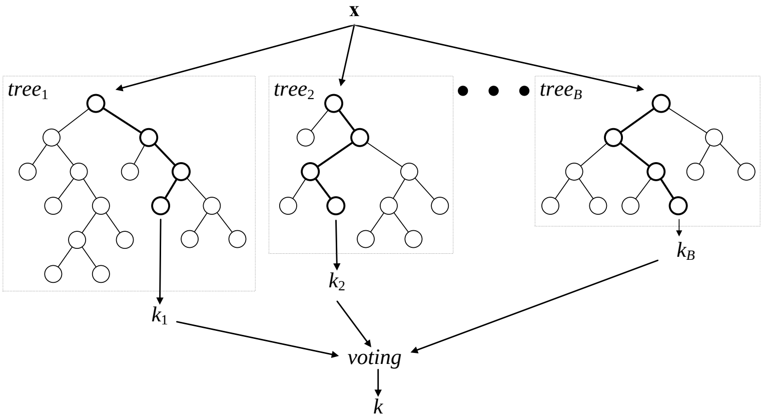


Figure 4.5 A general random forest architecture, where k stands for class label

Data proximity matrix Φ can be obtained from RF software. To obtain the matrix, data are run down for each tree that is grown. ϕ_{ij} is increased by one, if two observations - x_i and x_j - occupy the same terminal node of the tree. When the random forest is grown, the proximities are divided by the number of trees in the random forest.

A data proximity matrix, derived from RF, was used in this study for data analysis and visualization of data and decisions. To map data and decisions onto the 2D space, the t-distributed stochastic neighbor embedding (t-SNE) algorithm [154] was used, as it is capable of revealing both global and local structure in terms of clustering data with respect to similarity [76]. The t-SNE algorithm often outperforms other state-of-the-art techniques for dimensionality reduction and data visualization [154]. Fourteen separate RFs produce fourteen proximity matrices (Φ_i). To get a general mapping, a generalized proximity matrix Φ was obtained as:

$$\Phi = \frac{\sum_{i=1}^{N_F} w_i \Phi_i}{\sum_{i=1}^{N_F} w_i} \tag{4.32}$$

where $\mathrm{N_F}$ is the number of separate forests ($\mathrm{N_F}$=14) and w_i is a weight proportional to the average accuracy of the ith RF [157]. A sample t-SNE visualization of the generalized proximity matrix can be seen in Fig. 5.2. The algorithm has a compelling property allowing it to embed a new observation onto a previously-generated map.

4.5. Random Forest Classifier with information fusion

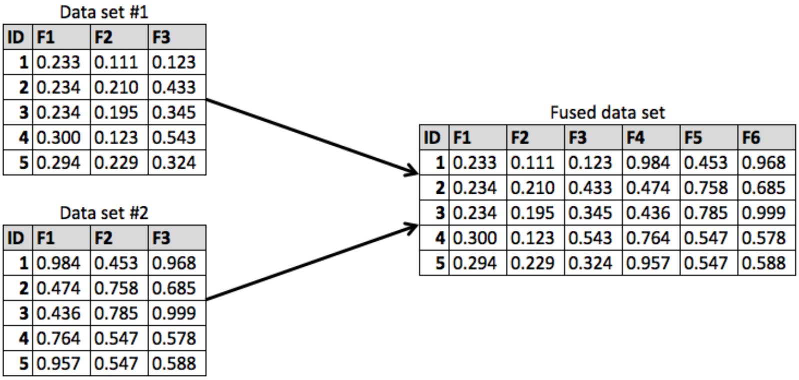
In this study, we propose a new data dependent random forest-based way to combine available data from multiple data sources. Multiple data fusion techniques were tested, and automated feature selection based on sequential backward elimination was applied to improve classification accuracy. As it was mentioned by [157], feature leading to maximum increase in OOB data classification accuracy is the criterion used for feature elimination. For all experiments of different techniques, the same set of audio recordings (Described in Chapter 3) was used.

The data set used for training and testing of this algorithm might raise some concerns because of its dimensionality, where number of observations is smaller than a number of features (273 subjects with 3 recordings each, compared to 927 features). According to [98], machine learning (ML) algorithms ability to learn is compromised when high dimensional data sets are used with small sample of observations. In such situations, it is common practice to reduce dimensionality in order to improve classification accuracy [97]. However, as mentioned in [18] and proved by many researches such as [17, 39, 48], RF shows great performance and accuracy when dealing with high dimensional data.

While constructing RF, such value as number of trees in RF - B, and the number of variables (randomly selected) used to split node - n, has to be selected. In this study B was set to 1000 trees per RF and n was set to $n = \sqrt{N}$, which is recommended by [18]. Selection of B and n values does not affect classification accuracy noticeably and may vary in extensive range.

4.5.1. Feature level fusion

The first of the data fusion techniques tested was the data-level fusion that can be explained as integration of data from multiple sources [60], which in our case are features from voice recordings and questionnaire answers. Combination is performed simply by taking required parameters from different data sets and combining them into single vector (one vector per each data record). In some cases, data fusion may result in missing data errors due to absent data in one or another data set. Solutions for such situations are analyzed in following sections of Chapter 4. Visual example of feature level fusion is provided in Figure **??**.



Data set #1

ID	F1	F2	F3
1	0.233	0.111	0.123
2	0.234	0.210	0.433
3	0.234	0.195	0.345
4	0.300	0.123	0.543
5	0.294	0.229	0.324

Data set #2

ID	F1	F2	F3
1	0.984	0.453	0.968
2	0.474	0.758	0.685
3	0.436	0.785	0.999
4	0.764	0.547	0.578
5	0.957	0.547	0.588

Fused data set

ID	F1	F2	F3	F4	F5	F6
1	0.233	0.111	0.123	0.984	0.453	0.968
2	0.234	0.210	0.433	0.474	0.758	0.685
3	0.234	0.195	0.345	0.436	0.785	0.999
4	0.300	0.123	0.543	0.764	0.547	0.578
5	0.294	0.229	0.324	0.957	0.547	0.588

Figure 4.6 Sample scheme of feature level fusion. ID is an identifier of observation and F stands for *Feature*.

Such data combination in this study was also used for data visualization and classification using Decision Tree, where 6 audio parameters are combined with GFI parameter from query data. In this case, as mentioned in [99], extra GFI parameter adds additional clarity for the user (one who is analyzing visualization) and improves classification accuracy, comparing it to the one achieved by using only voice parameters. In the present work such data level fusion is also called feature level fusion, because different subsets of audio features are fused. While using data fusion for comparison of different types of microphones, we concatenate audio features from 14 different feature sets into a vector of 927 components. This high dimensional data is used to build a separate RF.

4.5.2. Decision level fusion

Decision level fusion can be described as an ensemble of classifiers, where each of them has to be built independently and provides its own result. To obtain the final result, one of several fusion techniques has to be applied. In this study we used decision level fusion by forest voting, by averaging probabilities and by meta RF.

Decision level fusion by forest voting can be described as a technique, where all RFs vote for the class label of unknown observation. We use three different strategies of voting:

1. All single RFs vote.
2. All single RFs vote, but weights proportional to the accuracy of RFs are applied.
3. Selected single RFs vote.

For experiments of acoustic and contact microphone comparison we also used decision level fusion by averaging probabilities. This fusion was achieved by averaging estimates of the posteriori class probabilities obtained from single RFs.

To achieve fusion by meta RF, results of base classifiers have to be combined in a meta-learner fashion, where results of base classifiers become features for another (second level) classifier. When Random Forest is used as a base classifier, we provide inputs for the meta-learner by class posteriori probabilities obtained from base RFs. Given a trained RF, the posteriori probability for an observation x to belong to the qth class is estimated as:

$$p(t_1, ..., t_L, x, q) = \frac{\sum_{i=1}^{L} f(t_i, x, q)}{L} \tag{4.33}$$

where L is the number of trees in the random forest, x is the object being classified, q is a class label and $f(t_i, q, x)$ stands for the q^{th} class frequency in the leaf node, into which x falls in the i^{th} three t_i of the forest:

$$f(t_i, x, q) = \frac{n(t_i, x, q)}{\sum_{j=1}^{Q} n(t_i, x, q_j)} \tag{4.34}$$

were Q is the number of classes and $n(t_i, x, q)$ is the number of training data coming from the class q and falling into the same leaf node of t_i as x. According to this, if we have Q decision classes, then we have $Q \times M$ inputs to the meta learner (M is the number of base classifiers).

4.5.3. Data dependent random forest

Data dependent decision level fusion was our completely new fusion technique proposed for classification accuracy improvement. It is based on a construction of a separate RF for each data set, where final result is achieved classifying by RF, constructed of trees from the initial RF classifiers. Construction technique of the final Random Forest classifier is the main novelty of our proposed technique. As far as we know, there are no attempts made by other researchers to use this technique. The most popular techniques for classification of data from multiple data sets are data fusion before classification [67, 9] or classification by meta-classifier (2nd level classifier) [62, 111].

Visual representation of our proposed technique is provided in Figure 4.7. By using this method, each set of observation x extracted audio parameters is classified with separate random forest classifier. In a second phase, the neighborhood $\aleph_i(x)$ of m most similar to the x OOB observations are determined from each of the $\Phi_i, i = 1, ..., n$ (where n is the number of RF). Later, trees correctly classifying OOB data from neighborhood $\aleph_i(x)$ are selected from all RFs and a new RF is constructed from them, which is used for obtaining final decision. Assessment of similarity between x and x_j is done by measuring distance between two terminal nodes of a decision tree occupied by these observations. We suggest assessing similarity between x and x_j using the following equation:

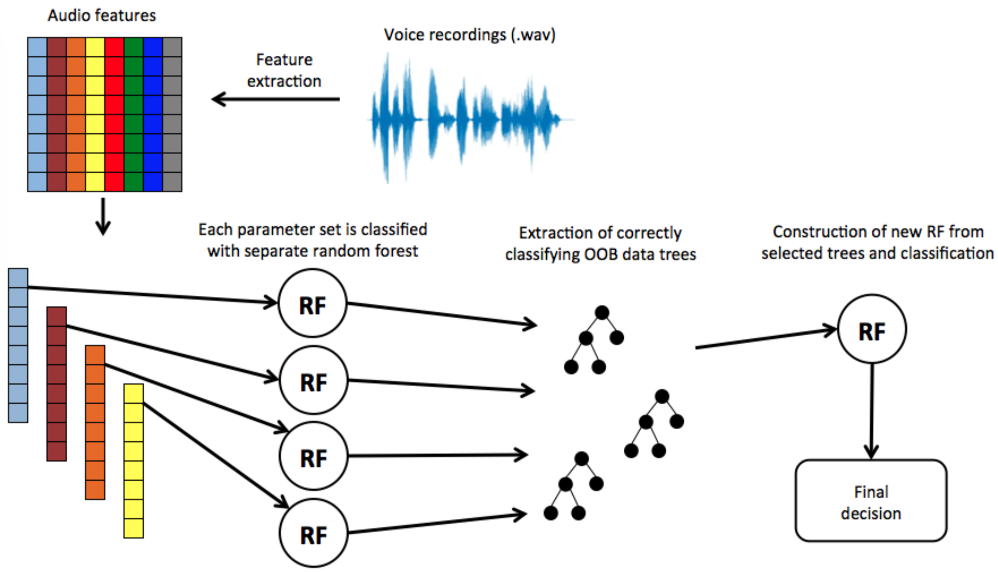

Audio features
Feature extraction
Voice recordings (.wav)
Each parameter set is classified with separate random forest
Extraction of correctly classifying OOB data trees
Construction of new RF from selected trees and classification
RF
RF
RF
RF
RF
Final decision

Figure 4.7 Proposed data dependent random forest classification technique

$$p_j = \frac{1}{K}\sum_{k}^{K} 1/(e^{w \cdot g_{jk}}) \tag{4.35}$$

where k runs over the K trees, for which x_j is among the OOB samples, w is a parameter and g is the number of tree branches between the two terminal nodes occupied by x and x_j. In the case where x and x_j occupy the same terminal node, $g = 0$ and p_j will be increased by value of one. Influence of g is controlled by the parameter w.

When making decision in RF, two options are taken into consideration: voting and weighted voting, where the k^{th} tree from the i^{th} single RF is given the weight, calculated by the following equation:

$$w_{ik} = \frac{1}{|\aleph_i^c(x)|}\sum_{x_j \in \aleph_i^c(x)} 1/(e^{w \cdot g_{jk}}) \tag{4.36}$$

where $|\aleph_i^c(x)|$ is the number of correctly classified OOB data in the neighborhood $\aleph_i(x)$, while meaning of the other variables are the same as in equation 4.35.

This proposed technique was used in this study for a comparison of acoustic and contact microphones, in order to determine the microphone (or a combination of microphones) which would provide the highest accuracy while screening for laryngeal disorders. To characterize audio recordings from both types of microphones for classification task, we used 14 ($14 \times 2 = 28$ feature sets in total) different sets of features, listed in Table 4.1. Effectiveness of determining two classes was analyzed for each separate and combined data set. For analysis of combined data sets, multiple data fusion techniques were applied, such as decision level with forest voting, decision level with averaging probabilities, decision level by meta RF and data dependent decision level fusion. Data sets of all fusion techniques (except decision level by meta RF) were used with all features and after additional processing - feature selection, forest weighting or forest selection.

4.6. t-Distributed stochastic neighbor embedding

t-Distributed stochastic neighbor embedding (t-SNE) is a prize-winning technique for dimensionality reduction. It allows to visualize high dimensional data in a low dimensional (2D) space. It is built upon earlier work on Stochastic Neighbor Embedding by [65, 28, 56]. t-SNE uses different cost function than SNE: it uses symmetrized version of the SNE cost function with simpler gradients and Student t-distribution rather than Gaussian to compute similarities [154]. t-SNE resolves crowding and optimization problems of SNE by the use of heavy-tailed distribution.

Input for t-SNE algorithm is a collection of N high dimensional data vectors $X = x_1, ..., x_N$. In this research data input for t-SNE is a proximity matrix obtained from RF. In high-dimensional space, similarity of x_j to x_i is defined as conditional probability $p_{j|i}$ where x_j is picked by x_i as its neighbor, when neighbors are picked in proportion to their probability density defined by a Gaussian centered on x_i [154]:

$$p_{i|j} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma^2)} \tag{4.37}$$

where σ is a parameter and the values of $p_{i|i}$ are set to zero. For low-dimensional counterparts y_i and y_j of x_i and x_j similar conditional probability $q_{j|i}$ is calculated. The joint probability is defined by the formula below:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}} \tag{4.38}$$

where $q_{i|i}$ are also set to zero. Minimization of Kullback-Leibler divergence between the joint probability distributions P and Q are used to find the desired mapping.

Similarity between observations is assessed by computing distance in equations 4.37 and 4.38. Yet, distance-based similarity of observations in multidimensional spaces suffers from irrelevant noisy variables, which have a really high influence. Therefore, in this work meta-learner RF of decision tree and association rules combination was used for assessment of similarities between observations.

Perplexity is an input parameter for t-SNE that determines configuration of the generated 2D map. The desired perplexity value can be set and algorithm itself determines the number of nearest neighbors, based on the data density [76]. This means, that the number of nearest neighbors is affected by the data and may vary after including new or removing existing data record.

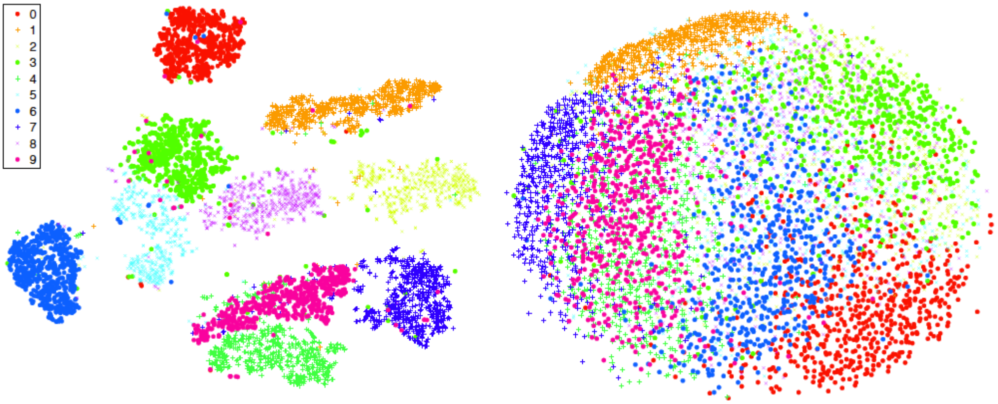


Figure 4.8 Visualization of 6000 handwritten digits from the MNIST data set [154]. *t*-SNE (left) and Sammon mapping (right).

Selection of *t*-SNE algorithm for dimensionality reduction and data visualization in this study is based on findings in [154]. Author compares *t*-SNE with Sammon mapping, Isomap and Local-linear embedding algorithms using MNIST data set. As the results shows, *t*-SNE clearly outperforms existing state-of-the-art techniques for visualization of real-world data. An example visualization of *t*-SNE compared to Sammon mapping is provided in Figure 4.8. Also, such features as ability to control the space between clusters and ability to embed a new observation into a previously calculated 2D map, were very appealing, considering our study.

4.7. Imputation techniques

In the database used in this study, each subject has up to 3 audio recordings and answers to a questionnaire. In some cases, query or audio data might be absent, so only data from one modality is available. In this situation, voice and query data fusion data set contains missing values. To avoid this situation, records with missing data have to be deleted before the data fusion, or missing values have to be filled with substitutes.

Various techniques, suggested in [83, 165, 112, 64] were considered, however, the same approach as in [151] was taken. Either complete case analysis with listwise deletion was performed, leaving only subjects having data from all modalities, or substitute values were used to fill missing data before fusion. We also analyzed and compared the following techniques:

1. *Median imputation.* In this method, all the missing values are filled out by calculated median of all observed decisions for the current modality.

2. *Linear regression.* For this technique, we used `mice` package which contains `mice.impute.norm.nob` function. `Mice` stands for multivariate imputation by chained equations [153].

3. *PCA-based imputation.* We used `imputePCA` function from `missMDA` package, which is described in [72]. For the selection of optimal number of components using generalized cross-validation criterion, the `estim_ncpPCA` function [73] was used.

4. *SVD-based imputation.* This technique is provided by the `pcaMethods` package [137]. We used the implementation of `SVDimpute` algorithm [145], where `kEstimate` function was used for the selection of optimal number of the components using cross-validation and $Q2$ distance.

5. *Iterative model-based imputation.* This technique is available in `VIM` package as `irmi` function [143]. In this work option `robust=FALSE` was set in order to use non-robust variant.

Finally, in data fusion situations, we used only records containing data from all modalities. In decision-level fusion situations, all data from all different data sets was used. Different classification algorithms use different data sets (voice or query data separately), therefore no missing data situations occur, unlike in data-level fusion, and no additional data preparation tasks have to be performed.

4.8. Classifier performance evaluation

In this work, we evaluated classifier performance as well, to show how well it performs. For this task multiple measures, such as detection error trade-off (DET) curve, equal error rate (EER), receiver operating characteristic (ROC) curve, cost of log-likelihood ratio (C_{llr}) and area under the curve (AUC), were calculated. These measures were estimated using an interpolated version of the ROC through pool adjacent violators algorithm (also called ROC convex hull method), which is available in the BOSARIS toolkit [23].

A more accurate evaluation of detection is possible because of score usage, instead of hard decisions. By varying decision threshold of obtained detector scores, corresponding reject and accept errors can be plotted and performance illustrated by ROC or DET curve. A quick way to compare the accuracy of detectors with different DET (ROC) curves is the equilibrium point, often known as equal error rate (EER). EER is the point where the DET (ROC) curve intersects the diagonal and:

a) false positive rate (miss rate) = false negative rate (false alarm rate), for DET.

b) true positive rate (sensitivity) = true negative rate (specificity), for ROC.

For estimation of DET and EER we used an interpolated version of the ROC through pool adjacent violators algorithm, which is called ROC convex hull (ROCCH) method (available in [23]). According to [128], DET curve tends to be more linear, due to logarithmic axes, so it allows easier comparison of several systems than ROC curve.

As mentioned in [99], various thresholds can be used to convert a soft decision into hard. ROC or DET curve can summarize the overall performance of a detector. DET curves allow easier comparison of several systems than ROC curves, due to the logarithmic scale [128]. Despite that, AUC measurements from ROC are valuable comparison parameters too. According to [49], DET curve can be represented as Function 4.39 (where p_i^F is false alarm probability), which is increasing and concave.

$$p_i^D = f_i(p_i^F) \tag{4.39}$$

In this study we also use another, more sophisticated, derivative of the ROC (DET) curve, which is called the cost of log-likelihood ratio (C_{llr}). The log-likelihood ratio can be described as the logarithm of the ration between two likelihoods: the likelihood that input is produced by pathological subject, and the likelihood that input is produced by healthy subject. In information theory terms, C_{llr} measures information loss, caused by the detector. Perfect detector, which makes no errors (also correctly notifies about the presence or absence of pathology), will have zero loss, while all others will have positive loss (lower is better). [23] provides more details about C_{llr}.

4.9. Confidence assessment of decisions

Our developed computer program provides user with confidence of a decision. This value is computed using information from RF made of $\{t_1, ..., t_L\}$ trees. For decision making, we formulate two types of rejections, which are based on computed posteriori probability values:

1. Similarity based rejection, where observation x is made if:

$$p^{max1}(\{t_1, ..., t_L\}, x, q) - p^{max2}(\{t_1, ..., t_L\}, x, q) < \varepsilon_s \tag{4.40}$$

where ε_s is a user defined similarity threshold and p^{max1} and p^{max2} are the first and the second largest posteriori class probabilities:

$$p^{max1}(\{t_1, ..., t_L\}, x, q) = \max_{q=1,...,Q} p(\{t_1, ..., t_L\}, x, q) \tag{4.41}$$

This type of rejection means that decision maker cannot make a reliable distinction between two most probable classes.

2. Dissimilarity based rejection, where observation x is made if:

$$p^{max1}(\{t_1, ..., t_L\}, x, q) < \varepsilon_d \tag{4.42}$$

where ε_d is the user defined dissimilarity threshold. In this type of rejection, none of the classes are similar enough to the observation x.

4.10. Similarity assessment of observations

In the experiments of this work, similarity between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ was assessed by measuring distance between two terminal nodes of the decision tree, which were occupied by those observations. As was suggested in [99], similarity for observations $\mathbf{x}_i$ and $\mathbf{x}_j$ was assessed by the equation 4.43, provided below:

$$p_{ij}^t = 1/(e^{w \cdot g_{ij}}) \tag{4.43}$$

where w is a parameter, and g_{ij} is the number of tree branches between the two terminal nodes occupied by $\mathbf{x}_i$ and $\mathbf{x}_j$. In case where single terminal node is occupied by both $\mathbf{x}_i$ and $\mathbf{x}_j$ observations, $g = 0$ and $p_{ij} = 1$. Parameter w is used to control the influence of the distance between two terminal nodes occupied by the observations on the similarity values.

Association rules-based similarity between observations $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed according to Eq. (4.44):

$$p_{ij}^r = \frac{2\sum_{n \in \mathcal{U} \cap \mathcal{V}} p_{in}}{\sum_{n=1}^{U} p_{in} + \sum_{n=1}^{V} p_{jn}} \tag{4.44}$$

where $\mathcal{U}$ and $\mathcal{V}$ are the sets of rules triggered by the observations. Respectively, U and V are number of rules in those sets, and p_{in} is the certainty of the nth rule triggered by $\mathbf{x}_i$.

4.11. Data visualization

Desktop computer application was created as one of the results of this study to ease patient vocal analysis. User interface (UI) is divided into 4 parts: audio file selection, graphical visualization of data (2D map of proximity matrix), textual view of analysis results and questionnaire data input window. After analysis of an unknown observation, user is provided with the determined class label, classification certainty obtained from the Affinity analysis and Decision tree, and the six voice parameters: F0, Jitter, Shimmer, NNE, HNR and SNR. Data visualization of statistical data view is also provided by using Probability Density Functions.

Program code is divided into two parts, each being responsible for separate tasks. The first part is responsible for audio feature extraction, classification and mapping proximity matrix to 2D space using t-SNE. Second part is responsible for storing audio recordings and questionnaire data, where *.wav*, *.mat* and *.data* file types are used.

The initial program window provides 2D map of audio database created by t-SNE algorithm. When classification is done, depending on the situation corresponding data set is visualized by t-SNE. Then unique t-SNE function is

Table 4.6 Parameters, available to user for monitoring. Abbreviations: F_0 - fundamental frequency; NNE - normalized noise energy; HNR - harmonic to noise ratio; SNR - signal to noise ratio.

Parameters	**Description**
Basic patient information	PatientID, Sex, Age, Diagnosis
Voice parameters	F_0, Jitter, Shimmer, NNE, HNR, SNR
Questionnaire data	Answers to query items (see Table 3.2)

used and subject of interest is mapped onto the same visualization. This allows to compare the analyzed patient to the others in the database and provides visual information to which class (healthy or pathological) this patient is closer. 2D map is created from similarities (as conditional probabilities) of data points.

Preliminary understanding of patient situations is very useful in the diagnosis process. Audio parameters, such as F0, Jitter, Shimmer, NNE, HNR, SNR and questionnaire data are very well known by otolaryngologists, this is why in this study, together with classification results, we provide audio parameters and questionnaire data distribution graphics. Probability density functions (PDFs) are used to visualize healthy ant pathological cases, where the Epanechnikoc kernel smoothing, as recommended in [133], was used to estimate PDFs. These distributions evolve over time with inclusion of new subjects (patients and controls), which provide doctors with information about trends and allow comparing different groups of patients. This also reveals the data variety in the dataset which allows spotting data deficiency.

Software program created in this study allows comparing audio parameters and query data of multiple concerned subjects. Comparison information is accessed by a click on a 2D map created using *t*-SNE. Click of a selected data point opens a new window, where information about the patient, linked to the point in a graph, is presented. Information which is provided in the patient's window was selected by otolaryngologists and is evaluated by them as most important. Table 4.6 provides a list of this data. As separate windows are opened for each click (patient), it is possible to compare multiple patient information, which allows to evaluate current patient more thoroughly. Information as such is very useful for preliminary diagnosis and teaching/learning purposes.

To facilitate reasoning, the user is provided with a decision tree (DT), which is visible in Figure 4.9. This tree is recalculated each time a new case is added to the database, so most recent and accurate DT is shown. Decision tree in Figure 4.9 is created using 6 audio parameters, which otolaryngologists are familiar with, and GFI parameter from query data, which improves classification accuracy. In the provided image, sample decision path is highlighted, which helps exploring patient diagnosis more thoroughly.

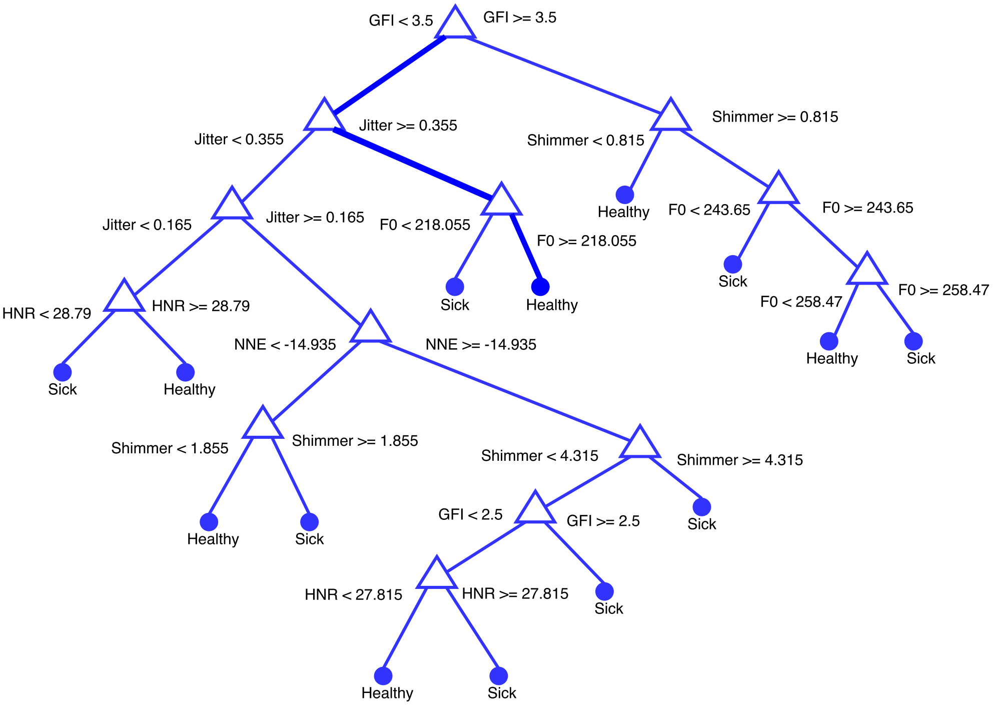


Figure 4.9 The decision tree created using the six audio parameters and the GFI parameter. Sample decision path is highlighted with a thicker line.

4.12. Ease of use evaluation

In this work, a computer program was created as an auxiliary tool for otolaryngologists. Before providing it to the end users, software ease of use and suitability for the task has to be evaluated to make sure it is user-friendly, self-explanatory, and meets the requirements. Evaluation was performed by taking the ISO-9241 standard part 11 (ISO 9241-11) into account, because it is impossible to evaluate the ease of use without taking into account user understanding [129]. The whole structure of ISO 9241 standard is provided in Figure 4.10.

As mentioned in [2, 118], user satisfaction is another important detail which greatly influences the success of software implementation. Evaluation of the developed software was done following seven principles of the standard 9241-11:

1. *Suitability for the task.* Software is suitable for the task if the user can easily understand what it can do.

2. *Self-descriptiveness.* This principle is evaluated by checking if software can be understood intuitively and no or very little additional information is needed. It also requires that any possible usage mistake would be followed by relevant information.

3. *Controllability.* Software controllability is achieved by creating user interface, which allows completing the task in one sequence of steps.

4. *Conformity with user's expectations.* Software conforms with user's expectations if it is consistent and complying with characteristics of the user.

5. *Error tolerance.* Computer program is admitted to be error tolerant if its usage requires no additional effort except in the events of obviously faulty usage.

6. *Suitability for individualization.* Software is suitable for individualization if it allows personal configuration for each user.

7. *Suitability for learning.* Software is suitable for learning if minimum effort for usage is required and help information is provided.

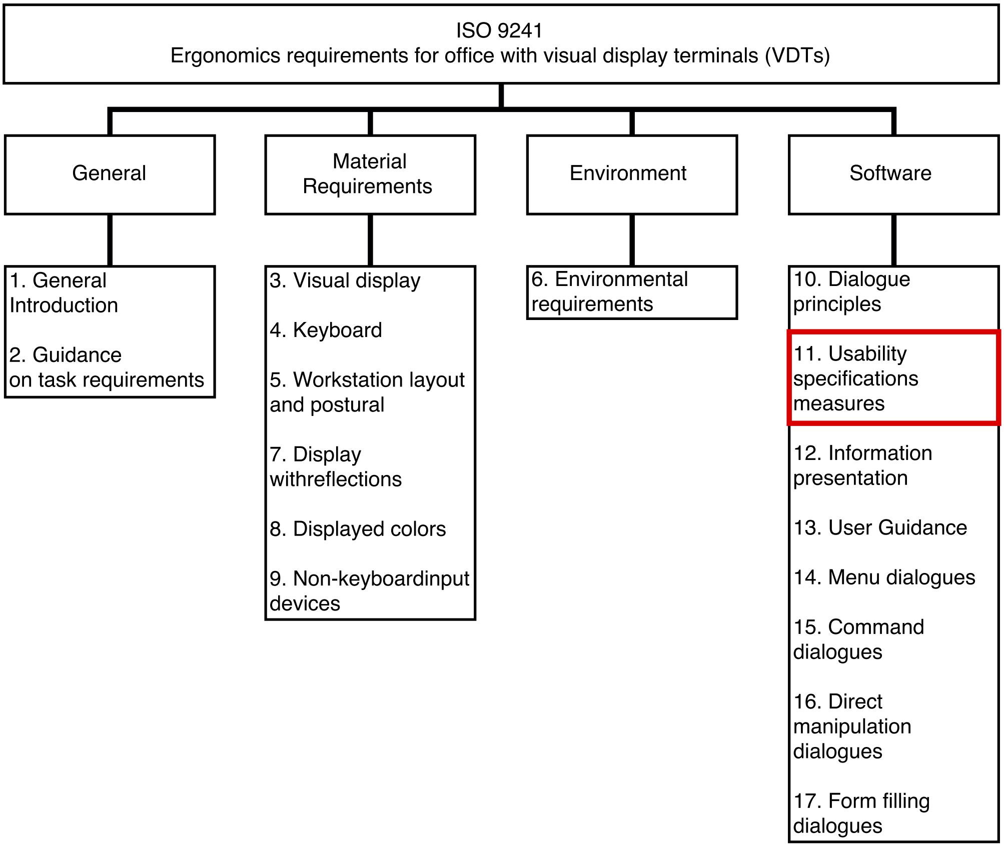

ISO 9241
Ergonomics requirements for office with visual display terminals (VDTs)
General
Material Requirements
Environment
Software
1. General Introduction
2. Guidance on task requirements
3. Visual display
4. Keyboard
5. Workstation layout and postural
7. Display withreflections
8. Displayed colors
9. Non-keyboardinput devices
6. Environmental requirements
10. Dialogue principles
11. Usability specifications measures
12. Information presentation
13. User Guidance
14. Menu dialogues
15. Command dialogues
16. Direct manipulation dialogues
17. Form filling dialogues

Figure 4.10 Structure of ISO 9241 standard. Part about software usability is marked by a red rectangle.

5. EXPERIMENTAL EVALUATION

The proposed methods in chapter 4 are designed to detect voice pathology from human voice and query data. To prove that these methods are usable in practice and can support the otolaryngologist decision, experimental evaluation has to be performed. Design and results of the experiments are provided in the following sections of this chapter.

5.1. Data used for experiments

For proper experimental testing, data should cover tested algorithm as much as possible. For the experiments in this research, three types of data sets are required: a set of audio parameters extracted by our used algorithms, a set of audio parameters extracted by using Dr. Speech software, and questionnaire answers filled out by patients. One of the requirements for data was that the same recording or questionnaire cannot appear in training and testing data sets. To fulfill the second requirement, data sets cannot contain records with missing data. If some audio parameters or answers to questions are missing, these records should be removed from training and testing data sets. In this research the data described in Chapter 3 (Voice & Query data) was used for experimental evaluation. Properties of the data sets are provided in Table 5.1.

Table 5.1 Three data sets used for the experiments

Dataset	Patients	Parameters	Comment
Voice data	273	927	Audio recordings were provided by the Department of Otolaryngology from Lithuanian University of Health Sciences and features were extracted before the experiments.
Voice data	273	6	Data set was provided by the Department of Otolaryngology from Lithuanian University of Health Sciences. Parameters were extracted using Dr. Speech software.
Query data	596	26	Data set was provided by the Department of Otolaryngology from Lithuanian University of Health Sciences. Questionnaires were filled by the patients themselves or by the doctors.

5.2. Experiment environment

To perform the experiments Matlab was used to design and develop program system. In order to separate logic from Graphical User Interface (GUI) and other parts, the multi-component approach was taken. The structure of the system is provided in Figure 5.1.

As it is visible in Figure 5.1, the system contains two main parts: GUI

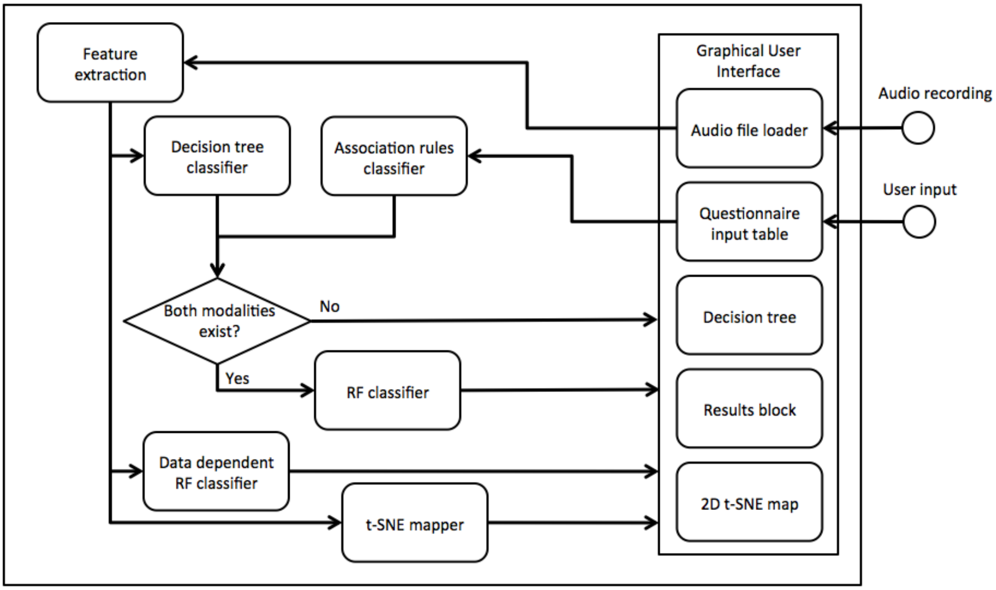
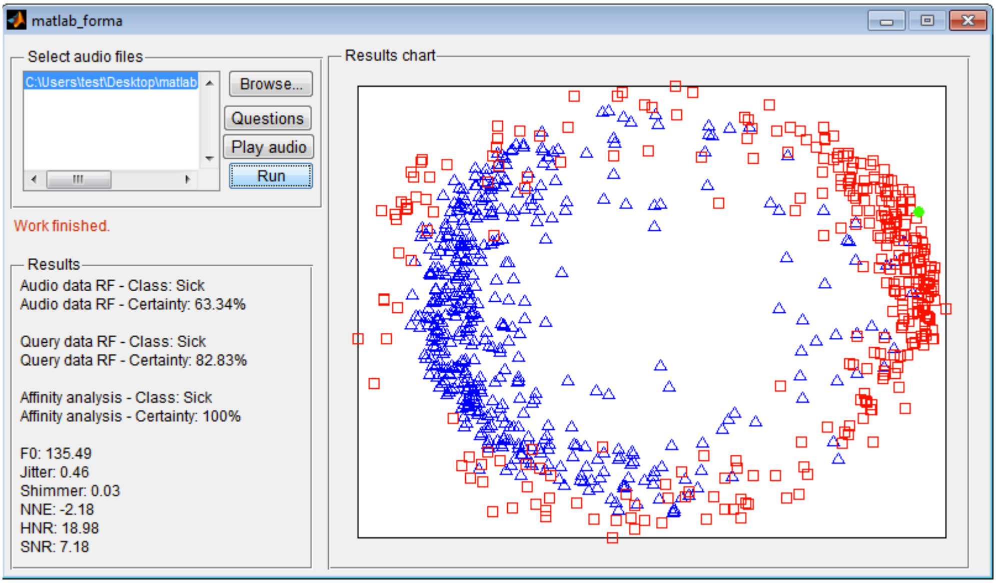


Figure 5.1 Architecture of experimental system

and backend. The latter one contains all the components for data analysis and software usage logic. *Feature extraction component* is responsible for feature extraction from the audio file. The voice recording file for this component is provided by the user interacting with GUI. This component contains multiple algorithms to extract audio features for two datasets: first containing 14 types of features and second containing the same features as provided by otolaryngologists using Dr. Speech.

Different classification algorithms are applied depending on what data is provided. *RF Classifier* for combined classification of voice and query data is executed only when both, voice recording and query data, are provided. This allows the avoidance of missing data errors and reduces time required for calculations. The *Association rules classifier* component is responsible for classification of questionnaire data and is executed only if query data is provided. *Decision tree classifier* is responsible for classification of the smaller audio features set and is executed only if voice recording is provided. *Data dependent RF classifier* component consists our newly proposed algorithm for voice data classification. It is executed every time when voice recording is provided and submits its results to the GUI. *t-SNE mapper* component is responsible for the employment of *t*-SNE algorithm to map training data and provided data to 2D map.

Graphical user interface was designed for the users interaction with the software and it is responsible only for the audio file selection, questionnaire answers imputation and representation of the results. Sample screenshots are provided in Figure 5.2 and Figure 5.3. As we can see, it consists of 5 different components. *Audio file loader* allows user to easily select voice recording file



Work finished.

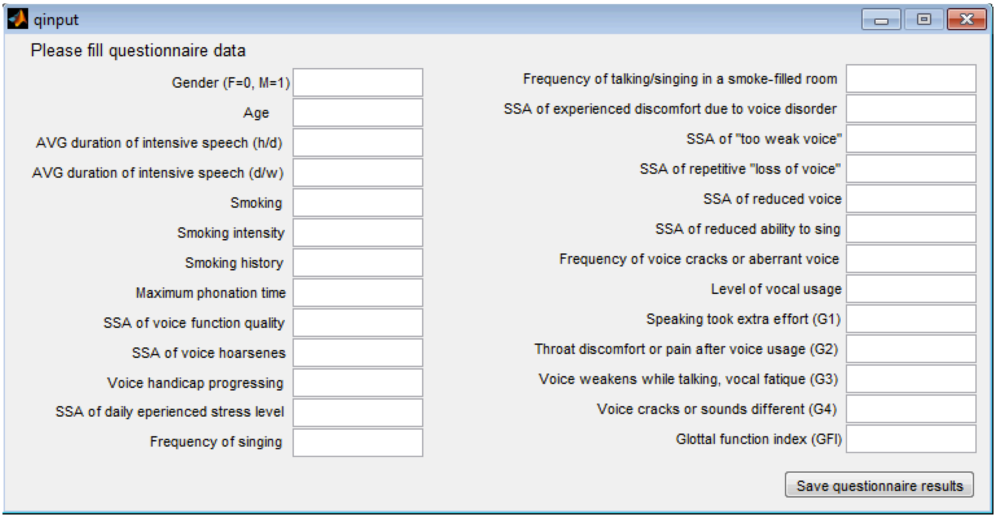
Audio data RF - Class: Sick
Audio data RF - Certainty: 63.34%

Query data RF - Class: Sick
Query data RF - Certainty: 82.83%

Affinity analysis - Class: Sick
Affinity analysis - Certainty: 100%

F0: 135.49
Jitter: 0.46
Shimmer: 0.03
NNE: -2.18
HNR: 18.98
SNR: 7.18

Figure 5.2 Screenshot of the main software window containing three UI parts: audio recording selector, 2D map and textual results viewer



Please fill questionnaire data

Gender (F=0, M=1)
Age
AVG duration of intensive speech (h/d)
AVG duration of intensive speech (d/w)
Smoking
Smoking intensity
Smoking history
Maximum phonation time
SSA of voice function quality
SSA of voice hoarsenes
Voice handicap progressing
SSA of daily eperienced stress level
Frequency of singing

Frequency of talking/singing in a smoke-filled room
SSA of experienced discomfort due to voice disorder
SSA of "too weak voice"
SSA of repetitive "loss of voice"
SSA of reduced voice
SSA of reduced ability to sing
Frequency of voice cracks or aberrant voice
Level of vocal usage
Speaking took extra effort (G1)
Throat discomfort or pain after voice usage (G2)
Voice weakens while talking, vocal fatique (G3)
Voice cracks or sounds different (G4)
Glottal function index (GFI)

Figure 5.3 Screenshot of software window, where user (doctor) has to enter answers to questionnaire

from computer disk. *Questionnaire input table* opens in a new window, where text fields are provided to enter query data. *Decision tree* component is responsible for the visualization of the constructed DT, however it is visible only when voice recording is provided. Visualization of 2D *t*-SNE map is done by the *2D t-SNE map* component. This requires voice recording to be provided, because 2D map is generated from proximity matrix, obtained during audio features classification. *Results block* contains textual information about the classification results. There are provided results from each algorithm, as well as certainty.

5.3. Evaluation of detection

Classification accuracy was evaluated using voice and query data from 48 unseen subjects (9 healthy and 39 pathological). Classification certainties were used as scores to evaluate how good the detection is. For detection using voice data, scores for each observation were given by the probability of the dominant class at the terminal node which the observation is assigned to. When query data is used for classification, a score is calculated by the Equation 4.30. To better evaluate accuracy improvement, the classifier was tested with voice and query data separately, as well as in decision-level fusion. Classification goodness analysis results are visible in Figure 5.4 and Figure 5.5, where DET and ROC curves can be seen together with EER and AUC values.

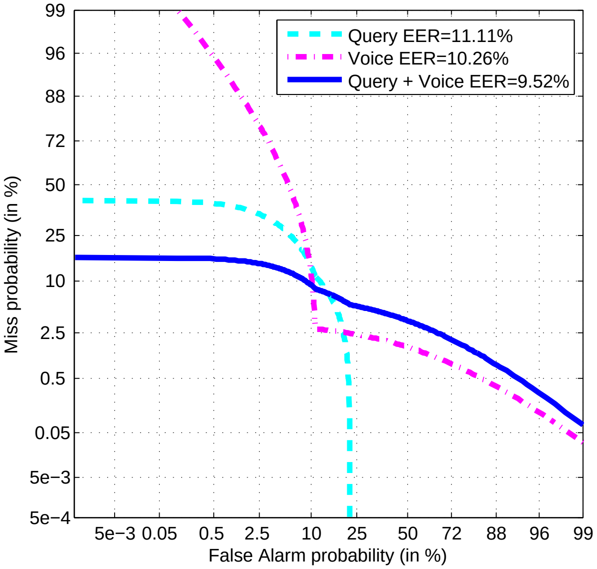


Figure 5.4 DET curves and EER for unseen voice and query data

Achieved EER values are very encouraging despite the simplicity of the techniques used. Using voice and query data separately, EER of 11.11% and 10.26% was achieved using association rules and decision tree respectively. Combination (via weighted averaging) of results, obtained from affinity rules and decision tree, resulted in EER of 9.52%. As expected, combination of two data modalities improved classification accuracy. As it can be seen in DET and ROC curves, combined classifier has the lowest false alarm probability (highest specificity) near the low miss probability (high sensitivity) mode of operation. Last property shows significant benefit for initial screening in preventive healthcare.

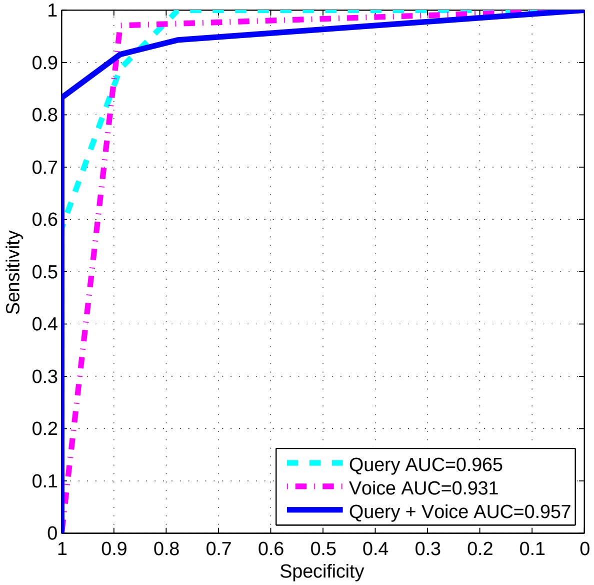


Figure 5.5 ROC curves and AUC for unseen voice and query data

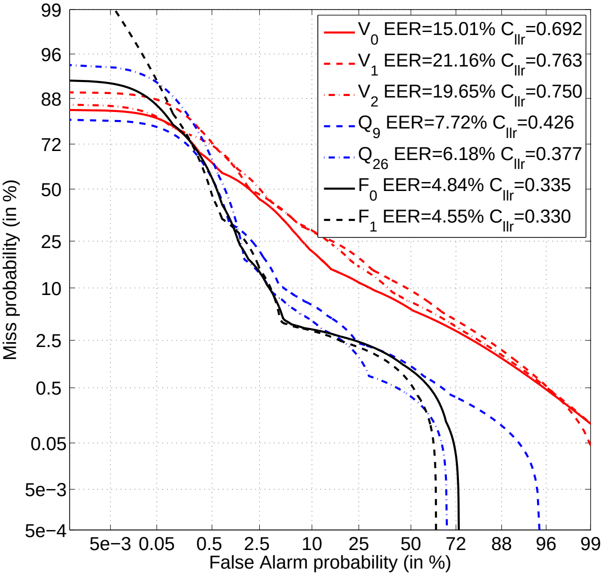


Figure 5.6 DET curve of classification performance of RF on OOB data. Systems compared: audio modalities (V_0, V_1, V_2), query modalities (Q_9, Q_{26}), all modalities fusion after listwise deletion (F_0), fusion of 3 modalities (V_0, Q_9, Q_{26}) after iterative model-based imputation (F_1).

Classification results, achieved through experiments with association rules and different data sets of voice and query data, provide other interesting findings. As can be seen in Figure 5.6, query data performs consistently better than

Table 5.2 Classification performance evaluation for various decision-level fusions

Type of imputation	2 modalities		3 modalities		4 modalities		5 modalities	
	EER	C_{llr}	EER	C_{llr}	EER	C_{llr}	EER	C_{llr}
Listwise deletion	5.37	0.3457	5.07	0.3385	4.75	0.3374	4.84	0.3355
Median	5.58	0.3479	4.99	0.3353	4.84	0.3330	4.82	0.3326
Regression	5.97	0.3468	5.15	0.3351	5.02	0.3346	4.88	0.3347
PCA-based	5.38	0.3424	4.76	0.3314	4.76	0.3311	4.91	0.3319
SVD-based	5.37	0.3447	5.09	0.3377	5.03	0.3359	5.07	0.3341
Model-based	4.99	0.3367	4.55	0.3303	4.52	0.3313	4.71	0.3317

Table 5.3 Details for the best fusions - RF parameter q and modalities which were used

Type of imputation	2 modalities	3 modalities
Listwise deletion	2 of $\{Q_9, Q_{26}\}$	1 of $\{V_2, Q_9, Q_{26}\}$
Median	1 of $\{V_0, Q_{26}\}$	1 of $\{V_2, Q_9, Q_{26}\}$
Regression	2 of $\{Q_9, Q_{26}\}$	3 of $\{V_0, Q_9, Q_{26}\}$
PCA-based	2 of $\{Q_9, Q_{26}\}$	3 of $\{V_0, Q_9, Q_{26}\}$
SVD-based	1 of $\{V_0, Q_{26}\}$	1 of $\{V_2, Q_9, Q_{26}\}$
Model-based	2 of $\{Q_9, Q_{26}\}$	3 of $\{V_0, Q_9, Q_{26}\}$

Type of imputation	4 modalities	5 modalities
Listwise deletion	3 of $\{V_0, V_1, Q_9, Q_{26}\}$	2 of $\{V_0, V_1, V_2, Q_9, Q_{26}\}$
Median	3 of $\{V_0, V_1, Q_9, Q_{26}\}$	2 of $\{V_0, V_1, V_2, Q_9, Q_{26}\}$
Regression	3 of $\{V_0, V_1, Q_9, Q_{26}\}$	2 of $\{V_0, V_1, V_2, Q_9, Q_{26}\}$
PCA-based	4 of $\{V_0, V_2, Q_9, Q_{26}\}$	4 of $\{V_0, V_1, V_2, Q_9, Q_{26}\}$
SVD-based	3 of $\{V_0, V_1, Q_9, Q_{26}\}$	2 of $\{V_0, V_1, V_2, Q_9, Q_{26}\}$
Model-based	4 of $\{V_0, V_2, Q_9, Q_{26}\}$	3 of $\{V_0, V_1, V_2, Q_9, Q_{26}\}$

voice data. However, the best overall result was achieved by using fusion of voice and query data. Assessing by C_{llr} values provided in Table 5.2, for fusion of 2 modalities only imputation of median and imputation by regression were a little inferior than the listwise deletion. When fusion of more than 2 modalities were used, all types of imputation outperformed listwise deletion. Best classification results were achieved by using PCA-based and model-based imputation techniques, where fusion of less than all available modalities was used.

More detailed information about the best combination of RF parameter q and selected modalities is provided in Table 5.3. A longer version of questionnaire (Q_{26}) proved to be more important. This is quite obvious, since in Figure 5.6 it can be identified as the best single modality. The second-best combination is the short version of questionnaire (Q_9) and raw recording (V_0) fusion. For PCA-based and model-based imputations the optimal number of components was found to be 2. The best fusions were found by setting RF parameter q to the maximum, which conforms as unpruned bagging.

5.4. Data dependent random forest

Data dependent random forest was our newly proposed data fusion and classification technique. During the experiments, multiple fusion techniques were tested to compare them with our method and effectiveness of each feature set was evaluated. As one of the results of these experiments, acoustic and contact microphones were evaluated, in order to determine which one of them or combination of both can help to achieve higher classification accuracy.

5.4.1. Comparison of acoustic and contact microphones

For classification accuracy, a comparison using data from both types of microphones, detection error tradeoff (DET) curve was used. It was obtained by using ROC convex hull approximation BOSARIS toolkit [22], and is shown in Figure 5.7.

In this study, results from meta RF were used to generate DET curve, as this was the most accurate fusion scheme for contact microphones. EER stands for equal error rate for both classes (normal and pathological). The difference between EERs of different microphones is not high, but it would be higher if decision level fusion by meta RF had been used, which was the most accurate for the acoustic microphone data.

As can be seen in Figure 5.7, acoustic microphone is superior to the contact one. The only region where contact microphone provides better results is the region of high miss probabilities. As it is mentioned in [157], fusion of information from acoustic and contact microphones for the data sets used in this study is not effective. Table 5.4 provides out of bag (OOB) data classification accuracy using different fusion schemes, which shows that even if small classification accuracy improvements are visible for some schemes, the difference in accuracy is not statistically significant. According to these results, a decision was made to use only acoustic microphone for further voice recordings.

All classification results presented in this section are averages of eight runs. Each time different bootstrap samples were selected from OOB data with different sets of randomly selected variables. In the database used in this research, each subject has three audio recordings, so OOB data set was generated in such way, that none or all three samples of one subject were in it.

5.4.2. Different feature sets

Usage of different feature sets provides different classification results. According to the data given in Table 5.5, the acoustic microphone was superior to the contact one with respect to 13 different feature sets out of 14, where for most of the feature sets, the difference was statistically significant. It is worth noting, that a single forest was created for each feature set. Numbers of total features and selected features of each set are also provided in Table 5.5, which

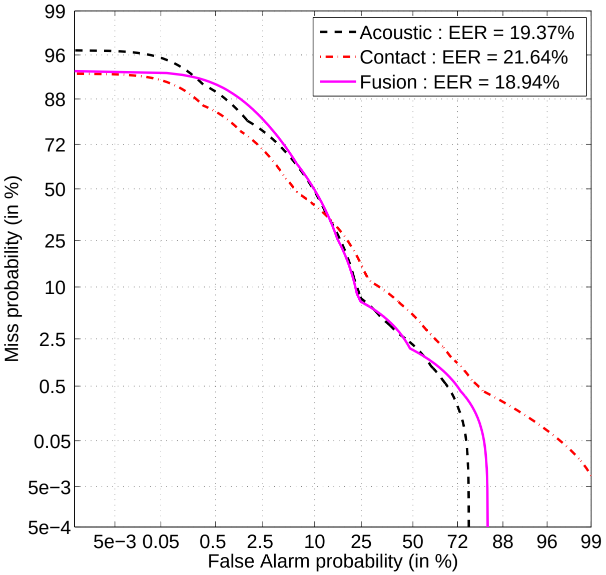


Figure 5.7 DET curves of different types of microphones. EER - equal error rate

Table 5.4 Data classification accuracy obtained by using different fusion schemes (OOB, %)

Fusion microphone	Acoustic	Contact	Combined
Feature level			
All features	78.67	73.81	73.87
Selected features	83.18	81.18	83.21
Decision level, forest voting			
All forests	79.93	75.46	79.06
All weighted forests	82.28	78.44	80.67
Selected forests	84.63	79.93	85.01
Decision level, averaging probabilities			
All forests	80.17	73.73	79.06
Selected forests	83.52	77.32	83.64
Decision level, by meta RF			
	84.77	81.38	85.66
Decision level, data dependent			
All forests	80.55	73.98	79.43
All forests, weighted trees	84.01	78.56	83.15
Selected forests, weighted trees	86.37	80.79	86.62

Table 5.5 Classification accuracy (OOB) and number of initial/selected features in different data sets

#	Features	# All	# Selected Acoustic	(%)	# Selected Contact	(%)
1	Perturbation	24	6	77.26	14	76.10
2	Frequency	100	13	71.30	12	70.07
3	Mel-frequency	35	7	69.16	8	70.03
4	Cepstral energy	100	27	72.51	20	69.95
5	Mel-coefficients	35	10	70.14	13	67.87
6	Autocorellation	80	13	65.09	10	64.08
7	HNR-spectral	11	8	62.17	10	59.70
8	HNR-cepstral	11	6	64.44	3	60.70
9	LP-coefficients	77	12	76.66	25	64.78
10	LPCT-coefficients	77	13	78.70	7	64.13
11	Signal shape	128	50	70.62	10	68.70
12	Reflection-coefficients	24	9	76.60	10	69.56
13	Tract irregularity	71	11	80.36	21	69.44
14	PLPC-coefficients	154	11	81.20	29	76.35
		Average	14.0	72.59	13.7	67.96

shows how many features were selected to get the best classification result on the OOB data (the same as in [157]).

Classification accuracy improved with all 14 feature sets and both types of microphones when feature selection was used. In some cases, higher classification accuracy was achieved by using less than 15% of features from a single feature set. After feature selection, 196 features were left for acoustic microphone and 192 for contact microphone, which is 14 and 13.7 features on average respectively. According to the classification results, Perceptual Linear Predictive Cepstral Coefficients (PLPCC) can be presented as the best feature set for both types of microphones. In acoustic microphone case, classification accuracy was significantly increased using tract irregularity features built upon the reflection-coefficients. Perturbation measurements were the second-best feature set for the contact microphone.

5.4.3. Different fusion schemes

Several fusion schemes were analyzed in this research in order to improve classification accuracy and compare acoustic and contact microphones. In feature level fusion scenario, RFs were built using all features (927 in total for each type of microphone and 1854 in the combined case) and features after selection. The results in Table 5.4 show that when all features are used, classification accuracy is lower than when using a single feature set. However, after feature selection, accuracy was significantly improved for both acoustic and contact microphone data, and it was achieved higher than the best accuracy obtained using single data set.

Another fusion scheme which was tested was a decision level-based fusion. In a forest voting scenario, the same pattern of behavior was observed as in the data fusion case. Random forest weighting, according to classification accuracy of single RF, improved overall classification accuracy. However, usage of only selected forests improved classification accuracy even more. The least effective decision level fusion scheme was by averaging class a posteriori probabilities, while fusion scheme based on meta RF provided higher accuracy than forest voting. Data dependent decision level fusion provided the best classification results. Three different variations (all presented in Table 5.4) were tested, where *selected forests with weighted trees* proved to be the most accurate. Only selected single RFs are used to create data dependent RF which uses weighted voting to obtain the decision.

5.5. Association rules

Association rules were filtered, and only relatively significant ones were left. For *healthy* class, only the rules with $support > 0.28$ and $confidence > 0.9$ were selected, while for *pathological* class limits were different: $support > 0.16$ and $confidence > 0.9$. Gravity by absolute frequency of each specific response is provided in Figure 5.8. 'Healthy rules' turned out to be more complex, however, more redundant, where the most persistent components had the highest gravity. 'Pathological rules' were found to be shorter and their components slightly more diverse. Detailed results of the affinity analysis are provided in Table 5.6 and Table 5.7. As it can be seen, 11 rules for each classification class (*healthy* and *pathological*) were created. Rule names mapping with questionnaire questions (see Table 3.2) are done like this (marker name - question number): Gender - 1, U - 4, C - 6, Y - 7, MFT - 8, VAS - 9, H - 11, D - 15, W - 16, R - 18, S - 19, X - 20, L - 21, G_0 - 22-25.

One of the benefits of affinity analysis performed in this work, is that it highlighted 17 most important questions from our questionnaire. This indicates that some questions provide very little useful information and can be eliminated. According to the results, the most useful parameter is GFI, which is a combination of questions 22 to 25. If the answer to any of these questions is "no problem", the patient does not smoke, and voice handicap progression is marked as 2, it is an indicator that a patient is healthy. On the other hand, any non-zero answer to GFI questions and short duration of MFT is a strong indicator of pathology.

5.5.1. Variable importance

As mentioned before, association rules helped to determine the most important variables from our data sets. Variable importance values were determined from RF and they are shown in Figure 5.9 for query data, and Figure 5.10 for voice data. As it can be seen in Figure 5.9, 10 out of 26 (38.5%) of

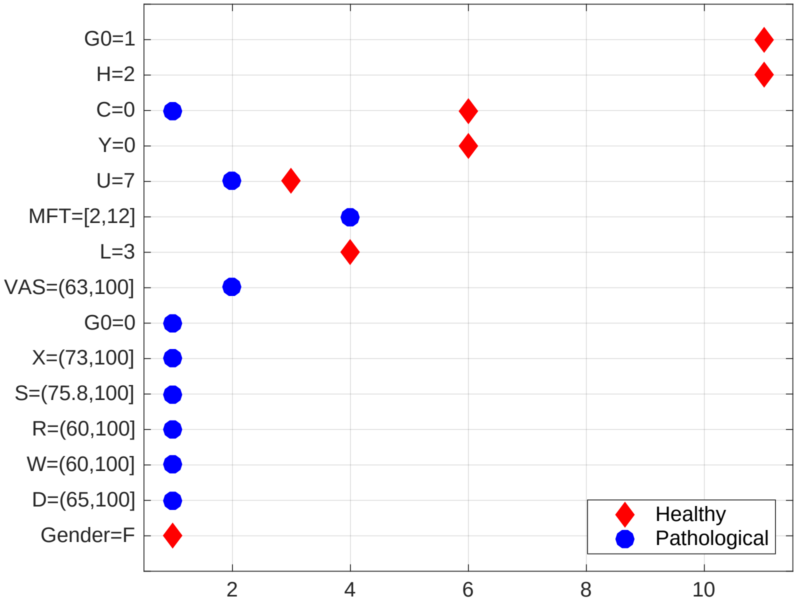


Figure 5.8 Rule coverage for mined antecedent items. Absolute frequency, as the number of rules the specific response participates in, is marked with "Healthy" and "Pathological" markers

Table 5.6 Association rules (*healthy* subject) used in this study and extracted in [151] from the same query database as in this study

#	Antecedent of the rule	Support	Confidence	Lift
1	H=2, G_0=1, Gender = F	0.302	0.928	1.691
2	H=2, G_0=1, Y=0, L=3	0.292	0.926	1.687
3	H=2, G_0=1, C=0, Y=0, L=3	0.289	0.925	1.685
4	H=2, G_0=1, C=0, L=3	0.297	0.922	1.680
5	H=2, G_0=1, Y=0	0.341	0.919	1.674
6	H=2, G_0=1, C=0, Y=0	0.337	0.918	1.673
7	H=2, G_0=1, C=0	0.351	0.917	1.671
8	H=2, G_0=1, Y=0, U=7	0.285	0.909	1.657
9	H=2, G_0=1, C=0, Y=0, U=7	0.284	0.909	1.656
10	H=2, G_0=1, L=3	0.381	0.908	1.655
11	H=2, G_0=1, C=0, U=7	0.294	0.907	1.653

the query variables can increase RF accuracy by more than 1%: questions # 2, 9, 10, 15, 16, 18, 20, 22, 25 and 26. The results that are shown in Figure 5.10 indicates that 61.5% (16 out of 26) of low-level descriptors (LLDs) can increase RF accuracy by more than 1%: MFCCs, LSP frequencies, pitch, pitch envelope, ZCR, probability of voicing loudness.

From the query data set the most important questions for RF are based on subjective self-assessment: experienced discomfort due to voice disorder (#15),

Table 5.7 Association rules (*pathological* subject) used in this study and extracted in [151] from the same query database as in this study

#	Antecedent of the rule	Support	Confidence	Lift
1	MFT=[2,12], G_0=0	0.168	1	2.216
2	R=(60,100]	0.191	0.983	2.177
3	MFT=[2,12], C=0	0.164	0.980	2.171
4	MFT=[2,12]	0.267	0.975	2.161
5	D=(65,100]	0.190	0.974	2.158
6	MFT=[2,12], U=7	0.211	0.969	2.147
7	X=(73,100]	0.191	0.966	2.141
8	VAS=(63,100]	0.210	0.947	2.098
9	VAS=(63,100], U=7	0.169	0.944	2.091
10	S=(75.8, 100]	0.185	0.940	2.083
11	W=(60,100]	0.188	0.933	2.068

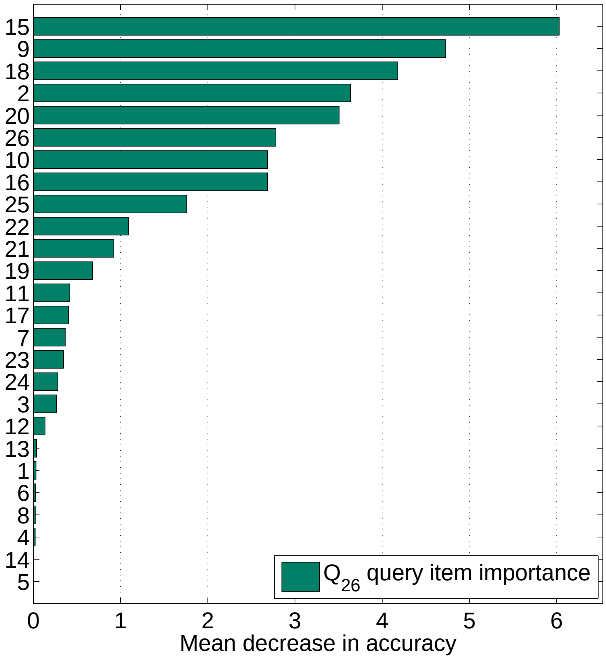


Figure 5.9 RF permutation-based variable importance for the query data

voice function quality on visual analogue scale (#9) and strength of reduced voice effect (#18). With respect to the voice data set, the most important audio features for RF are MFCCs.

When comparing discriminatory information in LLD (original) with its delta (velocity) for cases when increase of accuracy is higher than 1%, in 9 cases LLD and in 7 cases delta is more important. Original information is more discriminatory in first 2 MFCCs, LSP frequencies, probability of voicing and

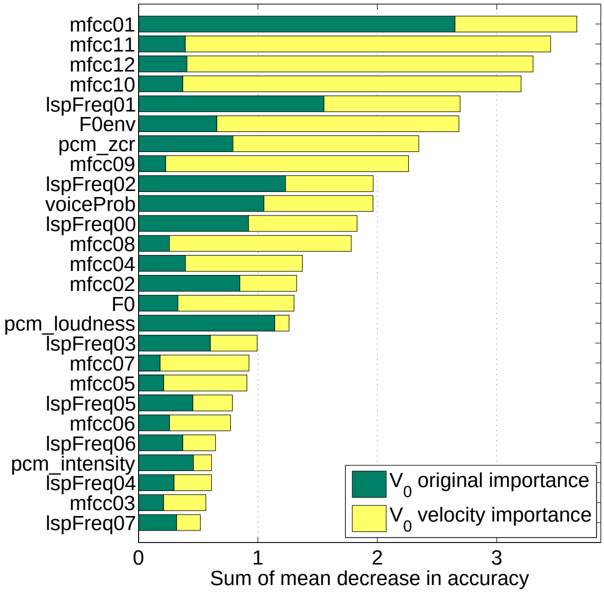


Figure 5.10 RF permutation-based variable importance for the voice data

loudness. The velocity information is more discriminatory in higher MFCCs, pitch, pitch envelope and ZCR.

5.6. Exploring data and decisions

To visualize data in a 2D map, a similarity/proximity matrix was created by averaging and collecting data similarity values, calculated using equations 4.43 and 4.44. As a mapper for 2D space, *t*-SNE algorithm was used, with a perplexity parameter set to 50, which was chosen empirically. The algorithm converged after 320 iterations. Resulting visualization from *t*-SNE algorithm is shown in Figure 5.11, where 'triangles' represent healthy cases and 'squares' represent pathological ones. For tracking purposes, two cases were selected from the 'healthy' class, which are marked as light-filled and dark-filled circles.

2D map created by *t*-SNE algorithm eases selection and more detailed comparison of different cases. As is seen in Figure 5.11, a map enables the detection of misclassified, incorrectly labelled data, or data mapped very closely, but coming from different classes. One example of such case is represented by light-filled circle in Figure 5.11. This observation was assigned to 'healthy' class by classifier, but was mapped to area, mostly occupied by observations from 'pathological' class. Observations like this are candidates for deeper analysis.

Figure 5.12 represents the proximity matrix mapped to a 2D space by a *t*-SNE algorithm, which was created by meta RF using voice database. All three voice recordings of each patient were used. It is expected that all three recordings would be close to each other in the 2D map, however, in Figure 5.12 it can be seen that it is not always true. All three recordings of four subjects

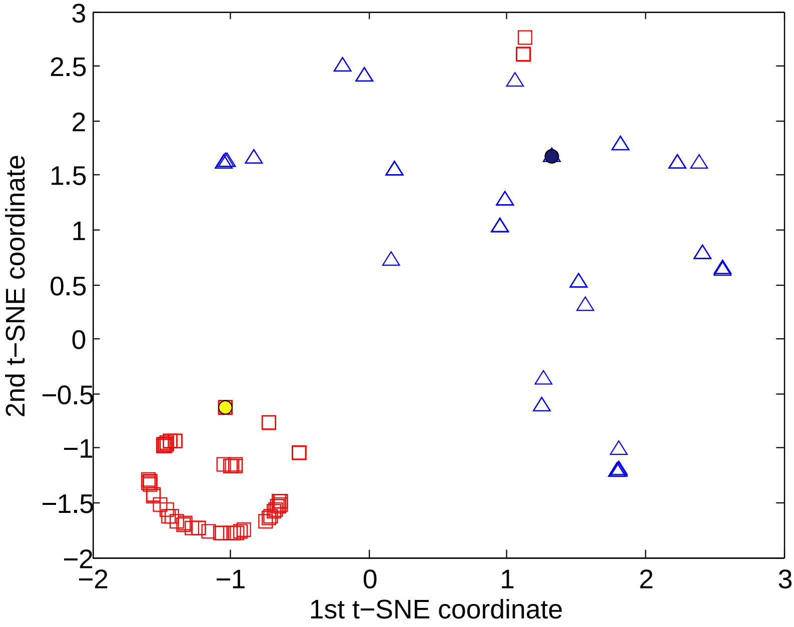


Figure 5.11 2D visualization of proximity matrix by the *t*-SNE algorithm (Voice and query data fusion)

are shown in distinct color and are encircled to visualize that. Most likely this is caused by the fact that meta RF operates with posteriori probabilities, so pairwise similarities between observations are not utilized directly.

To study classification confidence, we also created a 2D map of generalized proximity matrix, where we labelled data with a predicted class (Figure 5.13). The size of the marker indicates confidence in decision - larger marker means higher ($p^{max1} > 0.97$) confidence and smaller marker means lower ($p^{max1} < 0.65$) confidence. By comparing Figures 5.13 and 5.14, classification errors in the upper left part of the map can be easily indicated. It is worth noting that sometimes even erroneous decisions can be made with high confidence.

In cases where deeper analysis is required, class-conditional probability density functions of audio and query parameters are a great help. Examples of such functions, estimated using the Epanechnikov kernel smoothing are provided in Figure 5.15 and Figure 5.16. As it can be seen from probability density functions, dark-filled circle from Figure 5.11 represents a correctly classified 'healthy' subject.

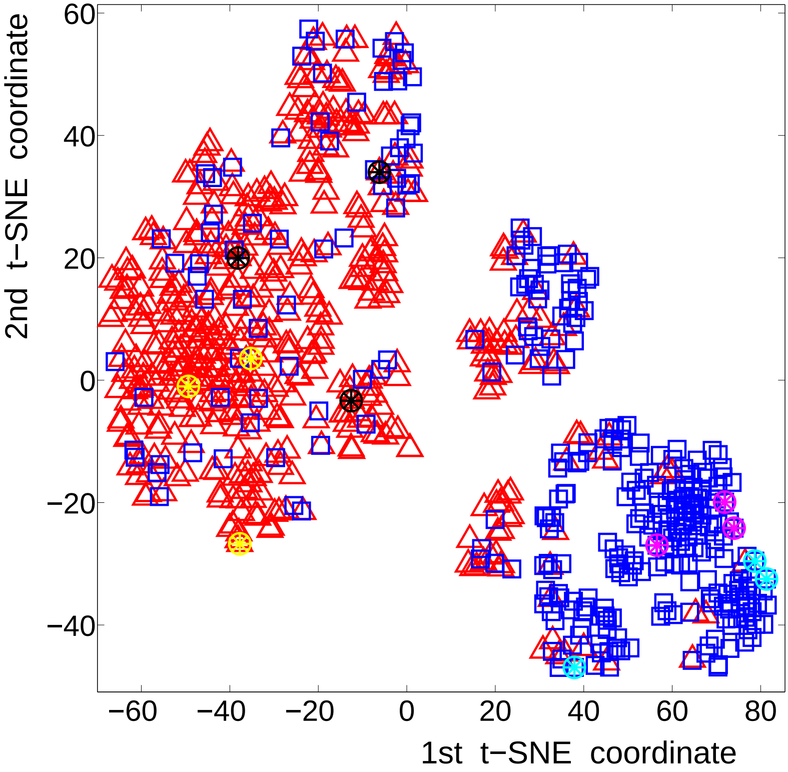


Figure 5.12 The proximity matrix created by meta RF from voice data set and mapped to a 2D space

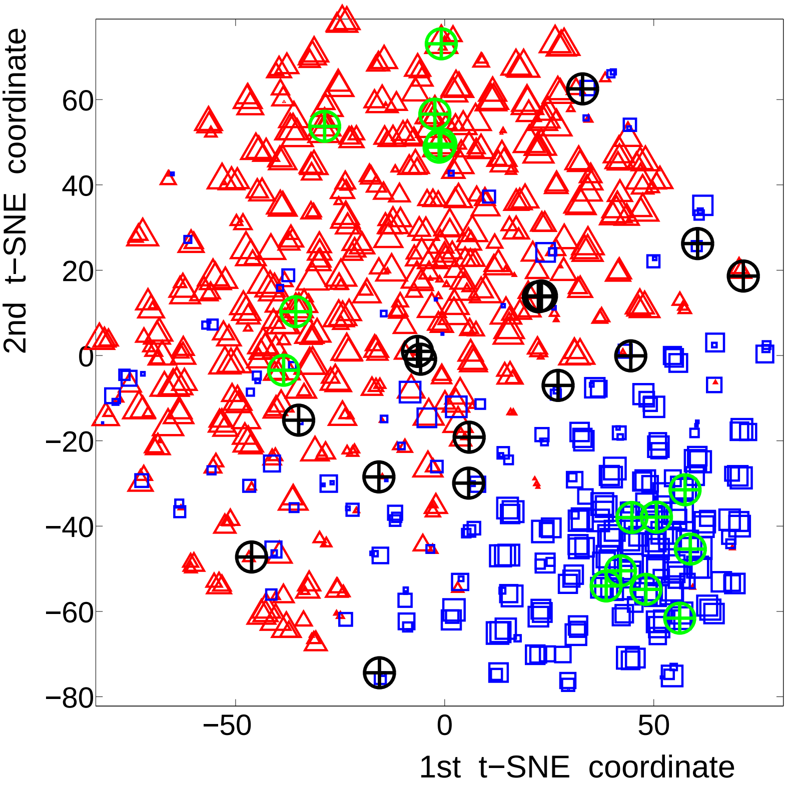


Figure 5.13 The generalized proximity matrix mapped to 2D space. The size of the marker reflects prediction confidence (larger marker - higher confidence). Black circles indicate predictions of low confidence and green circles - predictions of high confidence.

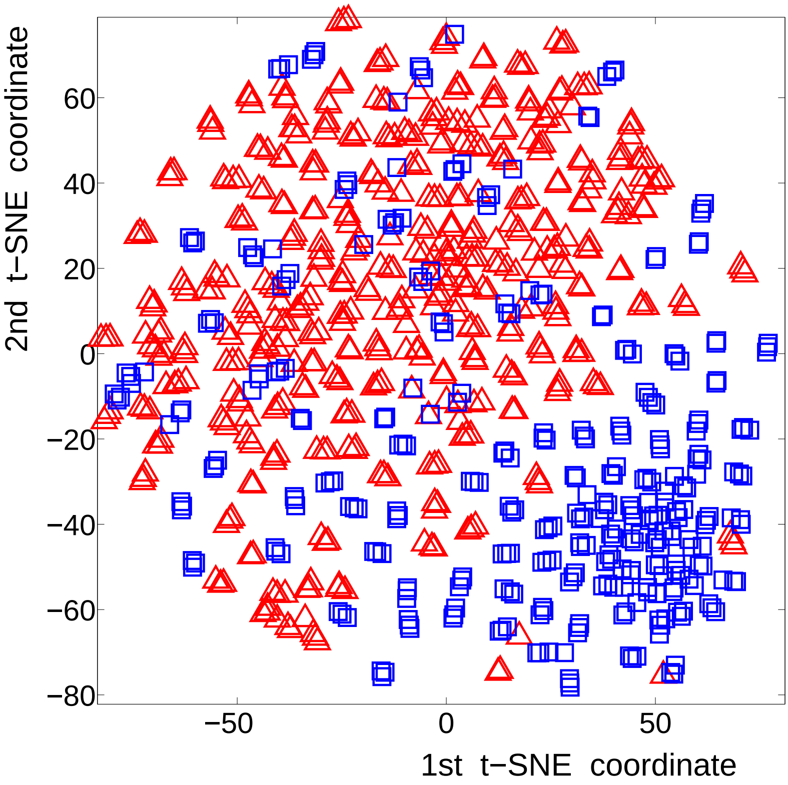


Figure 5.14 The generalized proximity matrix mapped to 2D space where marker color indicates healthy or pathological class

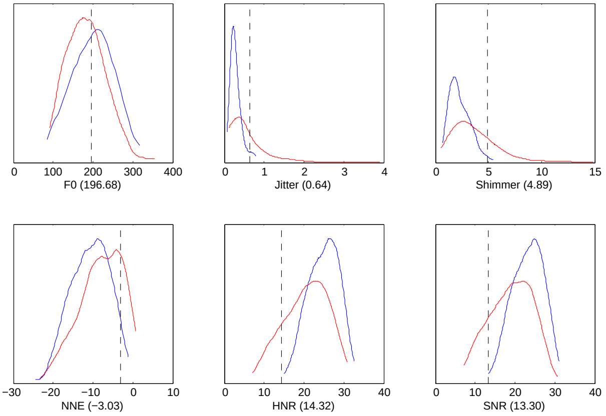


Figure 5.15 Probability density functions of various audio parameters for *healthy* (*blue*) and *pathological* (*red*) cases, estimated using the Epanechnikov kernel smoothing method. Dashed line and numbers in parenthesis correspond to an exact parameter value for a selected subject.

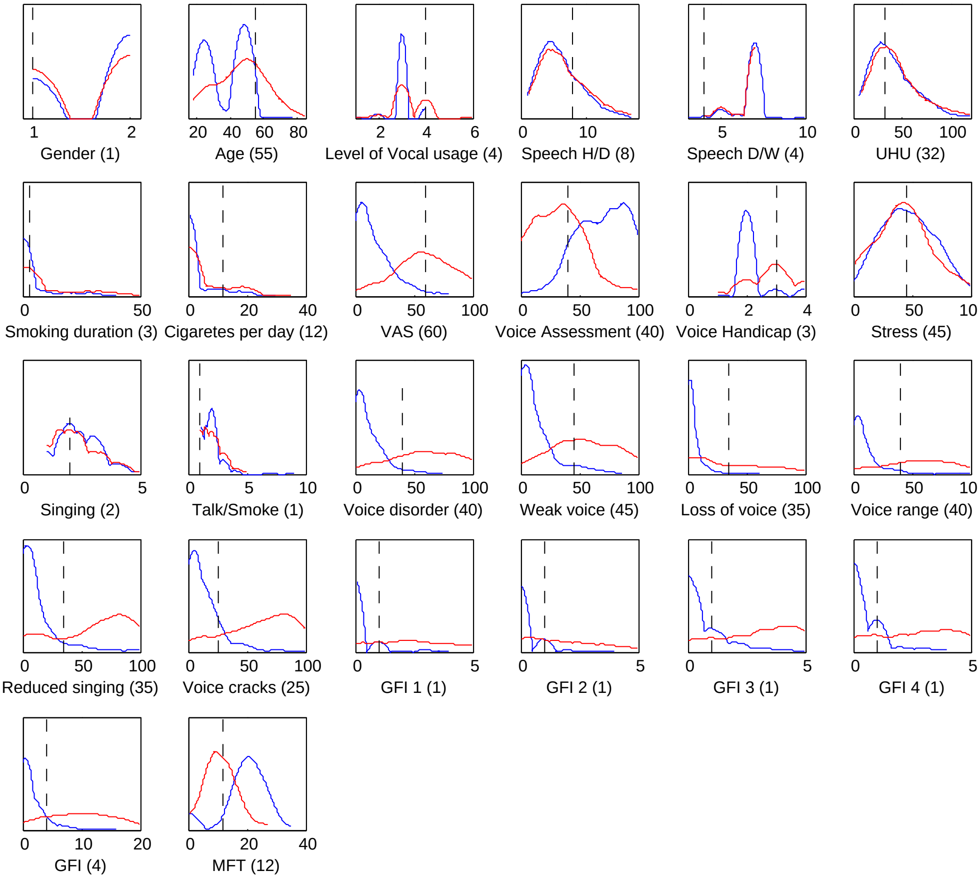


Figure 5.16 Probability density functions of various query data parameters for *healthy* (*blue*) and *pathological* (*red*) cases, estimated using the Epanechnikov kernel smoothing method. Dashed line and numbers in parenthesis correspond to an exact parameter value for a selected subject.

5.7. Threats to Validity

The experiments in this research were carried out in order to evaluate the capabilities of proposed techniques to solve the problems they were developed for. Threats to validity of the research must be assessed to determine validity threats and potential risks of the design and execution of proposed methods. Validity threats in this research were evaluated by separating them into several classes, as suggested in [15]. Assessment is presented in the Tables 5.8, 5.9, 5.10.

Table 5.8 Construct Validity Threats

Threat	Management
Improperly chosen evaluation measures	Performance and efficiency evaluation was not in the scope of this research. The selection of metrics was based directly on the problem the solution is designed to solve.
Poorly chosen effectiveness measures	To evaluate classification accuracy of each method used the out-of-bag (OOB) error was calculated. How good classification algorithm performs was also evaluated by calculating DET and ROC curves. Those provided False alarm probability versus Miss probability (DET) and Specificity versus Sensitivity (ROC).
Bugs in implementation	The solution was developed by a person with proper experience, separate parts of the solution were used and tested in other experiments, therefore risk of bugs in implementation is minimal.

Table 5.9 Internal Validity Threats

Threat	Management
Poor parameter settings	The experiment was described in detail with all parameter values provided, which increases the reproducibility.
Lack of discussion on instrumentation	The implementation was described and source code of the solution with audio recordings was provided in the DVD.
Lack of clear of data collection tools and procedures	Procedure of data collection was described in detail with data samples provided in the appendix. Collected data covered as many real-life cases as possible.
Instrumentation	Microphone for audio recording was selected by evaluating results of other researchers, so that device with high frequency sensitivity was selected, to capture all possible audio data.

Table 5.10 External Validity Threats

Threat	Management
Selection biases	Data for the experiments was selected to cover as many scenarios as possible. Data was collected from patients of different age, sex and voice pathology (including healthy).
Non-comparable experiment	The proposed techniques were described in detail; therefore experiments can be done to compare the approach with other research. Data used in this research is available only by directly contacting the source, however other publicly available data can be used for the experiments.
Multiple-treatment interference	There were 3 recordings done for each patient, which might raise concern that process of previous recording might influence the quality of the following one. However, patients were given a decent amount of time to rest their voices before proceeding.
Lack of evaluations for instances of growing size and complexity	Data for this research was provided by the Department of Otolaryngology from Lithuanian University of Health Sciences. It contained multiple data sets with a wide variety of properties. However, scalability evaluation was not the focus of this research.

5.8. Ease of use

The developed software was evaluated by otolaryngology specialists, who were using and testing it in their everyday work. Special forms, containing 7 evaluation questions from ISO-9241 standard, were printed and provided for our users to evaluate the software. Computer program was evaluated by three otolaryngology specialists, because the specificity of our tool limits the number of proper users. Collected paper forms were analyzed, and final evaluation according to the aforementioned ISO-9241 standard part 11 was obtained by generalization of this data. Combined evaluation is provided in Table 5.11.

Because of the specificity of the analyzed area, the software is very peculiar and did not obtain the highest scores. All of the users who were testing and evaluating the software, marked *suitability for the task*, *self-descriptiveness*, *controllability* and *conformity with user expectations* very high. This means that our created computer software is easy to use, requires no additional training, and all program functionality performs as is expected. The best possible evaluation of *conformity with user expectations* is probably the most important rating, which shows that program usability is very high. *Error tolerance* has received a bit lower evaluation, but not because it was not working well - some error explanations were mentioned as "somewhat unclear". However, the assessment of the error tolerance does not have a significant influence on the overall evaluation of the software usability.

Suitability for individualization and *suitability for learning* were evaluated worse than other items. However, it was expected, because there is no individualization, and no help file or online resource is available. It seems that in this kind of work, software individualization is not necessary and self-descriptiveness compensates the need for help.

Despite lower evaluation of a couple of criteria, the overall rating of software is quite high. None of the items were evaluated as wrong, invalid, incorrect or unfinished, so the software can be named as suitable for the task it was made for.

Table 5.11 Software evaluation by 7 principles of the ISO 9241 standard part 11. General evaluation from filled out forms.

ISO-9241 principle	**Evaluation**
Suitability for the task	Software has all the features necessary to achieve the task effectively and efficiently.
Self-descriptiveness	All functions are understood in an intuitive way and no additional usage information is necessary.
Controllability	Software is implemented in the way that a user can do only one task at a time, and steps order is controlled by disabling features that are not available at the current moment.
Conformity with user expectations	Software is designed with doctors as users in mind, so UI is simple and easily understandable.
Error tolerance	The software is equipped with an error reporting system, which indicates usage faults and shows error messages to the user. However, some of the messages are not very clear for first-time users.
Suitability for individualization	Personal configuration is not allowed, however UI is very minimal and does not require that.
Suitability for learning	The software is suitable for learning. User interface is self-descriptive, easy to use and requires no help. Provided results are useful for educational purposes, clinical decision-support allows comfortable comparison of various subjects. Unfortunately, there is no help file and online resources to read about software functionality.

6. DISCUSSION

There are around 200,000 annual deaths worldwide only from larynx-related cancer, and laryngeal disorders affect about 5-6% of human population. Preventive health care techniques are required, which would be easily accessible and would not require expensive medical equipment. Some studies show that classification of voice data can achieve classification accuracy as high as 100% [62, 121]. Query data analysis results are also promising, because they consistently outperform voice data based detection [151]. In this research we analyze voice and query data classification individually and in decision-level fusion.

Many researchers test different algorithms trying to improve voice data classification accuracy. Depending on the algorithm, amount of voice recordings and features used, classification results differ significantly - from 72.53% achieved by [8] up to 100% achieved by [62, 121]. We proposed a new data dependent random forest based technique for combination and classification of multiple data sets, which led to the achievement of 86.37% classification accuracy. This technique outperformed our other tested data-level and decision-level fusion techniques with 1.6% - 3.19% of accuracy improvement, however did not manage to achieve as high results as mentioned by some other researchers. Such a difference in classification accuracy may occur due to many reasons, such as differences between classification algorithms, features extracted and voice recording databases. As seen in the related work analysis, many voice recordings databases used by other researchers are not well balanced (the number of pathological and healthy patients differs several times), which might increase average classification accuracy, while accuracy for separate classes can be much lower. Our results are similar to the ones reported by [152, 120, 8], and prove that non-invasive voice data can be successfully used for pathology detection with relatively high accuracy.

Our findings also show that voice pathology detection can be successfully performed using only query data. We introduced new association rules-based classification algorithm, solely designed for query data, which achieved an EER of only 11.11%. This algorithm outperforms many others reviewed in related work analysis, however improvements are needed to achieve results as high as 96.48% achieved by [147]. This classification technique can be very helpful in preventive health care, as no invasive or medical equipment is required, and classification can be performed even remotely. Only a few attempts have been made previously to use query data for voice pathology screening [159, 156, 155, 13, 158, 147, 84], verifying the validity of our study, as well as the limited possibilities of our result comparison with the findings of different researchers. Association rules extraction provided an additional benefit, by revealing the 17 most important questions from our used questionnaire, which allows us to reduce the total number of questions by discarding non-important ones.

The voice and query data classification in decision-level fusion provided the best classification results compared to classification of this data separately. The achieved EER of 9.52% coincided with the findings of other researchers [99, 151, 159] and justified the theory that voice and query data fusion can improve classification accuracy. These results allowed us to make an assumption that even better classification accuracy may be achieved. This motivates the development of new techniques, so that doctors could be equipped with accurate non-invasive tools for laryngeal disorder treatment.

Apart from the major findings, the acoustic and contact microphone comparison revealed that in a controlled noise environment, the acoustic microphone should be used without any other consideration. These results are broadly consistent with the findings of [104]. However, in noisy environments, the contact microphone could be more beneficial, but additional research has to be done in order to determine the noise level when it becomes superior to the acoustic one.

As an additional result of the work of this research, a computer program utilizing all proposed and applied techniques was developed. It is now used by otolaryngology specialists in their everyday work and helps them to achieve decisions about patient diagnosis. As far as we know, currently there is no alternative software with similar capabilities.

Future studies with a larger amount of learning data would be of interest. The variety of subjects might allow the increase of classification accuracy or reveal the problematic cases where additional experiments or new techniques are required. Also, it would be useful to explore the decision-level fusion by meta-learner using other combination techniques.

7. CONCLUSIONS

1. The review of non-invasive techniques used for voice pathology detection showed that many techniques exist for this task, however classification accuracy remains an obstacle while trying to apply these techniques in everyday work of medical specialists. This indicates that more research is required to improve the accuracy of these techniques.

2. Proposed data dependent Random Forest based technique for combination and classification of data from multiple data sets increased classification accuracy by 1.6% - 3.19%, compared to other our tested techniques. The maximum achieved accuracy of 86.37% indicates high potential of our method, but further improvements are necessary for even higher accuracy improvement.

3. The newly constructed algorithm from our extracted association rules achieved Equal Error Rate of only 11.11% . This confirmed the finding of related work analysis that successful voice pathology detection can be performed using only query data which represents patient's voice quality and function evaluation.

4. Voice and query data classification by combination of meta-learner achieved Equal Error Rate of only 9.52%. This was a 1.5% improvement, compared to the best achieved result when classifying both types of data separately. Base classifiers, Association rules and Decision tree, being completely transparent, provide the required transparency and helps in preventive health care and for learning purposes.

5. Dimensionality reduction by T-distributed stochastic neighbor embedding allowed to visualize initial data and classifier decision in a single two-dimensional image, which helps to analyze data and decisions. Statistical representation of data by Probability Density Functions is useful for deeper patient analysis and allows to indicate data deficiency, serves as a learning material.

References

[1] Halim Abbas, Ford Garberson, Eric Glover, and Dennis P. Wall. Machine learning approach for early detection of autism by combining questionnaire and home video screening. *CoRR*, abs/1703.06076, 2017.

[2] Mai Abusair, Antinisca Di Marco, and Paola Inverardi. Context-aware adaptation of mobile applications driven by software quality and user satisfaction. In *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*. IEEE, jul 2017.

[3] J.J. García Adeva, J.M. Pikatza Atxa, M. Ubeda Carrillo, and E. Ansuategi Zengotitabengoa. Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, 41(4):1498–1508, mar 2014.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[5] Ahmed Al-nasheri, Zulfiqar Ali, Ghulam Muhammad, Mansour Alsulaiman, Khalid H. Almalki, Tamer A. Mesallam, and Mohamed Farahat. Voice pathology detection with MDVP parameters using arabic voice pathology database. In *2015 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW)*. IEEE, feb 2015.

[6] Ahmed Al-Nasheri, Ghulam Muhammad, Mansour Alsulaiman, Zulfiqar Ali, Khalid H. Malki, Tamer A. Mesallam, and Mohamed Farahat Ibrahim. Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. *IEEE Access*, 6:6961–6974, 2018.

[7] Ahmed Al-nasheri, Ghulam Muhammad, Mansour Alsulaiman, Zulfiqar Ali, Tamer A. Mesallam, Mohamed Farahat, Khalid H. Malki, and Mohamed A. Bencherif. An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *Journal of Voice*, 31(1):113.e9–113.e18, jan 2017.

[8] Ahmed Y. Al-nasheri, Zulfiqar Ali, Muhammad Ghulam, and Mansour Alsulaiman. An investigation of mdvp parameters for voice pathology detection on three different databases. In *INTERSPEECH*, 2015.

[9] Zulfiqar Ali, Irraivan Elamvazuthi, Mansour Alsulaiman, and Ghulam Muhammad. Detection of voice pathology using fractal dimension in a multiresolution analysis of normal and disordered speech signals. *Journal of Medical Systems*, 40(1), nov 2015.

[10] A. Alpan, J. Schoentgen, Y. Maryn, F. Grenez, and P. Murphy. Assessment of disordered voice via the first rahmonic. *Speech Communication*, 54(5):655–663, jun 2012.

[11] J D Arias-Londono, J I Godino-Llorente, N Saenz-Lechon, V Osma-Ruiz, and G Castellanos-Dominguez. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Transactions on Biomedical Engineering*, 58(2):370–379, feb 2011.

[12] Anders Askenfelt, Jan Gauffin, Johan Sundberg, and Peter Kitzing. A comparison of contact microphone and electroglottograph for the measurement of vocal fundamental frequency. *Journal of Speech Language and Hearing Research*, 23(2):258, jun 1980.

[13] M. Bacauskiene, A. Verikas, A. Gelzinis, and A. Vegiene. Random forests based monitoring of human larynx using questionnaire data. *Expert Systems with Applications*, 39(5):5506–5512, apr 2012.

[14] Kevin K. Bach, Peter C. Belafsky, Kathleen Wasylik, Gregory N. Postma, and Jamie A. Koufman. Validity and reliability of the glottal function index. *Archives of Otolaryngology–Head & Neck Surgery*, 131(11):961, nov 2005.

[15] Márcio Barros and Arilo Neto. Threats to validity in search-based software engineering empirical studies. 5, 01 2011.

[16] Hans Behrbohm, Oliver Kaschke, Tadeus Nawka, and Andrew Swift. *Ear, Nose and Throat Diseases: With Head and Neck Surgery*. Thieme Medical, 3rd edition, August 2009.

[17] Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25(2):197–227, apr 2016.

[18] Leo Breiman. *Random Forests*, volume 45. Kluwer Academic Publishers, Boston, October 2001.

[19] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees (Wadsworth Statistics/Probability)*. Chapman and Hall/CRC, 1984.

[20] William Brent. A timbre analysis and classification toolkit for pure data. 2010.

[21] Mike Brookes. Voicebox: Speech processing toolbox for matlab, 1997.

[22] N. Brummer and E. de Villiers. *The BOSARIS toolkit user guide: theory, algorithms and code for binary classifier score processing*. 2011.

[23] Niko Brummer and Edward de Villiers. The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF. *arXiv*, 1304(2865v1):1–23, April 10 2013. Presented at the NIST SRE'11 Analysis Workshop, Atlanta, December 2011. Available at http://sites.google.com/site/bosaristoolkit/.

[24] Dieu Tien Bui, Tien Chung Ho, Inge Revhaug, Biswajeet Pradhan, and Duy Ba Nguyen. Landslide susceptibility mapping along the national road 32 of vietnam using GIS-based j48 decision tree classifier and its ensembles. In *Cartography from Pole to Pole*, pages 303–317. Springer Berlin Heidelberg, aug 2013.

[25] Edward C. Carterette, J. D. Markel, and A. H. Gray. Linear prediction of speech. *Language*, 53(3):723, sep 1977.

[26] Yen-Chi Chen, Christopher R. Genovese, and Larry Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, pages 0–0, sep 2016.

[27] Yogesh C.K., M. Hariharan, Ruzelita Ngadiran, Abdul Hamid Adom, Sazali Yaacob, Chawki Berkai, and Kemal Polat. A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Systems with Applications*, 69:149–158, mar 2017.

[28] James Cook, Ilya Sutskever, Andriy Mnih, and Geoffrey Hinton. Visualizing similarity data with a mixture of maps. In *In AI and Statistics, 2007. Society for Artificial Intelligence and Statistics*, 2007.

[29] Hugo Tito Cordeiro, José Manuel Fonseca, and Carlos Meneses Ribeiro. LPC spectrum first peak analysis for voice pathology detection. *Procedia Technology*, 9:1104–1111, 2013.

[30] Leandro de Araújo Pernambuco, Marluce Nascimento de Almeida, Keliane Gomes Matias, and Erika Beatriz de Morais Costa. Voice assessment and voice-related quality of life in patients with benign thyroid disease. *Otolaryngology-Head and Neck Surgery*, 152(1):116–121, jan 2015.

[31] Carl de Boor. *A Practical Guide to Splines (Applied Mathematical Sciences)*. Springer, 2001.

[32] Karmele López de Ipiña, Jordi Solé-Casals, Harkaitz Eguiraun, J.B. Alonso, C.M. Travieso, Aitzol Ezeiza, Nora Barroso, Miriam Ecay-Torres, Pablo Martinez-Lage, and Blanca Beitia. Feature selection for spontaneous speech analysis to aid in alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, 30(1):43–60, mar 2015.

[33] Guus de Krom. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech Language and Hearing Research*, 36(2):254, apr 1993.

[34] Tomas Dekens, Yorgos Patsis, Werner Verhelstand Frederic Beaugendre, and Francois Capman. A multi-sensor speech database with applications towards robust speech processing in hostile environments. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias Stelios Piperidis, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[35] Tomas Dekens, Werner Verhelst, Francois Capman, and Frederic Beaugendre. Improved speech recognition in noisy environments by using a throat microphone for accurate speech detection. *Signal Processing Conference, 2010 18th European*, 2010.

[36] Francois Deliege, Chua Bee Yong, and Pedersen Torben Bach. *High-Level Audio Features: Distributed Extraction and Similarity Search*, pages 565–570. ISMIR, 2008.

[37] Dimitar D. Deliyski, Maegan K. Evans, and Heather S. Shaw. Influence of data acquisition environment on accuracy of acoustic voice quality measurements. *Journal of Voice*, 19(2):176–186, jun 2005.

[38] Dimitar D. Deliyski, Heather S. Shaw, and Maegan K. Evans. Adverse effects of environmental noise on acoustic voice quality measurements. *Journal of Voice*, 19(1):15–28, mar 2005.

[39] Ramón Díaz-Uriarte and Sara Alvarez de Andrés. *BMC Bioinformatics*, 7(1):3, 2006.

[40] Peijun Du, Alim Samat, Björn Waske, Sicong Liu, and Zhenhong Li. Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105:38–53, jul 2015.

[41] Stephane Dupont, Christophe Ris, and Damien Bachelart. Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise. In *COST278 and ISCA tutorial and research wrkshop (ITRW) on robustness issues in conversational interaction*, pages 1–4, 2004.

[42] Dan Ellis. Plp and rasta (and mfcc, and inversion) in matlab, 2005.

[43] Daniel Engel, Lars HÃijttenberger, and Bernd Hamann. A survey of dimension reduction methods for high-dimensional data analysis and visualization, 2012.

[44] Anna S. Englhard, Tom Betz, Veronika Volgger, Eva Lankenau, Georg J. Ledderose, Herbert Stepp, Christian Homann, and Christian S. Betz. Intraoperative assessment of laryngeal pathologies with optical coherence tomography integrated into a surgical microscope. *Lasers in Surgery and Medicine*, 49(5):490–497, feb 2017.

[45] E. Erzin. Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1316–1324, sep 2009.

[46] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia - MM13*. Association for Computing Machinery (ACM), 2013.

[47] Shih-Hau Fang, Yu Tsao, Min-Jing Hsiao, Ji-Ying Chen, Ying-Hui Lai, Feng-Chuan Lin, and Chi-Te Wang. Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*, mar 2018.

[48] M Faurholt-Jepsen, J Busk, M Frost, M Vinberg, E M Christensen, O Winther, J E Bardram, and L V Kessing. Voice analysis as an objective state marker in bipolar disorder. *Translational Psychiatry*, 6(7):e856–e856, jul 2016.

[49] Fangwen Fu, Deepak S. Turaga, Olivier Verscheure, Mihaela van der Schaar, and Lisa Amini. Configuring competing classifier chains in distributed stream mining systems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):548–563, dec 2007.

[50] Zhun ga Liu, Quan Pan, Jean Dezert, and Arnaud Martin. Adaptive imputation of missing values for incomplete pattern classification. *Pattern Recognition*, 52:85–95, apr 2016.

[51] Jurgen T. Geiger, Bjorn Schuller, and Gerhard Rigoll. Large-scale audio feature extraction and SVM for acoustic scene classification. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, oct 2013.

[52] A. Gelzinis, A. Verikas, and M. Bacauskiene. Automated speech analysis applied to laryngeal disease categorization. *Computer Methods and Programs in Biomedicine*, 91(1):36–47, jul 2008.

[53] Hamzeh Ghasemzadeh, Mehdi Tajik Khass, Meisam Khalil Arjmandi, and Mohammad Pooyan. Detection of vocal disorders based on phase space parameters and lyapunov spectrum. *Biomedical Signal Processing and Control*, 22:135 – 145, 2015.

[54] Amanda I. Gillespie and Jackie Gartner-Schmidt. Voice-specialized speech-language pathologist's criteria for discharge from voice therapy. *Journal of Voice*, 32(3):332–339, may 2018.

[55] Andrej Gisbrecht, Alexander Schulz, and Barbara Hammer. Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing*, 147:71–82, jan 2015.

[56] Amir Globerson and Sam Roweis. Visualizing pairwise similarity via semidefinite programming. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 139–146, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

[57] J.I. Godino-Llorente and P. Gomez-Vilda. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering*, 51(2):380–384, feb 2004.

[58] Jorge Andrés Gómez-García, Laureano Moro-Velázquez, Juan Ignacio Godino-Llorente, and César Germán Castellanos-Domínguez. An insight to the automatic categorization of speakers according to sex and its application to the detection of voice pathologies: A comparative study. *Revista Facultad de Ingeniería Universidad de Antioquia*, (79), jun 2016.

[59] J. Hun Hah, Songyong Sim, Soo-Youn An, Myung-Whun Sung, and Hyo Geun Choi. Evaluation of the prevalence of and factors associated with laryngeal diseases among the general population. *The Laryngoscope*, 125(11):2536–2542, jul 2015.

[60] D.L. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

[61] Pavol Harar, Jesus B. Alonso-Hernandezy, Jiri Mekyska, Zoltan Galaz, Radim Burget, and Zdenek Smekal. Voice pathology detection using deep learning: a preliminary study. In *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*. IEEE, jul 2017.

[62] Daria Hemmerling, Andrzej Skalski, and Janusz Gajda. Voice data mining for laryngeal pathology assessment. *Computers in Biology and Medicine*, 69:270–276, feb 2016.

[63] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, apr 1990.

[64] Elena M. Hernández-Pereira, Diego Álvarez-Estévez, and Vicente Moret-Bonillo. Automatic classification of respiratory patterns involving missing data imputation techniques. *Biosystems Engineering*, 138:65–76, oct 2015.

[65] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003.

[66] Yoshiyuki Horii. Jitter and shimmer differences among sustained vowel phonations. *Journal of Speech Language and Hearing Research*, 25(1):12, mar 1982.

[67] M. Shamim Hossain and Ghulam Muhammad. Healthcare big data voice pathology assessment framework. *IEEE Access*, 4:7806–7815, 2016.

[68] Verduyckt Ingrid, Morsomme Dominique, and Remacle Marc. Validation and standardization of the pediatric voice symptom questionnaire: A double-form questionnaire for dysphonic children and their parents. *Journal of Voice*, 26(4):e129–e139, jul 2012.

[69] Ergonomic requirements for office work with visual display terminals (vdts) — part 11: Guidance on usability. Standard, International Organization for Standardization, Geneva, CH, 198.

[70] Takayuki Itoh, Ashnil Kumar, Karsten Klein, and Jinman Kim. High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. *CoRR*, abs/1609.05268, 2016.

[71] Luis M. T. Jesus, Joana Martinez, Andreia Hall, and Aníbal Ferreira. Acoustic correlates of compensatory adjustments to the glottic and supraglottic structures in patients with unilateral vocal fold paralysis. *BioMed Research International*, 2015:1–9, 2015.

[72] Julie Josse and François Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2):79–99, 2012.

[73] Julie Josse and François Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6):1869–1879, jun 2012.

[74] S. Jothilakshmi. Automatic system to detect the type of voice pathology. *Applied Soft Computing*, 21:244–249, aug 2014.

[75] Silvia Tieko Kasama and Alcione Ghedini Brasolotto. Percepção vocal e qualidade de vida. *Pró-Fono Revista de Atualização Científica*, 19(1), apr 2007.

[76] Brian Kirk, Kyle Conroy, Andrej Prša, Michael Abdul-Masih, Angela Kochoska, Gal Matijevič, Kelly Hambleton, Thomas Barclay, Steven Bloemen, Tabetha Boyajian, Laurance R. Doyle, B. J. Fulton, Abe Johannes Hoekstra, Kian Jek, Stephen R. Kane, Veselin Kostov, David Latham, Tsevi Mazeh, Jerome A. Orosz, Joshua Pepper, Billy Quarles, Darin Ragozzine, Avi Shporer, John Southworth, Keivan Stassun, Susan E. Thompson, William F. Welsh, Eric Agol, Aliz Derekas, Jonathan Devor, Debra Fischer, Gregory Green, Jeff Gropp, Tom Jacobs, Cole Johnston, Daryll Matthew LaCourse, Kristian Saetre, Hans Schwengeler, Jacek Toczyski, Griffin Werner, Matthew Garrett, Joanna Gore, Arturo O. Martinez, Isaac Spitzer, Justin Stevick, Pantelis C. Thomadis, Eliot Halley Vrijmoet, Mitchell Yenawine, Natalie Batalha, and William Borucki. Kepler eclipsing binary stars. vii. the catalog of eclipsing binaries found in the entire kepler data set. *The Astronomical Journal*, 151(3):68, 2016.

[77] Barbara Kitchenham and Shari Lawrence Pfleeger. Principles of survey research part 6. *SIGSOFT Softw. Eng. Notes*, 28(2):24, mar 2003.

[78] Sophie A. C. Kraaijenga, Lisette van der Molen, Irene Jacobi, Olga Hamming-Vrieze, Frans J. M. Hilgers, and Michiel W. M. van den Brekel. Prospective clinical study on long-term swallowing function and voice quality in advanced head and neck cancer patients treated with concurrent chemoradiotherapy and preventive swallowing exercises. *European Archives of Oto-Rhino-Laryngology*, 272(11):3521–3531, nov 2014.

[79] Robbi A. Kupfer, Emily M. Hogikyan, and Norman D. Hogikyan. Establishment of a normative database for the voice-related quality of life (v-RQOL) measure. *Journal of Voice*, 28(4):449–451, jul 2014.

[80] Adidah Lajis and Normaziah Abdul Aziz. Competency assessment of short free text answers. In *2014 4th International Conference on Engineering Technology and Technopreneuship (ICE2T)*. Institute of Electrical and Electronics Engineers (IEEE), aug 2014.

[81] D Lavanya. Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services*, 2(1):17–24, feb 2012.

[82] Claudia Lindner, Paul A. Bromiley, Mircea C. Ionita, and Tim F. Cootes. Robust and accurate shape model matching using random forest regression-voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1862–1874, sep 2015.

[83] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, 2002.

[84] Leonardo Wanderley Lopes, Gyllyane Furtado Cabral, and Anna Alice Figueiredo de Almeida. Vocal tract discomfort symptoms in patients with different voice disorders. *Journal of Voice*, 29(3):317–323, may 2015.

[85] Leonardo Wanderley Lopes, Jocélio Delfino da Silva, Layssa Batista Simões, Deyverson da Silva Evangelista, Priscila Oliveira Costa Silva, Anna Alice Almeida, and Maria Fabiana Bonfim de Lima-Silva. Relationship between acoustic measurements and self-evaluation in patients with voice disorders. *Journal of Voice*, 31(1):119.e1–119.e10, jan 2017.

[86] Julia Lukaschyk, Meike Brockmann-Bauser, and Ulla Beushausen. Transcultural adaptation and validation of the german version of the vocal tract discomfort scale. *Journal of Voice*, 31(2):261.e1–261.e8, mar 2017.

[87] Dalton Lunga, Saurabh Prasad, Melba M. Crawford, and Okan Ersoy. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *IEEE Signal Processing Magazine*, 31(1):55–66, jan 2014.

[88] C. Manfredi and G. Peretti. A new insight into postsurgical objective voice quality evaluation: Application to thyroplastic medialization. *IEEE Transactions on Biomedical Engineering*, 53(3):442–451, mar 2006.

[89] Claudia Manfredi, Massimo D'Aniello, Piero Bruscaglioni, and Andrea Ismaelli. A comparative analysis of fundamental frequency estimation methods with application to pathological voices. *Medical Engineering & Physics*, 22(2):135–147, mar 2000.

[90] Maria Markaki and Yannis Stylianou. Voice pathology detection and discrimination based on modulation spectral features. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1938–1948, sep 2011.

[91] S. Lawrence Marple and William M. Carey. Digital spectral analysis with applications. *The Journal of the Acoustical Society of America*, 86(5):2043–2043, nov 1989.

[92] Richard Marreel and Janet McLellan. *Information Management in Health Care.* Delmar Cengage Learning, 1999.

[93] Youri Maryn, Marc De Bodt, Ben Barsties, and Nelson Roy. The value of the acoustic voice quality index as a measure of dysphonia severity in subjects speaking different languages. *Eur Arch Otorhinolaryngol*, oct 2013.

[94] Jiri Mekyska, Eva Janousova, Pedro Gomez-Vilda, Zdenek Smekal, Irena Rektorova, Ilona Eliasova, Milena Kostalova, Martina Mrackova, Jesus B. Alonso-Hernandez, Marcos Faundez-Zanuy, and Karmele Lopez de Ipina. Robust and complex approach of pathological speech signal analysis. *Neurocomputing*, 167:94 – 111, 2015.

[95] Leonardo Alfredo Forero Mendoza, Edson Cataldo, Marley Vellasco, Marco Aurelio Silva, Alvaro David Orjuela Canon, and Jose Manoel de Seixas. Classification of voice aging using ANN and glottal signal parameters. In *2010 IEEE ANDESCON*. Institute of Electrical & Electronics Engineers (IEEE), sep 2010.

[96] José P. Miguel, David Mauricio, and Glen Rodríguez. A review of software quality models for the evaluation of software products. *International Journal of Software Engineering & Applications*, 5(6):31–53, nov 2014.

[97] Koreen Millard and Murray Richardson. Wetland mapping with lidar derivatives, sar polarimetric decompositions, and lidar–sar fusion using a random forest classifier. *Canadian Journal of Remote Sensing*, 39(4):290–307, 2013.

[98] Koreen Millard and Murray Richardson. On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote Sensing*, 7(7):8489–8515, jul 2015.

[99] Jonas Minelga, Antanas Verikas, Evaldas Vaiciukynas, Adas Gelzinis, and Marija Bacauskiene. A transparent decision support tool in screening for laryngeal disorders using voice and query data. *Applied Sciences*, 2017.

[100] Dunja Mladenić and Marko Grobelnik. Automatic text analysis by artificial intelligence. *Informatica*, 37:27–33, 2013.

[101] David Moffat, David Ronan, and Joshua Reiss. An evaluation of audio feature extraction toolboxes, 2015.

[102] R.J. Moran, R.B. Reilly, P. de Chazal, and P.D. Lacy. Telephony-based voice pathology assessment using automated speech analysis. *IEEE Transactions on Biomedical Engineering*, 53(3):468–477, mar 2006.

[103] Joanna Morawska, Ewa Niebudek-Bogusz, Justyna Wiktorowicz, and Mariola Śliwińska-Kowalska. Screening value of v-RQOL in the evaluation of occupational voice disorders. *Medycyna Pracy*, dec 2017.

[104] Nafeesa Mubeen, A. Shahina, A. Nayeemulla Khan, and G. Vinoth. Combining spectral features of standard and throat microphones for speaker identification. In *2012 International Conference on Recent Trends in Information Technology*. Institute of Electrical and Electronics Engineers (IEEE), apr 2012.

[105] Ghulam Muhammad. Voice pathology detection using vocal tract area irregularity measures. 2014.

[106] Ghulam Muhammad, Mansour Alsulaiman, Zulfiqar Ali, Tamer A. Mesallam, Mohamed Farahat, Khalid H. Malki, Ahmed Al-nasheri, and Mohamed A. Bencherif. Voice pathology detection using interlaced derivative pattern on glottal source excitation. *Biomedical Signal Processing and Control*, 31:156–164, jan 2017.

[107] Ghulam Muhammad, Mansour Alsulaiman, Awais Mahmood, and Zulfiqar Ali. Automatic voice disorder classification using vowel formants. In *2011 IEEE International Conference on Multimedia and Expo*. Institute of Electrical & Electronics Engineers (IEEE), jul 2011.

[108] Ghulam Muhammad, Ghadir Altuwaijri, Mansour Alsulaiman, Zulfiqar Ali, Tamer A. Mesallam, Mohamed Farahat, Khalid H. Malki, and Ahmed Al-nasheri. Automatic voice pathology detection and classification using vocal tract area irregularity. *Biocybernetics and Biomedical Engineering*, 36(2):309–317, 2016.

[109] Ghulam Muhammad and Moutasem Melhem. Pathological voice detection and binary classification using MPEG-7 audio features. *Biomedical Signal Processing and Control*, 11:1–9, may 2014.

[110] Jacob B. Munger and Scott L. Thomson. Frequency response of the skin on the head and neck during production of selected speech sounds. *The Journal of the Acoustical Society of America*, 124(6):4001–4012, dec 2008.

[111] Mohammad Nassralla, Zeina El Zein, and Hazem Hajj. Classification of normal and abnormal heart sounds. In *2017 Fourth International Conference on Advances in Biomedical Engineering (ICABME)*. IEEE, oct 2017.

[112] Thuy Tuong Nguyen and Yury Tsoy. A kernel PLS based classification method with missing data handling. *Stat Papers*, jun 2015.

[113] Anuradha S. Nigade and J. S. Chitode. Throat microphone signals for isolated word recognition using lpc. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2012.

[114] M. Nolan, B. Madden, and E. Burke. Accelerometer based measurement for the mapping of neck surface vibrations during vocalized speech. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Institute of Electrical and Electronics Engineers (IEEE), sep 2009.

[115] Ifeyinwa Okoye, Steven Bethard, and Tamara Sumner. Cu : Computational assessment of short free text answers - a tool for evaluating students understanding. In *Second Joint Conference on Lexical and Computational Semantics (SEM)*. Association for Computational Linguistics, jun 2013.

[116] Liran Oren, Sid Khosla, and Ephraim Gutmark. Effect of vocal fold asymmetries on glottal flow. *The Laryngoscope*, 126(11):2534–2538, mar 2016.

[117] Juan Rafael Orozco-Arroyave, Elkyn Alexander Belalcazar-Bolanos, Julian David Arias-Londono, Jesus Francisco Vargas-Bonilla, Sabine Skodda, Jan Rusz, Khaled Daqrouq, Florian Honig, and Elmar Noth. Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1820–1828, nov 2015.

[118] Asil Oztekin, Dursun Delen, Ali Turkyilmaz, and Selim Zaim. A machine learning-based usability evaluation method for eLearning systems. *Decision Support Systems*, 56:63–73, dec 2013.

[119] Senthil Kumar Palanisamy. Association rule based classification. Computer science, Worcester Polytechnic Institute, Worcester, Massachusetts, USA, May 2006. Prof. Carolina Ruiz, Advisor.

[120] Daria Panek, Andrzej Skalski, and Janusz Gajda. Quantification of linear and non-linear acoustic analysis applied to voice pathology detection. In *Advances in Intelligent Systems and Computing*, pages 355–364. Springer International Publishing, 2014.

[121] Daria Panek, Andrzej Skalski, Janusz Gajda, and Ryszard Tadeusiewicz. Acoustic analysis assessment in speech pathology detection. *International Journal of Applied Mathematics and Computer Science*, 25(3), jan 2015.

[122] Nico Paolo Paolillo and Giuseppe Pantaleo. Development and validation of the voice fatigue handicap questionnaire (VFHQ): Clinical, psychometric, and psychosocial facets. *Journal of Voice*, 29(1):91–100, jan 2015.

[123] M. Petrakos, J. Atli Benediktsson, and I. Kanellopoulos. The effect of classifier agreement on the accuracy of the combined classifier in decision level fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 39(11):2539–2546, 2001.

[124] Rūta Pribuišienė, Migle Baceviciene, Virgilijus Uloza, Aurelija Vegiene, and Jelena Antuseva. Validation of the lithuanian version of the glottal function index. *Journal of Voice*, 26(2):e73–e78, mar 2012.

[125] Tania Rajput, Vikas Mittal, and Tarun Gulati. Voice parameter analysis for the detection of effect of stress on voice: A review. *International Journal of Research in Information Technology*, 3(4):463 − 467, 2015.

[126] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Global refinement of random forest. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Institute of Electrical and Electronics Engineers (IEEE), jun 2015.

[127] Carlo Robotti, Francesco Mozzanica, Ilaria Pozzali, Laura D'Amore, Patrizia Maruzzi, Daniela Ginocchio, Stafania Barozzi, Rosaria Lorusso, Francesco Ottaviani, and Antonio Schindler. Cross-cultural adaptation and validation of the italian version of the vocal tract discomfort scale (i-VTD). *Journal of Voice*, oct 2017.

[128] Nicolás Sáenz-Lechón, Juan I. Godino-Llorente, Víctor Osma-Ruiz, and Pedro Gómez-Vilda. Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control*, 1(2):120–128, apr 2006.

[129] Reza Safdari, Hussein Dargahi, Leila Shahmoradi, and Ahmadreza Farzaneh Nejad. Comparing four softwares based on ISO 9241 part 10. *Journal of Medical Systems*, 36(5):2787–2793, jul 2011.

[130] Ali Salih Mahmoud Saudi, Aliaa A. A. Youssif, and Atef Z. Ghalwash. Computer aided recognition of vocal folds disorders by means of RASTA-PLP. *CIS*, 5(2), feb 2012.

[131] Julia C Selby, Harvey R Gilbert, and J.W Lerman. Perceptual and acoustic evaluation of individuals with laryngopharyngeal reflux pre- and post-treatment. *Journal of Voice*, 17(4):557–570, dec 2003.

[132] Cheng-Ya Sha, Yi-Hsuan Yang, Yu-Ching Lin, and Homer H. Chen. Singing voice timbre classification of chinese popular music. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, may 2013.

[133] Bernard W. Silverman. *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. Chapman and Hall, London, UK, 1st edition, 1986.

[134] Ilse Smits, Piet Ceuppens, and Marc S. De Bodt. A comparative study of acoustic voice measurements by means of dr. speech and computerized speech lab. *Journal of Voice*, 19(2):187–196, jun 2005.

[135] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, jan 1999.

[136] Laurent Sorber, Marc Van Barel, and Lieven De Lathauwer. Structured data fusion. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):586–600, jun 2015.

[137] Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, mar 2007.

[138] J. G. Svec and S. Granqvist. Guidelines for selecting microphones for human voice production research. *American Journal of Speech-Language Pathology*, 19(4):356–368, jul 2010.

[139] Jan G. Švec, Ingo R. Titze, and Peter S. Popolo. Estimation of sound pressure levels of voiced speech from skin vibration of the neck. *The Journal of the Acoustical Society of America*, 117(3):1386–1394, mar 2005.

[140] Fei Tang and Hemant Ishwaran. Random forest missing data algorithms, 2017.

[141] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web - WWW16*. Association for Computing Machinery (ACM), 2016.

[142] Caitlin N. Teague, Sinan Hersek, Hakan Toreyin, Mindy L. Millard-Stafford, Michael L. Jones, Geza F. Kogler, Michael N. Sawka, and Omer T. Inan. Novel methods for sensing acoustical emissions from the knee for wearable joint health assessment. *IEEE Transactions on Biomedical Engineering*, 63(8):1581–1590, aug 2016.

[143] Matthias Templ, Alexander Kowarik, and Peter Filzmoser. Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, 55(10):2793–2806, oct 2011.

[144] Alaa Tharwat, Abdelhameed Ibrahim, Aboul Ella Hassanien, and Gerald Schaefer. Ear recognition using block-based principal component analysis and decision fusion. In *Lecture Notes in Computer Science*, pages 246–254. Springer International Publishing, 2015.

[145] Olga G. Troyanskaya, Michael Cantor, Gavin Sherlock, Patrick O. Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, jun 2001.

[146] Athanasios Tsanas. Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms. 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), dec 2013.

[147] B. Tunc and H. Dag. Generating classification association rules with modified apriori algorithm. In *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, AIKED'06, pages 384–387, Stevens Point, Wisconsin, USA, 2006. World Scientific and Engineering Academy and Society (WSEAS).

[148] S.K. Ueng, C.M. Luo, T.Y. Tsai, and H. Chang. Voice quality assessment and visualization. In *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*. Institute of Electrical & Electronics Engineers (IEEE), jul 2012.

[149] Virgilijus Uloza, Evaldas Padervinskis, Aurelija Vegiene, Ruta Pribuisiene, Viktoras Saferis, Evaldas Vaiciukynas, Adas Gelzinis, and Antanas Verikas. Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening. *Eur Arch Otorhinolaryngol*, 272(11):3391–3399, jul 2015.

[150] Virgilijus Uloza, Ruta Pribuisiene, and Viktoras Saferis. Multidimensional assessment of functional outcomes of medialization thyroplasty. *European Archives of Oto-Rhino-Laryngology*, 262(8):616–621, dec 2004.

[151] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, J. Minelga, M. Hållander, E. Padervinskis, and V. Uloza. Fusing voice and query data for non-invasive detection of laryngeal disorders. *Expert Systems with Applications*, 42(22):8445–8453, dec 2015.

[152] Evaldas Vaičiukynas. *Computational Intelligence Methods for Voice Function Assessment and Laryngeal Pathology Detection.* PhD thesis, Kaunas University of Technology, Kaunas, 2013.

[153] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations inR. *Journal of Statistical Software*, 45(3), 2011.

[154] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, November 2008.

[155] A. Verikas, A. Gelzinis, M. Bacauskiene, and V. Uloza. Towards noninvasive screening for malignant tumours in human larynx. *Medical Engineering & Physics*, 32(1):83–89, jan 2010.

[156] A. Verikas, A. Gelzinis, M. Bacauskiene, V. Uloza, and M. Kaseta. Using the patient's questionnaire data to screen laryngeal disorders. *Computers in Biology and Medicine*, 39(2):148–155, feb 2009.

[157] A. Verikas, A. Gelzinis, E. Vaiciukynas, M. Bacauskiene, J. Minelga, M. Hållander, V. Uloza, and E. Padervinskis. Data dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: Acoustic versus contact microphone. *Medical Engineering & Physics*, 37(2):210–218, feb 2015.

[158] Antanas Verikas, Marija Bacauskiene, Adas Gelzinis, Evaldas Vaiciukynas, and Virgilijus Uloza. Questionnaire versus voice-based screening for laryngeal disorders. *Expert Systems with Applications*, 39(6):6254–6262, may 2012.

[159] Antanas Verikas, Adas Gelzinis, Marija Bacauskiene, Magnus Hållander, Virgilijus Uloza, and Marius Kaseta. Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders. *Artificial Intelligence in Medicine*, 49(1):43–50, may 2010.

[160] Amritha Vijayan, Bipil Mary Mathai, Karthik Valsalan, Riyanka Raji Johnson, Lani Rachel Mathew, and K. Gopakumar. Throat microphone speech recognition using mfcc. In *2017 International Conference on Networks & Advances in Computational Technologies (NetACT)*. IEEE, jul 2017.

[161] Zhijian Wang, Ping Yu, Nan Yan, Lan Wang, and Manwa L. Ng. Automatic assessment of pathological voice quality using multidimensional acoustic analysis based on the GRBAS scale. *J Sign Process Syst*, 82(2):241–251, jun 2015.

[162] Kai-Pun Wong, Jung-Woo Woo, Jason Yu-Yin Li, Kyu Eun Lee, Yeo Kyu Youn, and Brian Hung-Hin Lang. Using transcutaneous laryngeal ultrasonography (TLUSG) to assess post-thyroidectomy patients' vocal cords: Which maneuver best optimizes visualization and assessment accuracy? *World J Surg*, 40(3):652–658, nov 2015.

[163] R. N. Wormald, R. J. Moran, R. B. Reilly, and P. D. Lacy. Performance of an automated, remote system to detect vocal fold paralysis. *Annals of Otology, Rhinology & Laryngology*, 117(11):834–838, nov 2008.

[164] Simin Xie, Nan Yan, Ping Yu, Manwa L. Ng, Lan Wang, and Zhuanzhuan Ji. Deep neural networks for voice quality assessment based on the grbas scale. In *Interspeech 2016*, pages 2656–2660, 2016.

[165] Jianfeng Xu, Yu Li, Yuanjian Zhang, and Azhar Mahmood. WSN missing data imputing based on multiple time granularity. *International Journal of Future Generation Communication and Networking*, 9(6):263–274, jun 2016.

[166] Tet Fei Yap, Julien Epps, Eliathamby Ambikairajah, and Eric H. C. Choi. Voice source features for cognitive load classification. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical & Electronics Engineers (IEEE), may 2011.

[167] Yu Zhang and Jack J. Jiang. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *Journal of Voice*, 22(1):1–9, jan 2008.

[168] Yu Zheng. Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data*, 1(1):16–34, mar 2015.

A. APPENDIX 1. SAMPLES OF VOICE AND QUERY DATA

The samples of data used in this research are provided here. Voice recordings, smaller voice parameters data set and query data set were provided by the Department of Otolaryngology from Lithuanian University of Health Sciences. Audio features set containing 14 different parameters set was extracted during this work. Data sets examples are visible in the tables of this section. Table A.1 represents the smaller audio features data set, extracted using Dr. Speech software. Table A.4 contains sample data of questionnaire answers. Due to the number of questions, column names are abbreviations, which are explained below the table. Tables A.2 and A.3 represents very small part of the bigger audio features data set, which contains 14 subsets. Healthy class in this data set is represented as 0, while pathological as 1. Because of the size of data vectors, both these tables represent only a fraction of two subsets. Other data samples are provided in the compact disk with the printed version of this work.

Table A.1 Sample of voice parameters data set obtained from the Department of Otolaryngology in Lithuanian University of Health Sciences

ID	F0	Jitter	Shimmer	NNE	HNR	SNR
A1478	184.32	0.17	1.23	-13.30	26.50	25.51
A1479	135.49	0.28	2.21	-5.84	23.84	22.03
A1480	294.11	0.17	2.56	-12.32	26.09	24.19
A1481	222.04	0.46	3.63	-6.65	19.76	17.33
A1482	209.31	0.20	2.14	-11.20	24.11	21.82
A1483	168.51	2.23	11.18	-2.18	11.88	11.11
A1484	123.18	0.37	4.39	-8.38	20.56	18.62
A1485	199.79	0.29	3.33	-6.88	20.79	20.37
A1486	212.15	1.24	6.13	-1.23	11.93	10.36
A1487	160.04	0.92	4.89	-0.64	13.75	12.69
A1488	221.84	0.20	1.16	-18.24	28.74	26.47
A1489	214.07	0.65	3.42	-8.86	20.29	17.61
A1490	161.54	0.53	2.07	-13.03	26.95	24.97
A1491	122.58	0.69	7.65	-5.90	11.73	10.72
A1492	146.69	0.69	6.75	-1.89	12.46	11.13
A1493	206.02	0.18	1.78	-17.54	27.99	27.03
A1494	180.03	0.11	0.96	-14.97	30.96	30.41
A1495	150.97	0.13	2.44	-12.32	22.85	20.00
A1496	236.64	0.17	1.41	-18.22	26.82	25.46
A1497	145.54	0.18	3.53	-2.38	18.38	17.13
A1498	236.20	0.27	1.80	-6.03	25.40	23.54

Table A.2 Pitch and amplitude perturbation measures

Patient ID	Class	24 values
1199	1	-0.42588087468409; 0.06295904442376; -0.32303881733636; -0.09946876700287; 0.75004554375730; -0.09095639953627; -0.14950021880513; (...)
1199	1	-0.45165462281808; 0.16204993069158; -0.67847076287199; -0.08305895374887; 1.38880502022426; -0.08266252003320; -0.09208425864879; (...)
1199	1	-0.43020400924416; -0.12970632825778; 0.32867946402321; -0.12970913558570; -0.50140231667013; -0.09490088047865; -0.17334666994302; (...)
1200	1	-0.67236024633712; -0.25183244510107; -0.35613000025748; -0.13771659498922; -0.84172321574243; -0.11685437401108; -0.29687110944609; (...)
1200	1	-0.65423436364500; -0.24727455159494; -0.40920742692198; -0.14181285387868; -0.74767178894007; -0.11631251609984; -0.29395264200610; (...)
1200	1	-0.66031372626529; -0.24985728281385; -0.31282339189441; -0.14065342910466; -0.86204370043884; -0.11654819932601; -0.29533629731144; (...)
1201	1	0.12853529078441; 0.13344637589262; -0.75473934009339; 0.00181012837375; 1.40882895577203; -0.02981411026518; 0.04287949694721; (...)
1201	1	0.27411287507322; 0.11260259734606; -0.67247039575851; -0.05392161034262; 1.25268364161567; -0.07065648806490; -0.14866078007015; (...)
1201	1	0.19177985467932; -0.08637415055892; -0.83974162678703; 0.11668146991554; 0.77304925747088; 0.05906141998845; 0.39356651454674; (...)
1202	0	-0.03682356083672; -0.13528272259397; -0.73766706491185; -0.08135573493329; 0.39939426862060; -0.08007791466683; -0.15084051187694; (...)
1202	0	0.04193197420634; -0.08156863847417; -0.08894620362835; -0.07195285378778; 0.02534183284374; -0.08831288879825; -0.19773104727753; (...)
1202	0	0.05694975441061; -0.13127940163870; 0.34254711888960; -0.10232465501476; -0.51847152842657; -0.08301244211684; -0.17570876666110; (...)

Table A.3 Frequency (0-5000 Hz)

Patient ID	Class	100 values
1199	1	0.35123106070114; -0.19087691084707; 2.69169231353402; 0.26284184543908; -0.77239268279710; 0.92565296852507; 0.88686848529351; (...)
1199	1	-0.01857455931662; -0.24083635400203; 2.24318835296675; -0.40574179029635; -0.77548759027342; 0.52846454718026; -0.27002166304126; (...)
1199	1	-0.45554134797792; -0.32388879984949; 1.32320907678020; -0.09147010061406; -0.76427353415719; 0.21588596275694; 0.27138665520437; (...)
1200	1	-0.51954312219819; -0.26638605190933; 0.01966876723238; -0.68040550062549; -0.44858223995407; -0.60953224912423; -0.31011394050109; (...)
1200	1	-0.40113877811176; -0.28510750452270; 0.29401302534230; -0.66185400477567; -0.27445618357101; -0.57009886589454; -0.25923258633567; (...)
1200	1	-0.58223982265605; -0.30047533715509; 0.31552096978702; -0.69061331657024; -0.34631429833247; -0.60107469178241; -0.28950362564914; (...)
1201	1	-0.03284429731602; -0.28610439672784; -0.42106531997562; -0.44829897240903; 0.88980800973292; -0.47735136067947; -0.08894834160417; (...)
1201	1	-0.52091825966780; -0.23578697793296; -0.26564888291571; -0.37802769051723; 2.32385875102039; -0.36885160251058; 0.04760751835851; (...)
1201	1	-0.57048897953561; -0.22078065620133; -0.30762294775041; -0.18454028958153; 0.17541857632231; 0.47491037132857; 0.61269386991323; (...)
1202	0	-0.39201805946287; -0.31289367201102; -0.45078551501341; 1.05438768468396; 0.61609672288636; -0.49558609026390; -0.45554163365168; (...)
1202	0	-0.47681677982497; -0.30982311550434; -0.45422027522851; 1.23752907215856; 2.49322050714764; -0.46919558411282; 0.17526855238644; (...)
1202	0	-0.44743221189594; -0.23478911239912; -0.40287344374541; 0.01914528026729; 2.72508530773459; -0.42657693620150; -0.00527595945739; (...)

Table A.4 Sample of query data set obtained from the Department of Otolaryngology in Lithuanian University of Health Sciences

ID	G	A	D	LVU	SHD	D/W	S	DUR	C/D	VA	VH	ST	SG	TS	VD	WV	LOV	VR	RS	VC	GFI1	GFI2	GFI3	GFI4	GFI	MFT
A1477	1	31	8	3	10	5	4	2	0	30	3	78	1	1	67	67	3	72	8	67	2	0	2	1	5	18
A1478	2	59	2	4	6	7	1	0	6	18	3	42	3	1	65	65	8	54	62	67	3	2	2	3	10	12
A1481	2	77	3	4	3	7	2	0	0	70	2	20	2	1	0	0	0	0	0	0	0	0	0	0	0	22
A1482	2	39	22	4	6	7	2	0	0	70	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	20
A1485	2	70	9	4	2	7	2	0	0	46	4	72	1	1	12	71	0	60	55	18	1	2	1	0	4	8
A1486	2	26	2	4	7	5	1	10	10	12	3	62	3	2	52	85	95	67	60	80	2	1	4	3	10	12
A1488	1	47	2	4	8	7	1	33	15	27	3	50	2	2	60	45	50	18	30	47	2	1	3	3	9	15
A1489	2	48	3	4	10	7	2	0	0	46	2	18	3	2	10	19	0	0	0	24	0	0	0	0	0	19
A1490	1	50	3	3	3	7	2	0	0	60	2	47	1	2	4	5	5	5	5	17	0	0	0	1	1	18
A1491	1	49	3	4	4	7	4	5	0	100	2	40	1	1	4	4	4	7	0	8	0	0	0	0	0	15
A1492	2	50	3	3	13	7	1	20	20	41	2	47	3	7	6	5	13	0	9	6	0	0	0	0	0	19
A1493	1	42	7	4	6	7	1	20	20	55	2	30	2	1	22	22	5	15	22	25	2	1	1	1	5	10
A1494	2	24	3	4	8	7	2	0	0	63	2	60	5	1	5	6	15	10	7	20	0	0	0	0	0	17
A1495	2	50	3	3	10	7	2	0	0	75	2	65	2	3	9	3	11	3	4	12	0	0	0	1	1	22
A1498	1	40	11	4	3	7	1	20	20	5	3	30	1	3	95	95	80	87	85	85	4	0	5	5	14	8
A1502	2	54	7	4	5	7	1	40	20	45	2	40	3	1	40	38	30	35	77	80	2	2	2	2	8	6
A1503	2	64	2	3	7	7	2	0	0	47	2	25	2	2	27	37	0	17	35	11	1	1	1	1	4	13
A1504	2	44	6	3	5	7	2	0	0	47	3	43	3	2	45	45	5	40	93	43	1	1	3	2	7	8
A1505	1	47	7	3	4	7	4	0	0	48	3	48	1	3	90	92	80	84	82	85	5	1	5	5	16	11
A1506	2	23	1	3	12	7	1	6	15	40	2	17	3	4	24	53	16	31	55	65	2	3	3	2	10	16
A1507	2	46	2	3	12	7	2	0	0	51	3	85	4	1	96	20	40	45	60	65	1	1	1	1	4	22
A1508	1	28	2	3	5	7	2	0	0	62	4	60	2	1	47	57	47	57	59	63	3	0	4	4	11	18
A1511	2	61	4	3	4	7	1	45	4	37	3	40	2	1	85	80	17	26	70	75	2	1	4	4	11	16
A1512	2	57	2	3	8	7	4	1	0	50	3	50	2	2	50	10	5	7	50	13	2	2	2	0	6	12

Table column name mapping to real names: ID - Patient ID, G - Gender, A - Age, D - Diagnose, LVU - Level of vocal usage, SHD - Speech (hours per day), D/W - Speech (days per week), S - Smoking, DUR - Smoking duration in years, C/D - Cigarets per day, VA - Voice assessment index, VH - Voice handicap index, ST - Stress, SG - Singing, TS - Talk time in smoke filled room, VD - Voice disorder, WV - Weak voice, LOV - Loss of voice, VR - Voice range, RS - Reduced singing, VC - Voice cracks, GFI1 - Speaking takes extra effort, GFI1 - Throat discomfort or pain after voice usage, GFI3 - Voice weakens while talking, voice fatique, GFI4 - Voice cracks or sounds different, GFI - (GFI1 + GFI2 + GFI3 + GFI4), MFT - Maximum phonation time.

UDK 004.89 + 616.22-008-07](043.3)
SL344. 2018-10-26, 14 leidyb. apsk. l. Tiražas 12 egz. Užsakymas 299.
Išleido Kauno technologijos universitetas, K. Donelaičio g. 73, 44249 Kaunas
Spausdino leidyklos “Technologija” spaustuvė, Studentų g. 54, 51424 Kaunas