SIGBio Record

Newsletter of the SIGBio
ACM Special Interest Group

Volume 8, Issue 3, Dec 2018 ISSN 2159–1210

Notice to Contributing Authors to SIG Newsletters

By submitting your article for distribution in this Special Interest Group publication, you hereby grant to ACM the following non-exclusive, perpetual, worldwide rights:

- to publish in print on condition of acceptance by the editor

- to digitize and post your article in the electronic version of this publication

- to include the article in the ACM Digital Library and in any Digital Library related services

- to allow users to make a personal copy of the article for noncommercial, educational or research purposes

However, as a contributing author, you retain copyright to your article and ACM will refer requests for republication directly to you.

SIGBIO Record - Submission Guidelines

Submission categories

Submissions to the newsletter can be either on a special issue topic or on topics of general interest to the SIGBIO community.

These can be in any one of the following categories:
- Survey/tutorial articles (short) on important topics.
- Topical articles on problems and challenges
- Well-articulated position papers.
- Review articles of technical books, products and .
- Reviews/summaries from conferences, panels and special meetings within 1 to 4 pages [1500-2500 words]
- Book reviews and reports on relevant published technical books
- PhD dissertation abstracts not exceeding 10 pages
- Calls and announcements for conferences and journals not exceeding 1 page
- News items on the order of 1-3 paragraphs

Brief announcements Announcements not exceeding 5 lines in length can simply be sent as ASCII text to the
editors by e-mail. SIGBIO Record publishes announcements that are submitted as is without review.

Announcements cannot be advertisements and should be of general interest to the wider community. The Editor reserves the right to reject any requests for announcements at his discretion.

Authors are invited to submit original research papers or review papers in all areas of bioinformatics and computational biology. The papers published in SIGBioinformatics Record will be archived in ACM Digital Library. Papers should follow the ACM format, and there is no page limitation.

http://www.acm.org/sigs/publications/proceedings-templates

Submissions should be made via email to one of the two editors Pierangelo Veltri (veltri@unicz.it) *(University Magna Graecia of Catanzaro, Italy)- Pietro Hiram Guzzi (hguzzi@unicz.it),or to ACM SIGBio email acmsigbiorecord@gmail.com*

Table of Contents

Research Papers:

Analysis of biomedical signal for patterns identification in complex diseases

Extended Abstract

Domenico Mirarchi
University Magna Graecia of Catanzaro
viale Europa, 88100
Catanzaro, Italy
d.mirarchi@unicz.it

ABSTRACT

The human body is made up of many different biological systems that give information about their health condition. These information are contained in signals generated by the biological systems. The analysis of these signals is important to extract the information supporting physicians in their daily activity and improving their knowledge on health status. We investigate on biomedical signals for data analysis, as well as for relevant information extraction. Specifically, we have investigated on biomedical signals such as vestibular signals, voice signals, biological data and EEG signals related respectively to vestibular apparatus, speech impairments and neurological diseases, human papilloma virus and geolocalization, and neurological activity. Data have been stored and organized in specific database. Finally, relevant information have been extracted by these data by using data mining techniques.

CCS Concepts

•**Computer systems organization** → **Biomedical signal;** *Signal analysis;* ; •**Networks** → signal;

Keywords

Biomedical signal, Vestibular signal, Geo-analysis, EEG, Vocal signal, Biological signal

1. INTRODUCTION

Human body can generate many types of signals that could be analyzed and used to extract relevant information on the people health's status [9]. These information can be used to monitor and check health condition as well as to implement new models for wellness and health care or for disease prediction and diagnosis. A correct acquisition and analysis of these signals are necessary to improve the knowledge on health human condition [4]. This contribution is focused on the acquisition and analysis of biomedical signals aiming to extract information useful in clinical practice to evaluate health status. In this case, the analysis has been addressed to the following signals: (i) vestibular signals (they are extracted by vestibular apparatus and they are analyzed to create a system that supports physician activity with relative information); (ii) biological signals for geolocalization (these signals are extracted by blood analysis and a system useful to store and geolocate the distribution of these data has been development); (iii) vocal signals (they are useful to evaluate vocal tract condition and they could be valid indicators for specific neurological disease's progression); (iv) electroencephalogram signals (they are used to evaluate and monitor brain activity); (v) biological signals (they are extracted by Human Papilloma Virus test and a dedicated framework has been developed to storage data and extract relevant information from them).

2. BIOMEDICAL SIGNAL ANALYSIS

This section briefly reports examples of analysis applied to the biomedical signals treated in this contribution and listed above.

Vestibular data analysis.

The vestibular system represents the sensory system that controls motor functions, contributing to the perceptions and coordination of the movements, with the aim of providing the sense of balance and spatial orientation. The damages of the Vestibular system cause specific diseases and these damages are generally estimated with the Vestibular Evoked Myogenic Potentials (VEMPs) test that produces a lot of data used to extract relavant information. The aim of the work in [5] is to design a specific algorithm and to develop a dedicated framework to support clinical activity through an easy-to-use Graphical User Interface (GUI) analyzing data produced by VEMPs tests. Two algorithms have been applied to extract a prediction regarding the distribution of a pathology based on patient gender. The Bayesian Naive classifier and K-Nearest Neighbor Classification algorithm have been chosen to extract the probability that a disease is more present for a gender; the results showed a female predisposition to lodge diseases of the auditory system compared to the male for both methods.

Biological data for geo-analysis.

Clinical records (EMRs) have been considered as a col-

ISBN 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

lection of patient's health information. EMRs store information regarding life style, activities, habits, environment, routine health screening like blood biological analysis [3]. Patients and physicians stand to benefit from Internet of Things (IoT) in healthcare. Some uses of healthcare IoT are mobile medical applications or wearable devices that allow patients to capture their health data [1]. These data can be also related to location information within a geographic context. We use clinical anonymized data extracted from a Biological Department of University Magna Graecia Hospital. We show how to include geographic operators to statistical methods and how to analyze environmental data and citizen habits to improve wellness. We report on developing and testing of a geo-based system to analyze biological data [2].

Vocal signal analysis for neurodegenerative disease.

Parkinson's Disease (PD) and Multiple Sclerosis (MS) are neurodegenerative disorders that are frequently correlated with vowel articulation difficulties. The phonation problem arises in patients affected by these diseases and it is commonly known as dysarthria. Dysarthria can be identified by vocal signal analysis [8]. This analysis is useful to: (i) perform an evaluation of the vocal signal in patients affected by neurodegenerative disease, and (ii) investigate on the correlation between dysarthria and neurodegenerative disease. The proposed method provides the acquisition and analysis of vocal signals aiming to perform feature extraction and to identify relevant patterns useful to associate impaired speech with disease [11]. Method integrates two well-known methodologies: acoustic analysis and vowel metric methodology to better define pathological voices compared to healthy ones. As a result, this method provides patterns that could be useful indicators for physicians in monitoring patients affected by neurodegenerative disease. Moreover, the proposed procedure is a valid support in early diagnosis as well as in monitoring treatment success thus improving patient's life quality.

EEG analysis to support neurological research.

Electroencephalography (EEG) is a technique for the acquisition of brain electrical signals. The increase of information acquired from signal analysis and the tools for these analysis has generated a large amount of data. In this context, mEEG prototype for EEG data managing has been presented [6]. It offers a user-friendly communication solution to exchange data between physicians and biomedical engineers. Features can be used for: (i) performing a fast diagnoses; (ii) showing reports about clinical information; (iii) storing and retrieving neurological data.

Biological data for HPV evaluation.

Human papilloma virus (HPV) is a type of infection that can be pathogenic for the human. In many cases, HPV infection can produce precancerous lesions in skin and mucous membranes in the body causing genital warts and cervical cancer [10]. The idea of the proposed contribution in [7] is to develop a dedicate framework to support clinical activity in HPV treatment. Data mining techniques have been proposed to analyze HPV data coming from the microbiology unit of Magna Graecia University. Bayesian and k-Nearest Neighbor (k-NN) algorithms have been applied to HPV data stored in a dataset aiming to extract relevant clinical information useful to support physician and biologist in HPV evaluation. Results show that k-NN algorithm allows a more accurate prediction in the gender affected by the infection compared to Bayesian algorithm. Another relevant result is that the high-risk type of virus $HPV16$ represents the most common genotype for male and female. Finally, the heat map method has been applied to observe the relevant correlation between HPV genotypes and their relative risk level.

3. CONCLUSIONS

The human body can give many information on its health status through biomedical signals. It can give relevant and useful information to improve activity on health. The topic of this study is to furnish a support for the physicians and extract relevant information by the analysis of biomedical signals. These information can improve the quality of the data and add important details for disease diagnosis and treatment.

4. REFERENCES

[1] M. Atzori, Iera. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010.

[2] G. Canino, M. Scarpino, F. Cristiano, D. Mirarchi, G. Tradigo, P. H. Guzzi, G. Cuda, and P. Veltri. Geoblood: a web based tool for geo-analysis of biological data. *Procedia Computer Science*, 98:473–478, 2016.

[3] K. Häyrinen, K. Saranto, and P. Nykänen. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304, 2008.

[4] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao. Big data application in biomedical research and health care: a literature review. *Biomedical informatics insights*, 8:BII–S31559, 2016.

[5] D. Mirarchi, C. Petrolo, G. Canino, P. Vizza, S. Cuomo, G. Chiarella, and P. Veltri. Applying mining techniques to analyze vestibular data. *Procedia Computer Science*, 98:467–472, 2016.

[6] D. Mirarchi, P. Vizza, P. Cinaglia, G. Tradigo, and P. Veltri. meeg: A system for electroencephalogram data management and analysis. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 1642–1646. IEEE, 2017.

[7] D. Mirarchi, P. Vizza, G. Tradigo, G. Di Fatta, and P. Veltri. On the use of mining techniques to analyze human papilloma virus dataset. In *Bioinformatics and Biomedicine (BIBM), 2018*, 2018.

[8] B. E. Murdoch and D. G. Theodoros. *Speech and language disorders in multiple sclerosis*. Whurr London, 2000.

[9] R. M. Rangayyan. *Biomedical signal analysis*. John Wiley and Sons, 2015.

[10] K. Valentino and C. B. Poronsky. Human papillomavirus infection and vaccination. *Journal of pediatric nursing*, 31(2):e155–e166, 2016.

[11] P. Vizza, D. Mirarchi, G. Tradigo, M. Redavide, R. B. Bossio, and P. Veltri. Vocal signal analysis in patients affected by multiple sclerosis. *Procedia Computer Science*, 108:1205–1214, 2017.

On the analysis of biomedical signals for disease classification

Patrizia Vizza
Magna Graecia University of Catanzaro
viale Europa, 88100
Catanzaro, Italy
vizzap@unicz.it

Giuseppe Tradigo
Magna Graecia University of Catanzaro
viale Europa, 88100
Catanzaro, Italy
tradigo@unicz.it

ABSTRACT

The analysis of biomedical signals and images is relevant for early diagnosis, detection and treatment of diseases. It represents the first step in the proper management of pathological conditions. Therefore, it is essential to support clinical practice during the diagnosis process by extracting relevant information and by classifying different diseases. This contribution outlines the methodologies of the most frequently used analysis techniques in biomedicine and their applications. The aim is to report about typical biosignals and bioimages and their analysis to enhance the importance of signal processing in the study and classification of specific diseases.

CCS Concepts

•**General and reference** → **Surveys and overviews;** •**Applied computing** → **Health informatics;** •**Information systems** → *Data management systems;*

Keywords

Biomedical signal, bioimage, signal analysis, disease classification

1. INTRODUCTION

Biomedical signals and images contain information about physiological events and their analysis is useful for early diagnosis, detection, management and treatment of many pathological conditions and diseases. Biomedical signal processing performs the extraction of relevant features improving the diagnosis process and the disease classification. Biomedical signal analysis is a rapidly expanding field that has become more important to the physicians due to the growing complexity of the biomedical examinations and data [9]. Quantitative, objective and automation analysis of physiological systems is essential to perform a significant challenge in disease diagnosis and classification. Methods and techniques (e.g. filtering enhancement, clustering, classification) in the process of signal analysis are in a continuous optimization to cope with ever-increasing data availability. Databases contain a large amount of data that must be analyzed to carry out relevant information through the application of these techniques. In this contribution, typical biosignals and bioimages have been treated and examples of methodologies and applications have been reported aiming to explain about the relevance of signal processing and feature extraction in medical environment for diagnosis and disease classification.

2. BIOMEDICAL SIGNALS AND BIOIMAGES

Living organisms are composed by many component systems that carry on many physiological processes [14]. These processes are complex phenomena and most of them are accompanied by or manifest themselves as biomedical signals reflecting their nature and activities. These signals could generally classified in biochemical, electrical and physical. Diseases or defects in human system cause alterations in its normal physiological processes affecting its performance, health and well-being. Biomedical signals associated to a pathological process are different from the corresponding normal signals. The analysis of biomedical signals is useful to support physician in daily activity to identify pathologies, to extract relevant information and to discover hidden pattern, to make diagnosis and follow-up procedures and to predict anomalous behavior [4]. Following, typical biomedical signals are illustrated describing their characteristics and their corresponding normal and pathological range's values.

Electrocardiogram.

The Electrocardiogram (ECG) represents the electrical activity of the heart. In a normal electrocardiogram shown in Figure 1, the amplitude of the waves is of the order of mV and the frequency varies in the range between 0.1-200 Hz [12, 5]; specifically: (i) P wave: (atrial activation with amplitude of 0.2-0.4 mV and time of 0.07-0.12 s); (ii) PR segment (total atria depolarization); (iii) PR interval (atrio-ventricular conduction time of 0.2 s); (iv) QRS complex (ventricular depolarization with amplitude of 1-2 mV and time of 0.06-0.10 s); (v) ST segment (total ventricular depolarization); (vi) T wave (ventricular repolarization with amplitude of 0.4-0.5 mV and time of 0.12-0.20 s); (vii) U wave (little positive deflection following T wave of 0.08 s). Any alteration or irregularity in the normal hearth rhythm and in its values that are not physiologically validated could be associated

DOI: 10.475/123_4

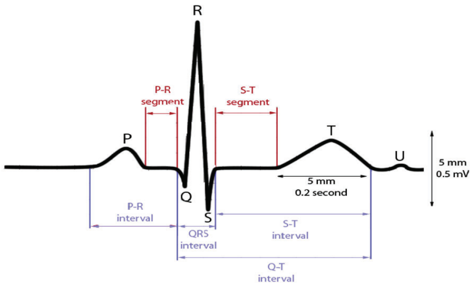

Figure 1: Normal ECG signal.

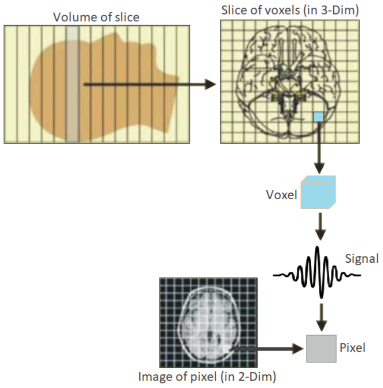

Figure 2: Spatial characteristics of bioimage.

with an anomalous and eventually pathological arrhythmia condition.

Bioimages.

Biomedical images are images of in vivo physiological processes for both diagnostic and therapeutic purposes [28]. They can be captured through advanced sensors and computer technology. CT, SPECT, PET, and MRI are examples of main biomedical imaging technologies to identify functional and anatomical condition of an organ or tissue and to monitor a patient for diagnostic and treatment evaluation. These imaging methods present some common spatial characteristics, as in Figure 2. Generally, an anatomical region is composed by slices and each slice is divided into a matrix (called array) of tissue voxels (called volume elements) [2]. A voxel (volumetric pixel or more precisely volumetric picture element) is a volume element that represents a value of signal intensity or color in a three-dimensional (3-Dim) space, similarly to the pixel (called picture elements) that represents a given two-dimensional digital image (2-Dim). The color or the brightness displayed in each pixel reflects the behavior and physical characteristic of organ or tissue in the voxel and allowing to identify any possible alterations.

Vocal signal.

The analysis of vocal signal is a non-invasive method to extract relevant information in clinical diagnosis and to detect and evaluate laryngeal pathologies. An example of vocal signal of normophonic subject it shown in Figure 3 and it represents the vocal signal of 48 years old female who pronounces the *a* vowel for few seconds. The most widely used objective parameters in vocal signal analysis to evaluate

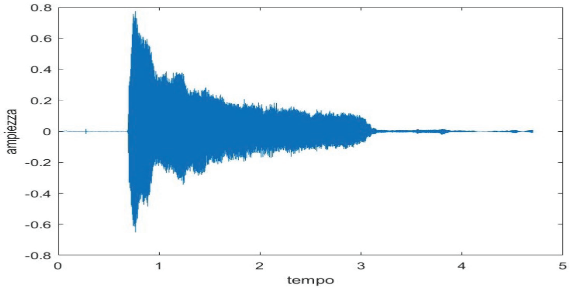

Figure 3: Example of vocal signal of normophonic female subject.

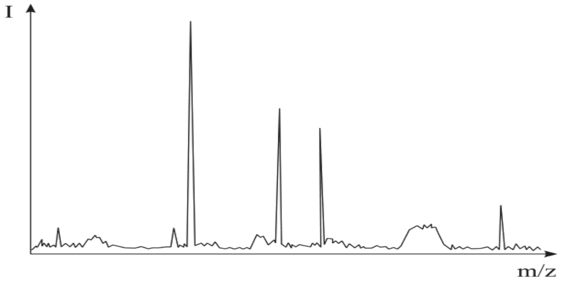

Figure 4: Example of mass spectrum of a biological sample.

normal or pathological condition are fundamental frequency (F0), jitter, shimmer and Harmonic to Noise Ratio (HNR) [17]. F0 is measured in Hertz (Hz) and it is defined as the number of times in which a sound wave produced by vocal cords is repeated during a specified time period. Jitter and shimmer are measurements of F0 disturbance defined as the frequency and the amplitude variation of the sound wave periodically, respectively. The HNR is a measurement of the voice pureness and it is expressed in decibel (dB). Pathological voice presents a higher percentage of jitter (generally, between 0.5% and 1%) and a lower percentage of shimmer (generally, less than 3%) compared to healthy voice and a less HNR (generally, less than 7 dB) [17].

Mass spectrometry data.

Mass spectrometry (MS) data are the results of an ionization process on biological samples produced by using a mass spectrometer. This instrument generates a spectrum of values for each sample and these values are represented as a set of couples [intensity, m/Z] representing a measure of quantity (intensity) of biomolecules having certain mass to charge ratio (m/Z) values, representing. An example of a generated spectrum is shown in Figure 4. Different expression of sample coming by similar biological samples helps in the identification of proteins/peptides that cause human diseases. MS-based proteomics is a widely used technique to identify different molecular targets in pathological contexts.

3. METHODOLOGIES AND APPLICATIONS

Biomedical signal and image analysis requires well defined methodologies. Generally, the common procedure provides for the following steps: acquisition, pre-processing, features extraction and results interpretation aiming to disease classification or prediction. Different methodologies can be applied to satisfy this procedure and in literature there are a lot of examples in this direction [4, 15]. Authors in [11, 16] propose the application of wavelet functions to analyze signal and image in biomedical context. Contributions in [10, 1] and in [18, 8] report on the use of artificial neural network and data mining techniques on biosignals and bioimages, respectively. Other examples can be found in [6, 7] about biomedical image classification and clustering methods.

The aim of this paper is to report some application of these methodologies for the analysis and classification of biomedical signals and images in specific clinical contexts. First, an algorithm applied on superficial ECG and intracavitary heart signals has been reported. Then, two methodologies for vocal signals analysis have been referred to classify neurological diseases. Moreover, an image analysis and disease classification based on regions of interest (ROIs) have been brief described. Lastly, mass-spectrometry based biological analysis is considered for discovering biomarkers or identifying patterns and associations in biological samples.

Atrial fibrillation prediction.

Many approaches have been proposed in literature for automated detection and classification of cardiac arrhythmias. Generally, signal processing techniques have been applied on ECG signals to extract relevant clinical information about heart activity and to identify and classify arrhythmias. An example of signal processing on ECG signals has been proposed in [24] about cardiac arrhythmia identification. In particular, Atrial Fibrillation (AFib) has been considered. AFib is one of the most common sustained arrhythmias. Superficial ECG and intracavitary signals have been extracted and analyzed to study patterns appearing in time interval preceding a fibrillation occurrence aiming to identify and predict AFib events [19]. In [21], a framework for AFib prediction during electrophysiological studies has been developed. The framework performs a workflow involving the following steps: (i) preprocessing (to acquire and pre-process the signals, e.g. through noise reduction); (ii) feature extraction (to identify waves in intracavitary signal); (iii) AFib prediction (to evaluate extracted waves and predict fibrillation events).

Neurodisease classification by vocal signal analysis.

Neurological diseases affect the brain areas involved in speech production and articulation, generating disrupted conditions as dysarthria. The analysis of vocal signals is useful to individuate, study and evaluate irregularities in speech domain and vocal performance degradation in order to predict a neurological disease or to furnish earlier indicators of its progression [22, 13]. In contribution proposed in [26], Parkinson's Disease (PD) and Multiple Sclerosis (MS) have been considered and two methods of acoustic vocal analysis and vowel metric have been applied. The aim of these methodologies is to describe pathological voices in a quantitative way identifying possible anomalies in dysarthria condition. Acoustic vocal analysis performs the extraction of four main parameters: fundamental frequency, jitter, shimmer and noise-to-harmonic ratio. Vowels metric measures vocal articulation impairment based on the values of two relevant acoustic parameters aiming to define the Vowel Space Area (VSA). Triangular VSA (tVSA) and quadrilateral VSA (qVSA) have been calculated to compare normal and pathological voices.

Neurodisease classification by bioimages analysis.

Neuroimaging analysis support clinicians in the diagnosis of neurological diseases. Positron Emission Tomography (PET) is a nuclear medicine imaging technique that produces three dimensional images of the human brain. These images are useful in neurological studies to functionally investigate brain processes. There are many software tools and statistical modules useful to perform image analysis and to extract relevant information in neurological images [27]. A framework for semi-automatic images analysis and disease classification based on regions of interest (ROIs) analysis has been proposed in [25]. The ROIs analysis has been carried out by using a Matlab tool called SPM (Statistical Parametric Mapping) and MarsBaR plugin. The proposed methodology for the analysis of PET image consists in: (i) spatial pre-processing, (ii) model specification and parameters estimation, (iii) inference. Statistical results have been used to define a ROI-based classification tool aiming to classify different neurological diseases (e.g. Parkinson disease, Alzheimer disease, Frontotemporal dementia and Multiple System Atrophy). Moreover, to automatize the preprocessing and preparation phases required to obtain statistical analysis and to extract ROIs, AutoSPET (for Automatic SPM invocation for PET images analysis) tool has been also designed and proposed as Matlab toolbox [23]. This tool includes ROI-based classification tool module for disease identification.

Diseases classification by spectrometry.

Data produced by mass spectrometry are used in proteomics experiments to identify proteins or patterns in clinical samples that may be responsible for human diseases. These samples contain a large amount of data that requires to be well retrieved and organized to reduce access time and to allow efficient knowledge extraction. Even if mass spectrometry data contain potentially huge amount of information, these data are often affected by noise and errors caused by sample preparation and instrument approximation. To reduce this problem, a preprocessing step must be applied to eliminate noise from spectra and to identify significant values. In this direction, contribution in [3] reports different techniques for mass spectrometry data preprocessing techniques applied either on a single spectrum or on an entire data set. These techniques aim to correct intensity and dedicated values in order to: (i) reduce noise, (ii) reduce amount of data, and (iii) make spectra comparable. After preprocessing application, cleaned spectra may be useful to be analyzed by using data mining techniques or they can be compared with known spectra in databases. Following the preprocessing step, another important issue in the analysis of mass-spectrometry data regards its management and querying. Contribution in [20] presents a software system to manage, share and query MS data in a distributed environment by using a spectra data management system and computational grid facilities to retrieve information.

4. CONCLUSIONS

Biomedical signal and image analysis is essential in diagnosis, treatment, management and classification of diseases. Different biosignals can be analyzed and many methods and

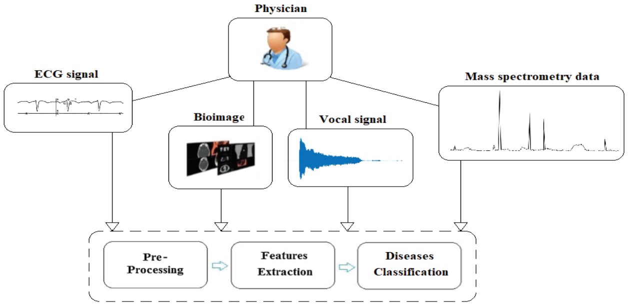

Figure 5: Block diagram of the system for biomedical signal and image analysis.

techniques can be applied in signal processing. This contribution represents a short excursus in this field, suggesting some example of application useful to identify, predict and classify pathological conditions. The idea is to used and apply the proposed methodologies and techniques in a dedicated system to support physicians in clinical practice, as shown in Figure 5. This system will be able to analyze specific types of biomedical signals and images by selecting the proper analysis and extracting relevant features.

5. REFERENCES

[1] Q. K. Al-Shayea. Artificial neural networks in medical diagnosis. *International Journal of Computer Science Issues*, 8(2).

[2] A. A. T. Bui and R. K. Taira. *Medical Imaging Informatics*. Springer Science and Business Media, 2000.

[3] M. Cannataro, P. H. Guzzi, G. Tradigo, and P. Veltri. Preprocessing of mass spectrometry proteomics data on the grid. *IEEE Symposium on Computer-Based Medical Systems*.

[4] S. Cerutti and C. Marchesi. *Advanced Methods of Biomedial Signal Processing*. John Wiley and Sons, 2011.

[5] M. B. Conover. *Understanding Electrocardiography*. Mosby, 8th. edition, 2002.

[6] S. N. Deepa and B. A. Devi. A survey on artificial intelligence approaches for medical image classification. *Indian Journal of Science and Technology*, 4(11):1583–1595, 2011.

[7] N. Dey. *Classification and Clustering in Biomedical Signal Processing*. IGI Global, 2016.

[8] G. H. Gonzalez, T. Tahsin, B. C. Goodale, A. C. Greene, and C. S. Greene. Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics*, 17(1):33–42, 2016.

[9] J. M. Gorriz, E. W. Lang, and J. Ramirez. *Recent Advances in Biomedical Signal Processing*. Bentham Science Publishers, 2011.

[10] J. Jiang, P. Trundle, and J. Ren. Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 34(8):617–631, 2010.

[11] A. Khatami, A. Khosravi, T. Nguyen, C. P. Lim, and S. Nahavandi. Medical image analysis usign wavelet transform and deep belief networks. *Expert Systems with Applications*, 86:190–198, 2017.

[12] D. Martin. *The complet study guide to learning the electrocardiogram*. Mind-Forge Education Services, 1st. edition, 2007.

[13] D. Mirarchi, P. Vizza, G. Tradigo, G. Arabia, and P. Veltri. Signal analysis for voice evaluation in parkinson's diseases. *IEEE International Conference on Healthcare Informatics (ICHI)*.

[14] R. Rangayyan. *Biomedical Signal Analysis: a Case-Study Approach*. J. Wiley Inc, 2002.

[15] J. L. Semmlow and B. Griffel. *Biosignal and medical image processing*. CRC press, 2014.

[16] S. P. Singh and S. Urooj. Wavelets: biomedical applications. *International Journal of Biomedical Engineering and Technology*, 19(1).

[17] J. P. Teixeira, C. Oliveira, and C. Lopes. Vocal acoustic analysis - jitter, shimmer and hnr parameters. *Procedia Technology*, 9.

[18] D. Tomar and S. Agarwal. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5).

[19] G. Tradigo, P. Vizza, P. H. Guzzi, A. Tagarelli, and P. Veltri. *Scientific Workflows for Healthcare*. Process Modeling and Managament for Healthcare. CRC Press, 2017.

[20] P. Veltri, M. Cannataro, and G. Tradigo. Sharing mass spectrometry data in a grid-based distributed proteomics laboratory. *Information processing and management*, 43(3).

[21] P. Vizza, A. Curcio, G. Tradigo, C. Indolfi, and P. Veltri. A framework for the atrial fibrillation prediction in electrophysiological studies. *Computer methods and programs in biomedicine*, 120(2).

[22] P. Vizza, D. Mirarchi, G. Tradigo, M. Redavide, R. B. Bossio, and P. Veltri. Vocal signal analysis in patients affected by multiple sclerosis. *Procedia Computer Science*, 108.

[23] P. Vizza, G. Tradigo, G. L. Cascini, and P. Veltri. Autospet: An spm plugin to automize neuroimages pet analysis. *Interdisciplinary Sciences: Computational Life Sciences*, 5(3).

[24] P. Vizza, G. Tradigo, A. Curcio, C. Indolfi, and P. Veltri. Intracavitary signal analysis for atria fibrillation prediction. *IEEE International Conference on Bioinformatics and Biomedicine Workshops*.

[25] P. Vizza, G. Tradigo, D. Messina, G. L. Cascini, and P. Veltri. Methodologies for the analysis and classification of pet neuroimages. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 2(4).

[26] P. Vizza, G. Tradigo, D. Mirarchi, R. B. Bossio, N. Lombardo, G. Arabia, A. Quattrone, and P. Veltri. Methodologies of speech analysis for neurodegenerative diseases evaluation. *International Journal of Medical Informatics*.

[27] P. Vizza, P. Veltri, and G. L. Cascini. A roi-based algorithm for the classification of pet neuroimages. *IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*.

[28] A. G. Webb. *introduction to Biomedical Imaging*. John Wiley and Sons, 2017.

Early radiomics experiences in predicting CyberKnife response in acoustic neuroma

Natascha Claudia D'Amico[1,2,5,*], Rosa Sicilia[2,5], Ermanno Cordelli[2,5], Giovanni Valbusa[1,5], Enzo Grossi[3], Isa Bossi Zanetti[4,5], Deborah Fazzini[1,5], Giuseppe Scotti[1,5], Giancarlo Beltramo[4], Giulio Iannello[2], and Paolo Soda[2,5]

[1]Imaging Department Centro Diagnostico Italiano S.p.A., Milan, Italy
[2]Unit of Computer Systems and Bioinformatics Università Campus Bio-Medico di Roma, Rome, Italy
[3]Bracco Imaging S.p.A., Milan, Italy
[4]CyberKnife Department Centro Diagnostico Italiano S.p.A., Milan, Italy
[5]Joint Laboratory on Precision Medicine and BioData Analytics, Università Campus Bio-Medico di Roma - Centro Diagnostico Italiano S.p.A., Italy

*Corresponding author: Natascha Claudia D'Amico, Via Saint Bon 20, 20147, Milan, Italy, nataschaclaudia.damico@cdi.it, (+39)0248317446

ABSTRACT

Vestibular schwannomas, also known as acoustic neuromas, are benign primary intracranial tumor of the myelin-forming cells of the 8^{th} cranial nerve. Stereotactic radiosurgery is one of the available therapies that can effectively control tumor growth, and it can be performed using the CyberKnife robotic device. However, this therapy may have side effects and its efficacy should be assessed up to two years. In this respect, being able to forecast the treatment response using the data collected during the initial and routinely MR images could be a valuable support when planning a personalised therapy. This manuscript therefore introduces a machine learning-based radiomics approach that first computes quantitative biomarkers from MR images and then predicts the treatment response, taking also into consideration the dataset class skewness.

Keywords

Radiomics, machine learning, acoustic neuromas, CyberKnife

1. INTRODUCTION

Acoustic neuroma is a benign, intracranical tumor caused by an overproduction of the Schwann cells surrounding the 8^{th} cranial nerve [2]. The acoustic neuroma is very rare and has an incidence rate of 1.1/100000 cases [10] every year in the US. Initial symptoms include hearing loss and tinnitus while advanced symptoms are related to the increase of the intracranial pressure.

One of the treatment possibilities is the *radiation/ stereotactic surgery (SRS)*, which can be performed using the CyberKnife device [1], a 6-degrees-of-freedom robotic radiosurgey system that minimizes the irradiation to adjacent healthy tissues due to its ability to delivery defined and adapted doses. Currently, after a CyberKnife treatment on an acoustic neuroma, it is necessary to wait up to two years to check if it has been effective or not.

Recent research in cancer treatment and precision medicine and the technological development brought to the community a promising field named as *radiomics*. Radiomics is defined as the high throughput extraction of quantitative features extracted from medical images such as CT, MR and PET, that can predict the prognosis, the tumor pathology or the heterogeneity, the therapeutic response for various conditions, thus providing valuable information for personalised therapy [13].

Radiomics, firstly presented in 2012, has been widely studied during the last years. To the best of our knowledge there are no studies focused on the acoustic neuroma, but different cancer cases have been studies with several objectives and techniques, such as the non small cell lung cancer (NSCLC) [5, 14] or the glioblastoma multiforme and the breast cancer [11, 6]. The interested readers may refer to [7, 12] for reviews on this topic.

On these premises, the objective of the study is to predict the response to CyberKnife treatment of patients suffering from acoustic neuroma by analyzing MR images acquired before the therapy. This could provide great advantages in terms of sparing radiation toxicity to patients who won't respond to the treatment.

2. MATERIALS

This study includes a cohort of 38 patients presenting an acoustic neuroma treated with CyberKnife at our center (Table 1). Patients have a follow-up up to 10 years and a mean age of 61 years. The mean volume of the included acoustic neuromas is 2999 mm^2 (range: from 251.05 mm^2 to 11788.9 mm^2).

All patients underwent a t1-w 3D MRI with contrast media (ProHance [4]) and, after a few days, a CyberKnife treatment on the monolateral acoustic neuroma. Images were acquired on a 1.5T Philips Achieva and a Signa GE 1.5T scanners. Figure 1 shows an example of patient data where the borders of the neuroma were highlighted.

An expert radiotherapist and a neuroradiologist were asked to compare the pre-treatment MRI and the last available MR image, in order to divide the population into three classes: patients with *volumetric reduction*, *stability* and pa-

Characteristic	Total	Class		
		volume reduction	no volume variation	volume increase
Number of Patients	38	25	10	3
Male/Female	18/20	11/14	5/5	2/1
Age (at treatment)				
< 50	10	5	4	1
50-70	14	10	3	1
70-90	14	10	3	1
Mean Age	61.2	63	58.3	56.3
Length of Follow-up (in years)				
y ≤ 1	3	1	1	1
1< y ≤3	10	5	3	2
3< y ≤5	10	7	3	0
5< y ≤7	10	8	2	0
7< y ≤10	5	4	1	0

Table 1: Study Population

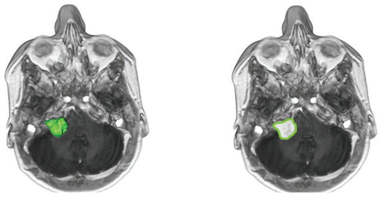

Figure 1: An example of acoustic neuroma whose borders are highlighted in green.

tients with *volumetric increase*. Tumor has been considered stable if there was no difference in the measurement (accuracy= 0.1mm). The a-priori class probabilities are 65%, 27% and 8%, respectively.

3. METHODS

Figure 2 shows a schematic representation of the proposed radiomics system. First it semi-automatically segments the images in the training set. Then it computes the pool of radiomics features (feature computation) and it applies a filter-based feature selection method. To deal with data imbalance we use an oversampling algorithm and we use the new balanced dataset to train the classification algorithm. The test phase, after image segmentation, gets as input the ROI of the test patient and it computes the descriptors selected during the training phase. Next, the trained classifier labels the patient images as *patients with volume reduction*, *patients with volume stability* or as *patients with volume increase*. In the rest of this section, we detail each of such steps.

3.1 Image Segmentation

In this study a semi-automatic tumor segmentation was carried out using the 3DSlicer image analysis software [3], which is an open source software platform for medical image processing and three-dimensional visualization.

After tumor segmentation, images containing the Volume Of Interest (VOI) were resampled with a linear resampling method to get an isotropic voxel with size equal to $1 \times 1 \times 1mm$.

3.2 Features computation and selection

Before radiomics feature extraction, for each tumor the personal data, age and gender and length of follow-up were collected. Besides semantic information, we gathered various quantitative biomarkers extracted from the VOI, i.e. the radiomic features. In total, 1135 shape-based, intensity-based and texture-based features were extracted, and they can be divided into the following categories: *Gradient Orientation Histogram*, *Gray level Cooccurrence Matrix*, *Gray Level Intensity*, *Intensity Histogram*, *Intensity Histogram Gaussian Fitting*, *Neighbor Intensity Difference* and *Shape*.

After the feature computation, a features selection step was performed using the Mann Whitney U test to pair-wisely compare the features belonging to each subgroup. We decided to differentiate between patients with a volumetric reduction (*1*, feature set F_1) from patients with a volumetric stability (*0*, feature set F_0) and a volumetric increase (*-1*, feature set F_{-1}).

Selected features fulfill the following *p-value* constrains: $p < 0.05$, in case of F_0 vs. F_1, $p < 0.05$, in case of F_{-1} vs. F_1, $p >= 0.05$, in case of F_{-1} vs. F_0. Hence we look for features which significantly differ from each other when comparing data coming from classes 1 and 0, or classes 1 and -1. When comparing samples coming from classes 0 and -1 we relaxed these constraints.

3.3 Data resampling and classification

Class imbalance refers to the problem that the training set has a disproportion among different classes, as happens in our dataset (section 2). To cope with this issue, we applied a resampling approach named as Synthetic Minority Over-sampling TEchnique (SMOTE) [9], which resizes the training set to make the class distribution more balanced, so as to match the size of the other class(es). It creates synthetic samples in the feature space along the line segments joining any/all of the k minority class nearest neighbors, randomly chosen. We set $k = 2$.

Once the dataset has been resampled we proceeded with the learning and classification block using a 10-fold cross validation and a Random Forest classifier, which has the beneficial property of avoiding overfitting [8].

4. RESULTS

The features selection step selected 427 features out of 1135. No features were extracted from the *Neighbour Intensity Difference* group, which evidences that the grey level difference does not discriminate. However texture information is detected in features belonging to the *Gray level Cooccurence Matrix*. Furthermore form the semantic features the *gender* and the *follow-up* were chosen.

The obtained results are measured in terms of accuracy, precision, recall and ROC AUC for each class (Table 2) and showed the best performance in the *volume reduction* class. This result is coherent with the features selection step. Moreover, in the *volume increase* class the precision has a higher value as the recall, which means that the system is more conservative in the prediction of results when they deviate from the original volume condition.

For completeness in Table 3, the results achieved without resampling step given by SMOTE were reported. The drop of accuracy and precision of prediction in all classes is conspicuous, such as the increased recall of the *volume increased* class shows the tendency of the classifier to prefer this group

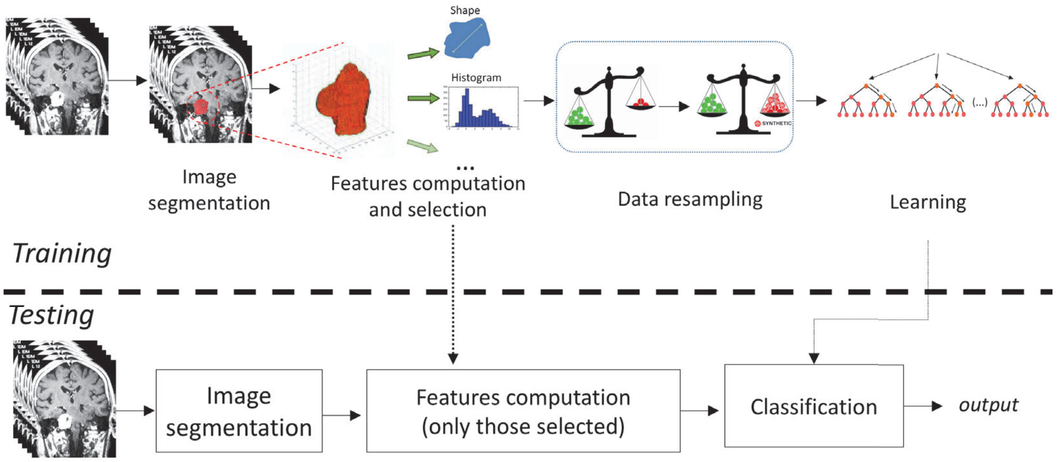

Figure 2: Schematic representation of the machine learning approach adopted.

Class	*Accuracy*	*Precision*	*Recall*	*AUC*
volume reduction (1)	85.33	1.00	1.00	1.00
no volume variation (0)		0.73	0.88	0.94
volume increase (-1)		0.85	0.68	0.92

Table 2: Results

Class	*Accuracy*	*Precision*	*Recall*	*AUC*
volume reduction (1)	63.16	0.67	0.67	0.94
no volume variation (0)		0.20	0.10	0.38
volume increase (-1)		0.70	0.84	0.66

Table 3: Results without resampling during classifier learning.

of patients. This can be expected as such class is the largest one, therefore confirming that the skeweness of the dataset biases the results.

5. CONCLUSION

This work proposed a machine-learning based radiomic pipeline that predicts the outcome of patients affected by acoustic neuroma and treated with the CyberKnife therapy. The method works with MR images routinely collected.

Although the achieved results are promising, further research is needed for several reasons, at first more patients are needed to permit to increase the robustness and investigate the effect of using images belonging to two different scanners since a small difference in pixel size is present (3%).

6. REFERENCES

[1] CyberKnife System from Accuray.
[2] Vestibular Schwannoma (Acoustic Neuroma) and Neurofibromatosis. www.nidcd.nih.gov/health/vestibular-schwannoma-acoustic-neuroma-and-neurofibromatosis.
[3] 3D Slicer. https://www.slicer.org/, Mar. 2018.
[4] ProHance® (Gadoteridol) injection, 279.3 mg/mL | Bracco Imaging, Mar. 2018.
[5] H. J. Aerts et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5:4006, 2014.
[6] A. Ahmed et al. Texture analysis in assessment and prediction of chemotherapy response in breast cancer. *Journal of Magnetic Resonance Imaging*, 38(1):89–101, 2013.
[7] M. o. Avanzo. Beyond imaging: The promise of radiomics. *Physica Medica: European Journal of Medical Physics*, 38:122–139, 2017.
[8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
[9] N. V. Chawla et al. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(3):321–357, 2002.
[10] T. J. Gal, J. Shinn, and B. Huang. Current epidemiology and management trends in acoustic neuroma. *Otolaryngology–Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, 142(5):677–681, May 2010.
[11] P. Kickingereder et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*, 280(3):880–889, 2016.
[12] G. Lee et al. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art. *European Journal of Radiology*, 86:297–307, 2017.
[13] G. Lim et al. Tumor regression and positional changes in non-small cell lung cancer during radical radiotherapy. *Journal of Thoracic Oncology*, 6(3):531–536, 2011.
[14] S. Ramella et al. A radiomic approach for adaptive radiotherapy in non-small cell lung cancer patients. *PloS one*, 13(11):e0207455, 2018.