BIOTECHNO 2019

The Eleventh International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

ISBN: 978-1-61208-717-7

June 2 - 6, 2019

Athens, Greece

BIOTECHNO 2019 Editors

Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany

BIOTECHNO 2019

Forward

The Eleventh International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2019), held between June 02, 2019 to June 06, 2019 - Athens, Greece, covered these three main areas: bioinformatics, biomedical technologies, and biocomputing.

Bioinformatics deals with the system-level study of complex interactions in biosystems providing a quantitative systemic approach to understand them and appropriate tool support and concepts to model them. Understanding and modeling biosystems requires simulation of biological behaviors and functions. Bioinformatics itself constitutes a vast area of research and specialization, as many classical domains such as databases, modeling, and regular expressions are used to represent, store, retrieve and process a huge volume of knowledge. There are challenging aspects concerning biocomputation technologies, bioinformatics mechanisms dealing with chemoinformatics, bioimaging, and neuroinformatics.

Biotechnology is defined as the industrial use of living organisms or biological techniques developed through basic research. Bio-oriented technologies became very popular in various research topics and industrial market segments. Current human mechanisms seem to offer significant ways for improving theories, algorithms, technologies, products and systems. The focus is driven by fundamentals in approaching and applying biotechnologies in terms of engineering methods, special electronics, and special materials and systems. Borrowing simplicity and performance from the real life, biodevices cover a large spectrum of areas, from sensors, chips, and biometry to computing. One of the chief domains is represented by the biomedical biotechnologies, from instrumentation to monitoring, from simple sensors to integrated systems, including image processing and visualization systems. As the state-of-the-art in all the domains enumerated in the conference topics evolve with high velocity, new biotechnologes and biosystems become available. Their rapid integration in the real life becomes a challenge.

Brain-computing, biocomputing, and computation biology and microbiology represent advanced methodologies and mechanisms in approaching and understanding the challenging behavior of life mechanisms. Using bio-ontologies, biosemantics and special processing concepts, progress was achieved in dealing with genomics, biopharmaceutical and molecular intelligence, in the biology and microbiology domains. The area brings a rich spectrum of informatics paradigms, such as epidemic models, pattern classification, graph theory, or stochastic models, to support special biocomputing applications in biomedical, genetics, molecular and cellular biology and microbiology. While progress is achieved with a high speed, challenges must be overcome for large-scale bio-subsystems, special genomics cases, bionanotechnologies, drugs, or microbial propagation and immunity.

We welcomed academic, research and industry contributions. The conference had the following tracks:

- Biomedical technologies
- Bioinformatics
- Bioenvironment

We take here the opportunity to warmly thank all the members of the BIOTECHNO 2019 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to BIOTECHNO 2019. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also thank the members of the BIOTECHNO 2019 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that BIOTECHNO 2019 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the areas of bioinformatics, biocomputational systems and biotechnologies. We also hope that Athens, Greece provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

BIOTECHNO 2019 Chairs

BIOTECHNO Steering Committee

Gilles Bernot, University Nice Sophia Antipolis, France
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA
Erliang Zeng, University of Iowa, USA
Y-h. Taguchi, Chuo University, Japan
Hunter Moseley, University of Kentucky, USA
Magnus Bordewich, Durham University, UK

BIOTECHNO Industry/Research Advisory Committee

Steffen Heber, North Carolina State University, USA
Alexandru Floares, SAIA Institute, Romania

BIOTECHNO 2019
Committee

BIOTECHNO Steering Committee

Gilles Bernot, University Nice Sophia Antipolis, France
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA
Erliang Zeng, University of Iowa, USA
Y-h. Taguchi, Chuo University, Japan
Hunter Moseley, University of Kentucky, USA
Magnus Bordewich, Durham University, UK

BIOTECHNO Industry/Research Advisory Committee

Steffen Heber, North Carolina State University, USA
Alexandru Floares, SAIA Institute, Romania

BIOTECHNO 2019 Technical Program Committee

Antonino Abbruzzo, Università degli Studi di Palermo, Italy
Hesham H. Ali, University of Nebraska at Omaha, USA
Joel P. Arrais, University of Coimbra, Portugal
Erich J. Baker, Baylor University, USA
Yoseph Bar-Cohen, Electroactive Technologies / NDEAA Lab - Jet Propulsion Laboratory (JPL), USA
Andrés Becerra Sandoval, Independent researcher, Colombia
Vincenzo Belcastro, Philip Morris International, Neuchatel, Switzerland
Kais Belwafi, King Saud University, Kingdom of Saudi Arabia
Boubaker Ben Ali, University of Bordeaux, France / University of Manouba, Tunisia
Gilles Bernot, University Nice Sophia Antipolis, France
Razvan Bocu, Transilvania University of Brasov, Romania
Magnus Bordewich, Durham University, UK
Klaus Brinker, Hamm-Lippstadt University of Applied Sciences, Germany
Rita Casadio, University of Bologna, Italy
Matthias Chung, Virginia Tech, USA
Peter Clote, Boston College, USA
Jean-Paul Comet, University Nice Sophia Antipolis, France
Maria Evelina Fantacci, University of Pisa, Italy
Alexandru Floares, SAIA Institute, Romania
Sebastian Fudickar, University of Oldenburg, Germany
Said Gaci, Sonatrach, Algeria
Xin Gao, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hamed Ghaffari, University of California, Santa Barbara, USA

Radu Grosu, Vienna University of Technology, Austria
Steffen Heber, North Carolina State University, USA
Asier Ibeas, Universitat Autònoma de Barcelona, Spain
Sohei Ito, National Fisheries University, Japan
Filip Jagodzinski, Western Washington University, USA
Xiaoqian Jiang, UC San Diego, USA
Nicolas Kemper Valverde, Universidad Nacional Autónoma de México, Mexico
Attila Kertesz-Farkas, National Research University - Higher School of Economics (HSE), Moscow, Russia
Valentinas Klevas, Lithuanian Energy Institute, Lithuania
Jan Kubicek, VSB-Technical University of Ostrava, Czech Republic
Man-Kee Lam, Universiti Teknologi PETRONAS, Malaysia
Antonio LaTorre, Universidad Politécnica de Madrid, Spain
Cedric Lhoussaine, lab CRIStAL | University Lille 1, France
Chen Li, Monash University - Melbourne, Australia
Yiheng Liang, Bridgewater State University, USA
Giancarlo Mauri, University of Milano-Bicocca, Italy
Chilukuri K. Mohan, Syracuse University, USA
José Manuel Molina López, Universidad Carlos III de Madrid, Spain
Hunter Moseley, University of Kentucky, USA
Constantin Paleologu, University Politehnica of Bucharest, Romania
Marco Pellegrini, Institute of Informatics and Telematics of CNR, Italy
Marianna Pensky, University of Central Florida, USA
Leif Peterson, Houston Methodist Research Institute, USA
Jürgen Pilz, Alpen-Adria-Universität Klagenfurt | Institut für Statistik, Austria
Yann Ponty, CNRS Ecole Polytechnique / Inria Saclay, France
Bhanu Rekepalli, BioTeam, USA
Vincent Rodin, University of Brest (UBO), France
Ulrich Rueckert, Bielefeld University, Germany
J. Cristian Salgado Herrera, University of Chile, Chile
Thomas Schmid, Universität Leipzig, Germany
Andrew Schumann, University of Information Technology and Management in Rzeszow, Poland
Jaime Seguel, University of Puerto Rico at Mayaguez, Puerto Rico
Patrick Siarry, Université Paris-Est Créteil, France
Christine Sinoquet, University of Nantes, France
Piotr Szczepaniak, Lodz University of Technology, Poland
Y-h. Taguchi, Chuo University, Japan
Bensellak Taoufik, ENSA of Tangier, System and Data Engineering Team (SDET), Morocco / IIB, UoL, Liverpool, UK
Arkadiusz Tomczyk, Lodz University of Technology | Institute of Information Technology, Poland
Sophia Tsoka, King's College London, UK
Marcel Turcotte, School of Electrical Engineering and Computer Science (EECS), Ottawa, Canada
Bing Wang, Anhui University of Technology, China
Yanshan Wang, Mayo Clinic, Rochester, USA

Copyright Information

Table of Contents

Nichrome Electrode Technology for Cell Culture Monitoring

Andrzej Kociubiński, Dawid Zarzeczny,
Maciej Szypulski, Tomasz Lizak, Krzysztof Muzyka
Institute of Electronics and Information Technology
Lublin University of Technology
Lublin, Poland
e-mail: akociub@semiconductor.pl,
dawid.adrian.zarzeczny@gmail.com,
szypulski.maciej@gmail.com, tomasz@lizak.pl,
krzysztof.muzyka.1990@gmail.com

Monika Prendecka, Dominika Pigoń,
Teresa Małecka-Massalska
Chair and Department of Human Physiology
Medical University of Lublin
Lublin, Poland
e-mail: monika.prendecka@umlub.pl,
dominika.pigon@umlub.pl,
teresa.malecka-massalska@umlub.pl

***Abstract*—The aim of this work was to present a method of tissue culture research by measuring the impedance of cells cultured in the presence of nichrome. For this purpose, the Electric Cell-substrate Impedance Sensing (ECIS) system was used with the substrate consisting of nichrome electrode arrays. The electrodes were made using a thin film magnetron sputtering. In the experimental part, the culture of cells of mouse fibroblasts on the prepared substrate was performed.**

Keywords-BioMEMS; ECIS; nichrome (NiCr); thin film.

I. INTRODUCTION

A local heating of biological substances in Biomedical MicroElectroMechanical Systems (BioMEMS) devices can be performed using a contact or noncontact method. For the contact method, the heating element is mainly fabricated using thin film deposition techniques (e.g., thermal evaporation or magnetron sputtering). The selected material should be biocompatible and should not react with the active substance. Due to the ability to withstand high temperatures, good chemical stability and biocompatibility, platinum is the most commonly used material for contact heating [1]. However, alternative metals are used for many biomedical applications, in particular for disposable structures, or for short-term applications. Some other metals offer different chemical, physical and electrical properties than platinum, which is also an expensive material. In biomedical microdevices, the heaters are also made of nickel, aluminum, tungsten, silver alloys, aluminum alloys and Indium-Tin Oxide (ITO) [2]. However, one of the most interesting materials is nichrome (Ni-Cr 80/20 wt. %), due to its high stability of electrical properties, high resistivity, low Temperature Coefficient of Resistance (TCR), adequate price and technological simplicity [3][4].

The main problem of the choice of material for the heater in biomedical applications is the assessment of the influence of its presence on the cells or substances tested [5]. The aim of this work is the presentation of the extension of the Electric Cell-substrate Impedance Sensing (ECIS) method [6] to study the activities of cells grown in tissue culture in the presence of nichrome.

The paper consists of 4 sections. Section II describes the fabrication approach of the nichrome electrode array. The results of monitoring the cells behavior in tissue culture are presented in Section III. We conclude the work in Section IV.

II. TECHNOLOGY OF NICHROME ELECTRODE ARRAY

The original ECIS method was used for the first time in 1984 by competing with microscopic methods. This impedance-based cell monitoring technology uses sterile, disposable arrays of gold electrodes placed on a biocompatible substrate [7][8]. Based on standard 8 well arrays, a mask was designed with eight electrodes located on a single substrate to work with ECIS instruments. Single electrodes were designed as comb capacitors in which the width of a single finger was 200μm.

A 2mm thick polycarbonate was used as the substrate. The key step in the sequence of technological processes was the fabrication of the metallization layer by magnetron sputtering using the Kurt J. Lesker NANO 36™ deposition tool. The next step was to obtain shapes in the lithography process and to etch the nichrome layer. Special polystyrene wells were placed on the electrodes and fixed using biocompatible silicone (Figure 1).

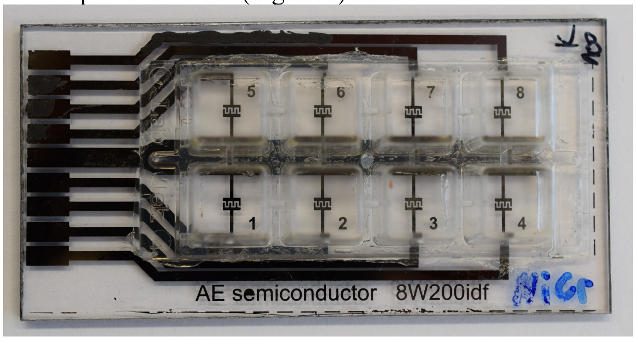

Figure 1. Polycarbonate substrate with 8 wells and electrodes made of nichrome

Each of the 8 wells with a volume of 600μL and a substrate area of 0.8cm^2 contains a single comb active electrode. For the whole setup to be sterile, ready-made substrates with attached wells are subjected to bactericidal ultraviolet radiation.

In the experiment, cells of mouse fibroblast cell line, - NCTC clone 929 [L cell, L-929, derivative of Strain L] (ATCC® CCL-1™) derived from ATCC organization were cultured according to the instruction manual in complete Eagle MEM medium (Sigma Aldrich) supplemented with 10% Fetal Bovine Serum (FBS) Good HI, in an Galaxy 170R incubator, under controlled growth conditions, constant humidity and air saturation of 5% CO_2. After (approx. 7–14 days) the culture reached at least 75% confluence, the next stage was culturing the cells on the tested nichrome electrode array. Inoculation of arrays was carried out by 300 microliters per well of cell (L929) suspension at ~1.2×10^5cell/ml. Every cell type has its characteristic adhesion and growth curve that can be manipulated by, e.g., varying seeding density or other stimuli like concentration of substances in the medium [9].

III. Results of Experiment

During the experiment, it was found that the resistance increased, reaching 4750 ohms, during the initial 10 hours of cells culturing (Figure 2). This indicates good cell viability and proliferative potential. The following drop in resistance indicates that the nichrome electrode used in the system makes it difficult to achieve stabilization in culture. After the time of 20 hours, resistance begins to fall, which should be interpreted as a progressive cell death. However, it should be noted that despite the difficulties, the fibroblast cells used in the study, as a result, maintain the growth and proliferation process (as evidenced by a stable resistance value of over 2500 ohms) in the environment of the nichrome electrode.

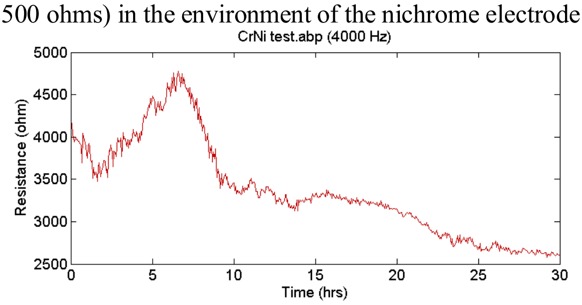

Figure 2. Resistance response measured by an ECIS sensor array at 4kHz.

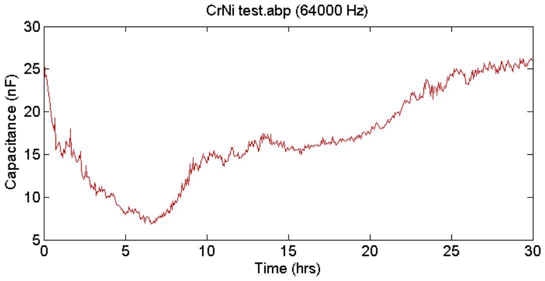

Figure 3. Capacitance response measured by an ECIS sensor array at 64kHz.

Similar fluctuation was observed for the high-frequency capacitance. The resistance represents the quality and function of the cell barrier and therefore takes into consideration the resistance towards para- and trans-cellular current flow. Capacitance provides an overall measure of electrode coverage [10]. The decrease in capacitance (Figure 3) reached during the initial 10 hours indicates fibroblast cell proliferation while the increase in the resistance should be interpreted as cell proliferation and for that matter both values complement each other and should be analyzed in parallel as a standard.

IV. CONCLUSION AND FUTURE WORK

In this paper, we demonstrated that the nichrome electrode had a significant effect on the resistance and capacitance of L929 cell line but did not kill them, that indicates the possibility to use the examined medium. However, as the studies have been carried out on cells characterized by stable growth, it is necessary to test the nichrome electrode for other types of cells, including cancerous ones.

References

[1] D. Trau et al., “Genotyping on a Complementary Metal Oxide Semiconductor Silicon Polymerase Chain Reaction Chip with Integrated DNA Microarray,” Analytical Chemistry, 2002, vol. 74, pp. 3168–3173.

[2] K. Scholten and E. Meng, “Materials for microfabricated implantable devices: a review,” Lab on a Chip, 2015, vol. 15, pp. 4256–4272.

[3] J. Rölke, “Nichrome Thin Film Technology and its Application,” ElectroComponent Science and Technology, 1981, vol. 9, pp. 51–57.

[4] R. Kilaru, Z. Celik-Butler, D. P. Butler, and I. E. Gonenli, “NiCr MEMS Tactile Sensors Embedded in Polyimide Toward Smart Skin,” Journal of Microelectromechanical Systems, 2013, vol. 22, pp. 349–355.

[5] S. S. Stensaas and L. J. Stensaas, “Histopathological evaluation of materials implanted in the cerebral cortex,” Acta Neuropathologica, 1978, vol. 41, pp. 145–155.

[6] I. Giaever and C. R. Keese, “A morphological biosensor for mammalian cells,” Nature, 1993, vol. 366, pp. 591–592.

[7] I. Giaever and C. R. Keese, “Monitoring fibroblast behavior in tissue culture with an applied electric field,” Proceedings of the National Academy of Sciences of the United States of America, 1984, vol. 81, pp. 3761–3764.

[8] ECIS Cell-based Assays from Applied BioPhysics. [Online]. Available from: http://www.biophysics.com [retrieved: 04.2019].

[9] M. Prendecka et al. “Effect of exopolysaccharide from Ganoderma applanatum on the electrical properties of mouse fibroblast cells line L929 culture using an electric cell-substrate impedance sensing (ECIS) – Preliminary study,” Annals of Agricultural and Environmental Medicine, 2016, vol. 23, pp. 280–284.

[10] R. Szulcek, H. J. Bogaard, and G. P. van Nieuw Amerongen, “Electric Cell-substrate Impedance Sensing for the Quantification of Endothelial Proliferation, Barrier Function, and Motility,” Journal of Visualized Experiments, 2014, vol. 85, pp. e51300-1–e51300-12.

Combination of Tsallis Entropy and Higutchi Fractal Dimension for Quantifying Changes in EEG signals in Alzheimer's Disease

Apostolos C. Tsolakis[2], Olga Kapetanou[1], George Petsos[1], Ioannis Nikolaidis[1], Elias C. Aifantis[1]
[1]Laboratory of Mechanics and Materials, School of Civil Engineering
[2]Laboratory of Materials for Electrotechnics, School of Electrical and Computer Engineering
Faculty of Engineering, Aristotle University of Thessaloniki
Thessaloniki, Greece
email: aptsolak@gmail.com, olgasemfe@hotmail.com, gpetsos@physics.auth.gr, iwavvns@gmail.com, mom@mom.gen.auth.gr

***Abstract*—Alzheimer's Disease (AD) is one of the challenges of modern medicine since no cure has been found yet, the scientific community still does not fully understand the pathogenesis behind it, and any interventions found can delay the progress for only a limited amount of time. Over the years, research has shifted from curing the disease to understanding the mechanisms behind it as well as finding tools that will speed up diagnosis many years before its clinical manifestation, when the decline begins. One of the many promising tools that have been explored towards that direction is the electroencephalogram (EEG), which holds many different measures that can be used as biomarkers for early diagnosis and differentiation from other neurodegenerative disorders by exploiting various bio-informatics techniques. Literature has presented a high correlation between EEG signals and structural abnormalities in AD. However, there is no analysis that can provide a clear result that binds the two and leads to early diagnosis, and very few studies have explored early stages of AD, such as Mild Cognitive Impairment. Moreover, most of the approaches applied do not adopt a multimodal methodology that combines different analysis methods. To that end, the present work proposes the combination of Tsallis Entropy and Higuchi Fractal Dimension, in a common framework for either the entire EEG or on each frequency separately, to examine the performance in Mild Cognitive Impairment (MCI) and AD subjects.**

Keywords- Alzheimer's Disease; Mild Cognitive Impairment, EEG; Tsallis Entropy; Higuchi Fractal Dimension.

I. INTRODUCTION

As stated by the International Alzheimer's Association [1], Alzheimer's Disease (AD) is the most common form of dementia. With 60% to 80% of dementia cases being diagnosed as AD, which practically means that one out of ten people over 65 has AD, it is one of the most severe diseases that affect mainly elderly people and is expected to affect roughly 131 million people by 2050. Although medicine and technology breakthroughs follow one after another, the mortality of AD keeps rising. From 2000 to 2014, an increase of 89% has been observed. To that end, an enormous amount of resources have been employed not only to postpone the progression of AD (which is still currently the only successful course of action) but to understand and thoroughly analyse the processes responsible for the brain degradation. As depicted in Figure 1, brain degradation originating from AD has severe effects both in terms of quality and volume.

As research has failed so far to grasp a cure for AD and solutions available only target symptoms and not the cause of the disease [2], effort has shifted towards better understating of the initial mechanisms that cause cognitive decline that could lead to an early diagnosis, especially at Mild Cognitive Impairment (MCI) level, which is considered a precursor stage of AD [3]. An early diagnosis may contribute not only to develop more effective interventions that could delay the progress or even inhibit it entirely but could also prevent some of the symptoms to evolve when dealt with at an early stage.

Towards the direction of early diagnosis, a handful of different methodologies have been proposed, some of which are invasive (i.e., blood) and dangerous (i.e., Cerebrospinal Fluid – CSF), others are expensive (i.e., Magnetic Resonance Imaging – MRI, Single-Photon Emission Computed Tomography – SPECT, or Positron Emission Tomography - PET), and with some still eluding significant results [4]. In contrast with these, a non-invasive, low-cost and with high resolution in terms of brain activity tool is the electroencephalogram (EEG).

A. EEG and AD

With research going back a few decades [5][6], many studies have been focused on researching the use of EEG in AD, revealing certain commonly agreed features and some other somewhat controversial [7]. The most interesting features that are commonly agreed upon in the literature regarding EEG and AD can be summarized as follows [8]-[11]: a) Overall retardation of specific rhythms, in particular, the observations so far present an increase in delta (0.1 - 4 Hz) and theta (4 – 8 Hz) activities and decrease in alpha (8 – 13 Hz) and beta (13 – 30 Hz) activities. Earliest changes are an increase in theta and a decrease in beta activities, followed by a decrease in alpha, while delta increases later during the progress of the disease. This is supported by the fact that patients with severe dementia exhibit a decrease in alpha and an increase in delta activity, whereas patients with mild dementia show a decrease in beta and an increase in theta activity, b) decreased complexity, and c) decreased coherence in general and among different brain regions. From a topographic perspective, observations indicate that slow activity is prominent in the left temporal area of AD patients, whereas differences between pre-senile patients and healthy controls are detected in the right posterior temporal area. Most significant differences between senile patients and the controls are found in the midfrontal and anterior frontal lobes bilaterally.

When evaluating complexity, significant effort has been focused on non-linear dynamics [12], under the assumption that EEG signals are generated by nonlinear deterministic processes with nonlinear coupling interactions between neurons. Studies employing such measures have found that AD patients have reduced values of the correlation dimension (D2) in the occipital EEG compared with those of healthy subjects, and with probable AD subjects [13]-[16]. In addition, it has been highlighted that AD patients exhibit reduced spatiotemporal brain activity in comparison with that in healthy controls [17], and in some cases, the former subjects are characterized by specific patterns of dysfunction in dementia [18].

Investing in the analysis of EEG complexity, a lot of novel biomarkers have been extracted from non-linear approaches (e.g., entropies, fractality, lacunarity) towards providing the necessary methods for accurate and early diagnosis of AD. This study is focused on two of them that hold promising potential and intends to combine them into a single biomarker for the intended purpose.

The manuscript is structured as follows: Section I summarizes the related work on the subjects discussed, whereas Section II presents the methodology designed to address the identified challenges. Section III describes in detail the dataset selected, and finally Section IV concludes this work with some initial findings.

II. METHODOLOGY

There are a lot of different complexity measures that have been employed in EEG signal analysis for many diseases including AD [19][20]. Two of them that have been found to hold much potential when used individually [21] are the Tsallis Entropy and the Higuchi Fractal Dimension.

A. Tsallis Entropy

The Tsallis Entropy (TE) [22] has been widely used in the analysis of EEG signals for over two decades now [23], with work on AD starting somewhere in between [24]. In multiple occasions, TE has been introduced as a possible biomarker for differentiating AD from Healthy and even MCI subjects [21][25][26].

Given a discrete set of probabilities {pi} with the condition $\sum$pi =1, and q any real number, then the Tsallis Entropy is defined as:

$$\mathrm{TE}_q = \frac{(\sum_{i=1}^{k} p_i - p_i^q)}{(q-1)} \qquad (1)$$

B. Higuchi Fractal Dimension

The Fractal Dimension (FD) as a nonlinear approach for analyzing EEG signal complexity in AD was introduced around the same time as the introduction of the TE [27]. By using it as an index of irregularity of a time series, thus evaluating time series with non-periodic and turbulent behavior [28], FD becomes a very suitable tool for EEG waveforms. Specifically in AD, FD has been found significantly lower in AD subjects when compared with healthy individuals [14][29][30]. As the basic FD can be quite processing-intense, the Higuchi Fractal Dimension (HFD) [31] has been introduced as a fast and efficient computational method that is able to successfully and accurately estimate the dimension also for segments shorter than 250 ms, thus enabling the study of brief EEG events and the identification of behavioral variations with a good temporal resolution [32][33].

For a N-sample EEG data sequence (1), (2), . . . , $x(N)$, the data is first divided into a k-length sub data set as:

$$\begin{gathered} x_k^m : \\ x(m), x(m+k), x(m+2k), \ldots, x(m+[\frac{N-m}{k}]k) \end{gathered} \qquad (2)$$

where [ ] is Gauss' notation, k is constant, and m=1,2,…,k. The length (k) for each sub data set is then computed as:

$$L_m(k) = \left\{ [\sum_{i=1}^{[(N-m)/k]} \begin{matrix} |\, x(m+ik) - x(m+(i-1)k \,|] \cdot \\ (\dfrac{N-1}{[(N-m)/k] \cdot k}) \end{matrix} \right\} \qquad (3)$$

The mean of Lm(k) is then computed to find the HFD for the data as:

$$< \mathrm{L_m(k)} >= \frac{1}{K} \sum_{M=1}^{K} L_m(k) \qquad (4)$$

In order to calculate the HFD, a least-squares linear best-fitting procedure as the angular coefficient of the linear regression of the log-log graph of $<L_m(k)> \sim k^{-HFD}$ is applied.

C. Complexity Measures Combination

As both of these measures have been tried individually in the analysis of EEG signals for AD diagnosis and are presenting rather promising results (sensitivity and specificity more than 90% when comparing only AD and normal subjects), the purpose of this study is to present an approach that will also focus on MCI stage, exploring not only the effect of these complexity measures individually but also combine them in formulating a new biomarker that is based on both for the indented purposes, as similarly suggested by [34] for depression. By employing Support Vector Machine algorithms after extracting features from applying TE and HFD on both the entire EEG bandwidth and each channel separately, we hypothesize that it will lead to enhanced performance in diagnosing and differentiating normal, MCI, and probable AD subjects.

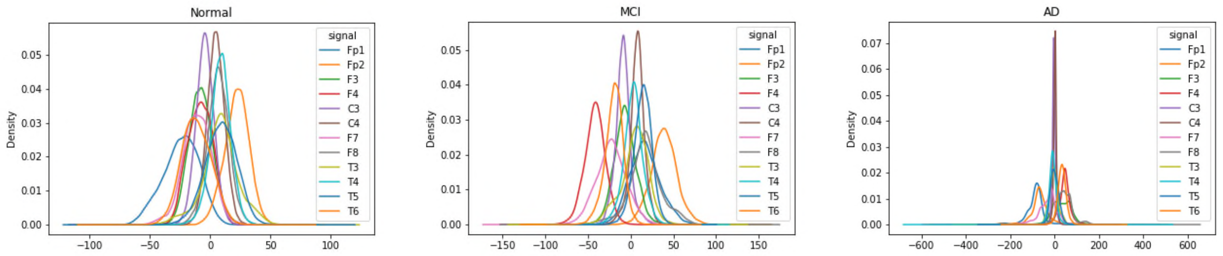

Figure 1. Whole EEG probability density functions: (a) Normal, (b) MCI and (c) probable AD subjects.

As most of the literature suggests changes of EEG signals on the frontal and temporal brain regions, initially only specific channels have been examined, namely Fp1, Fp2, F3, F4, F7, F8, T3, T4, T5, T6, C3, and T4. Since it is suggested that both MCI and AD have a different effect on the four main rhythms (delta, theta, alpha, and beta), these are also evaluated separately to identify any distinguished alterations on the proposed metrics.

The analysis of the EEG signals was performed using Python language and various open source libraries with the main ones being MNE (Minimum Norm Estimates) [35] and SciPy.

III. DATASET

This study is currently performed on EEG samples collected from a 10-20 electrode system placement [34], over 100 subjects (30 healthy, 16 probable AD, and 54 MCI) from an EEG setup with 21 electrodes, and in particular a NIHON KOHDEN Neurofax JE-921A, digitized and analysed with Neurofax EEG-1200.

Based on the available signals, the study aims to explore the potential of diagnosing and differentiating AD, starting from the initial stages of MCI.

All of the subjects were examined and diagnosed by experts at the Greek Association of AD and Related Disorders. A battery of neuro-psychometric tests has also been provided to evaluate future findings better.

The EEG signals were collected following the 10-20 placement system (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, Pz, A1, A2) [36] at 500Hz. The input impedance was set to Z < 10kΩ. The protocol used for the acquisition of the EEG signals refers to resting stage and lasts for 10 minutes with 5 minutes eyes closed and 5 minutes eyes open. An one minute window was used during eyes closed for the analysis of the EEG signals.

Low and high band pass filters have been applied to remove any artifacts prior to analysis, including a filter on 50Hz for noise from electrical equipment, whereas specific zone band filters were applied to retrieve the different EEG rhythms.

Finally, for calculating Tsallis Entropy, the probability density function was found and normalised for every examined signal.

IV. DISCUSSION & CONCLUSION

Initial findings of a first uniform sample from all three groups/classes (Healthy/Normal, MCI and probable AD) indicate significant changes between Healthy Vs. (probable) AD, and MCI Vs. AD, but only mild ones between Healthy Vs. MCI, even from the fundamental comparison of the Probability Density Functions (PDF), as can be observed in Figure 1.

By calculating the TE and HFD for each electrode and the different basic bands, for all three classes we have so far identified that it is extremely difficult for TEq and HFD (with the configuration parameters explored) individually to provide valuable insight for the differentiation between Normal/Healthy, MCI and AD subjects (Figures 2 and 3). Nevertheless, certain characteristics are in line with the literature, and thus more elaborated research is required.

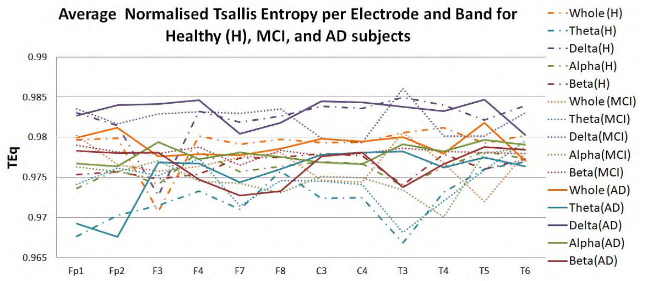

Figure 2. EEG Tsallis Entropy

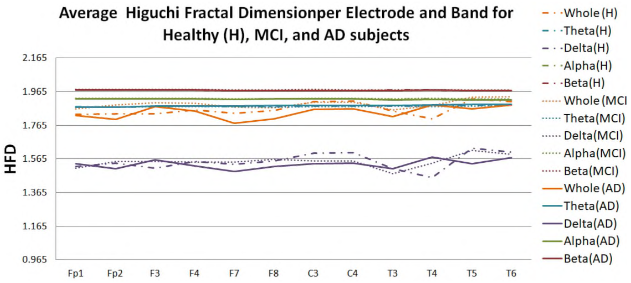

Figure 3. EEG Higuchi Fractal Dimension

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-717-7

As the presented work is an ongoing research endeavour, current and future steps involve the complete analysis of the subjects' pool with SVM and the extraction of the proper features (electrodes and bands) for maximising the accuracy of the suggested methodology.

Additional analysis of the signals is required in order to be able to provide more clear results, as well as applying additional machine learning algorithms towards evaluating the fusion of TE and HFD in a common biomarker for early diagnosis of both MCI and AD. Currently, a set of 8 subjects from each class have been used to train the models and its accuracy is being evaluated to the remaining subjects. To further enhance the reach of the analysis, the presented work will also be used for evaluating signals from a 256 electrodes set up that has many additional capabilities [14].

ACKNOWLEDGMENT

This study is co-financed by the European Union and Greek national funds through the National Strategic Reference Framework (NSRF) 2014-2020, under the call Researchers' Support with Emphasis on Young Researchers "ΕΔΒΜ34". We would also like to show our gratitude to the Greek Association of Alzheimer's Disease and Related Disorders (Alzheimer Hellas) [37] for providing us with the EEG samples and the neuro-psychometric tests.

REFERENCES

[1] International Alzheimer's Association. [Online]. Available from: https://www.alz.org/. Last Accessed 2019.04.06.

[2] A. Kumar and A. Singh, "A review on Alzheimer's disease pathophysiology and its management: an update," Pharmacological Reports, vol. 67, no. 2, pp. 195-203, 2015, doi: 10.1016/j.pharep.2014.09.004.

[3] R. C. Petersen et al., "Practice guideline update summary: Mild cognitive impairment: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology," Neurology, vol. 90, no. 3, pp. 126-135, 2018, doi: 10.1212/WNL.0000000000004826.

[4] K. Blennow and H. Zetterberg, "Biomarkers for Alzheimer's disease: current status and prospects for the future," Journal of internal medicine, vol. 284, no. 6, pp. 643-663, 2018, doi: 10.1111/joim.12816.

[5] H. Berger, "On the electroencephalogram of man," Twelfth report, Arch Psychiatr Nervenkr, vol. 106, pp.165-187, 1937.

[6] H. Wiener and D. B. Schuster, "The electroencephalogram in dementia: some preliminary observations and correlations," Electroencephalography and clinical neurophysiology, vol. 8, no. 3, pp. 479-488, 1956, doi: 10.1016/0013-4694(56)90014-1.

[7] J. Jeong, "EEG dynamics in patients with Alzheimer's disease," Clinical neurophysiology, vol. 115, no. 7, pp. 1490-1505, 2004, doi: 10.1016/j.clinph.2004.01.001.

[8] J. Dauwels, F. Vialatte, and A. Cichocki, "Diagnosis of Alzheimer's disease from EEG signals: where are we standing?," Current Alzheimer Research, vol. 7, no. 6, pp. 487-505, 2010, doi: 10.2174/156720510792231720.

[9] A. Alberdi, A. Aztiria, and A. Basarab, "On the early diagnosis of Alzheimer's Disease from multimodal signals: A survey," Artificial Intelligence in Medicine, vol. 71, pp. 1-29, 2016, doi: 10.1016/j.artmed.2016.06.003.

[10] N. Houmani et al., "Diagnosis of Alzheimer's disease with Electroencephalography in a differential framework," PloS one, vol. 13, no. 3, p.e0193607, 2018, doi: 10.1371/journal.pone.0193607.

[11] A. Tsolaki, D. Kazis, I. Kompatsiaris, V. Kosmidou, and M. Tsolaki, "Electroencephalogram and Alzheimer's disease: clinical and research approaches," International Journal of Alzheimer's Disease, 2014, doi: 10.1155/2014/349249.

[12] J. Jeong, "Nonlinear dynamics of EEG in Alzheimer's disease," Drug development research, vol. 56, no. 2, pp. 57-66, 2002, doi: 10.1002/ddr.10061.

[13] M. J. Woyshville and J. R. Calabrese, "Quantification of occipital EEG changes in Alzheimer's disease utilizing a new metric: the fractal dimension," Biological Psychiatry, vol.35, no.6, pp. 381–7, 1994, doi: 10.1016/0006-3223(94)90004-3.

[14] C. Besthorn et al., "Discrimination of Alzheimer's disease and normal aging by EEG data," Electroencephalography and Clinical Neurophysiology, vol. 103, no. 2, pp. 241–8, 1997, doi: 10.1016/S0013-4694(97)96562-7.

[15] J. Jeong, S. Y. Kim, and S. H. Han, "Non-linear dynamical analysis of the EEG in Alzheimer's disease with optimal embedding dimension," Electroencephalography and Clinical Neurophysiology, vol. 106, no.3, pp. 220-8, 1998, doi: 10.1016/S0013-4694(97)00079-5.

[16] B. Jelles, J. H. Van Birgelen, J. P. J. Slaets, R. E. M. Hekster, E. J. Jonkman, and C. J. Stam, "Decrease of non-linear structure in the EEG of Alzheimer patients compared to healthy controls," Clinical Neurophysiology, vol. 110, no. 7, pp. 1159-1167, 1999, doi: 10.1016/S1388-2457(99)00013-9.

[17] T. Yagyu et al., "Global dimensional complexity of multichannel EEG in mild Alzheimer's disease and age-matched cohorts," Dementia and geriatric cognitive disorders, vol. 8, no. 6, pp. 343-7, 1997, doi: 10.1159/000106653

[18] C. J. Stam, B. Jelles, H. A. Achtereekte, S. A. Rombouts, J. P. Slaets, and R.W. Keunen, "Investigation of EEG non-linearity in dementia and Parkinson's disease," Electroencephalography and Clinical Neurophysiology, vol. 95, no. 5, pp. 309-317, 1995, doi: 10.1016/0013-4694(95)00147-Q.

[19] N. Kulkarni and V. Bairagi, "EEG-based Diagnosis of Alzheimer Disease: A Review and Novel Approaches for Feature Extraction and Classification Techniques," Academic Press, 2018, doi: 10.1016/C2017-0-00543-8.

[20] A. Horvath, A. Szucs, G. Csukly, A. Sakovics, G. Stefanics, and A. Kamondi, "EEG and ERP biomarkers of Alzheimer's disease: a critical review," Front Biosci (Landmark Ed), vol. 23, pp. 183-220, 2018.

[21] A. H. H. Al-Nuaimi, E. Jammeh, L. Sun, and E. Ifeachor, "Complexity Measures for Quantifying Changes in Electroencephalogram in Alzheimer's Disease," Complexity, 2018, doi: 10.1155/2018/8915079.

[22] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," Journal of statistical physics, vol. 52, no. 1-2, pp. 479-487, 1988, doi: 10.1007/BF01016429.

[23] L. G. Gamero, A. Plastino, and M. E. Torres, "Wavelet analysis and nonlinear dynamics in a nonextensive setting," Physica A: Statistical Mechanics and its Applications, vol. 246, no. 3-4, pp. 487-509, 1997, doi: 10.1016/S0378-4371(97)00367-1.

[24] N. V. Thakor and S. Tong, "Advances in quantitative electroencephalogram analysis methods," Annual Review of Biomedical Engineering, vol. 6, pp. 453-495, 2004, doi: 10.1146/annurev.bioeng.5.040202.121601.

[25] P. Zhao, P. Van-Eetvelt, C. Goh, N. Hudson, S. Wimalaratna, and E. C. Ifeachor, "Characterization of EEGs in Alzheimer's disease using information theoretic methods," In Engineering in Medicine and Biology Society 29th Annual International Conference of the IEEE (EMBS 2007), pp. 5127-5131, 2007, IEEE, doi: 10.1109/IEMBS.2007.4353494.

[26] T. J. De Bock et al., “Early detection of Alzheimer's disease using nonlinear analysis of EEG via Tsallis entropy,” In Biomedical Sciences and Engineering Conference (BSEC), pp. 1-4, 2010, IEEE, doi: 10.1109/BSEC.2010.5510813.

[27] M. J. Woyshville and J. R. Calabrese, “Quantification of occipital EEG changes in Alzheimer's disease utilizing a new metric: the fractal dimension,” Biological Psychiatry, vol. 35, no. 6, pp.381-387, 1994, doi: 10.1016/0006-3223(94)90004-3.

[28] B. B. Mandelbrot, “Fractals: Form, Chance and Dimension,” New York: WH Freeman, 1979.

[29] G. Henderson et al., “Development and assessment of methods for detecting dementia using the human electroencephalogram,” IEEE Transactions on Biomedical Engineering, vol. 53, no. 8, pp. 1557-1568, 2006, doi: 10.1109/TBME.2006.878067.

[30] T. Staudinger and R. Polikar, “Analysis of complexity based EEG features for the diagnosis of Alzheimer's disease,” In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE pp. 2033-2036, 2011, IEEE, doi: 10.1109/IEMBS.2011.6090374.

[31] T. Higuchi, “Approach to an irregular time series on the basis of the fractal theory,” Physica D: Nonlinear Phenomena, vol. 31, no. 2, pp. 277-283, 1988, doi: 10.1016/0167-2789(88)90081-4.

[32] A. Accardo, M. Affinito, M. Carrozzi, and F. Bouquet, “Use of the fractal dimension for the analysis of electroencephalographic time series,” Biological Cybernetics, vol. 77, no. 5, pp. 339-350, 1997, doi: 10.1007/s004220050394.

[33] F. M. Smits et al., “Electroencephalographic fractal dimension in healthy ageing and Alzheimer’s disease,” PloS one, vol. 11, no. 2, pp. 1-16, e0149587, 2016, doi: 10.1371/journal.pone.0149587.

[34] M. Cukic et al., “EEG machine learning with Higuchi fractal dimension and Sample Entropy as features for successful detection of depression,” [Online]. Available from: https://arxiv.org/ftp/arxiv/papers/1803/1803.05985.pdf. Last Accessed: 2019.04.20.

[35] A. Gramfort et al., “MEG and EEG data analysis with MNE-Python”, Frontiers in Neuroscience, vol. 7, 2013, ISSN 1662-453X

[36] H. H. Jasper, “The ten-twenty electrode system of the International Federation”, Electroencephalography and clinical neurophysiology, vol. 10, pp. 371-375, 1958, doi: 10.1016/0013-4694(58)90053-1.

[37] Greek Association of Alzheimer’s Disease and Related Disorders (Alzheimer Hellas). [Online]. Available from: http://www.alzheimer-hellas.gr. Last Accessed: 2019.04.20.

Parallel Signal Acquisition by an Embedded System for Monitoring and Analysing Multimodal Signals of Aliveness and Non-aliveness of Biological Objects

Oliver Czepulowski, Stefan Weidemann, Anett Bailleu

Department I – Energy and Information
HTW Berlin, University of Applied Sciences
12459 Berlin, Germany
e-mail: Czepul@htw-berlin.de, Stefan.Weidemann@htw-berlin.de, Anett.Bailleu@htw-berlin.de

***Abstract*—The multimodal sensor system that is presented in this paper has been developed to measure and analyze typical information of biological objects. The heart of the system consists of a field of light emitting diodes combined with photodiodes. The light emitting diodes are characterized by several peak wavelengths in the visible and in the near infrared range. Primarily, these optoelectronic components have the function to measure the heart rate and the oxygen saturation of a person's finger. Furthermore, the special arrangement of the optoelectronic components is implemented to measure the remission of diffuse reflectance from the depth of a few millimeters of the measurement object. A hardware fusion with ancillary sensors to measure the temperature, the bio-impedance and the humidity of the measurement object is realized, too. All sensor signals are measured with a high sample rate in the presented system.**

Keywords-Multimodal sensor system; hardware fusion; additional tool for biometric application.

I. Introduction

With smart devices, watches and fitness tracker, there is a variety of utensils for the acquisition of individual data of personal health condition. Most devices use only a limited number of sensors and estimate from these data the actual variables of interest (e.g., fitness activity by pulse and Global Positioning System (GPS) coordinates; quality of sleep by the heart frequency). Often these devices only communicate online with a server, or send the resulting final value via an App to a mobile phone. Thereby, it is not possible to get a detailed insight of the original acquired data, not to mention a deeper and comprehensive analysis of these data.

In this work, we present the development of a bio-monitoring system, which makes it possible to make statements about the heart rate, the variability of the heart rate and the bio impedance, the ElectroDermal Activity (EDA) and other bio and vital parameters. Even other, non-human, biological objects, for example food, can be analyzed by this combined sensor system, so a rating about the aging, the decomposition level and a consumption recommendation could be created. Here again, other approaches use only optical analysis methods.

In Section II, we show the concept of the measurement system in principle and describe the details of the built-in hardware like the used sensors and other components of the embedded system. In Section III, we explain some features of the measurement system and sketch possible areas of application. In the end, in Section IV, we summarise the central points and give an outlook for further development and the future use.

II. System Development

A. Sytem concept

We developed a multifunctional measurement device for monitoring different parameters of biological objects. This device is used for acquisition, digital storage and for time-series visualization of the measured parameters.

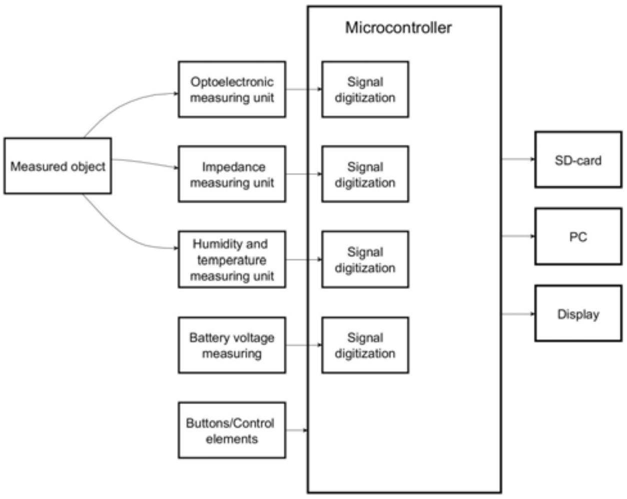

Figure 1. Block diagram of the device with the communication channels from the sensors/elements, via the microcontroller through to the data output/storage.

We designed the device as an embedded system (Figure 1). It consists of parallel channels to measure the remission, the bio-impedance, the humidity and the temperature of objects almost simultaneously.

By using a battery and an Secure Digital Memory Card (SD-card), the device operates as a stand-alone system in most cases. It can also operate as a PC-controlled unit.

B. Details of realising the sensitive hardware

The size of the measurement device should be kept small so the device is still handy (Figure 2). It contains a small chamber to measure the humidity, the temperature and the bio-impedance of a biological object, for instance a living person’s finger. Furthermore, there are Light Emitting Diodes (LEDs) and an area of photo diodes on the contact area of this chamber (Figure 3).

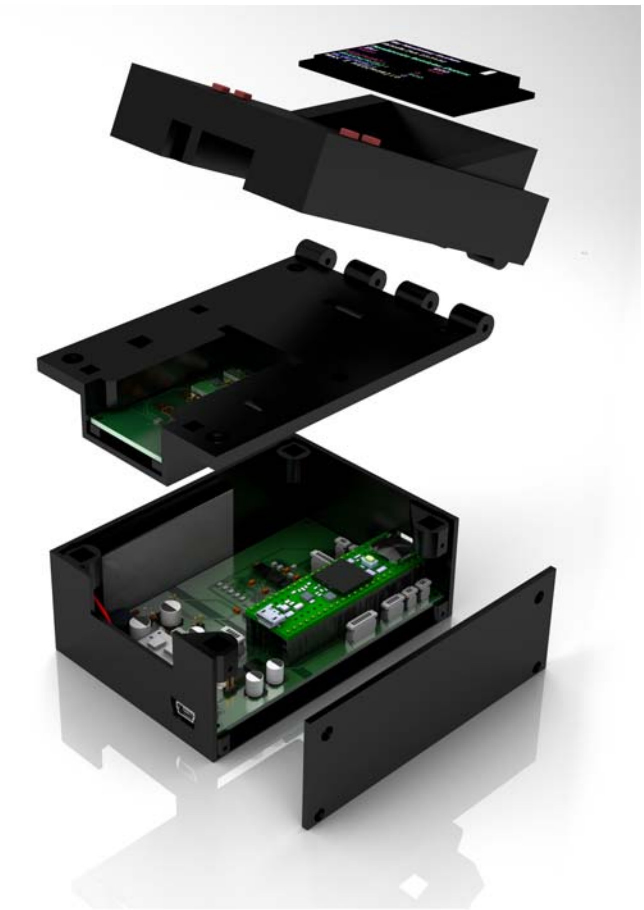

Figure 2. Design of the measurement device, useable as an about fist-sized, stand-alone, mobile monitoring and analysing system. Exploded view showing, top to bottom, the display and control elements, the measurement chamber and the electronic components, microcontroller-board, SD-card slot and USB port.

For some applications, e.g., for liveness detection in biometric applications, it is very important to get all the information of the object from a measurement of a small, spot-sized area [1]-[3]. That is why all sensors of the system are placed relatively close to each other, like illustrated in Figure 3.

The heart of the system is an optical field (upper part of Figure 3.(a)) consisting of three LEDs and five photo diodes. The LEDs are located in the center of this field, whereas the photo diodes are arranged around the center. Therefore, it is possible to detect phenomena of intrinsic absorption of the objects, in the lateral dimension and into the depth of the object [4][5].

To get information of the depth of biological objects, it is important to emit wavelengths of the so-called tissue-optical window in the range of approximately 650 – 950 nm [6][7]. We choose a 660 nm RED-LED (type: SML-LXFM0603SRC) and a 940 nm NIR-LED (type: VSMB1940X01) for this.

Furthermore, we integrated a RGB-LED (type: SMLVN6RGB) in the optical field. Hence, the user of the measurement system gets the possibility to tune an emitting wavelength of the visual light range by the software control of the RGB-LED. Optionally, an UV-LED can be placed instead of this, if it seems to be more advantageous for samples and special objects [8].

To detect the diffuse remission from the biological objects, high-sensitive photo diodes (type: ADPD2211) are used. Each of them has an integrated current-amplifier, so a separate transimpedance amplifier is not necessary.

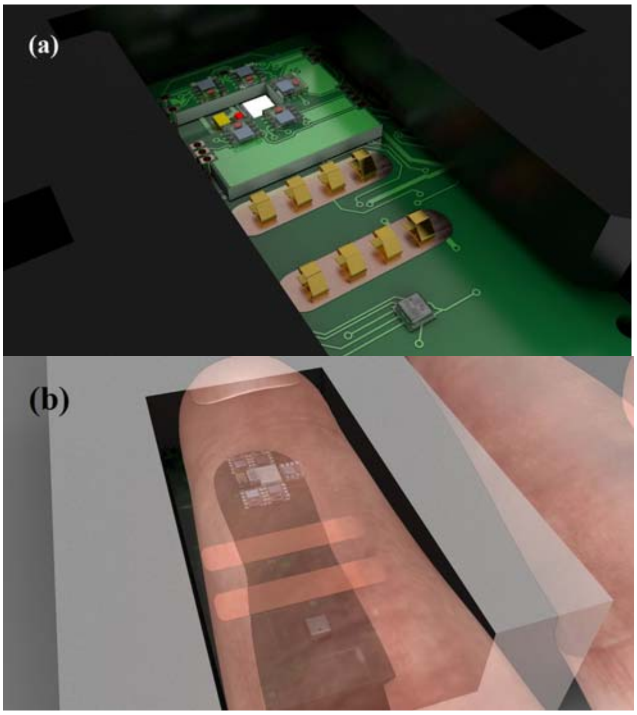

Figure 3. The heart of the measurement device. (a) measurement chamber with the optical field (three LEDs and five photo diodes), two electrodes with superimposed contact springs and a combined sensor for measuring the temperature, pressure and humidity (the grey square). (b) Finger superimposed on the contact area.

Near the optical field, two electrodes in the form of two broad metal strips can be seen (Figure 3.(a)), which are added with superimposed contact springs for a better contact to an applied finger, or sample. These electrodes are there to measure the bio-impedance [5][9]. In dependency of the distance of the electrodes on one hand and in dependency of the stimuli-voltages respective currents on the other hand, it is possible to get electrical impedance information of different depths of the biological measurement object [4][10]. In our case, the realized bio-impedance measurement works by using a constant current source. While injecting a constant current into the skin the system measures the voltage drop and the phase shift in relation to the current. Next, the system calculates the impedance from the measured values.

Finally, you find on the contact area (at the bottom of Figure 3 (b)) a combined sensor (type: BME280) to measure the humidity and the temperature of a presented measurement object (This combined sensor is able to measure the pressure, too. However, we do not use this functionality in our applications).

C. *Details of the embedded system hardware*

The measurement device, which is about fist-sized, consists of an embedded system hardware. The developed measurement device is based on a 32-bit ARM-processor, the NXP/Freescale MK64FX512VMD12Cortex-M4F processor with 120 MHz, 512 kB flash memory and 192 kB RAM. As microcontroller-board a Teensy 3.5 is used. This microcontroller is compatible to Arduino and its firmware was designed in the Arduino IDE.

The measurement device is equipped with an SD-card unit and is powered by a rechargeable battery. Therefore, this design enables a stand-alone use as a small mobile device.

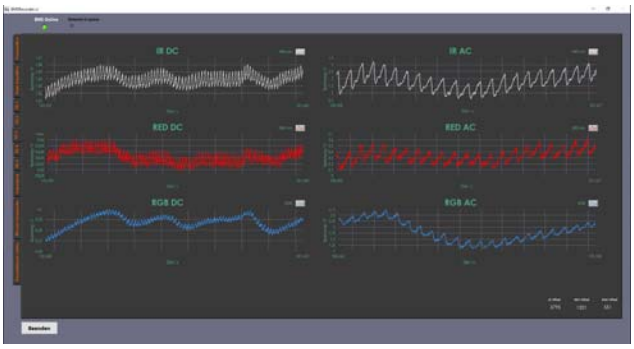

Figure 4. Graphical User Interface (GUI) of the LabVIEW-software-PC monitoring results of exemplary measuring signal courses of the reflections of the different LED wavelengths. Here, three pulse curves are shown. In addition other biological parameters, like, for instance, amplitude and phase shift of the bio impedance measurement can be displayed additionally. Also, the PC control elements are available in this LabVIEW GUI.

An USB port allows the measurement device to communicate to a PC and thereby the data export from the microcontroller to a PC is possible. So, the measurement signals can be visualized and analyzed directly and, at the same time, it is possible to store them as a base for the further development of multimodal analysis routines for various issues.

The PC Software is implemented with LabVIEW and the PC graphical user interface is shown in Figure 4. The signals of the photo diodes are shown for the three different LEDs and the DC- and AC-parts are presented separately.

III. Advantage and Benefit of the Measurement Device

In contrast to conventional pulse oximeters [1][6], which typically output averaged values of the heart rate and the oxygen saturation, our measurement device provides much more information about the status of the probed persons. For instance, the information received includes the time-resolved heart rate (and out of this the heart rate variability is identifiable), changes of the respiratory rate, the temperature and investigation of tonic and phasic EDA (variability of the bio-impedance). In addition, information about the transpiration and the peripheral blood flow through a person’s finger pad are also available. Different time regimes can be useful depending on the investigation purpose. Figure 5 shows the results of two different measurement options/variants. The figure shows the heart rate measurement results obtained with two different wavelengths. Through the respective pulse curves, it is possible to calculate the oxygenation of the blood in a noninvasive method [8]. Two alternatives to get the blood oxygenation level are presented as a function of time.

For the first method, shown in Figure 5 (a), one needs the acquisition of multiple, at least five, periods of the heart beat obtained by infrared radiation (940 nm) and save the AC and DC parts. Then one irradiate the tissue with red light (660 nm), again for at least five periods of heart beat and then calculate the oxygenation of the blood [7][8].

In the second method, shown in Figure 5(b), the system is alternating switching on and off the red (660 nm) and infrared (940 nm) LED with a high frequency, for example with 1 kHz, in equally spaced time intervals. By doing this, one can calculate the oxygenation of the blood again (theoretically after two periods of the heart beat) [6]. For both methods of oxygenation level determination, a low-pass filtering of the signals is necessary. Furthermore, the second method has the advantage to reduce the potential negative influence of environmental light/ambient light.

An additional operational scenario for this measurement device is the inspection and evaluation of food, e.g.[11]. Through the multimodal analysis of the measured signals, it is possible to obtain a lot more information about inherent material properties of biological measurement objects. For instance, it is possible to get insights about the freshness, respectively about the age of food. It is commonly known that the aging of meat, fish, cheese, eggs, vegetables and so on, causes generally the drying and colour change of these food products. Both can be measured by our device. In addition, a qualitative evaluation with respect of the composition of processed food products (ingredients, inferior food additives, preservatives, artificial colours and flavourings) can be possible. Through this analysis, a

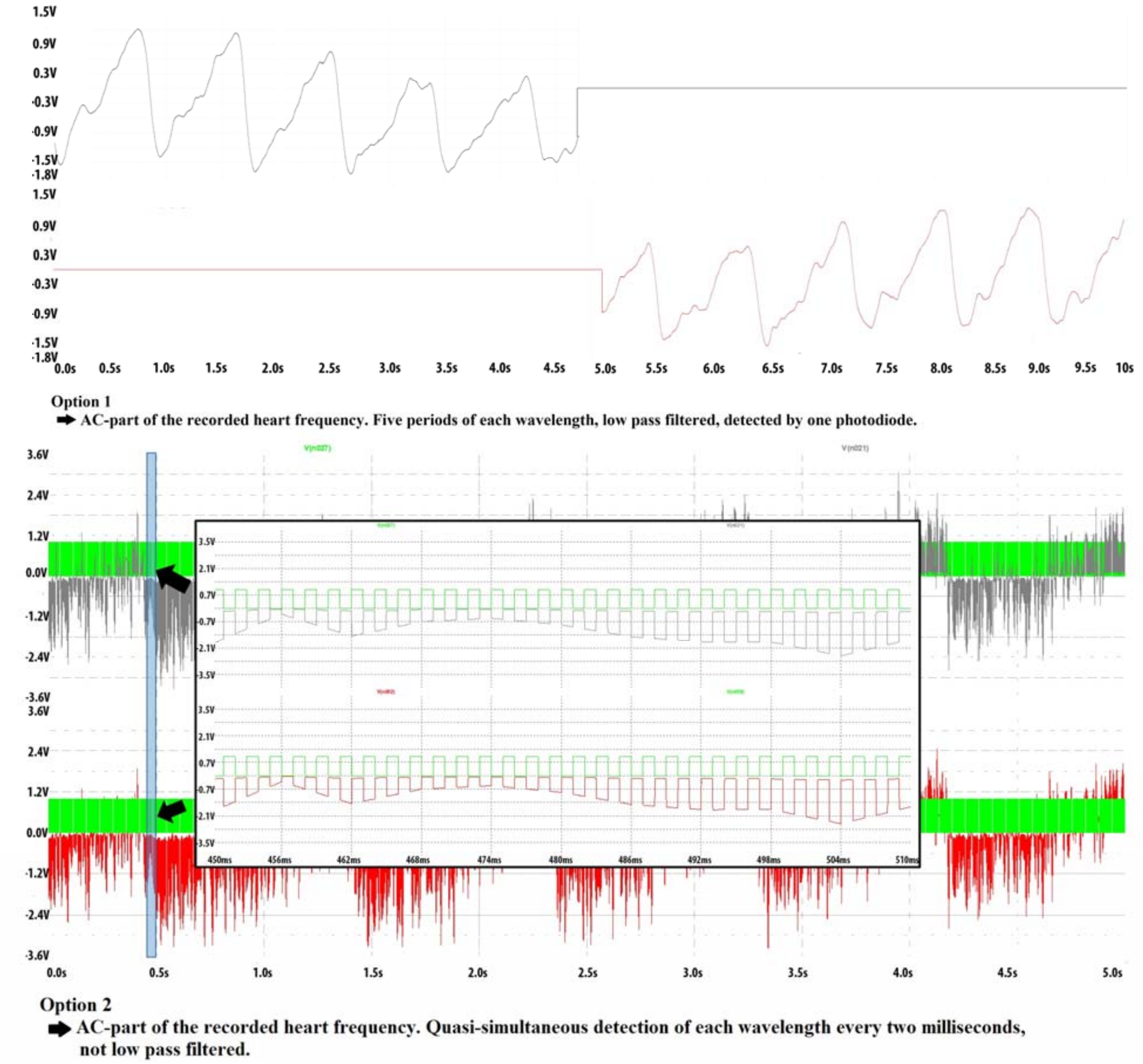

Figure 5. Two options of monitoring results- measurement graphs for different timescales of the measurement procedure.

consumption recommendation could be made, which is independent of the *best-before*-date or other manufacturer's specifications.

Within our department, the developed measurement device is used as a multifaceted measuring platform for the education of students of the "health electronics" study program, as well. Additionally, it will be used for the further development of intelligent algorithms for multimodal signal analysis, which aim to find yet unknown correlations of metabolic processes that could be useful for therapy-supporting or diagnostic procedures.

IV. Conclusions

We presented the development of a bio-monitoring system measurement device for the acquisition of a variety of information of human tissue or other biological objects. This measurement device detects the pulsatile and non-pulsatile parts by quasi-simultaneous, subsequent one after another, irradiation with light of three different wavelengths and both parts are measured and recorded quasi-parallel. Even without the pulsatile (dynamic) part, the static part of the light reflection is detected. Compared to a conventional pulse oximeter [1][6] our system is extended by electrodes for measuring the bio impedance (a constant current power supply is applied and the voltage drop and the phase shift is measured). A third essential measurement component is a combined sensor for measuring the relative humidity and the temperature on the surface of a measurement object. The bio-monitoring system is designed as a mobile, stand-alone device and therefore, it is equipped with a SD-card slot and a rechargeable battery. Data transfer is also possible via serial communication by a USB port to a PC for further analysis. We intend the development of analysis algorithms for detecting new correlations of the measured parameters on biological objects. A further development is focused to a miniaturization toward a *Lab-on-a-chip* design [12]. Hereby, a simple use for personal health data acquisition and/or a clinical use is thinkable.

ACKNOWLEDGMENT

The authors would like to thank the "Center for biomedical image and information processing (CBMI)", which enables joint research and development as a platform and promotes professional, interdisciplinary exchange and productive discussions. The project “Innovation Hub Digital Health” (IHDH), funded by the European Union (EFRE), aims at bringing together research groups of the HTW and companies in the health sector for creation, evaluation and assessment of digital data. It is linking highly specific competencies concerning medicine and health with information technology.

REFERENCES

[1] E. Kaniusas, “Biomedical Signals and Sensors I - Linking Physiological Phenomena and Biosignals”, Springer -Verlag Berlin Heidelberg, 2012, ISBN 978-3-642-24842-9; DOI: 10.1007/978-3-642-24843-6

[2] E. Kaniusas, “Biomedical Signals and Sensors II - Linking Acoustic and Optic Biosignals and Biomedical Sensors”, Springer -Verlag Berlin Heidelberg, 2015, ISBN 978-3-662-45105-2; DOI: 10.1007/978-3-662-45106-9

[3] A. Kumar, “Fingerprint spoof detection using blood flow analysis”, SPIE Newsroom, 2009, DOI:10.1117/2.1200909.1794

[4] S. Grimnes and Ø. G. Martinsen, "Bioimpedance", Wiley Encyclopedia of Biomedical Engineering, 2006; DOI: 10.1002/9780471740360.ebs0128

[5] S. Grimnes and Ø. G. Martinsen, "Bioimpedance and Bioelectricity Basics", 3rd Ed., Academic Press, ISBN: 9780124114708, 2014

[6] J. Webster (Ed.). ”Design of Pulse Oximeters”. Boca Raton: CRC Press, (1997) https://doi.org/10.1201/9781420050790

[7] E. Rabe *et al.,*: "Praktische Phlebologie", 2006, DOI: 10.1055/b-0034-8814

[8] E. Kochs and K. Zacharowski (Ed.): “anesthesia preparation and perioperative monitoring“ (“Anästhesievorbereitung und perioperatives Monitoring“), Stuttgart: Georg Thieme Verlag, pp. 113 - 115, 2015, DOI: 10.1055/b-0034-101489

[9] Data Input GmbH: "Das BIA-Kompendium", III. Ausg.; http://www.data-input.de/media/pdf-deutsch/Kompendium_III_Ausgabe_2009.pdf; Darmstadt, Germany, 2009, 2019.04.20

[10] Ø. G. Martinsen and S. Grimnes, „“Volume specific charcterization of human skin by electrical immitance“, International Patent Application WO03094724, 2003

[11] Ø. G. Martinsen, S. Grimnes, and P. Mirtaheri, “Non-invasive measurements of post mortem changes in dielectric properties of haddock muscles – a pilot study“, J. Food Eng., 43(3), pp. 189-192, 2000

[12] EP 2 435 993 B1, “Mikrosystem zur Erkennung vorbestimmter Merkmale von Wertpapieren, Sicherheitsdokumenten oder sonstigen Produkten” [In English: “microsystem for identification of predetermined features of securities, security documents and other products“] in Patentblatt 2016/27, 06.07.2016

Using Data Mining to Investigate Correlation between Traditional Chinese Medicine Body Constitution and Postnatal Problems

Winnie W. M. Lam
Department of Mathematics and Information Technology
The Education University of Hong Kong
Hong Kong
email: winnielam@eduhk.hk

Rebecca W. Y. Lee
Traditional Chinese Medicine (Private Practice)
Hong Kong
email: novembrelee@gmail.com

Regina W. S. Sit
The Jockey Club School of Public Health and Primary Care
The Chinese University of Hong Kong
Hong Kong
email: reginasit@cuhk.edu.hk

Abstract – **Many mothers experience maternal physical and mental health problems a few weeks after childbirth with unknown causes. To investigate this problem, Traditional Chinese Medicine (TCM) theory is applied to define individuals' body constitution. As constitution refers to the relatively consistent characteristics of individual body structure and body function which affect the susceptibility towards different causing factors and the tendency of disease development, it is useful to predict the postnatal problems with body constitutions. The objective of this paper is to investigate the correlation between prenatal conditions (particularly TCM body constitution) and postnatal health problems with data mining approaches. Data was collected from pregnant women aged 21 to 45 through a standardized Constitution in Chinese Medicine Questionnaire (CCMQ) before childbirth and a face-to-face assessment that was conducted by the TCM practitioner after childbirth. The collected data was analyzed by Pearson's Chi-square test and three benchmark data mining models to discover significant correlation. This study confirms a strong correlation between imbalanced TCM body constitution and postnatal problems. The discovered correlation can help to provide personalized TCM treatment to prevent the potential postnatal problems through an early TCM intervention during pregnancy to regulate the imbalance condition of pregnant women.**

Keywords—TCM body constitution; postnatal problems; data mining; chi-square test; correlation.

I. INTRODUCTION

Almost every mother experiences physical and psychological sicknesses after childbirth. Even though the results of prenatal check-up and tests, including ultrasound exams, blood tests, glucose challenge screening, etc., are normal, many women are still suffering from postnatal problems such as prolonged lochia, depression, tiredness, sleep disturbance, headache, constipation, etc. [2]-[4], while the reasons are unknown [6][7]. To ensure both mothers and babies are provided with the best care and protection, in addition to the Western medical check-up and tests, many mothers in Asian countries [8] adopt Traditional Chinese Medicine (TCM) as complementary maternal healthcare during their prenatal period, or even before becoming pregnant. They believe that TCM treatment before childbirth could help them to prevent miscarriage and sicknesses.

TCM is widely used in maternal healthcare in both Asian [8] and Western countries [9] nowadays. TCM practitioners perform predictive, preventive and personalized diagnosis [10][11] to individuals. Treatments could be different for individuals with the same symptom, but the different diagnosis of TCM syndrome differentiation is due to their different body constitution types. TCM body constitution [12] could be divided into nine types: gentleness, qi-deficiency, yang-deficiency, yin-deficiency, phlegm-damp, damp-heat, blood-stagnation, qi-stagnation and special diathesis, according to Q. Wang [18].

In the theory of body constitution in TCM, every individual has his/her body constitution. Balanced body constitution (i.e., gentleness type) is considered as the most desirable constitution which represents good health [22]. A person with balanced status is the least susceptible to illness [13] because the body can maintain homeostasis physically and psychologically, and people with imbalanced status are considered as unhealthy and prone to sickness or diseases [23] [31]. The imbalanced health status can be caused by the deficiency of "qi" (vital energy), "yang" (masculine or positive principle that is characterized by light/warmth/ dryness/activity) or "yin" (feminine or negative principle that characterized by dark/wetness/cold/passivity/disintegration), the build-up of dampness, phlegm or heat, stagnation of "qi" and blood, or special diathesis (sensitive to external stimulants such as drugs, food, smells or seasonal triggers) [24].

Prevention of disease is part of maintaining good health. The aim of this study is to discover the significant correlation between prenatal conditions (especially TCM body constitution) and postnatal problems from the personal profile for each pregnant woman from prenatal to postnatal period, so that postnatal problems can be predicted at an earlier stage. By matching the discovered correlation patterns with the prenatal conditions of pregnant women, it would be possible to predict if a mother will develop particular postpartum diseases or symptoms. This will allow TCM practitioners to provide an intervention procedure to adjust the imbalanced condition to prevent certain postnatal problems.

The rest of the paper is organized as follows. Section II describes the methods of the data collection, preparation and modelling. Section III presents the experimental results. Finally, Sections IV and V are the discussion and conclusion.

II. METHODS

A. *Selection of Participants*

The recruitment of participants was started in May, 2013 in community TCM clinics in Hong Kong. The selection criteria were women who are over 18 years old and living in Hong Kong for more than 10 years. To avoid outliers or external factors, women with serious health problems including HIV/AIDS, cancer, viral infections (e.g., Ebola, Zika virus, etc.) within a year were excluded.

B. *Building Health Profiles*

Data was collected from 1) prenatal period (i.e., the time between conception and childbirth) and 2) postnatal period (i.e., six weeks after childbirth) to build the health profiles that include the maternal health conditions of participants.

1) *Prenatal Profile*

In general medical practice, the TCM practitioner can determine the body constitution of patients by four diagnostic methods including observation (face, body form, tongue, etc.), smelling and hearing (odours from mouth/ body, loudness of speaking voice), inquiry (symptoms) and palpitation (pulse) [17]. To classify each participant into one or multiple body constitutions systematically, each patient needs to complete a standardized Constitution in Chinese Medicine Questionnaire (CCMQ) that was developed by Q. Wang [18] and China Association of Chinese Medicine [19] with the diagnosis and validation of the TCM practitioner. The CCMQ is a self-rating scale questionnaire that includes 60 questions related to a variety of prenatal conditions (e.g., cold/warm hands and feet, anxiety level, dark circles around the eyes, dry mouth, etc.). The participants should complete the questionnaire in *Google Forms* according to their physical and mental health condition in the prenatal period.

a) *CCMQ Score*

A set of CCMQ scores $S = \{s_1, s_2, s_3, \ldots, s_9\}$ that are corresponding to the nine types of TCM body constitution ($c_{1:}$ gentleness, $c_{2:}$ qi-deficiency, $c_{3:}$ yang-deficiency, $c_{4:}$ yin-deficiency, $c_{5:}$ phlegm-damp, $c_{6:}$ damp-heat, $c_{7:}$ blood-stagnation, $c_{8:}$ qi-stagnation and $c_{9:}$ special diathesis) respectively are calculated from the CCMQ by following the equations in [20]. After calculating the scores, all participants are classified into one or multiple TCM body constitution types based on Table I .

TABLE I. CLASSIFICATION OF TCM BODY CONSTITUTION TYPES

TCM body constitution	Conditions [a]	Results [a]
Balanced ($c_{1:}$ gentleness)	$s_1 \geq 60$ and $s_n < 30$	c_1
	$s_1 \geq 60$ and $s_n < 40$	Marginal c_1
	$s_n \geq 40$	Not c_1
Imbalanced (c_2: qi-deficiency c_3: yang-deficiency c_4: yin-deficiency c_5: phlegm-damp c_6: damp-heat c_7: blood-stagnation c_8: qi-stagnation c_9: special diathesis)	$s_n \geq 40$	c_n
	$30 \leq s_n \leq 39$	Tends to c_n
	$s_n < 30$	Not c_n

[a] where $n = 2$ to 9

b) *Classification of TCM Body Constitution*

In the classification of TCM body constitution types, we can identify the likelihood (in terms of CCMQ score) of each body constitution type of each participant according to Table I. Each participant may have one or more dominant body constitution types, or even have a conflict between balanced and imbalanced body constitutions. Here are three sample cases:

Case 1: Given a participant P_1 with CCMQ scores $S_1 = \{75, 56, 27, 25, 12, 15, 20, 18, 10\}$. s_1 is higher than 60, but s_2 is also higher than 40. Hence, P_1 is classified as c_2 instead of c_1.

Case 2: Given a participant P_2 with CCMQ scores $S_2 = \{75, 16, 27, 25, 32, 25, 10, 18, 10\}$. As s_1 is higher than 60, s_5 is higher than 30, and s_n is less than 40 where $n = 2$ to 9, P_2 is classified as c_1 marginally. Also, s_5 is between 30 and 39, so P_2 has a mixture of two body constitution types: marginal c_1 and tends to c_5.

Case 3: Given a participant P_3 with CCMQ scores $S_3 = \{59, 16, 27, 25, 12, 15, 20, 18, 10\}$. s_1 is less than 60, and s_n is less than 30 where $n = 2$ to 9, P_3 is classified as undefined body constitution.

2) *Postnatal Profile*

All participants who completed the CCMQ in the first stage should arrange a face-to-face assessment with a registered TCM practitioner after childbirth. The TCM practitioner needs to complete the postnatal profile that includes two parts: a) demographics and maternal health information, and b) physical and mental postnatal health problems within six weeks after childbirth.

a) *Demographics and Maternal Health Information*

In the first part, the basic demographic data, such as age, occupation and blood type, are collected, and the other questions are related to the menses cycle and volume, ovulation, delivery method, and the number of fetuses of the participants.

b) *Physical and Mental Postnatal Health Problems*

In the second part, postnatal health problems including prolonged lochia (i.e., vaginal discharge after giving birth containing blood, mucus, and uterine tissue over 21 days), tiredness, abnormal sweating, etc. are examined. Since all the questions are TCM-specific, some factors, such as tongue appearance and pulse patterns, should be examined face-to-face by a TCM practitioner.

C. *Data Preparation*

Firstly, the data of the prenatal and postnatal profiles was merged and stored into a database, which includes 132 rows (i.e., records) and 210 columns (i.e., attributes). Secondly, incorrect and inconsistent data was identified by the TCM practitioner, and the identified fields and records were either corrected or removed from the data set. Lastly, the continuous data, such as age, age of first menses, period cycle and week of childbirth, were discretized into intervals, as advised by the TCM practitioner.

D. *Identify Dependent and Independent Attributes*

Dependent and independent attributes [21] were identified from our collected data, as shown in Table II. The dependent attributes represent the outcome that we want to predict (i.e., postnatal problems), and the independent attributes represent the input of potential causes (e.g., demographics, prenatal conditions, TCM body constitution, etc.).

E. *Data Modeling*

The correlation between prenatal health conditions (especially TCM body constitution) and postnatal problems can be discovered by supervised data mining models. We used an open-source data mining tool named *Orange*, which includes these three supervised data mining models: decision tree [26], Support Vector Machine (SVM) [27] and Artificial Neural Networks (ANN) [28][29] for discovering significant patterns from the collected maternal data and comparing the effectiveness of different data mining models on TCM data analysis. The *Orange* version 3.14 was installed. To obtain the best accuracy and results, the parameters of the decision tree, SVM and ANN were set as described below.

In the decision tree modelling, the “Tree” widget was used. The parameters of the minimum number of instances in leaves, the smallest number of instance of splitting subsets and maximal tree depth were set to 2, 5 and 100, respectively. In the SVM modelling, Radial basis function (RBF) kernel was used, and the parameters of C and ε were set to 1.00 and 0.1, respectively to obtain the best accuracy. In the ANN modelling, the number of neurons per hidden layer was set to 100, and the parameters of activation, solver and alpha were set to ReLu (rectified linear unit function), Adam (stochastic gradient-based optimizer) and 0.0001 respectively.

TABLE II. DEPENDENT AND INDEPENDENT ATTRIBUTES

Dependent attributes *(postnatal problems)*	**Independent attributes**	**Remarks**
1. Tiredness	1. Age	
2. Unhealthy face colour	2. Occupation	
3. Prolonged lochia	3. Living environment	
4. Excessive sweating	4. Blood type	
5. Dry mouth	5. Genetic/ congenital disease	
6. Feeling annoyed	6. Chronic disease	
7. Joint pain	7. Gynecological disease	
8. Feeling of anger	8. Age of first menses	Related to menses
9. Feeling hot/ Hot flash	9. Regular period	
10. Emotional depression	10. Days of period cycle	
11. Fear of cold/ wind	11. Days of menustration	
12. Excessive dreaming	12. Menses volume	
13. Body pain	13. Menses colour	
14. Shortness of breath	14. Menses with blood clot	
15. Dizziness	15. Dysmenorrhea	
16. Headache	16. Number of pregnancy	Related to pregnancy and delivery
17. Flat feeling in mouth	17. Number of abortion	
18. Bright yellow urine	18. Number of miscarriage	
19. General fatigue	19. Number of fetuses	
20. Tinnitus	20. Expected date of delivery (EDD)	
21. Dry/ hard stool	21. Date of childbirth	
22. Insatiable hunger	22. Week of childbirth	
23. Constipation	23. Way of delivery	
24. Excessive urine	24. Curettage	
25. Excessive belching	25. Medication	Taken during pregnancy
26. Loose stool	26. Supplement	
27. Poor appetite	27. Number of days in hospital	After childbirth
28. Fever	28. Follow traditional rituals	
29. Palpitation	29. Breastfeeding	
30. Insomnia	30. TCM body constitution types	Obtained from CCMQ
31. Bitter taste in mouth		
32. Epigastric distemison		
33. Rectal tenesmus		
34. Stomachache		
35. Gastric excretion (from mouth)		
36. Frequent nighttime urination		

To prevent over fitting, the validation was carried out with 5-fold cross validation due to the small sample size. The data was split into a training set (80%) for building the model and test set (20%) for validating the built model. The data mining steps are summarized here:

Step 1: Import the collected data with the "File" widget.
Step 2: Define the input and output attributes by the "Select Columns" widget.
Step 3: Feed the selected data to different data mining models: "Tree" (i.e., decision tree), "SVM" and "Neural Network" (i.e., ANN).
Step 4: Perform cross-validation accuracy estimation with the "Test & Score" widget.
Step 5: Predict the unknown output with the "Predictions" widget.
Step 6: Repeat Step 2 to remove the outliers and adjust the parameter settings to obtain better accuracy.

III. RESULTS

A. *Distributions*

1) Demographics

A total of 132 pregnant women aged 21 to 45 were recruited from May, 2013 to May, 2017. Participants aged 31 to 35 was the majority group (55.30%), followed by the age groups 26 to 30 (16.67%) and 36 to 40 (18.18%). Around half of the participants (56.8%) were working indoors, and some of them were not working (16.67%).

2) Maternal Health Information

According to the pregnancy information, 59.09% of the participants were at their first pregnancy, and over 90% of them had one fetus. Although most of the mothers were reported as healthy and had regular menses cycle and normal menses volume, there were 69.70% of them who mentioned that they were suffering from different kinds of discomforts during pregnancy. To reduce the risk of miscarriage or discomfort during pregnancy, most participants (88.64%) took Western supplements such as folic acid, Materna, and a few of them (37.12%) also took Chinese medicine. When they were sick, half of them did not take either Western or Chinese medicine because they were afraid the medicine would influence the health of their fetuses.

3) Delivery Information

The delivery dates of most of our participants (39.39%) were in winter, while the others who delivered in summer were the least (13.64%). Over 90% of them had full-term pregnancy (i.e., delivered within 37 to 42 weeks), and natural childbirth was the main type (61.36%) of delivery. Most of them (78.79%) were able to leave the hospital after childbirth within 4 days, and only a few (4.55%) were staying in the hospital for more than 6 days. After the delivery, 72.73% of the participants followed the traditional rituals of TCM (e.g., no hair wash, use TCM or ginger for body wash, avoid all cold and raw food and drinks, etc.) during their postnatal period.

4) Physical and Mental Postnatal Health Problems

Within the postnatal period, our participants were suffering from different postnatal health problems. We found that 92.42% of them suffered from four or more different health problems after childbirth, and the occurrence of 16 common postnatal TCM symptoms are shown in Table III (in descending order).

5) TCM Body Constitution

In the prenatal profile, participants were classified into nine types of TCM body constitution: gentleness, qi-deficiency, yang-deficiency, yin-deficiency, phlegm-damp, damp-heat, blood-stagnation, qi-stagnation and special diathesis according to the equations in Table I. Except for the gentleness type that is considered as balanced body constitution, the other types are considered as imbalanced. In our data, 100 women were classified as imbalanced body constitution, 28 were balanced, and 4 were unclassified.

By considering the participants who were belonging to a certain body constitution type obviously (i.e., c_n is "Yes"), the top three TCM body constitution types are Yin-deficiency (37.12%), Yang-deficiency (35.61%) and blood-stagnation (35.61%).

It is important to note that not all individuals would have a single body constitution type; it is common to have a mixed body constitution (i.e., two or more body constitution types). Among 104 participants who were considered as having imbalanced body constitution, 30.77% participants had single body constitution type, 65.38% of them had two or more, and only a small proportion (3.85%) were unclassified.

TABLE III. COMMON POSTNATAL TCM SYMPTOMS

Postnatal symptoms	%	Postnatal symptoms	%
Tiredness	80.30	Feeling hot/ Hot flash	44.70
Unhealthy face colour	68.18	Emotional depression	44.70
Prolonged lochia [a]	66.67	Fear of cold/ wind	43.94
Excessive sweating	60.61	Excessive dreaming	43.94
Dry mouth	59.09	Body pain	34.85
Feeling annoyed	59.09	Shortness of breath	31.82
Joint pain	53.79	Dizziness	31.82
Feeling of anger	48.48	Headache	31.06

[a] The number of days of lochia clearance is more than 21 days

B. *Correlation between TCM Body Constitution and Postnatal Problems*

1) Statistical Test for Independence

Pearson's Chi-square test [30] was used to measure the correlations between TCM body constitution types and postnatal TCM symptoms, and the null hypothesis was applied to prove whether the occurrence of these two attributes was statistically independent. After carrying out the test in a statistical software package *SPSS*, we discovered that some postnatal symptoms from Table IV did not occur by chance, but were statistically dependent on certain TCM body constitution types. The significant pairs of attributes with a 95% confidence interval (i.e., p-value < 0.05) were listed in Table IV with the results of Cramer's V that indicates the strength statistic. The coefficient of Cramer's V ranges from 0 (no association) to 1 (strongest association). With the degree of freedom $df = 2$, the strength is considered as small, medium and large for the values 0.07, 0.21 and 0.35, respectively [32].

Within the 45 significant pairs of attributes, 16 pairs were statistically significant (i.e., significance level < 0.01) with medium/ strong strength in Cramer's V. However, the Chi-square test with Cramer's V can only determine the statistical significance between two attributes, but how exactly they are related to each other is unknown (e.g., Is qi-deficiency/ tends to qi-deficiency/ non-qi-deficiency related to emotional depression or not?). Thus, we need to use decision tree modelling to discover detailed relationships.

2) Decision Tree Modeling

From the results of Pearson's Chi-square test in Table IV, we observed that some TCM body constitution types are highly correlated to certain postnatal symptoms. To further investigate the detailed relationships between these two attributes, we selected a significant pair of attributes (qi-deficiency and dizziness) with 99% confidence interval (i.e., p-value is less than 0.01) to demonstrate how the decision tree is used to discover the correlations in the form of tree and rules.

First, the collected data was imported into *Orange*. Next, it was connected to the other widgets (i.e., Select Columns, Tree, Test & Score and Tree Viewer). In the "Select Columns" widget, the TCM body constitution "qi-deficiency" and the postnatal symptom "dizziness" were set as input and output, respectively. Finally, we obtained the decision tree model (in Figure 1) with an overall classification accuracy of 72.7%.

TABLE IV. SIGNIFICANT RESULTS OF PEARSON'S CHI-SQUARE TEST

Attributes		Significance level (< 0.01)	Cramer's V Result (> 0.35)
TCM body constitution types	**Postnatal symptoms**		
gentleness	Excessive sweating	0.045	
qi-deficiency	Tiredness	0.026	
	Shortness of breath	0.013	
	Feeling hot/ Hot flash	0.036	
	Headache	**0.004**	0.292
	Dizziness	**0.000065**	**0.382**
	Emotional depression	**0.002**	0.309
	Excessive urine	0.011	
	General fatigue	0.04	
	Joint pain	0.012	
	Body pain	**0.009**	0.267
Yang-deficiency	Dizziness	0.041	
	Fear of cold/wind	**0.000245**	**0.355**
	Stomachache	0.044	
Yin-deficiency	Bitter taste in mouth	0.031	
	Dry mouth	**0.002**	0.301
	Dry/hard stool	0.035	
phlegm-damp	Dizziness	0.013	
	Bitter taste in mouth	**0.000011**	0.345
	Flat feeling in mouth	0.012	
	Feeling annoyed	0.03	
	Bright yellow urine	0.043	
	Frequent nighttime urination	**0.00004**	**0.392**
damp-heat	Dizziness	0.024	
	Bitter taste in mouth	**0.00004**	0.345
	Emotional depression	0.042	
	Frequent nighttime urination	0.01	
blood-stagnation	Shortness of breath	0.035	
	Dizziness	**0.003**	0.299
	Bitter taste in mouth	0.033	
	Dry mouth	0.027	
	Flat feeling in mouth	**0.006**	0.276
	Feeling annoyed	**0.001**	0.317
	Feeling of anger	0.013	
	Face colour	0.049	
qi-stagnation	Fever	**0.001**	0.325
	Dizziness	**0.001**	0.337
	Excessive belching	0.014	
	Bright yellow urine	0.025	
	Body pain	0.04	
special diathesis	Dizziness	**0.005**	0.283
	Bitter taste in mouth	0.012	
	Feeling annoyed	0.033	
	Rectal tenesmus	0.024	
	Prolonged lochia	**0.003**	0.295

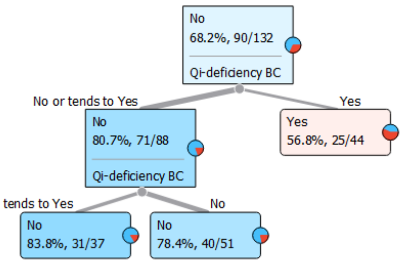

Figure 1. Relationships between qi-deficiency and dizziness in the decision tree

In this output of the decision tree model, we discovered three relationships between qi-deficiency and dizziness by following the paths from the root node to the leaf nodes. These paths are represented in the form of IF-THEN rules, where *n* refers to the number of participants who matched the following rules:

Rule 1: IF qi-deficiency = "tends to Yes" THEN dizziness = "No" (*n* = 31, accuracy = 83.8%)

Rule 2: IF qi-deficiency = "No" THEN dizziness = "No" (*n* = 40, accuracy = 78.4%)

Rule 3: IF qi-deficiency = "Yes" THEN dizziness = "Yes" (*n* = 25, accuracy = 56.8%)

In addition to the one-to-one correlation between qi-deficiency and dizziness discovered by Pearson's Chi-square test, a decision tree can define multiple inputs to find whether the output attribute (i.e., dizziness) is correlated to one or more input attributes. From Table IV, we found that dizziness is not only strongly correlated to qi-deficiency, but it is also correlated to the other six imbalanced TCM body constitution types. Instead of setting one input attribute in the decision tree, we set all eight imbalanced TCM body constitution types as input to predict the postnatal dizziness.

Among the eight imbalanced TCM body constitution types, qi-deficiency was selected as the root node (i.e., best predictor) with the highest information gain. The leaf node with the highest classification accuracy and number of samples was extracted from the rightmost part of the tree and shown in Figure 2. The leaf node shows that all women with multiple TCM body constitution types: qi-deficiency, special diathesis and tends to phlegm-damp were suffering from dizziness in the postnatal period. This proved that dizziness is related to multiple TCM body constitution types instead of one only. Since there were over 60% of women with multiple body constitution types and different prenatal conditions, a decision tree with multiple input attributes is the better way to predict postnatal problems.

In order to discover more possible prenatal factors that cause the postnatal problems, a decision tree and two other supervised data mining algorithms were applied to our data, including the attributes of TCM body constitution, demographics, maternal health information and delivery information, collected from the prenatal period to predict the postnatal problems. The independent (i.e., input) and dependent (i.e., output) attributes are listed in Table II.

C. *Significant Patterns with Data Mining Algorithms*

Supervised data mining was used to discover significant patterns from prenatal and postnatal data. The classification accuracies of predicting 36 postnatal problems, as shown in Table II, are reported in Figure 3. The classification accuracy of each postnatal problem (target attribute) in the decision tree, SVM and ANN are indicated by blue triangle dot, red circle dot and grey cross, respectively.

Half of the postnatal problems can be predicted with classification accuracy over 70%, and the postnatal problem no. 36 (i.e., frequent nighttime urination) has the highest classification accuracy of 92.27%. Overall, the average accuracies of the three supervised data mining algorithms are above 64%, and SVM has the highest average accuracy of 73.07%.

To predict the postnatal problem of frequent nighttime urination, the ROC curves of the decision tree, SVM and neural network are given in Figure 4. It shows the performance of neural network is the best among three with higher AUC, followed by decision tree, and lastly SVM.

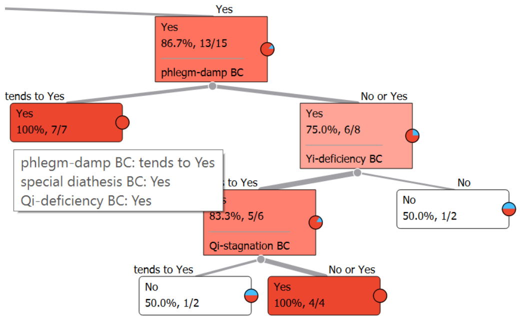

Figure 2. Leaf node with the highest accuracy rate to predict dizziness

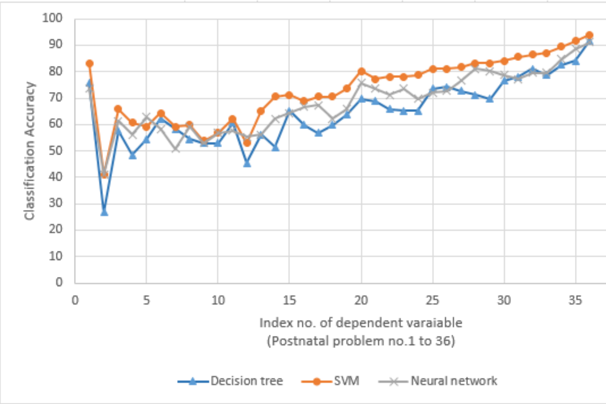

Figure 3. Classification accuracies of 36 postnatal problems

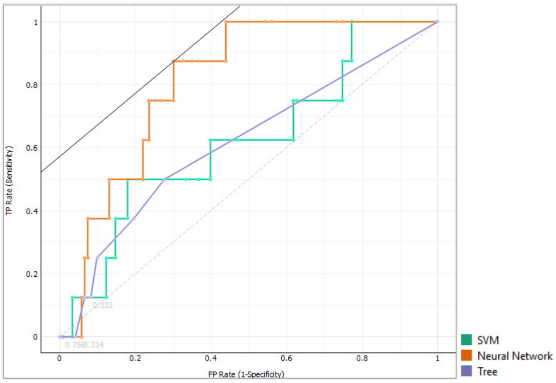

Figure 4. ROC curves of decision tree, SVM and neural network

After consulting the TCM practitioner, four postnatal problems, including dizziness, prolonged lochia, tiredness, and emotional depression, were identifisssed as important to clinical care after pregnancy. We used a decision tree to discover the correlations between these postnatal problems and prenatal factors because it can show the results in terms of readable rules which are understandable by TCM practitioners.

1) *Classification Rules of Dizziness*

As mentioned earlier, dizziness is highly correlated to multiple TCM body constitution types. From *Rule 1.1* to *Rule 1.4*, we can predict if a woman will suffer from dizziness after pregnancy related to five prenatal factors: 1) the number of TCM body constitution type, 2) TCM body constitution types (including qi-stagnation, Yang-deficiency, Yi-deficiency and special diathesis), 3) blood type, 4) living environment, and 5) number of days of menstruation before pregnancy.

Rule 1.1: IF the number of TCM body constitution types > 5 THEN postnatal problem = "dizziness" (*n* = 13, accuracy = 86.7%)

Rule 1.2: IF the number of TCM body constitution types ≤ 5 AND blood type = "AB, O or unknown" AND living environment = "Above the hill" THEN postnatal problem = "dizziness" (*n* = 5, accuracy = 100%)

Rule 1.3: IF the number of TCM body constitution types ≤ 5 AND days of menstruation ≤ 7 AND living environment ≠ "above the hill", THEN postnatal problem ≠ "dizziness" (*n* = 85, accuracy = 80.2%)

Rule 1.4: IF the number of TCM body constitution types ≤ 5 AND TCM body constitution type = "qi-stagnation, Yang-deficiency, Yi-deficiency or special diathesis" AND days of menstruation ≤ 7 AND living environment ≠ "above the hill", THEN postnatal problem ≠ "dizziness" (*n* = 18, accuracy = 100%)

By comparing *Rule 1.3* and *Rule 1.4*, we observed that when the additional factor, TCM body constitution, is considered when predicting dizziness, it can yield higher accuracy. Hence, the TCM body constitution is an important factor that can affect the accuracy of predicting dizziness.

2) *Classification Rules of Prolonged Lochia*

Prolonged lochia is the third most common postnatal problem in this study, and the abnormal duration of bleeding can be suspected as infection. According to *Rule 2.1* to *Rule 2.4*, we can detect this problem in the prenatal period by observing seven prenatal factors: 1) TCM body constitution type (special diathesis and qi-stagnation), 2) days of period cycle, 3) traditional rituals (avoid all cold/raw food), 4) tongue coating, 5) EDD, 6) blood type and 7) dysmenorrhea before or after menstruation.

Rule 2.1: IF TCM body constitution type (special diathesis) ≠ "tends to Yes" AND avoid all cold/raw food = "No" THEN postnatal problem = "prolonged lochia" (*n* = 53, accuracy = 81.5%)

Rule 2.2: IF TCM body constitution types ≠ "qi-stagnation or special diathesis" AND avoid all cold/raw food = "No" AND days of period cycle > 31 THEN postnatal problem = "prolonged lochia" (*n* = 25, accuracy = 100%)

Rule 2.3: IF TCM body constitution type (special diathesis) = "tends to yes" AND EDD = "spring or summer" THEN postnatal problem ≠ "prolonged lochia" (*n* = 7, accuracy = 100%)

Rule 2.4: IF TCM body constitution (special diathesis) ≠ "tends to yes" AND avoid all cold/raw food = "Yes" AND blood type = "O or unknown" AND dysmenorrhea = "occurred before or after menstruation" THEN postnatal problem ≠ "prolonged lochia" (*n* = 7, accuracy = 100%)

3) *Classification Rules of Emotional Depression*

Emotional depression is one of the top 10 postnatal problems, and it is a type of mood disorder associated with childbirth. It can negatively affect the mother, her family and newborn child. This problem can be detected by monitoring five factors: 1) TCM body constitution type, 2) blood vessel under the tongue, 3) days of menstruation, 4) menses, and 5) dysmenorrhea based on *Rule 4.1* to *Rule 4.4*:

Rule 4.1: IF days of menstruation > 4 AND age of the first menses ≠ "11–12" AND TCM body constitution type (qi-deficiency) = "yes or tends to yes" THEN postnatal problem = "Emotional depression" (*n* = 31, accuracy = 81.6%)

Rule 4.2: IF menses with blood clot = "no" AND menses colour ≠ "bright red or pale red" AND age of first menses = "11-12" AND TCM body constitution type (qi-deficiency) = "yes or tends to yes" THEN postnatal problem = "Emotional depression" (*n* = 7, accuracy = 87.5%)

Rule 4.3: IF TCM body constitution type = "complex, unclassified, Yang-deficiency or Yi-deficiency" AND qi-deficiency = "no" THEN postnatal problem ≠ "Emotional depression" (*n* = 29, accuracy = 90.6%)

Rule 4.4: IF Menses colour = "bright red or pale red" AND Age of first menses = "11–12" AND TCM body constitution type (qi-deficiency) = "yes or tends to yes" THEN postnatal problem ≠ "Emotional depression" (*n* = 5, accuracy = 100%)

IV. Discussion

In this study, we found that 92.42% of the women suffered from four or more different health problems after childbirth. Women with "gentleness" body constitution type tended to have less postnatal discomfort; the finding was consistent with the TCM theory that individuals with balanced body constitution are more likely to stay healthy even in the stressful postnatal period. Contrarily, women with imbalanced TCM body constitution were found to have more postnatal problems, among which qi-deficiency and blood stagnation had the worse postnatal profiles.

In our analysis, several factors were found to predict the outcome of dizziness. However, some factors, such as the week of childbirth, duration of period cycle and menses volume, were unmodifiable. Other factors, namely traditional rituals and the intake of calcium supplements could be changed easily through prenatal education and advice. Additionally, although TCM body constitution is constant and not easily changed at once, it can be adjusted in the long run through lifestyle modification, nutrition, exercise and herbal medications. Therefore, we believe that by knowing individuals' body constitution and their prenatal habits could potentially improve women's postnatal health.

There are several limitations to our study. First, the study was conducted in Hong Kong, and only women with Chinese ethnicity were invited; therefore, the findings may not be generalizable. Second, the correlations between TCM body constitution and postnatal problems were not one-to-one relationships; in both Pearson's Chi-square test and the three supervised data mining algorithms (SVM, decision tree and artificial neural networks), we can only set one attribute (one of the postnatal problems) as output at one time. Third, TCM body constitution is potentially adjustable before pregnancy; a future study can be conducted on the prediction of the other disease entities [13][14], especially emotional depression. Additional factors can be collected from Western medical records for building a better predictive model.

V. CONCLUSION

In this study, we discovered that women with imbalanced TCM body constitution were more likely to suffer from postnatal health problems. Some body constitution types and prenatal habits were predictive of certain common and important postnatal symptoms, which were believed to be modifiable. Supervised data mining has been demonstrated as a useful way to discover correlations between multiple prenatal conditions (especially body constitutions) and postnatal problems. After discovering the significant correlations in terms of classification rules, the future step is to predict postnatal problems before childbirth. By using the associations between body constitution and postnatal problems in Table IV and the classification rules in Section III C), the TCM practitioners can apply early interventions, such as medicine, acupuncture, suggesting nutrition or supplements, to adjust individuals' body constitution and prevent the postnatal problems during the prenatal period, or even before pregnancy.

REFERENCES

[1] D. Neagos, R. Cretu, R. C. Sfetea and L. C. Bohiltea, "The Importance of Screening and Prenatal Diagnosis in the Identification of the Numerical Chromosomal Abnormalities," Mædica, 6(3), pp. 179-184, 2011.

[2] S. Brown and J. Lumley, "Maternal health after childbirth: results of an Australian population based survey," An International Journal of Obstetrics and Gynaecology, vol. 105 (2), pp. 156-161, February 1998.

[3] J. Y. Lee and J. Y. Hwang, "A study on postpartum symptoms and their related factors in Korea," Taiwanese Journal of Obstetrics and Gynecology, vol. 54 (4), pp. 355-363, August 2015.

[4] S. Aksu, F. G. Varol and N. H. Sahin, "Long-term postpartum health problems in Turkish women: prevalence and associations with self-rated health,", Contemporary Nurse, 53:2, pp. 167-181, 2017.

[5] A. Rudman and U. Waldenström, "Critical views on postpartum care expressed by new mothers," BMC Health Services Research, 7, pp. 178, 2007.

[6] N. T. Hatfield, "Introductory Maternity and Pediatric Nursing," Lippincott Williams & Wilkins, pp. 405, Nov 2013.

[7] L. I. Alasoom and M. R. Koura, "Predictors of Postpartum Depression in the Eastern Province Capital of Saudi Arabia," Journal of Family Medicine and Primary Care, vol. 3(2), pp. 146-150, 2014.

[8] Q. Y. Jiang, J. Li, L. Zheng, G. H.Wang, and J. Wang, "Constitution of traditional chinese medicine and related factors in women of childbearing age," Journal of the Chinese Medical Association, 81(4), pp. 358-365, 2018.

[9] S. Lukman, Y. He, and S. C. Hui, "Computational methods for traditional Chinese medicine: a survey," Computer methods and programs in biomedicine, 88(3), pp. 283-294, 2007.

[10] W. Wang, A. Russell, Y. Yan, "Traditional Chinese medicine and new concepts of predictive, preventive and personalized medicine in diagnosis and treatment of suboptimal health," EPMA Journal, 5(1), pp. 12, 2014.

[11] Y. Wang, Y. Dai, F. Guo, S. Li, "Sensitivity-based data selection for predicting individual's sub-health on TCM doctors' diagnosis," IEEE International Conference on Systems, Man, and Cybernetics, Anchorage, AK, pp. 896-901, 2011.

[12] Y. B. Zhu, Q. Wang, H. Origasa, "Evaluation on reliability and validity of the constitution in chinese medicine questionnaire (CCMQ)," Chinese Journal of Behavioral Medical Science, 16(7), pp. 651–654, 2007.

[13] H. You, T. Zhang, W. Feng, and Y. Gai, "Association of TCM body constitution with insulin resistance and risk of diabetes in impaired glucose regulation patients," BMC complementary and alternative medicine, 17(1), pp. 459, 2017.

[14] Y. Zhu, et al., "Association between nine types of TCM constitution and five chronic diseases: a correspondence analysis based on a sample of 2,660 participants," Evidence-Based Complementary and Alternative Medicine, 9439682, 2017.

[15] C. I. Tsai, Y. C. Su, S. Y. Lin, I. T. Lee, C. H. Lee & T. C. Li, "Reduced health-related quality of life in body constitutions of Yin-Xu, and Yang-Xu, stasis in patients with type 2 diabetes: taichung diabetic body constitution study," Evidence-Based Complementary and Alternative Medicine, 309403, 2014.

[16] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer & A. Geissbuhler, "Clinical Data Mining: a Review," Methods of Information in Medicine, pp. 121-133, 2009.

[17] L. Zhang, "A TCM Way to be Healthy, Inside and Out," FriesenPress, 2017.

[18] Q. Wang, "Classification and diagnosis basis of nine basic constitutions in chinese medicine," Journal of Beijing University of Traditional Chinese Medicine, 28(4), pp. 1-8, 2005.

[19] M. Li, S. Mo, Y. Lv, Z. Tang and J. Dong, "A Study of Traditional Chinese Medicine Body Constitution Associated with Overweight, Obesity, and Underweight," Evidence-Based Complementary and Alternative Medicine, 7361896, 2017.

[20] W. Wong, C. L. K. Lam, V. T. Wong, Z. M. Yang, E. T. C. Ziea & A. K. L. Kwan, "Validation of the Constitution in Chinese Medicine Questionnaire: Does the Traditional Chinese Medicine Concept of Body Constitution Exist?," Evidence-Based Complementary and Alternative Medicine, 481491, 2013.

[21] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," European Conference on Machine Learning, pp. 171-182, 1994

[22] A. Al-Shura, "Integrative Cardiovascular Chinese Medicine: A Prevention and Personalized Medicine Perspective," Academic Press, pp. 39-56, 2014

[23] R. Y. Chan, and W. T. Chien, "Concepts of body constitution, health and sub-health from traditional Chinese medicine perspective," World J Transl Med, 2(3), pp. 56-66, 2013.

[24] G. Maciocia, "The Practice of Chinese Medicine," New York, Churchill Livingstone, 1994.

[25] J. Han, J. Pei and M. Kamber, "Data mining: concepts and techniques," Elsevier, 2011.

[26] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Transactions on System, Man, and Cybernetics, 21(3), pp. 660-674, 1991.

[27] H. Bhavsar and M. H. Panchal, "A review on support vector machine for data classification," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(10), pp. 185, 2012.

[28] F. Rossi and N. Villa, "Support vector machine for functional data classification," Neurocomputing, 69(7-9), pp. 730-742, 2006.

[29] R. J. Schalkoff, "Artificial Neural Networks," McGraw-Hill, 1997.

[30] H. Scheffé, "The Relation of Control Charts to Analysis of Variance and Chi-Square Tests," Journal of the American Statistical Association, 42(239) pp. 425-431, 1947.

[31] R.Y. Chan and W.T. Chien, "Concepts of body constitution, health and sub-health from traditional Chinese medicine perspective," World J Transl Med, 2, pp. 56-66, 2013.

[32] J. Cohen, "Statistical power and analysis for the behavioral sciences," 2nd ed. Hisdale, NJ: Lawrence Erlbaum Associates, pp. 79-80, 1988.

Multiclass Disease Predictions Based on Integrated Clinical and Genomics Datasets

Moeez M. Subhani
College of Engineering and Technology
University of Derby
Derby, England
Email: m.subhani@derby.ac.uk

Ashiq Anjum
College of Engineering and Technology
University of Derby
Derby, England
Email: a.anjum@derby.ac.uk

***Abstract*—Clinical predictions using clinical data by computational methods are common in bioinformatics. However, clinical predictions using information from genomics datasets as well is not a frequently observed phenomenon in research. Precision medicine research requires information from all available datasets to provide intelligent clinical solutions. In this paper, we have attempted to create a prediction model which uses information from both clinical and genomics datasets. We have demonstrated multiclass disease predictions based on combined clinical and genomics datasets using machine learning methods. We have created an integrated dataset, using a clinical (ClinVar) and a genomics (gene expression) dataset, and trained it using instance-based learner to predict clinical diseases. We have used an innovative but simple way for multiclass classification, where the number of output classes is as high as 75. We have used Principal Component Analysis for feature selection. The classifier predicted diseases with 73% accuracy on the integrated dataset. The results were consistent and competent when compared with other classification models. The results show that genomics information can be reliably included in datasets for clinical predictions and it can prove to be valuable in clinical diagnostics and precision medicine.**

Keywords–Clinical; Genomics; Data Integration; Machine Learning; Disease Prediction; Classification; Bioinformatics.

I. Introduction

The medical science is rich with various types of datasets ranging from clinical to genomics datasets. The clinical datasets are diverse in terms of their nature, format and the information they contain. On the other hand, genomics datasets are intrinsically enormous in size and dimensions, and so is the information contained in them [1]. The genomic information can be considered as the backbone of clinical information since the genomic structure derives the physical characteristics of any organism. If the two pieces of information are connected, it may help to improve the overall medical research by finding more accurate and advanced clinical diagnostic solutions. The connection essentially means to integrate clinical and genomics datasets. This is also a way forward in precision medicine studies, where medical practitioners want to make clinical decisions based on both clinical and genomics parameters and not just one of them [2][3].

However, the research to establish or explore this connection is not very commonly sought in the state-of-the-art [1][4]–[6]. The datasets from clinical and genomics sources are mainly used independently in their respective research domains. From literature review, it has been observed that most clinical prediction studies have been limited to either clinical datasets [7]–[13] or genomics datasets [14]–[20]. One common factor among these studies is that almost all of them are prediction studies, which establishes the fact that the trend for clinical predictions has long prevailed in research.

Although there are some studies which have attempted towards the inter-domain research, the trend does not seem to be very progressive. For example, [21] used decision trees to predict breast cancer outcomes. Similarly, [22] employed multiple regression and statistical methods to infer associations, and [3] used a graph-based approach to predict cancer clinical outcomes from multi-omics data. All these studies used integrated datasets for prediction or association studies using various approaches. However, most of these approaches are now outdated due to limitations in terms of their performance or accuracy [21][22]. The approach in [3] (combination of regression, Bayesian networks, and evolutionary neural networks) is more advanced and promising but this study is limited to binary classifications and multi-omics data only [23][24].

The research work mentioned above show that prediction based studies are common in the literature. The most popular or commonly sought predicting factors are survival rate and disease recurrence rate. However, we could not find any disease prediction model in the literature based on combined clinical and genomics data information. A typical disease prediction model, as we define, takes information from both clinical and genomics datasets and predicts disease(s) in a patient. This can be achieved when we have both clinical and genomics datasets available for a variety of diseases. Hence, we are attempting to design a disease prediction model which aims to predict possible medical condition(s) in a patient using information from both clinical and genomics datasets.

From ClinVar and Expression Atlas databases, we have been able to construct such dataset which contains both clinical parameters as well as gene expression values in a single dataset for several patients. Since the data retrieved from these databases is in eXtensible Markup Language (XML) format, we can create a very flexible schema for this dataset. Using this dataset, we can train a model to learn the diseases in various subjects. As an initial attempt to prove the concept, we have used the k-Nearest Neighbours (kNN) algorithm for the learning model, which is an instance based learner [25]. Considering the size and complexity of the dataset, kNN appears to be a reasonable choice of learning method since it learns the classification function only locally.

Genomics based clinical diagnosis does not exist in clinical environments. Traditionally, disease predictions are made using regular clinical practices only. Our disease prediction model can provide a genomic signature to verify the disease existence or possible occurrence. Hence, this model not only will help

TABLE I. CLINVAR DATASET.

Gene	Condition	Clinical Significance	Chromosome No.	Location	Variation ID	Allele ID
AKAP	Long QT syndrome	Benign/Likely benign	7	92001306	136347	140050
AKT2	Colorectal Neoplasms	Likely pathogenic	19	40236313	376039	362918
APC	Hereditary cancer-predisposing syndrome	Pathogenic	5	N/A	181836	181126
...	...	...	...	...	...	...

the medical practitioners to gain another step of confidence in terms of clinical diagnosis, but also help advance the precision medicine research.

The rest of the paper is arranged as follows. Section II discusses the challenges for data integration. Section III explains the data integration model. Section IV gives details of the prediction model and the algorithm along with the implementation details. Section V presents the results, followed by discussion in Section VI and conclusion in Section VII.

II. Clinical and Genomics Data Integration Challenges

The integration of clinical and genomics datasets is crucial to move towards precision medicine. The medical conditions of each person are transcribed from the underlying genomics structure. Hence, it is critical to bring forward the genomic information to play part in the clinical diagnostics [2][3]. The main challenge is to find a way to integrate datasets which are completely different from each other in terms of their nature, size, and properties.

Most biological databases have standardised the data storage in XML formats. European Molecular Biology Laboratory (EMBL) took an initiative in 2000 to provide access of all the flat files data in XML format [26]. XML provides more flexibility in terms of storage, transport and integration of complex biological datasets [27]. The format also provides the advantage that the schema of datasets is extensible and multiple datasets can be mapped together. Our datasets from both sources, ClinVar and Expression Atlas, are accessed in XML formats.

The scope of data integration models is vast, as mentioned in the literature review in the previous section. Various data integration models have been discussed by various authors including [1], [4] and [6]. For our study, we have adopted a meta-dimensional approach model, which refers to using multiple datasets simultaneously in the analysis [6]. This involves building a model on top of multiple datasets, which are combined or integrated either before or after building the data model. The approach facilitates the advantage of fetching information from multiple datasets and including it in the analysis model. However, the integration may also yield complex datasets resulting in less robust models.

There are multiple methods within the meta-dimensional approach as mentioned by [1] and [6]. We have adopted a concatenation-based integration method, where different matrices are combined into a large single matrix before building a model. One advantage of this method is that once it is determined how to concatenate the variables from different datasets into a single matrix, it is relatively easier to build any statistical analysis model on it. For example, on a combination of genomics datasets, [8] used a Bayesian model to predict phenotypes, and [28] used Cox Lasso model to predict time to recurrence.

It may be important to mention here that the integration attempt in this paper is only at the data level. Since the data being retrieved from public repositories is in XML format, we do not need to pre-build a structure to store data, and we are not dealing with databases either. Therefore, this method provides the advantage to avoid the data structure and storage issues. Hence, the data integration here must not be confused with the traditional database level data integration.

TABLE II. GENE EXPRESSION DATASET.

Gene	GSM452573	GSM452571	GSM452642	...
AKAP9	3.563587736	3.45243272	3.535150355	...
AKT1	10.8863402	10.34918494	9.129441853	...
AKT2	5.005896122	4.463927997	4.993673626	...

III. Data Integration Model

We have used completely anonymised clinical and genomics datasets obtained from public sources. The clinical dataset (ds1) has been obtained from ClinVar [29], which is an open source database that contains information about the genomic variation and links it with phenotype information. For each gene, it provides the diseases it causes and their clinical significance. In addition, it also includes the whereabouts of the gene, such as, chromosome number, location, variation ID etc. A snapshot of the data is illustrated in Table I. The database was searched for 'colorectal cancer', and all the search results were downloaded and saved as XML files.

The genomics dataset (ds2) is a Gene Expression dataset of primary colorectal tumours (E-GEOD-18105), obtained from the Expression Atlas of European Bioinformatics Institute (EBI), which is a public resource for gene expression datasets [30]. Gene expression data, as the name indicates, contains information for the expression of gene(s) in a particular biological sample(s). The expression data is obtained via microarray technology, which provides parallel processing and monitoring

TABLE III. INTEGRATED DATASET.

DISEASE	CLINICAL SIGNIFICANCE	CHROMOSOME NO.	LOCATION	VARIATION ID	ALLELE ID	GENE	GSM452573	GSM452571
LONG QT SYNDROME	BENIGN/ LIKELY BENIGN	7	92001306	136347	140050	AKAP	3.563587736	3.45243272
COLORECTAL NEOPLASMS	LIKELY PATHOGENIC	19	40236313	376039	362918	AKT2	5.005896122	4.463927997
...	...	...	...	...	...	...	...	...

of tens and thousands of genes, producing tons of valuable data [31]. A typical gene expression dataset contains a matrix with genes in rows and samples in columns. The number in each cell of the matrix characterises the expression level of a specific gene in the given sample [32]. Table II shows an example of how gene expression data looks like. After the first column, which is gene name, the rest of the columns represents samples, and the values represent the expression levels.

The primary reason for selecting these two datasets was that they fulfill the information requirement for this study. The ClinVar data provides the information about clinical condition against each gene present in the dataset. It also provides the clinical significance of these conditions [33]. The gene expression data brings the information about the activity of those genes in different samples. Hence, the two datasets provide the required information to create an integrated dataset for this model.

TABLE IV. STATISTICS OF INTEGRATED DATASET.

Output Classes	I	II
Unique Classes	80	76
Feature set	117	117
Training Examples	258	281

As mentioned previously, we are using a meta-dimensional approach based integration, and specifically the concatenation method. The datasets were concatenated via gene names. It has to be noted that there were multiple examples for each gene in both datasets. The examples in the ds1 with no feature sets available in ds2 were removed. On the contrary, the examples in the ds2 for which there were no feature sets in ds1, the data was extrapolated in ds1 so that the examples for that gene can be increased. Since each parameter in the feature set is independent, therefore, extrapolating some points does not affect the accuracy.

Table III shows an example of the integrated dataset, where the clinical and genomics parameters are concatenated via gene names. The statistics of the dataset is shown in Table IV. The data was trained with two different output classes: genes (class-I) and diseases (class-II). There are 80 unique genes, and 76 unique diseases in the dataset after removing the outliers.

It can be argued that predicting genes as output class does not provide much meaning. Predicting disease has a more clinical value since this information is not available in the gene expression data. The reason behind this selection is only to provide an example that the classifier can be used to predict any feature from an integrated dataset without any restriction.

The resulting schema includes clinical and genomics parameters in columns, while each row represents a gene. Hence, each row tells the possible medical condition for a gene if it is active in a sample. This schema is completely flexible and scalable. It can be expanded by adding data from different sources, as long as the new data can be mapped to existing schema. More data brings more information that will only help to improve the performance of the classifier by increasing the feature set and the training examples.

IV. PREDICTION MODEL

In this section, we will talk about the multiclass classification challenges, followed by the details of our prediction model, comprising the algorithm and the experimental environment.

A. *Multiclass Classification*

When we talk about disease classification, we are talking about a complicated multiclass classification problem. From classification perspective, it is relatively easier to classify binary problems or even few classes, but with increasing number of classes, the complexity of the dataset gets very high [34]. The data under consideration in this study contains more than 75 different classes. When the number of output classes is that high, the variance in the data is very high as well. In such a case, it is best to have as much data as possible so that every class has a sufficient representation in training data. This is a minor limitation in our study because of the limited number of examples available from public datasets.

There is no single classification method that can be suggested to be best suited for multiclass classification [34]. Any algorithm can perform better than the rest based on the characteristics and properties of the data. In this study we have used the k-Nearest Neighbours (kNN) algorithm. The reason for selecting the kNN instead of Support Vector Machines (SVM), which is a more popular classification algorithm, is our large number of output classes and the random distribution of data (Figure 1). Unlike SVM, which uses kernels for optimization, kNN determines the label for a given data point based on nearest data points on the distance metric. Since kNN is a non-parametric algorithm, it does not assume any explicit functions for the input data (such as Gaussian) [25]. This works well in our case when the data has no particular distribution and is widespread (Figure 1). Hence, we can avoid the algorithmic complexity by using an algorithm which uses local optimization only. Also, kNN performs well on small to medium sized datasets [25].

B. *Classification Algorithm*

The kNN is a non-parametric supervised learning algorithm [25]. For a given dataset X, with labels Y, the algorithm calculates the distances between a new data point z and all data points in X to create a distance matrix. Euclidean distance is the most common method for calculating this distance. Euclidean distance between point x_i and y_i can be calculated by:

$$D(x, y) = \sum_{i=1}^{k} (x_i - y_i)^2 \tag{1}$$

Let $R = (X_i, Y_i)$, where $i = 1, 2...N$, be the training set, where X_i is the $p * q$ feature vector, and Y_i is the q-dimensional vector which represents m output class labels, as we are considering multiclass classification problem. We presume that the training data has random numeric variables with unknown distribution.

From the training set R, the kNN algorithm narrows down to a local sub-region $r(x)$ of the input space, which is centered on an estimation point x. This predicting sub-region $r(x)$ contains the training points (x') nearest to x, which can be expressed as:

$$r(x) = \{x' \mid D(x, x') \leq d(k)\} \tag{2}$$

where, $D(x, x')$ is the distance metric between x' and x, and $d(k)$ is the k^{th} order statistic. $k[y]$ denotes the k samples in the sub-region $r(x)$, which are labelled y. The kNN algorithm estimates the posterior probability $p(y \mid x)$ of the estimation point x:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} \cong \frac{k[y]}{k} \tag{3}$$

Generally, when the kNN is used for binary classification, the label assignment is relatively easier since the algorithm has to select between two classes only, such as :

$$g(x) = \{\, 1 \ , k[y = 1] \geq k[y = -1] - 1, k[y = 1] \leq k[y = -1] \tag{4}$$

We have improvised this functionality for our study, where the output class is non-binary. In this case, for any estimation point x, the decision $g(x)$ for a given label y is estimated by:

$$g_k(x) = y_k \mid minD_k \tag{5}$$

where, D_k is represented by 1 Hence, the decision that will maximise the posterior probability will be assigned for the output label. For a multiclass classification problem, where $y \in \{1 \ldots k\}$, the kNN algorithm uses the following decision rule:

$$F(x) = argmax[g_k(x)] \tag{6}$$

Thus, for the selected nearest k neighbours, the algorithm calculates the posterior probability for each class, and the class with highest probability is assigned to x. Euclidean distance is the most common method, but there are other distance calculation methods as well, such as seuclidean, mahalanobis, spearman, etc [25].

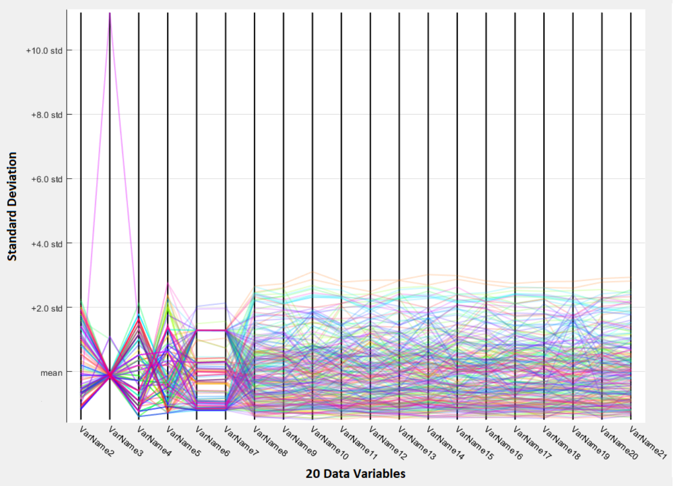

Figure 1. Distribution of variance in the integrated dataset.

C. *Performance Measurement*

Generally, the performance of machine learning classifiers is measured using various parameters, such as, accuracy, sensitivity, specificity, and Receiver Operator Curve (ROC). These parameters are calculated based on the true positives, true negatives, false positives and false negatives of classifier. For binary classes, these parameters are easier to calculate because there is only one positive and one negative class. However, for multicalss classification, the problem is more complicated and it is not easy to calculate each parameter for each class. Especially ROC, which is a standard measure to represent performance of a classifier, is very complicated to calculate for a very large multiclass problem. This problem has been discussed in further detail by Fawcett in [35].

Therefore, calculating each parameter for every class will not only be laborious, but will also produce loads of results that will be difficult to ensemble and explain. To simplify that, we have only used confusion matrices to represent the performance of the classifier and used the accuracy for each classifier to compare the results for the two classes.

D. *Experimental Environment*

We have used Matlab (R2018a) for all the experiments, which provides built-in libraries for machine learning classifiers. We used the machine learning toolbox to train the classification model using kNN. The toolbox takes the data as input and process the classification itself using the built-in library functions and selected features. The classification toolbox uses the Euclidean distance by default to compute the distance metrics. The tool box can be used to reproduce the results.

At first, we perform the Principal Component Analysis (PCA) for dimensionality reduction. Since, our data is multivariate, ranging from gene expression data to phenotypic data, the data points are widespread in the data space. Figure 1 shows the standard deviation distribution of the first 20 data variables from the integrated dataset. It can be seen that the data distribution is very random and does not follow any standard distribution function. Therefore, it is important to reduce the dimension of the integrated dataset. We performed PCA to explain 95% variance in the data.

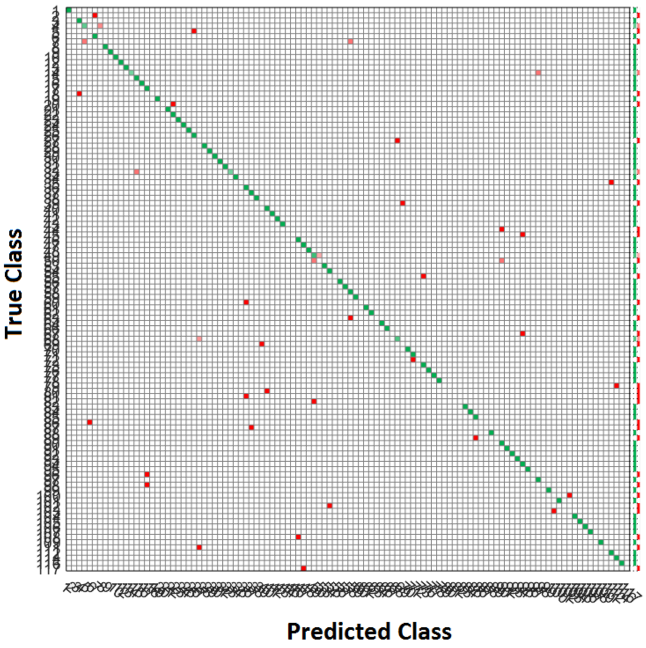

Figure 2. Confusion matrix for class-I.

The results of classification depend highly on the dimensions of the dataset. The correlation between the number of examples and feature sets is very critical in this case to avoid over-fitting [36][37]. The clinical dataset has only 5 features, which is not a large enough set to be used stand-alone for prediction model. With a feature set of 5, the prediction is neither reliable nor comparable with other datasets. The genomics dataset is large enough in this respect, but it does not contain the class-II so we cannot predict diseases. Therefore, we have only used the integrated dataset to train with the prediction model explained in the previous section (IV), and then compared it with other classifiers.

The results are validated using 10-fold cross-validation. This means, the dataset is divided into 10 parts; one part is held out as a test data and the rest of the 9 parts are used as training data. This step is repeated 10 times using a different part every time to holdout as a test data. This way every example from data is used both as training and test data. The resulting accuracy is an average of the 10-fold process.

V. Results

The performance of a classification model is analysed using a confusion matrix. Figure 2 shows a confusion matrix for class-I prediction. The rows in a confusion matrix represent the true output class, and the columns represent the predicted class. The diagonal cells indicate the true positives (green) and the false negatives; and the off-diagonal cells indicate the false positives and the true negatives (red). The bottom right cell shows the overall accuracy and the loss of the classifier.

A. *Classification with our Classifier*

The number of neighbours (NN) is a variable in the algorithm, which can be tuned to change the performance of the algorithm. We tested the performance of the algorithm over 10 different neighbours, from 1 to 10.

As mentioned previously, we trained the integrated dataset for two different classes: genes (class-I) and diseases (class-II). The results are shown in Figure 3. At NN=1, the trained model predicts class-I with 86% accuracy, and class-II with 73% accuracy.

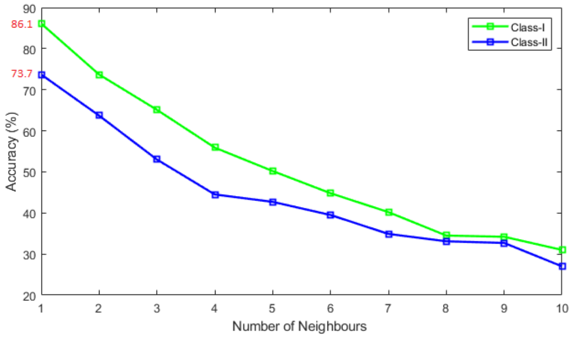

Figure 3. Accuracy of classification model for both classes.

Initially, the accuracy drops almost linearly with the increasing number of neighbours. The drop in accuracy can be attributed to the variation in data. As the algorithm considers more number of neighbours, each neighbour brings more variation that affects the prediction accuracy. However, as it can be seen in Figure 3, accuracy remains above 50% for up to 3 neighbours for both classes which can be regarded as a good accuracy considering a multivariate training data. Following NN=4, the accuracy drops almost exponentially.

This variation over neighbours may be avoided by introducing a weighted parameter in the algorithm. This parameter weighs the contribution of each neighbour under consideration based on its distance. The nearest neighbours gets higher weights than the distant ones. Matlab's classification tool uses the squared inverse method to calculate the weights, which can be expressed as:

$$w_n = \frac{1}{d(x_n - x_i)^2} \tag{7}$$

where, x_n is the neighbour to point x_i. To accommodate this weight parameter, the eq1 is adjusted as follows:

$$D(x, y) = \sum_{i=1}^{k} w_i (x_i - y_i)^2 \qquad (8)$$

We tested this updated version by training the integrated dataset, and we observed that the accuracy was raised to the maximum (86.1% for class-I and 73.7% for class-II) for all NN's. The results are shown in Figure 4. This is perhaps because the weighted version predicts based on the neighbour with the highest weight. Since the nearest neighbour is most likely to have highest weight out of all neighbours, the classification result is the same every time. This result seems to be not very helpful for our dataset.

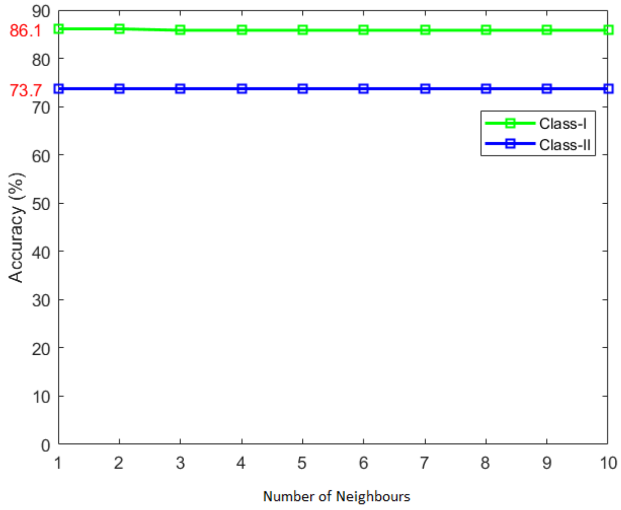

Figure 4. Accuracy for both classes with weighted kNN.

However, our model has predicted the diseases with up to 73% accuracy. The accuracy is not as good as for the class-I (86%). There can be multiple reasons behind this. The representation of each class label in the data varies, which affects the prediction accuracy. Some classes have sufficient examples in the data, while others have only few examples. The higher the representation of a class label in the training data, the better is the prediction accuracy for that class. The distribution of class-I labels in dataset is comparatively more uniform than class-II; hence, higher accuracy. Still, achieving 73% accuracy for class-II is a very good result considering the size, shape, and multivariate nature of the dataset.

B. Comparing with other Classifiers

We trained the same integrated dataset with other classifiers in order to compare the performance. Using PCA of 95%, we trained all the classifiers available on Matlab's classification toolbox, and then selected bluethe top 10 models (out of 22) to compare the classification accuracy for both classes. NN=1 for all the models in the classification toolbox. 10-fold cross-validation was used to avoid over-fitting. The results are shown in Figure 5.

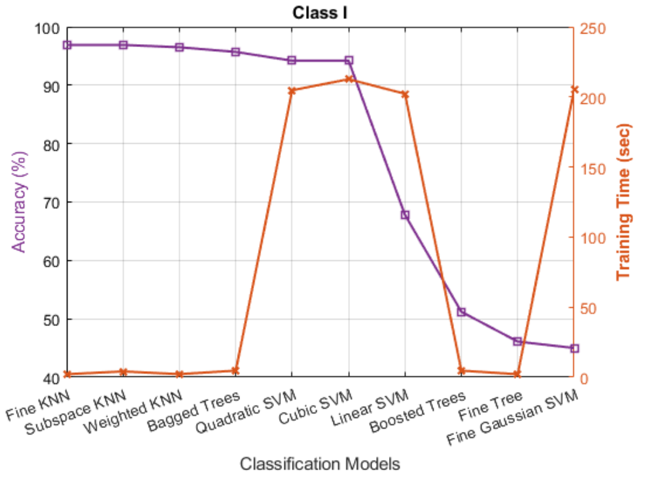

Figure 5. Performance of other classification models for class-I predictions.

For class-I, the kNN models provided the highest accuracy of 96.9%. kNN was followed by the Tree and SVM models. As we can see, the top three models are all kNN models providing accuracy of above 95%. The accuracy of the SVM models (Quadratic and Cubic) is almost in the same range (95-96%), however, the training time for the SVM models is 200 times higher than the kNN models. This is because the SVM uses the cost minimization functions, such as gradient descent or kernel functions, which take much longer to converge. Since kNN does not use any of those functions, it is more robust and provides with the same, rather better accuracy. To summarise, although both kNN and SVM models have predicted about the same accuracy, the kNN models are much more robust than the SVM models in terms of performance.

The tree models, except for bagged trees, performed poorly providing accuracy of about 50% or under. The training time of the tree models is as good as that of kNN models (few seconds), but the accuracy is poor. Bagged trees, which is a bootstrapping method, performed quite well. On the other hand, boosted trees provides an accuracy of just about 51%. Although both of them are ensemble methods, which means they provide an average of multiple models trained on a subset of data, bagged trees provided much better result.

The accuracy of these kNN models (Fine kNN, Subspace kNN, and Weighted kNN) is slightly higher than our prediction model (Figure 3). The reason for this is that the models in the toolbox are set on different defaults and use different functions than the ones we used. The classification function that we used is primarily for multiclass classification problems. On the other hand, the function used by the toolbox models are mainly designed for binary problems, hence, the difference of accuracy.

Similar results are seen for class-II. The results are shown in Figure 6. The top 10 models selected here are slightly different than those for class-I, but majority are the same. The highest accuracy achieved for class-II is 73.3%, which is just about the same as achieved by our model (Figure 3). The top 3 models are all kNN models, with bagged trees standing at 4th position with 73% accuracy. All SVM models provide accuracy of less than 50% with training times as high as over 200 times of the kNN models. The same is the case for tree

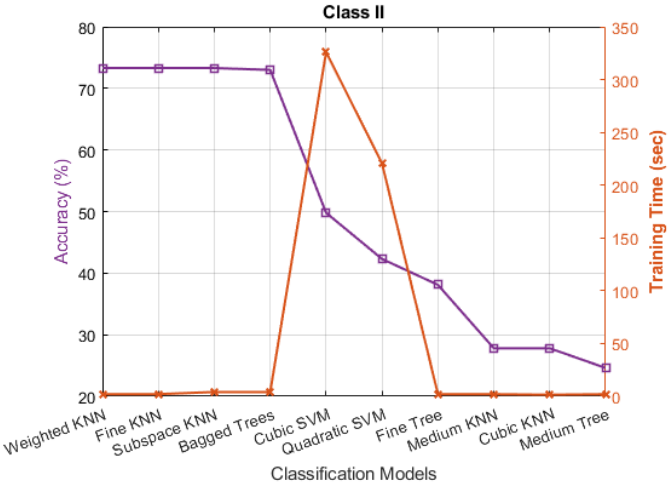

Figure 6. Performance of other classification models for class-II predictions.

models except for bagged trees; same result as for class-I. A plausible explanation for good performance of bagged trees could be that they perform better on high dimensional data.

VI. Discussion

We have demonstrated a novel way for multiclass classification based on integrated clinical and genomics datasets. We have used concatenation-based data integration model for this purpose, which has been discussed by various researchers before ([1][6]), but not implemented in the area of health care. Hence, this is the first time that we have attempted to use this meta-dimensional approach to integrate datasets.

In the past, people have used various other methods for data integration such as tree-based models [21], statistical models [22], and graph based models [3][18][20]. All these models require considerable amount of effort and time to build the data models first, before creating the data analysis model, such as building the binary trees, or creating graphs models from datasets. Our method does not involve any of those complex models; it only requires concatenation of all the datasets into a single matrix. Once concatenated, the model transfers the dataset directly to the analysis model and starts training the learning algorithm. Hence, it is way more efficient in terms of time and computational costs as compared to other methods.

In terms of analysis, from our knowledge, none of the previous models have been used for multiclass disease classification problems in health care. They have only been demonstrated for binary classifications; and, therefore, their results cannot be compared with our model, which is a multiclass classification model.

In terms of data models, it will be very difficult to perform multiclass classification based on the previously mentioned models because they will require to build a separate data model (trees of graphs) for each output class before the analysis model. Having multiple output classes, the analysis models will get extremely complicated with several input data models. With our proposed model, as there is only single concatenated dataset, the multiclass classification is less complicated and manageable because the dataset has only one data model with a single schema.

Since, we could not compare our results with any other previous results from other researchers, we have demonstrated comparison with other classification models. The results shown in Figures 5 and 6 demonstrate that the kNN models can outperform the rest of the classification models in terms of prediction accuracy and performance.

Our proposed approach provides a very flexible and scalable model, along the lines of our previous work as reported in [38]–[41], which can be scaled to adjust any new dataset and accommodate any analysis model. As long as there is a relational dataset, it can be concatenated to the existing dataset within the same data model and schema. Any analysis model or algorithm, including prediction, classification, regression models, can be built on top of the dataset. This flexibility enables this approach to be adapted for any research purpose in any domain.

VII. Conclusion and Future Directions

The way forward in precision medicine is to use all available data from clinical and genomics domains in order to provide the best clinical solutions. The datasets need to be intelligently integrated for this purpose. In this paper, we have performed clinical predictions based on clinical and genomics information. We have attempted to integrate a clinical (ClinVar) and a genomic (gene expression) dataset, and performed classification for disease predictions. We have designed a multiclass classification model that predicts diseases from integrated datasets. The model, which is validated by 10-fold cross-validation, has predicted diseases with up to 73% accuracy. We also predicted genes as an extra variable, from the same dataset, and achieved up to 86% accuracy. We have compared the results with other classification models and demonstrated that our model outperforms the rest. We can conclude that constructing the learning classifiers on top of large-scale inter-domain integrated datasets can provide very good clinical predictions. This can prove to be very beneficial and a stepping-stone towards the precision medicine.

This research study shows that diseases can be predicted with good accuracy from a patient's dataset if it has both clinical and genomics parameters present. The accuracy will further improve if we train the model with a much larger size of training data. The reliability and confidence in results will increase by incorporating more clinical and genomics information. We have demonstrated with a gene prediction example, that, when the dataset is more uniformly distributed among different classes, the prediction accuracy goes high even on a multiclass classification task.

This study has great potential to expand including achieving analysis provenance [13]. The more information a dataset will contain, higher the accuracy can be achieved. The dataset can be expanded to include more multivariate clinical and genomics datasets, such as clinical trials and multi-omics datasets, respectively. Including clinical information from clinical trials or laboratory tests will have a significant impact in the clinical prediction studies.

Acknowledgment

The authors would like to thank and acknowledge the help and support received from Usman Yaseen, Sanna Aizad, Bilal Arshad and Craig Bower in the successful completion of this study.

References

[1] M. M. Subhani, A. Anjum, A. Koop, and N. Antonopoulos, "Clinical and genomics data integration using meta-dimensional approach," in 2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC). IEEE, 2016, pp. 416–421.

[2] R. Higdon et al., "The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders," Omics: a journal of integrative biology, vol. 19, no. 4, 2015, pp. 197–208.

[3] D. Kim et al., "Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction," Journal of the American Medical Informatics Association, vol. 22, no. 1, 2014, pp. 109–120.

[4] J. S. Hamid, P. Hu, N. M. Roslin, V. Ling, C. M. Greenwood, and J. Beyene, "Data integration in genetics and genomics: methods and challenges," Human genomics and proteomics: HGP, vol. 2009, 2009.

[5] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, "Data integration and genomic medicine," Journal of biomedical informatics, vol. 40, no. 1, 2007, pp. 5–16.

[6] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype–phenotype interactions," Nature Reviews Genetics, vol. 16, no. 2, 2015, p. 85.

[7] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial intelligence in medicine, vol. 34, no. 2, 2005, pp. 113–127.

[8] B. L. Fridley, S. Lund, G. D. Jenkins, and L. Wang, "A bayesian integrative genomic model for pathway analysis of complex traits," Genetic epidemiology, vol. 36, no. 4, 2012, pp. 352–359.

[9] D. Kim, R. Li, S. M. Dudek, and M. D. Ritchie, "Athena: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network," BioData mining, vol. 6, no. 1, 2013, p. 23.

[10] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," PloS one, vol. 8, no. 6, 2013, p. e66341.

[11] R. Palaniappan, K. Sundaraj, and S. Sundaraj, "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals," BMC bioinformatics, vol. 15, no. 1, 2014, p. 223.

[12] J. Raikwal and K. Saxena, "Performance evaluation of svm and k-nearest neighbor algorithm over medical data set," International Journal of Computer Applications, vol. 50, no. 14, 2012.

[13] R. McClatchey et al., "Providing traceability for neuroimaging analyses," International journal of medical informatics, vol. 82, no. 9, 2013, pp. 882–894.

[14] U. D. Akavia et al., "An integrated approach to uncover drivers of cancer," Cell, vol. 143, no. 6, 2010, pp. 1005–1017.

[15] M. P. Brown et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," Proceedings of the National Academy of Sciences, vol. 97, no. 1, 2000, pp. 262–267.

[16] D. Kim, H. Shin, Y. S. Song, and J. H. Kim, "Synergistic effect of different levels of genomic data for cancer clinical outcome prediction," Journal of biomedical informatics, vol. 45, no. 6, 2012, pp. 1191–1198.

[17] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein, "Assessing the limits of genomic data integration for predicting protein networks," Genome research, vol. 15, no. 7, 2005, pp. 945–953.

[18] E. E. Schadt et al., "An integrative genomics approach to infer causal associations between gene expression and disease," Nature genetics, vol. 37, no. 7, 2005, p. 710.

[19] J. Zhu et al., "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," Nature genetics, vol. 40, no. 7, 2008, p. 854.

[20] X. Yang et al., "A network based method for analysis of lncrna-disease associations and prediction of lncrnas implicated in diseases," PloS one, vol. 9, no. 1, 2014, p. e87797.

[21] J. R. Nevins, E. S. Huang, H. Dressman, J. Pittman, A. T. Huang, and M. West, "Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction," Human molecular genetics, vol. 12, no. suppl_2, 2003, pp. R153–R157.

[22] E. Lee, S. Cho, K. Kim, and T. Park, "An integrated approach to infer causal associations among gene expression, genotype variation, and disease," Genomics, vol. 94, no. 4, 2009, pp. 269–277.

[23] D. Kim, R. Li, S. M. Dudek, and M. D. Ritchie, "Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer," Journal of biomedical informatics, vol. 56, 2015, pp. 220–228.

[24] S. S. Verma, D. Kim, M. D. Ritchie, A. Lucas, R. Li, and S. M. Dudek, "Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma," Journal of the American Medical Informatics Association, vol. 24, no. 3, 2016, pp. 577–587.

[25] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification and scene analysis. Wiley New York, 1973, vol. 3.

[26] A. Robinson, J.-J. Riethoven, and L. Wang, "XEMBL: distributing EMBL data in XML format," Bioinformatics, vol. 18, no. 8, 2002, pp. 1147–1148.

[27] Z. Lacroix, "Biological data integration: wrapping data and tools," IEEE Transactions on Information Technology in Biomedicine, vol. 6, no. 2, 2002, pp. 123 – 128.

[28] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, and C. Sander, "Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles," PLOS ONE, vol. 6, no. 11, 2011, pp. 1–12.

[29] "NCBI ClinVar Database," URL: https://www.ncbi.nlm.nih.gov/clinvar [accessed: January 2018].

[30] "Gene Expression Atlas," URL: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-18105/?query=E-GEOD-18105_A-AFFY-44 [accessed: January 2018].

[31] N. Raghavachari, "Microarray technology: basic methodology and application in clinical research for biomarker discovery in vascular diseases," in Lipoproteins and Cardiovascular Disease. Springer, 2013, pp. 47–84.

[32] A. Brazma and J. Vilo, "Gene expression data analysis," FEBS Letters, vol. 480, no. 1, 2000, pp. 17–24.

[33] M. J. Landrum et al., "Clinvar: public archive of relationships among sequence variation and human phenotype," Nucleic acids research, vol. 42, no. D1, 2013, pp. D980–D985.

[34] C. Zhang, M. Ogihara, and T. Li, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," Bioinformatics, vol. 20, no. 15, 2004, pp. 2429–2437.

[35] T. Fawcett, "An introduction to roc analysis," Pattern recognition letters, vol. 27, no. 8, 2006, pp. 861–874.

[36] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," Bioinformatics, vol. 21, no. 8, 2004, pp. 1509–1515.

[37] A. K. Jain and W. G. Waller, "On the optimal number of features in the classification of multivariate gaussian data," Pattern recognition, vol. 10, no. 5-6, 1978, pp. 365–374.

[38] A. Anjum, R. McClatchey, A. Ali, and I. Willers, "Bulk scheduling with the diana scheduler," IEEE Transactions on Nuclear Science, vol. 53, no. 6, 2006, pp. 3818–3829.

[39] F. van Lingen et al., "The clarens web service framework for distributed scientific analysis in grid projects," in 2005 International Conference on Parallel Processing Workshops (ICPPW'05). IEEE, 2005, pp. 45–52.

[40] F. Van Lingen et al., "Grid enabled analysis: architecture, prototype and status," 2005.

[41] S. L. Kiani, A. Anjum, M. Knappmeyer, N. Bessis, and N. Antonopoulos, "Federated broker system for pervasive context provisioning," Journal of Systems and Software, vol. 86, no. 4, 2013, pp. 1107–1123.

A Mathematical Model for Predator–Prey Ecosystems Facing Climate Changes

Katsumi Sakata
Department of Life Science and Informatics
Maebashi Institute of Technology
Maebashi, Japan
e-mail: ksakata@maebashi-it.ac.jp

Toshiyuki Saito
Department of Radiation Effects Research
National Institute of Radiological Sciences
Chiba, Japan
e-mail: saito.toshiyuki@qst.go.jp

Hajime Ohyanagi
Computational Bioscience Research Center
King Abdullah University of Science and Technology
Thuwal, Kingdom of Saudi Arabia
e-mail: hajime.ohyanagi@kaust.edu.sa

Setsuko Komatsu
Department of Environmental and Food Sciences
Fukui University of Technology
Fukui, Japan
e-mail: skomatsu@fukui-ut.ac.jp

***Abstract*—We developed a mathematical model for predator–prey ecosystems undergoing climate-related changes. The model introduces the amount of information transferred between the number of individuals of the predator and prey categories, and the regulation performance in a predator–prey ecosystem is measured by a reduction of Shannon entropy, which is achieved by predation events and decay in the ecosystem. We examine the model with a computer simulation for a well-studied bass–crayfish predator–prey ecosystem in a closed lake.**

Keywords-mathematical model; predator–prey ecosystem; climate change.

I. Introduction

Over the last three decades, environmental changes, such as global warming, desertification, and air pollution have worsened, and their effects on life systems is a serious concern [1]. Previous studies on the environmental responses of life systems have been conducted for specific networks, including genetic regulatory networks and ecological networks [2][3].

Schrödinger suggested that a life system takes orderliness from its environment and sustains itself at a fairly high level of orderliness, or at a fairly low level of thermodynamic entropy [4]. Kauffman investigated how the dynamic behavior of a Boolean network suddenly becomes orderly. He made the analogy that the behavior approximates cell fate which is characterized by expression patterns of multiple genes in an organism [5][6]. Barabási and Albert found that generic mechanisms form an ordered network structure with a scale-free property [7]. However, we could not find a mathematical model that clarifies the varying orderliness of biological systems undergoing environmental changes.

In this study, we quantify the environmental stimuli and orderliness achieved in state variables in life systems with Shannon entropy based on their probability distributions. The state variables represent the state of the system, such as expression levels in a genetic network. We then hypothesize a relationship between environmental changes and orderliness in the life systems. We validate the hypothesis on an ecosystem using numerical experiments on a computational model of differential equations for the ecosystem with the climate-shift model [8]. In the model, a climate-attribute change is modeled as a shift in the probability distribution of the climate attribute. We evaluate control performance by a difference of Shannon entropy as $\Delta H \equiv H(X) - H(X')$, where X and X' represent the state variable X at t_0 and at t_1 (unit time after t_0), respectively [9]. The Shannon entropy $H(X)$ indicates the uncertainty of X [10]. Section II includes our results and discussion, and Section III states our conclusion and future work.

II. Results and Discussion

We consider a predator–prey ecosystem in a closed lake (Figure 1a). The probability distribution of the number of viable predators, which we call “capacity”, varies according to the climate shift of a climate attribute against a range of climate attributes (survival region) in which the predator is viable. The predator capacity decreases with an increase in climate shift (Figure 1b). We derived (1), which shows that the Shannon entropy ($H(Y)$) of the number of predators decreases with an increase in the climate shift (Figure 1c):

$$H(Y)_{e+\delta e} \le H(Y)_{e}\,, \tag{1}$$

where e and δe indicate the level of the climate attribute and its increment. Generally, $I(X;Y) \le \min\{H(X), H(Y)\}$, thus

$$I(X;Y)^{U}_{e+\delta e} \le I(X;Y)^{U}_{e}\,, \tag{2}$$

where $I(X;Y)^U (\equiv H(Y))$ denotes an upper bound of the mutual information between X and Y. We merged (2) with an information–theoretic limit for general control systems [9], and thus obtained (3):

$$\Delta H^{U}_{e+\delta e} \le \Delta H^{U}_{e}\,, \tag{3}$$

where ΔH is the Shannon entropy reduction of the state variable X (the number of prey individuals) over the transition $X \rightarrow X'$ between t_0 and t_1 (unit time after t_0). It represents the control performance of the predator–prey ecosystem. Equations (2) and (3) suggest that the mutual information between the number of prey individuals (X) and predators (Y), as well as the control performance of the predator–prey ecosystem, decreases with an increase in climate shift. Furthermore, the control performance of the predator–prey ecosystem appears to degrade from the level of a closed-loop control system to an open-loop control system, based on the information–theoretic limits of control [9].

Numerical experiments on a well-studied bass–crayfish predator–prey ecosystem in a closed lake [11] validate the degradation of the control performance suggested by the model mentioned above.

The derived inequalities, (1), (2) and (3), are independent of the dynamics of the target ecosystem. Thus, our model can be applied to analyses of ecosystems in which the dynamics are unknown. Furthermore, our model and the numerical experiment results suggest that the maintenance of predator numbers is effective for protecting predator–prey ecosystems against climate-related changes.

III. Conclusion and Future Work

We developed an information–theoretic predator–prey ecosystem model that is independent of the dynamics of the ecosystem, and validated the model through numerical experiments.

Acknowledgment

This study was supported by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (Grant No. 16K00399). We thank Natalie Kim, PhD, from Edanz Group (www.edanzediting.com/ac) for editing a draft of this extended abstract.

References

[1] T. Stocker et al. “The Physical Science Basis – Summary for Policymakers,” IPCC WGI AR5 (Report), p. 4, 2013.

[2] G. Cramer, K. Urano, S. Delrot, M. Pezzotti, and K. Shinozaki, “Effects of Abiotic Stress on Plants: A Systems Biology Perspective,” BMC Plant Biol., vol. 11, p. 163, 2011, doi: 10.1186/1471-2229-11-163.

[3] R. Horan, E. Fenichel, K. Drury, and D. Lodge, “Managing Ecological Thresholds in Coupled Environmental-Human Systems,” PNAS, vol. 108, pp. 7333–7338, May 2011, doi: 10.1073/pnas.1005431108.

[4] E. Schrödinger, “What is Life?: With Mind and Matter and Autobiographical Sketches,” Cambridge University Press, 1992.

[5] S. Kauffman, “The Origins of Order: Self-Organization and Selection in Evolution,” Oxford University Press, June 1993.

[6] F. H. Westhoff, B. Yarbrough, and R. Yarbrough, “Complexity, Organization, and Stuart Kauffman’s the Origins of Order,” J. Econ. Behav. Organ., vol. 29, pp. 1–25, 1996, doi: 10.1016/0167-2681(95)00049-6.

[7] A. Barabási and R. Albert, “Emergence of Scaling in Random Networks,” Science, vol. 286, pp. 509–512, Oct. 1999, doi: 10.1126/science.286.5439.509.

[8] Q. Schiermeier, “Climate and Weather: Extreme Measures,” Nature, vol. 477, pp. 148–149, 2011, doi:10.1038/477148a.

[9] H. Touchette and S. Lloyd, “Information-Theoretic Limits of Control,” Phys. Rev. Lett., vol. 84, pp. 1156–1159, 2000, doi: 10.1103/PhysRevLett.84.1156.

[10] D. Robinson, “Entropy and Uncertainty” Entropy, vol. 10, pp. 493–506, 2008, doi: 10.3390/e10040493.

[11] K. Drury and D. Lodge, “Using Mean First Passage Times to Quantify Equilibrium Resilience in Perturbed Intraguild Predation Systems,” Theor. Ecol., vol. 2, pp. 41–51, 2009, doi: 10.1007/s12080-008-0027-z.

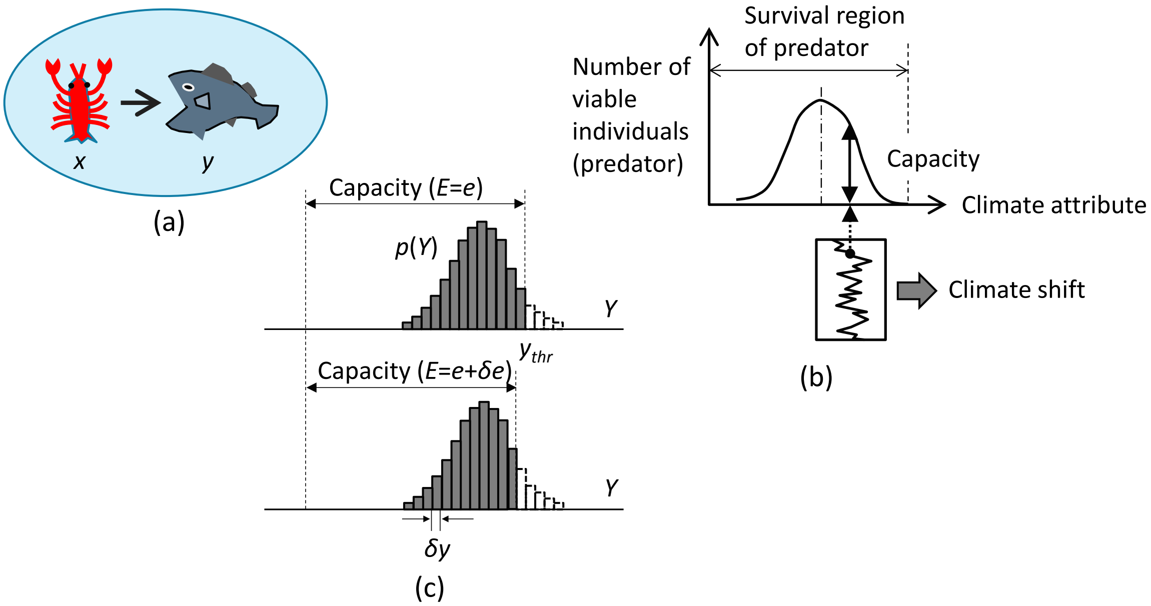

Figure 1. An ecosystem model: (a) Predator–prey ecosystem in a closed lake. The arrow denotes feeding relationship. (b) Number of viable predators and climate shift. (c) Probability distribution of the number of predators before (upper panel) and after (lower panel) an increase in climate shift by δe.