Zbornik 25. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2022

Zvezek H

Proceedings of the 25th International Multiconference

INFORMATION SOCIETY – IS 2022

Volume H

Vseprisotne zdravstvene storitve in pametni senzorji
Pervasive Health and Smart Sensing

Uredniki / Editors

Nina Reščič, Oscar Mayora, Daniel Denkovski

http://is.ijs.si

13. oktober 2022 / 13 October 2022
Ljubljana, Slovenija

DRAFT – NOT FOR PUBLICATION

Uredniki:

Nina Reščič
Department of Intelligent Systems
Institut »Jožef Stefan«, Ljubljana

Oscar Mayora
Digital Health Lab, Fondazione Bruno Kessler
Trento, Italy

Daniel Denkovski
Computers Science and Computer Engineering
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Lana Zemljak
Oblikovanje naslovnice: Vesna Lasič

Dostop do e-publikacije:
http://library.ijs.si/Stacks/Proceedings/InformationSociety

Ljubljana, oktober 2022

Informacijska družba
ISSN 2630-371X

DRAFT – NOT FOR PUBLICATION

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2022

Petindvajseta multikonferenca *Informacijska družba* je preživela probleme zaradi korone. Zahvala za skoraj normalno delovanje konference gre predvsem tistim predsednikom konferenc, ki so kljub prvi pandemiji modernega sveta pogumno obdržali visok strokovni nivo.

Pandemija v letih 2020 do danes skoraj v ničemer ni omejila neverjetne rasti IKTja, informacijske družbe, umetne inteligence in znanosti nasploh, ampak nasprotno – rast znanja, računalništva in umetne inteligence se nadaljuje z že kar običajno nesluteno hitrostjo. Po drugi strani se nadaljuje razpadanje družbenih vrednot ter tragična vojna v Ukrajini, ki lahko pljuskne v Evropo. Se pa zavedanje večine ljudi, da je potrebno podpreti stroko, krepi. Konec koncev je v letu 2022 v veljavo stopil nov raziskovalni zakon, ki bo izboljšal razmere, predvsem bo leto za letom povečeval sredstva, namenjena za znanost.

Letos smo v multikonferenco povezali enajst odličnih neodvisnih konferenc, med njimi »Legende računalništva«, s katero postavljamo nov mehanizem promocije informacijske družbe. IS 2022 zajema okoli 200 predstavitev, povzetkov in referatov v okviru samostojnih konferenc in delavnic, ter 400 obiskovalcev. Prireditev so spremljale okrogle mize in razprave ter posebni dogodki, kot je svečana podelitev nagrad. Izbrani prispevki bodo izšli tudi v posebni številki revije Informatica (http://www.informatica.si/), ki se ponaša s 46-letno tradicijo odlične znanstvene revije. Multikonferenco Informacijska družba 2022 sestavljajo naslednje samostojne konference:

- Slovenska konferenca o umetni inteligenci
- Izkopavanje znanja in podatkovna skladišča
- Demografske in družinske analize
- Kognitivna znanost
- Kognitonika
- Legende računalništva
- Vseprisotne zdravstvene storitve in pametni senzorji
- Mednarodna konferenca o prenosu tehnologij
- Vzgoja in izobraževanje v informacijski družbi
- Študentska konferenca o računalniškem raziskovanju
- Matcos 2022

Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi ACM Slovenija, SLAIS, DKZ in druga slovenska nacionalna akademija, Inženirska akademija Slovenije (IAS). V imenu organizatorjev konference se zahvaljujemo združenjem in institucijam, še posebej pa udeležencem za njihove dragocene prispevke in priložnost, da z nami delijo svoje izkušnje o informacijski družbi. Zahvaljujemo se tudi recenzentom za njihovo pomoč pri recenziranju.

S podelitvijo nagrad, še posebej z nagrado Michie-Turing, se avtonomna stroka s področja opredeli do najbolj izstopajočih dosežkov. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejel XXXXXXXXXXXXX. Priznanje za dosežek leta pripada XXXXXXXXXXXX. »Informacijsko limono« za najmanj primerno informacijsko potezo je prejela storitev XXXXXXXXXXXXX, »informacijsko jagodo« kot najboljšo potezo pa XXXXXXXXXXXXXXX. Čestitke nagrajencem! Opomba: imena nagrajencev bodo objavljena po proglasitvi. Rok za manjše popravke referatov v zborniku je 21.10.2022.

Mojca Ciglarič, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

DRAFT – NOT FOR PUBLICATION

FOREWORD - INFORMATION SOCIETY 2022

The 25th *Information Society Multiconference* (http://is.ijs.si) survived the COVID-19 problems. The multiconference survived due to the conference chairs who bravely decided to continue with their conferences despite the first pandemics in the modern era.

The COVID-19 pandemic from 2020 till now did not decrease the growth of ICT, information society, artificial intelligence and science overall, quite on the contrary – the progress of computers, knowledge and artificial intelligence continued with the fascinating growth rate. However, the downfall of societal norms and progress seems to slowly but surely continue along with the tragical war in Ukraine. On the other hand, the awareness of the majority, that science and development are the only perspective for prosperous future, substantially grows. In 2020, a new law regulating Slovenian research was accepted promoting increase of funding year by year.

The Multiconference is running parallel sessions with 200 presentations of scientific papers at eleven conferences, many round tables, workshops and award ceremonies, and 400 attendees. Among the conferences, “Legends of computing” introduce the “Hall of fame” concept for computer science and informatics. Selected papers will be published in the Informatica journal with its 46-years tradition of excellent research publishing.

The Information Society 2022 Multiconference consists of the following conferences:

- Slovenian Conference on Artificial Intelligence
- Data Mining and Data Warehouses
- Cognitive Science
- Demographic and family policy
- Cognitonics
- Legends of computing
- Pervasive health and smart sensing
- International technology transfer conference
- Education in information society
- Student computer science research conference 2022
- Matcos 2022

The multiconference is co-organized and supported by several major research institutions and societies, among them ACM Slovenia, i.e. the Slovenian chapter of the ACM, SLAIS, DKZ and the second national academy, the Slovenian Engineering Academy. In the name of the conference organizers, we thank all the societies and institutions, and particularly all the participants for their valuable contribution and their interest in this event, and the reviewers for their thorough reviews.

The award for life-long outstanding contributions is presented in memory of Donald Michie and Alan Turing. The Michie-Turing award was given to XXXXXXXXXXXXXX for his life-long outstanding contribution to the development and promotion of information society in our country. In addition, the yearly recognition for current achievements was awarded to XXXXXXXXXXXXXXXXX. The information lemon goes to XXXXXXXXXXXXXX. The information strawberry as the best information service last year went to XXXXXXXXXXXXXXXXXX. Congratulations!
The recipients of awards will be included after formal announcement. Small modifications of papers are acceptable till 21.10.2022.

Mojca Ciglarič, Programme Committee Chair
Matjaž Gams, Organizing Committee Chair

DRAFT – NOT FOR PUBLICATION

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia
Sergio Campos-Cordobes, Spain
Shabnam Farahmand, Finland
Sergio Crovella, Italy

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Mitja Lasič
Blaž Mahnič

Programme Committee

Mojca Ciglarič, chair
Bojan Orel,
Franc Solina,
Viljan Mahnič,
Cene Bavec,
Tomaž Kalin,
Jozsef Györkös,
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Mitja Luštrek
Marko Grobelnik

Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček

Andrej Ule
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah
Niko Zimic
Rok Piltaver
Toma Strle
Tine Kolenik
Franci Pivec
Uroš Rajkovič
Borut Batagelj
Tomaž Ogrin
Aleš Ude
Bojan Blažica
Matjaž Kljun
Robert Blatnik
Erik Dovgan
Špela Stres
Anton Gradišek

KAZALO / TABLE OF CONTENTS

Zbornik 25. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2022

Zvezek H

Proceedings of the 25th International Multiconference

INFORMATION SOCIETY – IS 2022

Volume H

Vseprisotne zdravstvene storitve in pametni senzorji
Pervasive Health and Smart Sensing

Uredniki / Editors

Nina Reščič, Oscar Mayora, Daniel Denkovski

http://is.ijs.si

13. oktober 2022 / 13 October 2022
Ljubljana, Slovenija

DRAFT – NOT FOR PUBLICATION

PREDGOVOR

Pomen digitalnih zdravstvenih storitev v zadnjih desetletjih nenehno narašča. Staranje prebivalstva je neposredno povezano s povečevanjem števila kroničnih bolnikov, ki jim razvoj medicine sicer omogoča zdravstveno oskrbo in posledično tudi podaljševanje življenjske dobe, hkrati pa je zdravstveni sistem zaradi tega dodatno obremenjen. Razvoj digitalne tehnologije je prinesel vse več dostopnih orodij za stroškovno učinkovito vzdrževanje in izboljševanje zdravja in kakovosti življenja ter obenem pripomogel k razbremenitvi zdravstva. Nedavna pandemija COVID-19 je dodatno poudarila potrebo po zagotavljanju zdravstvenih storitev na daljavo. Tehnološki napredek je sicer nekoliko upočasnjen zaradi zakonodaje, saj digitalne tehnologije ne morejo nositi odgovornosti zaradi napačnih zdravstvenih odločitev, prav tako je zelo pomembno tudi varstvo podatkov in spoštovanje zasebnosti pacientov. Sodelovanje vseh pomembnih družbenih, zdravstvenih in pravnih akterjev tako pomaga postaviti stabilnejše in zanesljivejše temelje za razvoj, uvajanje in uporabo digitalnih zdravstvenih tehnologij in storitev. Vseprisotne zdravstvene storitve in uporaba pametnih senzorjev so tako ključni deli digitalnega zdravja. Pametni senzorji in razne nosljive naprave omogočijo spremljanje na daljavo in in tako dodatno podprejo spremljanje zdravstvenega stanja bolnikov v klinikah in izven njih. Dodatno lahko pametni in vseprisotni sistemi za spremljanje zdravja zmanjšajo določena tveganja in odkrijejo težave v zgodnejših fazah bolezni.

Konferenco »Vseprisotni zdravstveni sistemi in pametni senzorji« organizira EU projekt WideHealth, t.i. »widening« projekt, katerega glavni namen je vzpostavljanje trajnostne mreže raziskav med vključenimi partnerji. Konzorcij projekta sestavlja pet partnerjev (trije »widening« in dva »non-widening«), ki preko izmenjav in drugih raziskovalnih sodelovanj poglabljajo znanje na treh glavnih področjih: »data-driven healthcare«, »human factors in pervasive health« in »federated learning«. Namen konference »Vseprisotne zdravstvene storitve in pametni senzorji« je izmenjava strokovnega znanja in napredka raziskav na omenjenih področjih. Na konferenci bo predstavljenih 12 prispevkov, ki se osredotočajo na različne vidike pametnega zaznavanja in vsesplošnega zdravja. V prvem delu konference so vključeni prispevki, ki se osredotočajo na prepoznavanje človeških aktivnosti z uporabo nosljivih naprav (vključno z novejšimi tehnologijami, npr. pametnimi očali). Prispevki drugega dela konference se osredotočajo na objektivno in subjektivno spremljanje duševnega zdravja. V zadnjem, tretjem, delu so zbrani prispevki, ki predlagajo nove aplikacije, metodologije in IKT rešitve za vseprisotne zdravstvene sisteme ter izboljšanje varnosti in zasebnosti v takih sistemih.

FOREWORD

The importance of digital health is constantly growing in recent decades. The reasons are well known: on the one hand, the aging of the population is producing an increasing number of chronic patients, and the progress of medicine is keeping them alive and in need of care; on the other hand, the progress of digital technology is creating an increasing number of available tools to maintain and/or increase health and quality of life cost-effectively. The recent COVID-19 pandemic has further emphasized the need to provide remote medical services to patients, which has boosted the emergence and adoption of digital technologies, especially in telehealth and telemedicine. Technological advances have been slowed mainly due to legislation since bad medical decisions cannot be blamed on digital technologies, and security and privacy issues also cannot be neglected. However, the involvement of all the important social, medical, and legal actors helps set up a more stable and reliable foundation for developing, deploying, and using digital health technologies and services. Pervasive health and smart sensing are crucial parts of digital health. Smart sensors and wearables can augment the healthcare system, enabling remote monitoring and supporting the patient's medical condition in and out of the clinics. Furthermore, smart and pervasive health monitoring systems can reduce death risks, identifying the issues at earlier stages of the diseases. They are the main focus of our "Pervasive Health and Smart Sensing" conference, as the name suggests.

The conference is organized by the EU WideHealth project, a widening project that aims to conduct research on pervasive eHealth and establish a sustainable network of research and dissemination across Europe. It connects five partners (3 widening and two non-widening) to share and develop their research on three main topics: data-driven healthcare, human factors in pervasive health, and federated machine learning. The Pervasive Health and Smart Sensing conference aims to share expertise and research advancements in these areas. The 12 papers we have accepted at the conference focus on different aspects of smart sensing and pervasive health. Several works utilize wearable devices (including new types, i.e., smart glasses) and machine learning for human activity recognition. Several others focus on objective and subjective monitoring of mental health. Finally, there are papers proposing new applications, methodologies, and ICT solutions for pervasive health and improving the security and privacy in such systems.

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Oscar Mayora

Daniel Denkovski

Nina Reščič

Orhan Konak

Hristijan Gjoreski

Valentin Rakovic

Diogo Branco

Monika Simjanoska

Martin Gjoreski

Tiago Guerreiro

Tome Eftimov

Vito Janko

Venet Osmani

Junoš Lukan

Eftim Zdravevski

Optimized Method for Walking Detection by Wristband with Accelerometer Sensor

Aleksander Hrastič
alekshrastic@gmail.com
University of Ljubljana, Faculty of Electrical Engineering
Ljubljana, Slovenia

Matej Kranjec
matejkranjec04@gmail.com
University of Ljubljana, Faculty of Electrical Engineering
Ljubljana, Slovenia

Primož Kocuvan
primoz.kocuvan@ijs.si
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia

ABSTRACT

This paper presents the part of the gait impairment measurement algorithm, which consists exclusively of the walking detection algorithm. The purpose of the optimized algorithm is to improve the detection of walking. Today's embedded devices (like wristbands) have low-level interrupts that detect steps and, consequently, walking. The problem is that these could be inaccurate in some cases. For example, a person can swing with a hand while sitting, and the device will detect steps. The importance of walking detection is crucial for gait impairment measurements, as gait data should only be collected when a person is walking in a "normal" manner and not performing any other walking-like activities. An algorithm to measure gait impairment will be developed in the later stages of this study. We focused on improving the walking detection algorithm with statistical methods in both time and frequency domains in contrast to computationally expensive algorithms that use machine learning. The walk detection algorithm has been optimized based on data collected by a wristband with a 3-axis accelerometer sensor. With our optimized algorithm, we got an average accuracy of 89.4%. We can conclude that our proposed method works well for detecting when a person is walking normally. The algorithm successfully detects "not natural walking" scenarios when the person is sitting and swinging their hand or walking with extreme hand movements.

KEYWORDS

wristband, walking detection, FFT, periodogram, activity recognition, hamming window

1 INTRODUCTION

Every year number of older adults fall and injure themselves. For example, in Western Europe, in 2017 alone, 13840 per 100,000 older adults over the age of 70 are known to have fallen and injured themselves to the extent of medical assistance [1]. To prevent such phenomena, measurement and monitoring of gait deterioration in the elderly must be developed. One part of the such algorithm must consist of a walking detection algorithm that detects whether a person is walking or not in a non-invasive way.

Wristbands with various sensors (e.g., accelerometer, gyroscope) have proven to be an excellent technology for automatic and non-invasive detection of daily activities. In this case, we can use the acceleration vector data from the accelerometer sensor to detect whether the person is walking or not. However, many studies have focused on using machine learning algorithms, which provide high accuracy but are computationally expensive to implement in embedded systems (wristbands).

We present to you a computationally inexpensive algorithm for detecting whether a person is walking or not. Furthermore, the algorithm can detect walking and other daily activities similar to the walking pattern and can be used on a low-power wristband system. In our case, the most crucial aspect of our gait detection algorithm should be to detect as minimal cases as possible where the algorithm predicts that the person is walking naturally. Still, in the actual case, the person is performing other activities.

An algorithm to measure gait deterioration (our next step) will help the elderly prevent falls. The algorithm will monitor a person's gait daily, and when a person's gait deteriorates dramatically, it will notify caregivers of increased chances of falling. Accordingly, caregivers can take the person to rehabilitative walking therapy or give them more care.

2 RELATED WORK

Advances in the accuracy and accessibility of wearable sensing technology (e.g., fitness bands and smartwatches) has allowed researchers and practitioners to utilize different types of wearable sensors to detect walking.

In [2] the authors explored the possibility of detecting activity from a smartphone-based accelerometer sensor. They used smartphones placed in different positions(backpack, pocket, in hand) to collect data when doing an activity (walking, fast walking, slow walking, running). To reduce complexity, they computed the magnitude of the 3-axis accelerometer. The magnitude vector is then processed using time and frequency domain statistical techniques. Finally, the statistical methods on the time-domain measures are applied for state recognition, while the statistical techniques on the frequency-domain features are implemented for walking movement distinction.

In [3], they use a smartphone with a gyroscope to collect data. They propose a new algorithm based on Fast Fourier Transform (FFT) [4] to identify the walking activity of a user who can perform different activities and hold the smartphone differently. The proposed algorithm (FFT) was able to achieve superior overall performance compared to the other two best-performing algorithms (Short Time Fourier Transform (STFT) and Standard Deviation Threshold (STD TH)).

The authors in [5] propose an algorithm that classifies human activity in real time based on data from an accelerometer attached to the subject. The algorithm uses dynamic linear discriminant analysis (LDA), which can dynamically update classifier matrices without storing all training samples in memory. LDA is used to find a transformation of extracted features that separate data distribution into different classes while minimizing the distribution of data of the same class in the newly transformed space.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia

© 2022 Copyright held by the owner/author(s).

Compared to the state-of-the-art algorithm, our paper aims to combine the FFT and threshold algorithm from [2] and axis selection algorithm from [3] while adding an upper bound threshold to detect exaggerated hand movements and excluding them from false positives.

3 METHODOLOGY

The main goal of the research was to improve, or rather optimize, the gait detection algorithm based on statistical methods and frequency coefficients obtained from the measurements of the Empatica E4 bracelet accelerometer. To achieve this, we had to record data with the wristband while performing various activities and test the performance of our algorithm on the collected data. The data was collected using the Empatica E4 wristband [6]. The sampling frequency for the 3-axis accelerometer is 32 Hz. It has an 8-bit resolution and a default range of ± 2 g with sensitive motion detection along three axes: x, y and z.

3.1 Data collection

An Empatica E4 bracelet was used for data collection and placed on the subject's left wrist. The wristband was connected to a smartphone via Bluetooth and streamed real-time data that was uploaded to the Empatica server. We have designed various routes and defined actions on these routes, which the subjects should carry out. Data was then collected from different individuals who wore the bracelet and followed the planned route. Various walking styles were performed on the designed paths, such as normal walking, slow walking, fast walking, and walking with random hand movements. Some actions involved sitting in a chair and performing arm swings that are similar in motion to arm swings if the subject were walking.

In [2], data was gathered from 7 individuals doing different walking styles (slow walking, fast walking, normal walking). They collected 27 samples. In our case, the data was collected from 4 individuals shown in Table 1. We also collected a total of 27 samples.

Table 1: Table of participants

Participant	Gender	Age	Disability
A	Male	22	None
B	Male	24	None
C	Male	83	Difficulty walking
D	Female	79	None

Figure 1 shows all three axes of raw accelerometer data collected from the Empatica wristband. During an interval between 20 seconds and 70 seconds, the subject wearing the Empatica walked in a straight line.

3.2 Algorithm

Our optimized algorithm combines aspects from two papers [2][3]. From the first paper, we used the modified periodogram thresholding algorithm to detect walking only when the minimum required hand activity is reached in frequency ranges that correspond to human walking activity. From the other paper, we implemented this on the 3-axial accelerometer. For each time window, we select and process only the data on an axis with the most variance. Our contribution to the algorithm for walking detection is a combination of the two, with added upper bound threshold to prevent false walk detection when a subject is swinging a hand uncontrollably, shown in the main algorithm 2 on line (18). This is more thoroughly described below.

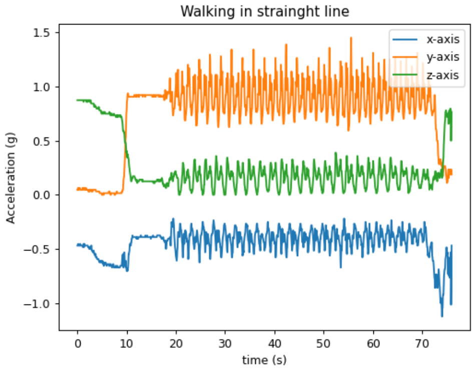


Figure 1: Example of raw signal from accelerometer sensor

First, we use the time windowing algorithm (Algorithm 1) to process the data in a shorter time frame. Then, we need to divide the data into time windows (W). We found empirically that it is best if the data window length (w_t) is 5 seconds with a 2.5-second overlap (o_t).

The time windows are then filtered (x, y, and z axes are filtered separately) with a high-pass Butterworth filter to capture the signal proportionally (symmetrically) with respect to the time axis. The general shape of the frequency response of a Butterworth filter is defined as equation (1). Where f_c is the cutoff frequency, ϵ is the passband gain, and n is the order of the filter. We chose the order of n to be 5. We chose it heuristically. For our example, the cutoff frequency was set to 1 Hz.

$$H(f) = \frac{1}{\sqrt{1 + \epsilon^2 \left(\frac{f}{f_c}\right)^{2n}}} \tag{1}$$

In the next step, we detect which of the three axes is the most sensitive for each time window. This step is accomplished by calculating each filtered axis's standard deviation (STD) separately and selecting the one axis with the highest STD value.

Afterward, we compute modified periodogram coefficients from the most sensitive axis for each window. To calculate the modified periodogram in the algorithm 2 we multiplied signal windows with Hamming window, which is defined as (2). The Hamming window is an extension of the Hamming window and is a semi-cosine bell-shaped curve.

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq M-1 \tag{2}$$

Where N represents total length of the window.

For each time window, two main conditions had to be met for it to be classified as "walking."

Modified periodogram coefficients are computed using equation (3). Time windows that met the first condition (4), need to have computed modified periodogram coefficients that are on the interval 0.6 to 2 Hz ($S_{xx}(f_i)$ where f_i represents all the frequencies inside the interval) and had higher mean than the mean of coefficients in the interval outside 0.6 to 2 Hz ($S_{xx}(f_o)$ where f_o represents all the frequencies outside the interval).

$$S_{xx}(f) = |F(f)|^2 \tag{3}$$

Where F(f) is output from FFT at desired frequency f.

$$\overline{S_{xx}(f_i)} > \overline{S_{xx}(f_o)} \tag{4}$$

The second condition (5) that had to be met for the time window is that the STD of the vector norm of the unfiltered signal must be between 0.3 g and 0.7 g. The lower limit (0.3 g) ensures that walking is not falsely detected when the subject is not moving. The higher limit (0.7 g) prevents walking detection when subjects move their arms uncontrollably. Both limits were determined empirically based on our collected data set. The norm is calculated using equation (6) where x, y, and z are the time-windowed accelerometer signal vectors, each representing an axis. "i" means the same index on all three axes, ranging from 1 to the length of the time window (this is calculated from the raw signal using the (7) where N is a number of samples in a time window). Time windows that satisfy both conditions are classified as "walking"; all other window cases are classified as "not walking."

$$0.3 < \sigma_{norm} > 0.7 \tag{5}$$

$$norm_i = \sqrt{x_i^2 + y_i^2 + z_i^2} \tag{6}$$

$$\sigma_{norm} = \sqrt{\frac{\sum_{i=1}^{N}(norm_i - \overline{norm})^2}{N}} \tag{7}$$

Algorithm 1 for windowing

```
Require: $(acc_x, acc_y, acc_z), w_t, o_t$        ▷ $o_t$ is the overlap $w_t$ = length of the window
Ensure: $(W_x, W_y, W_z)$
  $W \leftarrow []$
  $s_t \leftarrow 0$                          ▷ $s_t$ = start index of windowl
  $e_t \leftarrow s_t + w_t$                  ▷ $e_t$ = end index of window
  for all $(acc_x, acc_y, acc_z)$ do
      while $s_t \leq N$ do        ▷ N is the number of samples in a window, i represents index of current sample in a loop
          $acc_i' \leftarrow acc_i[s_t : e_t]$
          $W \leftarrow W + [acc_i']$
          $s_t \leftarrow s_t + o_t$
          $e_t \leftarrow e_t + o_t$
      end while
  end for
```

Algorithm 2 for detection of walking

```
  function STATIONARY($d$)
      $n_2 \leftarrow norm(d)$
      $m \leftarrow n_2[:] - mean(n_2)$
      $sd \leftarrow std(m)$
  end function
Require: $W$
Ensure: $boolean[]$
  for all $(W_i)$ do    ▷ i represents index of current window in a loop
      if $length(Stationary(W_i)) \geq 0$ then
          $W_x \leftarrow ButterworthFilter(W_i(x))$
          $W_y \leftarrow ButterworthFilter(W_i(y))$
          $W_z \leftarrow ButterworthFilter(W_i(z))$
          $n_{meanx2} \leftarrow avg(norm(W_x))$
          $n_{meany2} \leftarrow avg(2norm(W_y))$
          $n_{meanz2} \leftarrow avg(2norm(W_z))$
          $am \leftarrow argmax\{n_{meanx2}, n_{meany2}, n_{meanz2}\}$
          $pg \leftarrow periodogram(am, hamming)$    ▷ hamming is the windowing function
          if $(max(am) - min(am) > 0.3)$ and $pg(f > 0.6$ and $f < 2)$ then
              $boolean \leftarrow boolean + [1]$
          else
              $boolean \leftarrow boolean + [0]$
          end if
      end if
  end for
```

4 RESULTS

We ran the algorithm on different recordings taken with the Empatica wristband. Slow and fast straight walking, stair climbing, and sitting involving arm swing.

Figure 2 shows a dot plot where zero (on the y-axis) represents "no walking," and one represents "walking." The x-axis represents time (in seconds). Dots on the x-axis are linearly spaced by 2.5 seconds. During the first 8 seconds, the subject was standing, so for this part of the signal, the algorithm correctly classified it as "not walking." After 8 seconds, the subject started to walk in a straight line, and the algorithm correctly detected this activity as "walking." For our example, we can confirm that the algorithm works correctly under normal walking conditions.

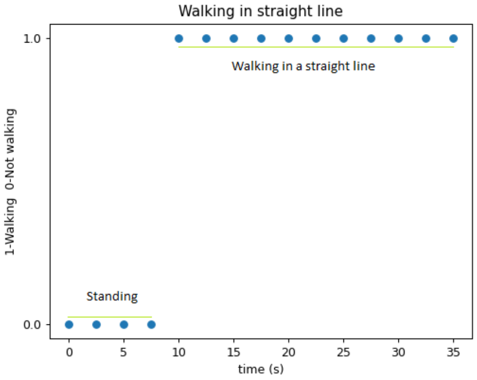


Figure 2: Proposed algorithm used on straight walking activity, recorded Empatica E4 wristband

Table 2: Table of activities and their accuracy

	Activity	Detected as "Walking"	Detected as "Not walking"	Accuracy
Walking	Walking and swinging hand	47.5s	7.5s	86.4%
	Slow walking	52.5s	2.5s	95.5%
	Stair climbing	17.5s	37.5s	68%
Not Walking	Fast walking	5s	52s	91.2%
	Standing	7.5s	47.5s	86.4%
	Sitting	0s	55s	100%
	Sitting and swinging hand	2.5s	52.5s	95.5%
	Walking and uncontrollably moving hand	0s	55s	100%

Figure 3 shows an example of a recording where several human activities are present, such as standing, sitting on a chair and performing random hand movements, walking and performing hand movements, and walking and performing exaggerated hand movements. It can be seen that the algorithm had difficulty in gait classification when high-amplitude arm movements were present during the subject's gait. This is because gait characteristics are lost in the noise of high-amplitude hand movements. For our purposes, the issue is not critical because the future end goal is to measure the subjects' gait impairment, so there is no problem in discarding the parts of the signal where the person does not walk in a "natural" way. However, we can also observe that there was deviation when the subject sat down and started swinging his arm (One instance at the 42nd second where the algorithm should predict "not walking" but instead, it predicted "walking"). On Figure 4 at about 78th second, we can observe that the algorithm detected sitting as if it were walking.

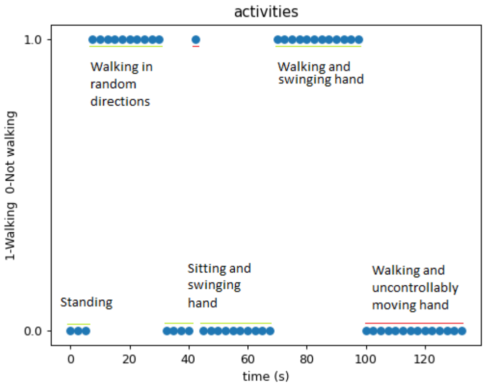


Figure 3: Proposed algorithm used on multiple activities, recorded on Empatica E4 wristband

We require that we have the least amount of false positives in our data set because we want to detect only the scenarios where a person is walking the most naturally. This is a typical binary classification problem, where the final results are shown in Table 2. The first three activities (walking and swinging hand, slow walking...) are considered natural walking and should be detected as walking. The next 6 (Fast walking, standing, sitting, sitting and swinging hand, walking and uncontrollably moving hand) activities should be considered as "not walking" because they are less optimal for feature collection for the algorithm that will be implemented in the next stages of this study. The study we are conducting is primarily meant for the elderly, so we categorized the "fast walking" scenario as not walking, as it is not common for the elderly to walk fast. In the stairs climbing case, the algorithm did not perform very well, but that is not relevant in our case. More importantly, in the last 6 cases algorithm performs well in detecting true negatives.

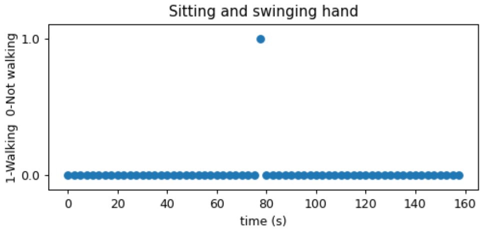


Figure 4: Proposed algorithm used when sitting and swinging hand, recorded on Empatica E4 wristband

5 CONCLUSION

In the related work, we described the state-of-the-art algorithms used in today's many applications. For this research, we selected two algorithms from many of them and expanded (optimized) the work for our purposes. The results of our algorithm were able to detect when a person was walking normally, slowly, and quickly. In addition, the algorithm correctly detected cases when a person does not walk while sitting but swings his arm.

To measure gait impairment, we only want to use time windows of the signal where we are certain that the person is walking and that there are no additional "unnecessary" hand movements. In the future, we will further improve the algorithm so that the deterioration of walking, our final goal, can be measured correctly.

REFERENCES

[1] Juanita A Haagsma et al. 2020. Falls in older aged adults in 22 european countries: incidence, mortality and burden of disease from 1990 to 2017. *Injury Prevention*, 26, (Feb. 2020), i67–i74. DOI: 10.1136/injuryprev-2019-0433 47.

[2] Chalne T ornqvist. 2017. *Walking movement detection using stationary stochastic methods on accelerometer data.* MA thesis. Lund University.

[3] Guodong Qi and Baoqi Huang. 2018. Walking detection using the gyroscope of an unconstrained smartphone. In (Jan. 2018), 539–548. ISBN: 978-3-319-66627-3. DOI: 10.1007/978-3-319-66628-0_51.

[4] E. O. Brigham and R. E. Morrow. 1967. The fast fourier transform. *IEEE Spectrum*, 4, 12, 63–70. DOI: 10.1109/MSPEC.1967.5217220.

[5] Yen-Ping Chen, Jhun-Ying Yang, Shun-Nan Liou, Gwo-Yun Lee, and Jeen-Shing Wang. 2008. Online classifier construction algorithm for human activity detection using a tri-axial accelerometer. *Applied Mathematics and Computation*, 205, 2, 849–860. Special Issue on Advanced Intelligent Computing Theory and Methodology in Applied Mathematics and Computation. DOI: https://doi.org/10.1016/j.amc.2008.05.099.

[6] [n. d.] Medical devices, ai and algorithms for remote patient monitoring. empatica. https://www.empatica.com/.

Android Integration of a Machine Learning Pipeline for Human Activity Recognition

Viktor Srbinoski, Daniel Denkovski, Emilija Kizhevska, Hristijan Gjoreski
Faculty of Electrical Engineering and Information Technologies,
Ss. Cyril and Methodius University in Skopje, N. Macedonia, Jozef Stefan Institute, Slovenia
viktor_srbinoski@hotmail.com, danield@feit.ukim.edu.mk, emilija.kizhevska@ijs.si, hristijang@feit.ukim.edu.mk

ABSTRACT

In the last decade, smartphones have seen a serious growth in the processing power. Coupled with greater affordability this has led to a worldwide smartphone ubiquity. Alongside the advances in processing and battery technology, there are great advances in sensor technology as well, and every smartphone today comes equipped with multiple sensors: accelerometer, gyroscope, magnetometer etc. The sensory data is already being used in a variety of applications, among which several focus on the human activity recognition. In this paper, we propose a smartphone Android integration of a machine learning pipeline for recognizing human activities. The proposed approach uses the 3-axis accelerometer in the smartphone, processes the data in real time, and then a machine learning model recognizes the user's activities in real time: walking, running, jumping, cycling and standing still. The proposed Recurrent Neural Network model and its machine learning pipeline are developed on a publicly open activity dataset, which are then implemented into the Android application and once again validated on a dataset recorded with a smartphone itself.

KEYWORDS

Human activity recognition, machine learning, Android integration, Tensorflow Light, recurrent neural network, accelerometer, magnitude.

1 INTRODUCTION

Human Activity Recognition (HAR) is the process of examining data from one or multiple sensors and determining which (if any) activity is being performed. The sensors are traditionally placed on key points on the human body and contain composite data (accelerometer, gyroscope, magnetometer data, etc.). Advances in sensor technology have made sensors more compact and precise over the years, but most importantly more affordable. Today these sensors can be found in the standard package of any smartphone.

The purpose of this paper is to leverage these smartphone sensors to perform HAR in real time, by utilizing an Android application which continuously reads its own sensor data, instead of using the traditional dedicated wearable sensors. The premise is that the smartphone sensors have reached the required quality to be comparable to the wearable sensors in accuracy [1]. The benefit of this approach is that it is much more convenient to use smartphone sensors for the common user, as smartphones have become ubiquitous.

Human activity recognition is a popular topic, which has been worked on extensively in the recent years [2]. Practical applications for HAR are mainly in improvement of the quality of life and medicine. A great example of HAR models being used in medicine can be found in paper [3], which focuses on fall detection mainly for the elderly population.

Using dedicated wearable sensors to recognize activities is the most common approach. Smartwatch is usually equipped with the same sensors as the smartphones and has a much more fixed position on the body (tightly around the wrist). The drawback is that the arms are more prone to random movement which introduces noise into the system and makes HAR more difficult. A detailed analysis on these issues can be found in paper [4].

Using data from smartphone sensors to train models for HAR has also been explored recently in [5], where a deep neural network is trained on the data from multiple sensors on the smartphone. In our study we go a step further and analyze and compare a simplified subset of the sensor data (only accelerometer magnitude) - which allows us to have a model that will work regardless of the smartphone orientation and to have a simple yet effective method of integrating a model into an Android application.

We propose an Android integration of a Machine Learning (ML) pipeline for recognizing human activities in real time on a smartphone. In particular, the proposed approach uses the 3-axis accelerometer in the smartphone, processes its data in real time, and then the ML model recognizes the user's activities: walking, running, jumping, cycling and standing still. The proposed Recurrent Neural Network (RNN) model and its machine learning pipeline are developed on a publicly open activity dataset, then implemented into an Android application, which finally, is once again evaluated on a dataset recorded with a smartphone itself. Additionally, as part of this study we release an Android application [6], which can be used by other researchers to easily gather data with a smartphone and as a practical demonstration of how to integrate an ML model with an Android application and use the built-in accelerometer data.

2 DATASET

The models were trained on a publicly available dataset which was originally used to evaluate the impact of sensor placement in activity recognition [7]. The dataset consists of wearable sensor readings from 17 healthy subjects which perform any of 33 different activities. There are a total of 9 wearables placed on the body: two on each arm and leg, and one on the back. Each wearable sensor reads 13 values with a frequency of 50Hz: three for acceleration, three for rotation, three for magnet flux vector and four for orientation in quaternion format. This brings the total amount of readings to 117 per frame (9 wearable sensors with 13 values each). Out of all these measurements only six are used: the **3 accelerometer values** from each of the two upper leg sensors (left and right). These sensors are chosen as they are approximately at the location where a smartphone would be (in a side pocket). Additionally, the magnitude of each sensor is added as an additional feature, calculated as:

$$magnitude = \sqrt{acc_x^2 + acc_y^2 + acc_z^2} \qquad (1)$$

Due to the position of the sensors, recognizing motion mainly expressed with the upper torso and arms is impossible, so the dataset is truncated to only activities that are dependent on the legs: walking, running, jumping, cycling and standing still.

3 METHODOLOGY

In order to adapt the dataset to fit the needs of this application, certain preprocessing and feature extraction is performed, described in detail in the following subsections.

3.1 Preprocessing and segmentation

The dataset contains a disproportionate number of readings for standing still in comparison to all other activities. To correct this a random under-sampling is performed (only 5% of the standing still data is used). Additionally, similar activities are grouped together, namely jogging and running are grouped together as running, and jumping upwards, jumping front and back, jumping side to side, and jump rope are grouped as jumping. The resulting distribution of data is illustrated on Figure 1, with running having the most amount of data (1760s), and cycling having the least (860s).

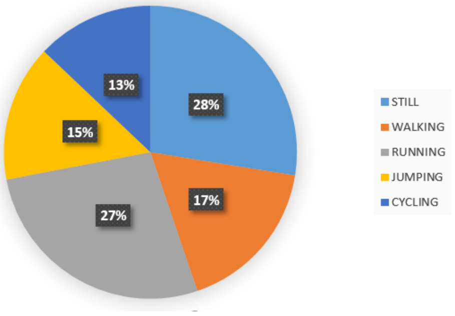


Figure 1 Activity distribution after selection

Once selection has been performed, the data is grouped into **3-second windows**. Since the data is collected at a frequency of 50Hz, each window contains 150 records.

3.2 Feature extraction

After the data has been split into 3-second windows, five statistical features are calculated per window. The first two are the **mean** and the **standard deviation** of the 150 values in the window. The three additional statistical features are:

- **Mean first-order difference**: average difference between consecutive values in the window. Computed by first creating a list of first-order differences between consecutive values in the window and then calculating the mean of this list.
- **Mean second-order difference**: average difference between consecutive elements in the list of first-order differences.
- **Min-max difference**: difference between the minimum and maximum value in the window.

The feature extraction is performed on every sensor (x, y, z axis and magnitude on both accelerometers, left and right), which gives a total of 40 features. The features are then separated into three datasets: *left accelerometer, right accelerometer* with 20 features each, and *combined accelerometers* which contains the data from both the left accelerometer and right accelerometer datasets, by matching the respective features (e.g., x-axis on the left accelerometer and x-axis on the right accelerometer are treated as the same feature: x-axis), thus the combined accelerometers dataset also contains 20 features, but it is twice as long.

To compare the effectiveness of a simplified version of the model that is orientation independent, a second version of the dataset is created. This dataset uses only the features extracted from the **magnitudes** of both accelerometers (5 features each). It is further split into three parts: *magnitude-only left*, *magnitude-only right* and *magnitude-only combined*, each containing five features.

3.3 ML Models

Multiple ML models were evaluated, such as K-NN, Linear SVM, Random Forest, Naïve Bayes and Neural Networks (DNN and RNN).

Ultimately the **RNN model** had the best performance. A simple RNN was chosen as the ML model for this application. The model is created using Keras and contains two RNN layers with 512 nodes each and tanh activation function. The final decision layer is a Dense layer with 5 nodes and a softmax activation function. It is trained for 100 epochs with a sparse categorical cross entropy activation function.

4 EXPERIMENTS

With the dataset prepared, the following experiments were conducted:

- Accuracy comparison between magnitude-only and full-featured versions of the dataset.
- Evaluation of models trained on data from the left accelerometer and evaluated on data from the right, and vice-versa.

4.1 Evaluation and metrics

The models were evaluated using **K-fold Cross-Validation**, where K is equal to the number of subjects, and in each iteration a different subject's data is used as the validation set. Splitting the data this way ensures that the test data and train data do not both contain windows from the same subject (as consecutive windows from the same subject are very similar). Instead, when using the data from a separate subject as a validation set, a good estimate can be made of how the model will behave when a never seen before person's data needs to be evaluated.

In every iteration of the K-fold Cross-Validation a confusion matrix is generated from the predicted values. From there the precision and recall are calculated for every activity as well as the overall accuracy. These metrics are compiled for every iteration and the average values across all iterations form the overall evaluation of the model.

4.2 Results

Initially nine models were considered and evaluated on both the full-featured dataset and the magnitude-only dataset (for combined accelerometers). The results are illustrated on Figure 2, sorted by accuracy.

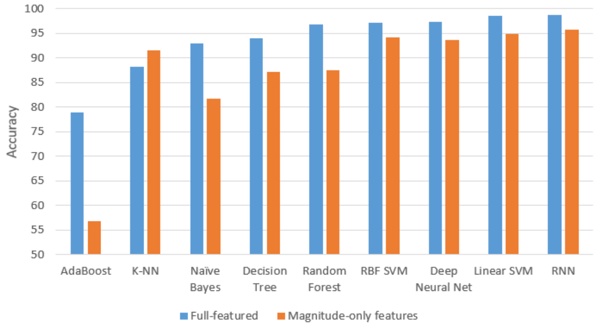


Figure 2 Accuracy comparison of all inspected ML models

The accuracy of the models with full features was expectedly higher than the magnitude-only version, with the drop in accuracy being on average 7% (K-NN being the exception with an increase in accuracy of 2%). The RNN had the highest accuracy in both cases, with 98.8% on the full-featured dataset and 95.8% on the magnitude-only dataset. Therefore, the following results focus on the RNN model.

The comparison in accuracy between the full-featured and magnitude-only versions was made on all three datasets (left, right and combined). The results for the RNN are displayed on Figure 3.

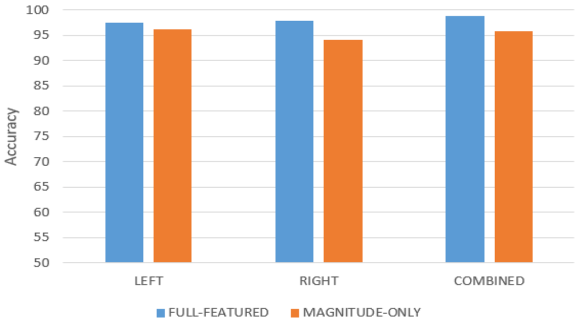


Figure 3 Comparing full-featured and magnitude-only datasets

The average drop in accuracy for the RNN was 3% which is well within acceptable boundaries. As a side note, the right side in general seems to show slightly weaker results, however at most this is 1.5% (when comparing the left simplified and right simplified sets) which could be due to random noise.

In order to evaluate if the model takes in a bias from the side on which it is trained or if the sides carry an intrinsic difference, the model was trained on one side and evaluated on the other. This was done twice, trained on left and evaluated on right, and trained on right and evaluated on left. The results are displayed on Figure 4, along with a control set which was trained and evaluated on the same side.

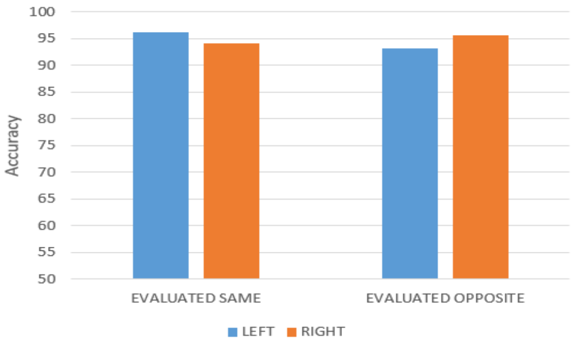


Figure 4 Comparison between same and opposite side evaluation

The accuracy differences are within 2% which is negligible, and in the case of the right accelerometer dataset, evaluating on the left actually increased the overall accuracy. This is due to the slight difference in quality between the left and right sides, and not due to switching sides when evaluating.

These results suggest that there is no significant side bias in the models and thus the activity recognition will work regardless of on which side the smartphone is located. This in addition to the simplified model's independence from orientation make it the ideal choice for integrating with a smartphone.

5 ANDROID INTEGRATION

In order to integrate with an Android smartphone device, the magnitude-only model with combined accelerometers was converted into a tflite format using the Tensorflow Lite library, which is the most commonly used library for artificial intelligence in Android. The converted models are then added in the file structure of an Android application which reads them into memory when it starts up and uses them in real time to recognize activities.

All Android devices come equipped with accelerometers (along with many other sensors) and they can be accessed with the built-in class SensorManager, which is part of the default library: android.hardware. The data read by the SensorManager is on a by-axis basis and in the standard unit of m/s^2. The orientation of the x, y and z axis is illustrated in Figure 5.

The frequency with which the sensor records data is adjustable, with the tradeoff being higher quality data vs lower battery consumption. In our implementation, the sensor delay is set to 20ms between reads (50Hz frequency).

Since there is no way to predict which way the smartphone will be oriented in the pocket, the magnitude of the accelerometer is the only thing that is used in the feature calculation. The magnitude readings are kept in memory until 150 samples are accumulated (exactly 3s), which is the size of the window used in the training of the models. Then the same statistical features are calculated on the collected window: mean, std. deviation, mean first-order and second-order differences, min-max difference. These values are then placed in a tensor and it is sent as the input into the model, which is also kept in memory (in the form of an object). The output of the model is also a tensor (the output layer which has a softmax activation function), which is then converted into a single result (the node with the highest value) and is displayed on screen.

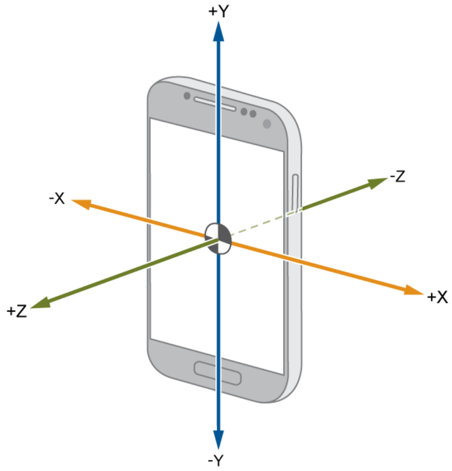
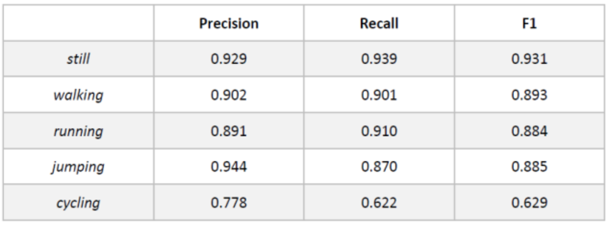


Figure 5 Accelerometer axis orientation in smartphones

Since 150 samples need to be accumulated before the features are calculated and the model is called to make the prediction, there is the side effect that the displayed value on screen is 3s behind (in other words the current activity the user is doing will be displayed in 3s). All the data read by the accelerometer along with the prediction and a timestamp and is kept in memory (a single entry will contain all the calculated features from the 3-second window, the model prediction and a timestamp). The user can choose to export this data to csv and use it as a dataset.

The model was evaluated on a practically collected dataset with a Samsung Galaxy s20 smartphone (5 minutes of each activity). The predicted value was compared to the actual activity by cross-referencing the timestamps (the activities were performed at specific times), and a confusion matrix was created, from which the precision, recall and f1 score, as well as overall accuracy, was calculated. The results are displayed on Figure 6.



	Precision	Recall	F1
still	0.929	0.939	0.931
walking	0.902	0.901	0.893
running	0.891	0.910	0.884
jumping	0.944	0.870	0.885
cycling	0.778	0.622	0.629

Figure 6 Precision, recall and f1 score results on the practically collected dataset on a Galaxy s20 smartphone

The overall accuracy of the model was **90.2%,** which is a noticeable drop from the 95.8% evaluated from the original training dataset. This is expected, as there is a certain amount of noise introduced to the system from the fact that the smartphone is not fixed in place as rigidly as the wearables.

6 CONCLUSION

This paper presented a practical way of training and implementing a HAR model in an Android application, along with solving the practical issues of reading smartphone accelerometer data such as unpredictable orientation and whether it is kept on the left or right side.

To determine whether there is an intrinsic difference between the left and right side or whether the models develop a side bias, an experiment was conducted where models were evaluated on the opposite side of where they were trained, and it was determined that no such bias existed.

To gain independence from orientation, a simplified dataset was created which used only the magnitude readings. Training on this dataset resulted in an expected drop in accuracy, but within an acceptable margin.

An RNN was trained on the magnitude-only dataset and integrated into an Android application which reads the accelerometer data and calculates the features in real time. The calculated features are used as an input for the model, which then outputs the predicted activity, and is subsequently shown on screen.

The sensors in the used smartphone did prove to be of a comparable quality to the wearable sensors as the model successfully recognized activities recorded with smartphone sensors with a solid accuracy of 90.2%, even though it was trained on a dataset from wearable sensors.

ACKNOWLEDGEMENT

This research was partially supported by the WideHealth project - EU Horizon 2020, under grant agreement No 952279.

REFERENCES

[1] Patima Silsupadol, Kunlanan Teja, Vipul Lugade, “Reliability and validity of a smartphone-based assessment of gait parameters across walking speed and smartphone locations: Body, bag, belt, hand, and pocket”, Gait & Posture, Volume 58, 2017,

[2] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," in IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1192-1209, Third Quarter 2013

[3] Kozina, S., Gjoreski, H., Gams, M., & Luštrek, M. (2013, September). Efficient activity recognition and fall detection using accelerometers. In International competition on evaluating AAL systems through competitive benchmarking (pp. 13-23). Springer, Berlin, Heidelberg.

[4] Gjoreski, M.; Gjoreski, H.; Luštrek, M.; Gams, M. How Accurately Can Your Wrist Device Recognize Daily Activities and Detect Falls? Sensors 2016, 16, 800. https://doi.org/10.3390/s16060800

[5] Charissa Ann Ronao, Sung-Bae Cho, Human activity recognition with smartphone sensors using deep learning neural networks, Expert Systems with Applications, Volume 59, 2016, ISSN 0957-4174

[6] https://github.com/ViktorSrbinoski/SmartphoneActivityRecognition

[7] Oresti Banos, Miguel Damas, Hctor Pomares, Ignacio Rojas, Mt Attila Toth, and Oliver Amft. A benchmark dataset to evaluate sensor displacement in activity recognition. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp ’12, pages 1026–1035, New York, NY, USA, 2012. ACM.

Speaking Recognition with Facial EMG Sensors

Antonio Nikoloski[1], Petar Poposki[1], Ivana Kiprijanovska[2, *], Simon Stankoski[2], Martin Gjoreski[3], Charles Nduka[1], Hristijan Gjoreski[1, 2]

[1] Ss. Cyril and Methodius University in Skopje, N. Macedonia
[2] Emteq Ltd, Sussex Innovation Centre, Science Park Square, Brighton, UK
[3]Università della Svizzera Italiana, Switzerland
ivana.kiprijanovska@emteqlabs.com[*]

ABSTRACT

With the advent of interactive virtual reality (VR) applications, the interest in tools that allow users to engage with VR environments unobtrusively and intuitively is growing. One such interfacing tool for VR applications is speech recognition, which can contribute to enhanced human-computer interaction. In this study, we explore the usage of a novel VR facial mask equipped with seven surface electromyography (sEMG) sensors to recognize if the user is speaking or not using machine learning. We collected speaking and non-speaking data from 30 participants. The machine learning pipeline that was developed included data preprocessing, de-noising, filtering, segmentation, feature engineering, and training of a binary classification model. The experimental results indicate that the mask can be used to recognize the speaking activity. On the test data of five unseen participants, the best-performing model achieved an accuracy of 89% and an F1-macro score of 91. Additionally, by removing each sensor from the dataset, we analyzed the individual influence each sensor has on the models' outcomes. We did not observe a significant drop in the accuracy of the models, indicating that using the mask speaking can be detected even if some of the sensors are not used.

KEYWORDS

speaking recognition, machine learning, classification, wearable sensors, surface EMG, facial muscles.

1 INTRODUCTION

Virtual reality (VR) is an emerging technology that has introduced immersive user experience in virtual environments and is expected to revolutionize the way we interact with the digital world. VR applications have already been widely used in many different disciplines, ranging from research and training facilities to entertainment and healthcare. With the emergence of interactive VR applications, there is an increasing interest in new immersive tools that enable users to interact with VR surroundings in an unobtrusive and intuitive manner. One such interfacing tool for VR applications is speech recognition. Its incorporation with VR provides users with increased flexibility for interfacing with VR environments and can contribute to improved human-computer interaction.

In recent years, surface electromyography (sEMG)-based interfaces have been utilized for unobtrusive interaction in a VR environment. sEMG is used to measure muscle contractions using sensors applied directly on the skin by detecting changes in surface voltages on the skin when muscle activation occurs. In part due to its ability to be applied non-invasively, facial sEMG has been used to detect the activation of facial muscles that are activated during speaking. However, most sEMG sensors used in conventional speaking recognition systems have been attached around the user's lips and neck. This poses a number of practical issues, including the need for extra wearable devices in addition to the VR headset, limited facial muscle movement, and user discomfort.

To overcome these issues, in this study we explore the usage of a novel facial mask equipped with sEMG sensors. The mask is incorporated into a VR headset to recognize if the user is speaking or not. Our approach is based on signal processing and machine learning (ML), which are used to develop a binary classification model.

2 RELATED WORK

The first studies with sEMG sensors were performed by Piper[1]. Since then, researchers have been widely using sEMG sensors to measure the electrical signal that emanates from contracting muscles. The usefulness of the sEMG signal for measuring human performance was demonstrated by Inman [2] who investigated the technical aspects of human locomotion. By the early 1960s, the improvements in signal quality and convenience made the sEMG sensors a common tool in clinical and research laboratories. Despite their popularity, current recording methods can be problematic in maintaining signal fidelity when vigorous or long-duration activities are monitored [4] [3] .

Speech recognition by using sEMG was first used in the 80s [4] [6] . The results in these studies were preliminary but important for the further progress of the field. Jorgensen and Binsted [6] showed that it is possible to recognize speaking even if the words are spoken silently and/or without any actual sounds. Jou et al. [7] showed that it is possible to recognize not just the words but also the phonemes to a certain degree. Additional works include direct synthesis of speech via sEMG – which aids people who have problems with their vocal cords or airways [8] [9] .

Compared to the previous studies, we differ in the sense that we are using a novel facial mask – emteqPRO^{tm}, which is equipped with seven sEMG sensors. The sEMG sensors may be more error-prone compared to the intramuscular EMG sensors, and thus here we study their utility. Additionally, the location of our sEMG sensors makes the task of speaking recognition more challenging because the facial mask is placed on the upper part of the face (as part of the VR headset) and not the mouth and the lips – which would be more convenient for speech recognition.

3 DATASET

The data collection protocol included healthy participants that were asked to read a pre-defined text (news article). Additionally, we recorded a segment where the participants were sitting still, i.e., we recorded a baseline session with a neutral face. This data was recorded while the participants were watching a neutral video, without moving their facial muscles or speaking. A total of 30 participants were recorded, of which 18 were male and 12 were female, with a mean age between 19 and 25 years. The native language of all the participants was Macedonian.

During the data collection protocol, we were using the emteqPRO^{tm} mask [10] [11] to record sEMG sensor data. The mask has seven EMG sensors (Figure 1): two frontalis sensors (6 and 0 in Figure 1) used to monitor eyebrow movement; two orbicularis sensors (4 and 2 in Figure 1) used to monitor eye movements; two zygomaticus sensors (5 and 1 in Figure 1) used to monitor mouth and cheek movements; and one corrugator sensor (3 in Figure 1) used to monitor forehead movements.

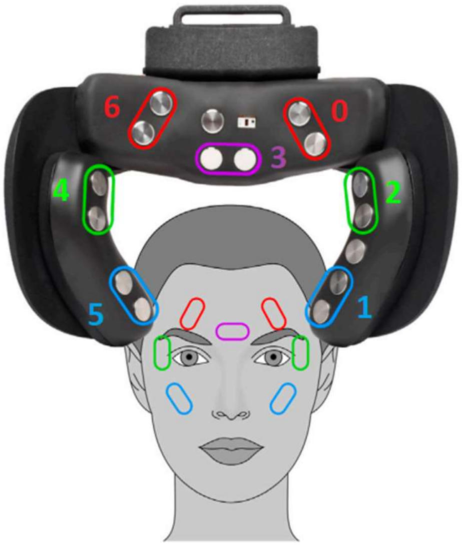


Figure 1: emteqPRO face mask with all 7 EMG sensors

4 DATA PREPROCESSING AND MODELING

The sEMG data were continuously recorded at a fixed rate of 1000 Hz. These data underwent a data preparation process, which included data filtering, segmentation, and feature engineering. To improve the quality of the sensor data, we performed signal de-noising and filtering. The EMG signals were initially filtered with a Hampel filter to eliminate sudden peaks in the signals that emerge as a result of quick movements. Additionally, we also applied a frequency-based filtering method based on spectrum interpolation [12] to reduce the noise caused by electromagnetic interference. [12] A sliding window technique was utilized for data segmentation. Specifically, the data were segmented into windows of size of 0.5 seconds with 0.4 seconds overlap (0.1 seconds slide). Finally, for each sEMG channel, we extracted 34 features, including various amplitude-based features, amplitude derivatives, auto-regressive coefficients, frequency-based features, and statistical features. The feature extraction procedure resulted in a total of 238 features.

The extracted features were used as input to four classification algorithms: (i) K- Nearest Neighbors [13] - a simple statistical algorithm where a datapoint is assigned a class according to the most numerous class of its k nearest neighbors; (ii) Support Vector Machine Classifier (SVM) [14] – an algorithm that works along the principle of finding a hyperplane in N-dimensional space to separate two classes of data points; (iii) Random Forest [15] - an ensemble learning method that trains N decision trees using random subsets of data and features and determines the instance's class by majority voting among the trained decision trees; and (iv) Extreme Gradient Boosting [16] - a gradient boosting algorithm which trains decision tree models sequentially, and each subsequent model strives to correct the errors of its predecessors.

5 EXPERIMENTS

5.1 Evaluation Setup

The recorded data was split into training (20 of the participants), validation (5 of the participants) and test datasets (5 of the participants). The train dataset was used to train the models, the validation was used to optimize hyperparameters, and the test dataset was used to report the accuracy. The evaluation metrics we used to test the performance of our models were accuracy and F1 score.

Additionally, the experiments were performed so that the training validation and test subsets do not have overlapping participants - i.e., each participant's data is found only in one of the three subsets. This is done so that we replicate a scenario where the model is used in practice on participants that are not in the training dataset.

5.2 Default Hyperparameters Results

Figure 2 presents the results (accuracy and F1-score) achieved by each of the algorithms with their default hyperparameters. We additionally included the Dummy classifier as a reference (which predicts the majority class). The results show significant improvement by all the algorithms compared to the Dummy classifier. The Random Forest and the SVM achieved similar results, while the XGBoost classifier achieved the best results overall (87% accuracy and 89% F1-score). Apart from this, this classifier also scaled efficiently with the size of the datasets, as it was able to quickly and efficiently create and train models. This was also beneficial for the hyperparameter optimization – explained in the next subsection.

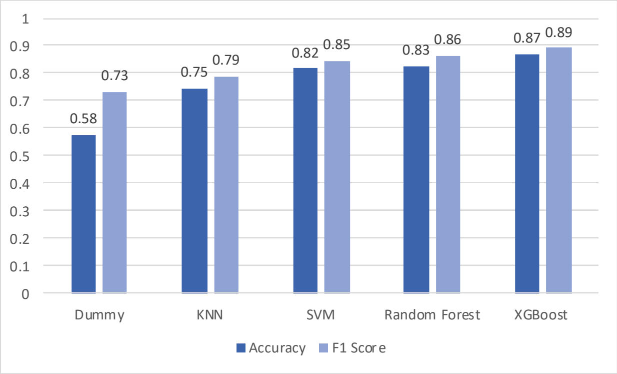


Figure 2: Algorithm comparison (accuracy and F1-score) using default hyperparameters

5.3 Optimized Hyperparameters Results

In the next step, we performed hyperparameter optimization. This process involves iterative changes of certain parameters of a classifier. During this process, an interval for every hyperparameter is defined, and afterward, each parameter is iteratively updated, and the performance of the models is monitored. During this step, all 238 features of the datasets were used, and a large number of numerical and other parameters (such as kernel for SVM, booster for XGB, etc.) were tuned.

Figure 3 presents the results (accuracy and F1-score) achieved by each of the algorithms after the hyperparameter optimization. The results show slight improvement for the KNN, SVM, and XGBoost algorithms, the latest one achieving 89% accuracy and 91% F1-score – which was the best score that we achieved on this dataset.

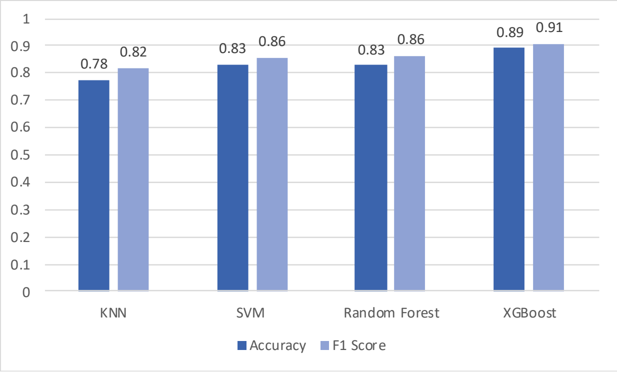


Figure 3: Algorithm comparison (accuracy and F1-score) using optimized hyperparameters

5.4 Continuous Recognition Results

Figure 4 illustrates the continuous recognition results for the five subjects from the test set achieved by the best-performing XGBoost classifier. A comparison was made between the true and the predicted class on a time scale, i.e., with a blue line, the true classes are presented (1 represents speaking, 0 represents not speaking). Additionally, the orange color presents the speaking predictions by the model. Each subject's data is separated with black dashed lines in the figure. The results show that a large portion of the error is down to the baseline sessions of the last two subjects in the test dataset, marked with red circles. In a large portion of the baseline sessions, the model is falsely predicting speaking activity. We speculate that the reason might be that these two subjects were moving their head during the baseline session, which may have caused the sensors to shift from their original position and deteriorate their contact with the skin.

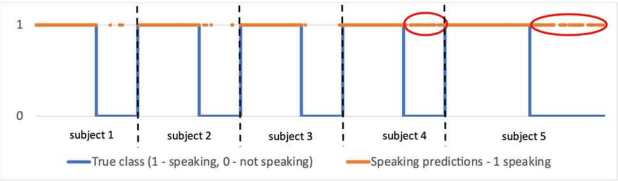


Figure 4: Continuous recognition results for the XGBoost algorithm. The blue line represents true classes (1 – speaking, 0 – not speaking), and the orange line represents the predictions (1 – speaking)

5.5 Sensor Analysis Results

We additionally analyzed the results achieved by the models if a certain sensor is missing. This way, we were able to check the importance of each sensor for the given task. Knowing the positions of the sensors on the face, we wanted to learn how the data would change if we were to drop data from a certain sensor while keeping the rest.

The results are shown in Figure 5, which in general, show that the drop in accuracy and F1 score is not significant for all the sensors. The accuracy drops from 87% to 85% at most. A more detailed analysis shows that the sensors placed on left and right orbicularis, corrugator, and left frontalis have the most impact on accuracy, i.e., the accuracy drops the most when one of these sensors is missing. One of the reasons for this is that while the participants were speaking, they were actually reading – which means they activated their eyes which is recorded by the orbicularis muscles. This analysis shows us that certain muscles activate more while speaking compared to others, so that is why the model itself gains or loses accuracy more, depending on which sensor is dropped.

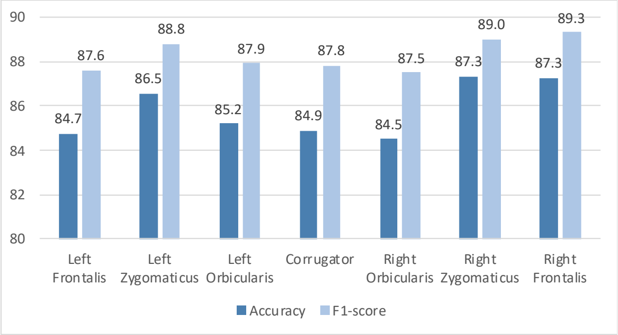


Figure 5: Sensor analysis showing the performance when a particular sensor is missing.

6 CONCLUSION

In this work, we presented a ML approach for speaking recognition using facial sEMG sensors integrated into a VR headset. The dataset was collected with 30 healthy participants while reading a news article and watching videos. The results show that the best performing model is XGBoost, which achieved 89% accuracy. Additionally, the error analysis per participant showed that most of the misclassifications were incorrect speaking predictions in the baseline (non-speaking) sessions of two participants. We speculate that this is caused by the head movement of the participants and we plan to tackle this using the IMU sensor on the emteqPRO^{tm} mask.

An additional problem was that while the participants were reading, they were making small breaks, which were automatically labeled as speaking – but in fact were not speaking. This labeling problem will be tackled in future by using audio to exactly label the speaking segments.

Finally, we plan to implement person-specific normalization on the EMG data. This is an important step given that different participants have different facial muscles, and even more, those muscles are activated differently while doing the same facial expressions or speaking.

ACKWNOLEDGEMENT

Part of this study was supported by the Innovate UK Project no. 81376: Virtual Reality rehabilitation tailored to older brain injury patients (Healthy Ageing), and part by the WideHealth project no. 952279 - European Union’s Horizon 2020 research and innovation programme.

REFERENCES

[1] Piper H (1912) Elektrophysiologie menschlicher Muskeln. Springer, Berlin, pp 1–163.

[2] Inman, V. T., Saunders, J. B., & Abbot, L. C. (1944). Observations on the function of the shoulder joint. Journal of Bone and Joint Surgery, 26, 1-30.

[3] M. Wand, M. Janke, and T. Schultz, “Investigations on Speaking Mode Discrepancies in EMG-based Speech Recognition,” in Proc. Interspeech, 2011, pp. 601–604.

[4] N. Sugie and K. Tsunoda, “A speech prosthesis employing a speech synthesizer—Vowel discrimination from perioral muscle activities and vowel production,” IEEE Trans. Biomed. Eng., vol. BME-32, no. 7, pp. 485–490, Jul. 1985.

[5] M. S. Morse and E. M. O’Brien, “Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes,” Comput. Biol. Med., vol. 16, no. 6, pp. 399–410, 1986.

[6] C. Jorgensen and K. Binsted, “Web browser control using EMG based sub vocal speech recognition,” in Proc. 38th Annu. Hawaii Int. Conf. Syst. Sci., 2005, p. 294c.

[7] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, “Towards continuous speech recognition using surface electromyography,” in Proc. Interspeech, 2006, pp. 573–576.

[8] J. Freitas, A. Teixeira, and M. S. Dias, “Towards a silent speech interface for portuguese,” in Proc. Biosignals, 2012, pp. 91–100. [23] A. Toth, M. Wand, and T. Schultz, “Synthesizing speech from electromyography using voice transformation techniques,” in Proc. Interspeech, 2009, pp. 652–655.

[9] K.-S. Lee, “Prediction of acoustic feature parameters using myoelectric signals,” IEEE Trans. Biomed. Eng., vol. 57, no. 7, pp. 1587–1595, Jul. 2010.

[10] Gjoreski, H., I. Mavridou, I., Fatoorechi, M., Kiprijanovska, I., Gjoreski, M., Cox, G., & Nduka, C. EmteqPRO: Face-mounted Mask for Emotion Recognition and Affective Computing. In Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (pp. 23-25).

[11] Gnacek, Michal & Broulidakis, John & Mavridou, Ifigeneia & Fatoorechi, Mohsen & Seiss, Ellen & Kostoulas, Theodoros & Balaguer-Ballester, Emili & Kiprijanovska, Ivana & Rosten, Claire & Nduka, Charles. 2022. emteqPro-Fully Integrated Biometric Sensing Array for Non-Invasive Biomedical Research in Virtual Reality. Frontiers in Virtual Reality. 3. (Mar. 2022)

[12] Mewett, D. T., Reynolds, K. J., & Nazeran, H. Reducing power line interference in digitised electromyogram recordings by spectrum interpolation. Medical and Biological Engineering and Computing, 42(4), 524-531, (2004).

[13] D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66.

[14] Zhang, Yongli. (2012). Support Vector Machine Classification Algorithm and Its Application. 179-186.

[15] Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

[16] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794.

Machine-learning models for MDS-UPDRS III Prediction: A comparative study of features, models, and data sources

Vitor Lobo[1], Diogo Branco[1], Tiago Guerreiro[1], Raquel Bouça-Machado[2,3], Joaquim Ferreira[2,3,4]
and CNS Physiotherapy Study Group[2]
[1]LASIGE, Faculdade de Ciências, Universidade de Lisboa
[2]CNS—Campus Neurológico, [3]Instituto de Medicina Molecular João Lobo Antunes,
[4]Faculdade de Medicina, Universidade de Lisboa
vitormarqueslobo@gmail.com;djbranco@fc.ul.pt;tjvg@di.fc.ul.pt;raquelbouca@gmail.com;jferreira@medicina.ulisboa.pt

ABSTRACT

Parkinson's disease is the second most common neurodegenerative disease worldwide. Symptoms tend to fluctuate during the day and through disease progression. Clinical evaluations tend to occur spaced in time. Further, the assessments used are mostly subjective. The gold standard for evaluating disease severity is MDS-UPDRS. The increase in sensor usage enabled objective evaluation and continuous monitoring of the disease fluctuations. One of the symptoms that most affect mobility are gait disorders. The use of gait characteristics started to become popular to monitor the disease. However, the approaches used lack in-depth knowledge of machine learning models for disease staging. In our work, we try to estimate the MDS-UPDRS part III score from accelerometer data. We collected data from 74 patients using the Axitvity AX3 device both on the wrist and lower back. We did experiments with different models, features, and windows size. We achieved a 4.26 Mean Absolute Error on the on left out 10% data using both devices with a 2.5-second sliding window and a random forest model for prediction. We contribute with a comparison of the performed experiments and provide, according to our experiments, the optimal models for MDS-UPDRS part III estimation using only accelerometer data.

KEYWORDS

gait, accelerometer, mds-updrs, Parkinson's disease, features, machine learning, models

1 INTRODUCTION

Parkinson's Disease (PD) is a neurodegenerative disease that affects around 1% of the world's population. This disease is characterized by motor and non-motor symptoms [15]. Motor symptoms include bradykinesia, tremor, rigidity, and gait impairment. These are present in the early stages of the disease and worsen as the disease progresses.

Although there is no cure, the available pharmacological and non-pharmacological therapeutic interventions effectively control symptoms. However, as the disease progresses their efficacy tends to reduce and motor complications, such as motor fluctuations and dyskinesia, appear [11]. These have been labeled as 'ON' and 'OFF' stages [4]. To minimize the impact of these fluctuations and inform better the clinicians there is the need to periodically assess the symptoms. Generally, these evaluations require a visit to a clinic or hospital. Clinicians use validated assessments for PD to characterize a patient's current disease stage [9]. These assessments occur spaced in time and can be hard to capture all the fluctuations that may have happened between appointments. Further, instruments used in clinical practice focus on subjective evaluations. Namely, visual assessments during clinical visits that are supported by clinical scales.

The gold standard for evaluating disease severity in PD is the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS). This is a comprehensive rating scale that assesses both motor and non-motor symptoms associated with Parkinson's [7]. To optimize disease management, close monitoring of symptom fluctuations is crucial. However, today this monitoring is usually performed through medical appointments, every six months, with a mean duration of 30 minutes. Additionally, what published evidence suggests is that patients perform differently during these moments, providing only information about their best capacity, rather than their usual performance in their daily lives.

The democratization of sensors' usage, namely the body-worn devices, that measure acceleration, and angular velocity allowed the increase of objective evaluations [10]. These devices passively monitor patients during clinical evaluation and in free-living environments. Furthermore, allow movement metrics and feature extraction that can be related to motor symptoms or clinical scales used for disease assessments [6]. Gait disorders are one of the symptoms that most affect mobility. Inertial measuring units can help to identify fluctuations. There have been studies that leverage the identification of walking bouts to extract gait metrics like step length or step variability [1, 4].

Research using these gait characteristics as a marker for PD has demonstrated the potential for monitoring the disease in several ways [2]. While the use of these gait characteristics has become a popular approach for monitoring PD, novel research has started to analyze signal processing metrics that could also be of use for this purpose. In a 2019 study, the contributions of signal-based features and gait characteristics for the classification of PD were analyzed [13]. Another emerging method to stage PD is the use of total scores of the entire MDS-UPDRS or subparts of the scale. Specifically, MDS-UPDRS III scores have been empirically demonstrated as a good metric for monitoring the progression of PD [12]. As such, several studies have focused on the prediction of this score to monitor disease progression. A recent example of this approach for the monitoring of PD progression is the 2021 study that leveraged a convolutional neural network (CNN) model trained using inertial data collected from the lower back during gait to estimate MDS-UPDRS III scores [14]. While these results are promising, the authors suggest that a comparison with traditional feature-engineered machine learning models could be an avenue for future work, towards

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2021 Copyright held by the owner/author(s).

the deployment of such technologies for continuous monitoring of PD. Other studies have revealed that it is possible to estimate PD progression using gait data collected with accelerometers [8]. However, the relative efficacy and effect of different approaches to data collection and processing, and machine learning pipeline design still lack consensus and clear comparisons that could help inform future research in this field.

In our work, we try to estimate the MDS-UPDRS part III from accelerometer data. We collected the data using the Axitvity AX3 device both on the wrist and lower back [3]. Our dataset contains data collected from 74 patients (HY between 2 and 4) at Campus Neurológico (CNS), a tertiary specialized movement disorders center in Portugal. The final subset of data contained 267 instances of gait from 104 evaluation sessions. We did different experiments with 4 models (Random Forest, XGBoost, SVM, Linear Regression), and 59 features from the statistical, spectral, and temporal domains. Furthermore, we used non-overlapping window sizes of 2.5 and 5 seconds. To validate the trained models we used Leave One Subject Out (LOSO) cross-validation.

Our results showed that the best configuration, with the lowest prediction error on the left out of 10% data, achieved a 4.26 MAE, with the Random Forest model, and a 2.5-second sliding window using combined data from the wrist and lower back. For all of the selected models, the configurations that achieved the best results using either of the validation schemes used data collected from the lower back or both sensors. Most models performed better using a 5-second window length, with the exception of the xgboost model. The best-performing linear regression and SVM-based models used the SURF and relieF feature selection methods.

Therefore, we contribute with the comparison of different models, features, sensor placement, and window sizes. We provide, according to our experiments, the optimal models for MDS-UPDRS part III estimation using only accelerometer data.

2 METHODS

The MDS-UPDRS III estimation was performed using different approaches to data collection, signal processing, and using different machine learning pipelines. In this section, we describe the steps taken together with the variables for each step, in order to enable a comparison between different design decisions and their effect on the estimation of the disease stage.

2.1 Data Collection

We collected data from 74 patients with PD at CNS from periodic evaluations conducted by trained physiotherapists. Each participant wore an Axivity AX3 on the wrist and lower back during a set of clinical assessments. Accelerometer data was set to record at 100 Hz. Our dataset includes 267 instances of gait from 104 evaluation sessions of the 10-meter walk. MDS-UPDRS were also applied for each patient in each session. Among these patients, 49 were male and 23 were female, while the gender of the remaining 2 patients was not reported. The average patient age was 70.4 years (SD=13.12). The average weight was 71.76 kg (SD=13.89) and the average height was 166.49 cm (SD=9.26). Finally, the average MDS-UPDRS III score was 40.92 (SD=14.31) and 2.57 (SD=0.97) for the H&Y scale.

2.2 Data Pre-Processing

In order to isolate gait instances, the selected data files were segmented using the annotated timestamps for the 3 trials of the 10-meter walk test. Visualization of each of the segmented gait instances was then created in order to exclude session data that contained sensor failures and misalignment, or mismatched timestamps. During this step, the vector magnitude of the accelerometry signal was computed and appended to each segment using the traditional euclidean vector norm formula $\sqrt{x^2 + y^2 + z^2}$. To avoid the possible temporal drift associated with the process, a resampling step was performed after segmentation to ensure even sampling, as required for the extraction of some of the used Time and Frequency domain features. Finally, all segments were filtered using a fourth-order, digital low pass Butterworth filter with a cut-off frequency of 20 Hz in order to remove possible ”machine noise” [5].

2.3 Evaluated Models and Features

We used 16 statistical, 26 temporal, and 17 spectral domain features, with a total of 59. They were computed from all accelerometry axes and vector magnitude. A sliding window technique was used to segment the signal into non-overlapping windows from which the features were extracted. Different feature data frames were then created using 2.5 and 5-second windows, both of which were previously used in the literature [14], in order to assess the effect of window size on the estimation task. During this feature extraction process, MDS-UPDRS III scores were also computed and appended to the corresponding windows for both data frames. The first step toward feature selection was to use a variance filter to exclude features with low (<0.025%) or zero variance which lowered the feature space from 2081 to 266 in the 2.5-second window and 3081 to 452 in the 5-second window. While this reduction may seem drastic, it is to be expected because of the way Time Series Feature Extraction Library works, computing the same feature several times for different frequencies for example which results in a large number of feature columns with hardly any variability, and thus, descriptive power. A further feature selection step was performed using four different feature selection methods that implement different strategies for feature ranking. Each of these feature selection algorithms was used to rank and select the top 10/25/50 features to be used for the regression task using the linear regression algorithm, and with the support vector-based model. The complete feature subset was also used for these models, in order to establish a baseline comparison with the remaining tree-based models that are less affected by the number of features due to their capability to perform intrinsic feature selection.

For each model, a set of parameters were selected and used in a grid search procedure to test all possible combinations. This procedure was then carried out for each sensor placement and the combined sensors, and for the different sliding window lengths used during feature extraction, in order to compare the effect of these variables for the estimation task. Leave One Subject Out (LOSO) cross-validation was used during the grid search procedures in order to avoid overfitting and optimize the models for generalizability. Finally, the optimal models for each combination of these variables were saved and used for the ensuing validation tasks. To validate the trained models, the original dataset was split into training and testing subsets. The training subset comprised 90% of the data and was used during the grid-search procedure for training the models using LOSO cross-validation. The remaining 10% of the data was then used as a validation set to test the model’s performance on unseen data from patients whose data the model had already seen, providing information on

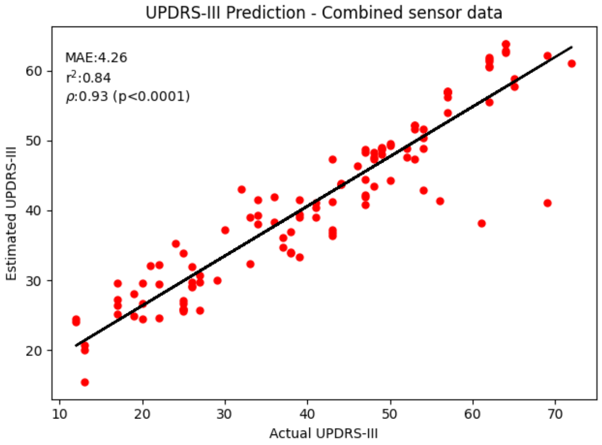


Figure 1: Overall optimal predictions on the 10% of left out data using a Random Forest model on data collected from both sensors and a 2.5s sliding window. Each point represents a window.

the model's ability to estimate MDS-UPDRS III scores for patients that were already known to these models. These steps yield two different scores for each of the optimal models using the same Mean Absolute Error (MAE) evaluation metric: the average MAE for all LOSO splits during training and the MAE for the held-out validation set. For the purpose of this study, this metric is defined as the mean absolute difference between real (x) and estimated (y) MDS-UPDRS III scores over the number of samples used for estimation.

3 RESULTS AND DISCUSSION

This section lays out the results from all of the steps taken toward UPDRS III estimation, including data processing, feature extraction and selection, and finally model training and validation results.

3.1 Optimal configurations

The configuration with the lowest prediction error on the left out 10% of data used data from both devices processed using a 2.5-second sliding window and a Random Forest model for prediction, achieving 4.26 MAE and strong correlation ($\rho = 0.93$) as illustrated in Figure 1. The best performing configuration when performing LOSO CV was a Support Vector-based model, using data from both sensors but a 5-second feature extraction window, achieving a MAE of 9.99. While predictions using this model on the validation set were less accurate than some of the other options at 7.94 MAE, it achieved the best balance when considering both of the validation schemes. Table 1 summarizes the optimal results achieved by each model along with the used data sources and sliding window length for the 10% left out and LOSO validation tasks.

3.2 Sensor placement and windows size

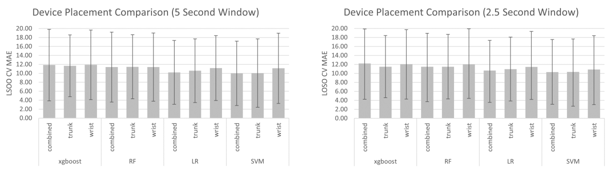
Both device placement and window length used during feature extraction significantly impacted the performance of all models. For all of the selected models, the configurations that achieved the best results using either of the validation schemes used data collected from the lower back or both sensors combined. Specifically, all of the non-tree-based models performed better in both validation schemes using data from both sensors, with the exception of the SVM-based model using a 2.5-second window, which compared to the other options using the same window length achieved lower, albeit negligible, validation MAE using data from the wrist. As for the tree-based models, optimal validation MAE was attained by models using both sensors with the 2.5-second sliding windows, and data from the lower back for the same models using the 5-second window. Figures 2a and 2b illustrate the intra and inter-model comparison for both of the validation schemes, using different window lengths. While the fluctuations were relatively low using LOSO CV, most models performed better using a 5-second window length, with the exception of the xgboost model. MAE using the left out 10% of validation data fluctuated more considerably but was also lowest using 5-second windows for all models except RF.

val_m	model	device_placement	win_length	ft_sel	num_fts	loso_mae	val_mae
1	rf	combined	250	-	266	11.50	**4.26**
1	xgboost	trunk	500	-	229	11.67	4.39
1	svm	combined	500	SURF	25	9.99	7.95
1	lin_reg	combined	500	reliefF	25	10.21	8.98
2	rf	combined	500	-	452	11.39	11.39
2	xgboost	trunk	250	-	133	11.49	5.74
2	svm	combined	500	SURF	25	**9.99**	7.95
2	lin_reg	combined	500	reliefF	25	10.21	8.98

Table 1: Optimal configurations used by each model to achieve optimal MAE on the left out 10% of data (val_m => 1) and LOSO (val_m => 2).

3.3 Optimal parameters

As for model parameters, excluding linear regression, the remaining models had different parameters to achieve the best performance during LOSO CV. For Random Forest (criterion: mae ; max_features: 0.333 ; n_estimators: 250), for xgboost (colsample_bynode: 1; eta: 0.1 ; importance_type: total_gain; max_depth: 3 ; num_parallel_tree: 100 ; tree_method: gpu_hist), and for svm (C: 10 ; epsilon: 0.3 ; gamma: auto ; kernel: rbf). The xgboost was the one that used only the trunk sensor. The others models used both devices. We used a Grid Search procedure that exhaustively tested all parameter combinations for each model, independently of the used device placements and sliding window lengths. The exhaustive nature of the grid search procedure makes this method of parameter optimization computationally expensive. For this reason, and considering that the procedure was used for several models, the used parameter space for each model was not as comprehensive as those used in some other works with a smaller scope and narrower focus. However, the present results should still serve as a good starting point for model tuning in future research.



(a) LOSO CV MAE values (Y-axis) for different device placements using 5-second windows.

(b) LOSO CV MAE values (Y-axis) for different device placements using 2.5-second windows.

Figure 2

3.4 Feature importance

For the models that benefited from it, several feature selection methods were tested, along with different numbers of features to select. The best performing linear regression and SVM-based models used the SURF and relieF feature selection methods respectively, both selecting 25 as the optimal number of features. We then selected the top 20 for each model. Among the 8 top performing models across the two tested window lengths, no model used data exclusively from the wrist, and only 3 models used data exclusively from the trunk. As for the remaining models, the majority of top-ranking features were extracted from devices mounted on the lower back. In some cases, no wrist features were ranked among the top 20, which suggests that although these were used for the estimation task, their contribution is minimal, which is in line with the minimal performance gain in these models when compared to their counterparts using data exclusively from the lower back. Features from the anteroposterior plane of movement (z-axis) were the most prevalent among the top 20 extracted from the trunk sensor, consisting of 50 out of the 140 features considered for this analysis. The vertical plane of movement (x-axis) produced the least amount of features among those considered here, with only 22 ranking among the top contributing features. Spectral-domain features were the most prevalent among these, making up almost half of the 140 considered features, with temporal domain features coming in second by a small margin, and temporal features last consisting of a quarter of this total.

3.5 Limitations

The dataset used in this study consisted of data collected from 74 patients. While this number of patients is significant for preliminary results, a larger sample size could improve the estimation task and further validate the present findings. Beyond the volume of data used to train the models, a wider range of MDS-UPDRS III and Hoehn and Yahr scores could also possibly improve the results, by including a wider variety of walking patterns that in smaller sample sizes could be considered outliers and negatively affect performance. Furthermore, the inclusion of a healthy cohort in the dataset could provide a baseline for the models to recognize healthy gait, exacerbating the difference between data from healthy and affected subjects. Therefore, in future work a longitudinal study in free-living environments with a larger sample size to address our limitations and extend our conclusions.

4 CONCLUSIONS

This paper presents a study that compares the different models, features, and window sizes to estimate MDS-UPDRS part III using accerometer data. One of the most common disorders for people with PD is gait. The increase in sensor usage opened the opportunity for increasing objective evaluations. However, there is a lack of knowledge of the current machine learning approaches. In our work, we compare 4 machine learning models (random forest, xgboost, svm, and linear regression), 59 features (16 statistical domain, 26 spectral domain, and 17 temporal domain), and windows size (2.5 and 5 seconds). To validate our models we used LOSO cross-validation. We showed that the configuration with the lowest prediction error on the left out 10% of data used data from both devices processed using a 2.5-second sliding window and a Random Forest model for prediction, achieving 4.26 MAE. This work opens the opportunity to improve the knowledge of machine learning approaches. However, in future work, there are opportunities for longitudinal studies in free-living environments with larger datasets.

ACKNOWLEDGEMENTS

We would like to thank all the participants that kindly participated in the studies. This project was partially supported by FCT through LASIGE Research Unit funding refs. UIDB/00408/2020 and UIDP/00408/2020 and SFRH/BD/144242/2019 to Diogo Branco., and the WideHealth project which has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 952279.

REFERENCES

[1] Raquel Bouça-Machado, Diogo Branco, Gustavo Fonseca, Raquel Fernandes, Daisy Abreu, Tiago Guerreiro, Joaquim J Ferreira, and CNS Physiotherapy Study group. 2021. Kinematic and clinical outcomes to evaluate the efficacy of a multidisciplinary intervention on functional mobility in Parkinson’s disease. *Frontiers in neurology* 12 (2021), 637620.

[2] Raquel Bouça-Machado, Constança Jalles, Daniela Guerreiro, Filipa Pona-Ferreira, Diogo Branco, Tiago Guerreiro, Ricardo Matias, and Joaquim J Ferreira. 2020. Gait kinematic parameters in Parkinson’s disease: a systematic review. *Journal of Parkinson’s disease* 10, 3 (2020), 843–853.

[3] Clare L Clarke, Judith Taylor, Linda J Crighton, James A Goodbrand, Marion ET McMurdo, and Miles D Witham. 2017. Validation of the AX3 triaxial accelerometer in older functionally impaired people. *Aging Clinical and Experimental Research* 29, 3 (2017), 451–457.

[4] Silvia Del Din, Alan Godfrey, Brook Galna, Sue Lord, and Lynn Rochester. 2016. Free-living gait characteristics in ageing and Parkinson’s disease: impact of environment and ambulatory bout length. *Journal of neuroengineering and rehabilitation* 13, 1 (2016), 1–12.

[5] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T Van Hees, Michael I Trenell, Christoper G Owen, et al. 2017. Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PloS one* 12, 2 (2017), e0169649.

[6] Alberto J Espay, Paolo Bonato, Fatta B Nahab, Walter Maetzler, John M Dean, Jochen Klucken, Bjoern M Eskofier, Aristide Merola, Fay Horak, Anthony E Lang, et al. 2016. Technology in Parkinson’s disease: challenges and opportunities. *Movement Disorders* 31, 9 (2016), 1272–1282.

[7] Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. 2008. Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society* 23, 15 (2008), 2129–2170.

[8] Murtadha D Hssayeni, Joohi Jimenez-Shahed, Michelle A Burack, and Behnaz Ghoraani. 2021. Ensemble deep model for continuous estimation of Unified Parkinson’s Disease Rating Scale III. *Biomedical engineering online* 20, 1 (2021), 1–20.

[9] Anthony E Lang, Shirley Eberly, Christopher G Goetz, Glenn Stebbins, David Oakes, Ken Marek, Bernard Ravina, Caroline M Tanner, Ira Shoulson, and LABS-PD investigators. 2013. Movement disorder society unified Parkinson disease rating scale experiences in daily living: longitudinal changes and correlation with other assessments. *Movement Disorders* 28, 14 (2013), 1980–1986.

[10] Walter Maetzler, Josefa Domingos, Karin Srulijes, Joaquim J Ferreira, and Bastiaan R Bloem. 2013. Quantitative wearable sensors for objective assessment of Parkinson’s disease. *Movement Disorders* 28, 12 (2013), 1628–1637.

[11] C Warren Olanow, Yves Agid, Yoshi Mizuno, Alberto Albanese, U Bonucelli, Philip Damier, Justo De Yebenes, Oscar Gershanik, Mark Guttman, F Grandas, et al. 2004. Levodopa in the treatment of Parkinson’s disease: current controversies. *Movement disorders* 19, 9 (2004), 997–1005.

[12] Antoine Regnault, Babak Boroojerdi, Juliette Meunier, Massimo Bani, Thomas Morel, and Stefan Cano. 2019. Does the MDS-UPDRS provide the precision to assess progression in early Parkinson’s disease? Learnings from the Parkinson’s progression marker initiative cohort. *Journal of neurology* 266, 8 (2019), 1927–1936.

[13] Rana Zia Ur Rehman, Christopher Buckley, Maria Encarna Micó-Amigo, Cameron Kirk, Michael Dunne-Willows, Claudia Mazzà, Jian Qing Shi, Lisa Alcock, Lynn Rochester, and Silvia Del Din. 2020. Accelerometry-based digital gait characteristics for classification of Parkinson’s disease: what counts? *IEEE open journal of engineering in medicine and biology* 1 (2020), 65–73.

[14] Rana Zia Ur Rehman, Lynn Rochester, Alison J Yarnall, and Silvia Del Din. 2021. Predicting the Progression of Parkinson’s Disease MDS-UPDRS-III Motor Severity Score from Gait Data using Deep Learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 249–252.

[15] Ole-Bjørn Tysnes and Anette Storstein. 2017. Epidemiology of Parkinson’s disease. *Journal of neural transmission* 124, 8 (2017), 901–905.

Elements of a System for Automatic Monitoring of Specific Mental Health Characteristics at Home

Kristina Kirsten, Bert Arnrich
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
{kristina.kirsten,bert.anrich}@hpi.de

ABSTRACT

Addressing one's mental health has never been more important. The incidences of mental diseases, such as depression or anxiety disorders, have drastically increased in recent years. The longer an adequate treatment is delayed, the greater the impact on the severity of the illness which often results in long absences from work. With the development of smart devices and wearables, it is already possible to measure many physiological parameters in everyday life. In addition, monitoring people in their natural environment offers many advantages, e.g. it is not based on retrospective feelings and memories but can measure and reflect the momentary state. This conceptual paper presents an overview of possible elements of a system for automated monitoring of mental health characteristics in the home. We describe examples of typical parameters for various mental disorders and present different systems and methods to measure them. Furthermore, we show how the individual components of a system can be connected to get a holistic view of specific mental health characteristics. Finally, we also discuss challenges and limitations.

KEYWORDS

mental health, wearables, ubiquitous sensing, monitoring concept

1 INTRODUCTION

Being mindful of mental health is more important than ever. In 2019, according to the World Health Organization (WHO), one in eight people worldwide suffered from a mental disorder [20]. That is associated with significant impairments in thinking, emotion regulation, or behavior. The WHO also states that in 2020, the number of people with depression and anxiety disorders increased significantly, due to the COVID-19 pandemic.

The most common mental illnesses include depression, anxiety disorders, bipolar disorder, and obsessive-compulsive disorder (OCD), among others. Often, initial symptoms are not recognized and, consequently, diagnoses are made late, which in many cases leads to a worsening of the symptoms [6]. Nevertheless, mental illnesses have, partly overlapping, typical characteristics. For example, fatigue, and lack of energy are among the most common symptoms of depression, or checking things over repeatedly are signs of OCD. Some of these characteristics are measurable and interpretable with modern sensors, devices, and machine learning models especially when it comes to behavioral or determining physiological parameters. In addition, studying people in their natural environment respectively at home, so-called ambulatory assessment, has many advantage as it can minimize retrospective bias. On the one hand, it enables long-term monitoring which makes it easier to detect small changes. On the other hand, data can be collected at the time of occurrence and do not have to be remembered and described retrospectively when the actual condition has already passed [17].

This paper presents a collection of elements that can be included in a system for automatic monitoring of mental health characteristics in the home environment. These approaches go beyond conventional questionnaires and refer to technical possibilities for measuring individual characteristics. For this, we look at various characteristics of individual mental disorders and present ways in which these can be measured in an automatic way. However, questionnaires, for example in the form of ecological momentary assessments (EMAs), can always be considered as an additional tool for comparison with the automatic measurements. Finally, we also review different solutions for measurability and propose a potential system overview.

2 BACKGROUND

Mental illnesses are disorders that are very diverse and individual and can affect thinking, mood, and behavior. In 2019, 280 million people were living with depression, 301 million people had an anxiety disorder, 40 million people had a bipolar disorder and 14 million people suffered from an eating disorder [20]. But also lesser-known disorders, such as OCD, which affects about 2.3% of people at least once in their lifetime [11], should not be ignored.

There are characteristics or behavioral patterns that can be observed in various mental illnesses and also generally indicate a bad mental health state. These include, but are not limited to, sadness and dejection, excessive anxiety or worry, decreased ability to concentrate, significant fatigue, low energy, sleep problems, and inability to cope with everyday problems or stress [2].

Nevertheless, each mental disorder also has very specific characteristics. Depressed patients, for example, often describe feeling empty and worthless inside and experiencing hopelessness, sadness, and restlessness. Sleep is also affected in most patients, but it can go both ways, with insomnia or excessive need for sleep as symptoms. Furthermore, a loss of interest in hobbies and social activities may also indicate depression. Sometimes patients even report unexplained physical problems such as back pain or headaches [1]. People with bipolar disorder also experience the above symptoms during the depressive phase. But in addition to this, patients also go through manic episodes. In this phase, many characteristics of the depressive episode reverse. Patients often experience an energetic and euphoric phase where their motivation is increased, concentration is improved, less sleep is required, and they feel the drive to be active [3].

There are several types of anxiety disorders, including generalized anxiety disorder (GAD), panic disorder, social anxiety disorder, and phobia-related disorders [4]. They have in common

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

that people suffer from anxiety over a long period of time, which also often increases and interferes with daily activities ranging from the job to personal relationships. In anxiety disorders, individuals often experience physical symptoms. GAD often comes with headaches, muscle and stomach pain, or other unexplained aches. During a panic attack, affected people may feel a racing heart, sweat intensely, tremble, experience loss of control, or feel chest pain. In addition, people with social anxiety disorder tend to blush, adopt a rigid posture, or speak with an overly soft voice.

For OCD, patients suffer from recurrent obsessive thoughts or compulsive acts. Obsessive thoughts are ideas, images, or impulses that repeatedly appear in the mind of the affected person. The patient cannot successfully suppress these thoughts. Further, more obvious symptoms of OCD are compulsive acts or rituals. They are closely related to the obsessions and serve to alleviate them and the anxiety that is constantly present. The patient is aware of the unusualness of these actions. Most compulsive acts involve cleaning (especially hand washing), repetitive checking to ensure that a potentially dangerous situation does not occur, or order and cleanliness [5].

For any mental illness, not every patient needs to experience all of the characteristic symptoms. Because symptoms can overlap between disorders, it can be difficult to clearly assign them to a single mental illness. By having a system that automatically monitors a range of characteristics, a more holistic picture of mental status can be created, and changes can be detected early.

Diagnoses for mental illness can only be made by professionals. Experts often use various forms of questionnaires and scales to determine the severity of an illness (e.g. Beck Depression Inventory for depression or Yale-Brown Obsessive Compulsive Scale for OCD). However, collecting and analyzing sensor data to monitor mental health in general, is a topic that has been studied a lot in recent years but is still very relevant and has great potential. The majority of studies are related to the analysis of smartphone data, but wearables are also increasingly used for mental health studies. When it comes to the specific monitoring of certain mental illnesses, the vast majority of these studies relate to anxiety disorders, depression, bipolar disorder or stress in general [14]. This paper focuses on technical possibilities to unobtrusively measure certain mental health characteristics in the home environment by using the latest technologies.

3 MONITORING SYSTEM ELEMENTS

To monitor certain mental health characteristics in the home environment, it is possible to use various new wearable devices, human activity recognition (HAR), indoor positioning systems (IPSs) and already derived parameters from consumer devices.

3.1 Smart Devices and Wearables

The smartphone is an integral part of everyday life and almost all of us carry it with us all the time. Although it is the most common everyday smart device, the use of so-called wearables has also been rising rapidly in recent years [19]. The term Internet of Things (IoT) is shaping the technological development of the last decade. It includes devices such as activity trackers, smartwatches, and smart rings. Since these are worn on the body and therefore often called wearables, they can measure physiological parameters such as heart rate variability (HRV), blood oxygen level, or skin conductivity. The modern smart devices contain a variety of sensors, such as oximetry sensors, skin temperature, and ambient temperature sensors, electrodermal activity sensors, heart rate sensors but also Global Positioning System (GPS) and inertial measurement units (IMUs). The latter is a combination of several inertial sensors such as a 3D accelerometer and a 3D gyroscope. However, the term IoT covers many more areas and intelligent devices, such as connected personal scales, smart ovens, and stoves, or smart lighting systems which can be grouped together under the term smart home.

3.2 Human Activity Recognition

The topic of HAR has been widely researched as it offers enormous potential and numerous use cases [8, 9, 12]. It comprises the research field of automatic detection and differentiation of various everyday activities and can be divided into video-based and sensor-based HAR. With the development of new and increasingly powerful smart devices and wearables, HAR is becoming less expensive, easily accessible, and unobtrusive. Research shows that when combining data from different devices, such as smartphone and smartwatch, the results become even more accurate [13]. These days, HAR goes far beyond simple classifications, such as the distinction between sitting, standing, and walking. Among others, HAR also finds great application in the healthcare sector, e.g. through gait analyses that indicate diseases such as Alzheimer's [18] or in systems that focus on elderly care to detect falls [10], for example.

3.3 Indoor Positioning Systems

The ability to determine a person's exact location in a home can help better identify activities that are connected to specific locations, for example, compulsive or eating behavior. Although GPS offers high coverage, it is not suitable for indoor localization because the receiver and satellite have to be in the line of sight, and walls, roofs, and other objects prevent this. That is why in recent years approaches for IPS have been designed which use various available technologies such as radio-frequency identification (RFID), Wireless Local Area Networks (WLAN), Bluetooth Low Energy (BLE) beacons, and more recently Ultra Wideband (UWB) [15, 21, 22]. Localization techniques can be divided into triangulation algorithms (e.g. Time of Arrival (ToA), Time Differences of Arrival (TDoA), Received Signal Strength Indicators (RSSI)-based, Angle of Arrival (AoA)), scene analysis (e.g. Fingerprinting-based techniques) and proximity detection algorithms [21]. The latter is the process of determining whether a user is close to a certain range. This concept is often found in combination with BLE beacons, which are installed stationary at points of interest and send Bluetooth packets that are picked up and processed by the user's smartphone, calculating the distance. In a scene analysis with using Fingerprints, measurements as e.g. RSSI-values, are collected in an offline phase for different positions and stored in a map. For position determination in real-time, the current measurements are then compared with offline measurements to determine the user's location [22].

Different localization techniques have advantages and disadvantages and it depends on the use case which methods are suitable. Most triangulation techniques (e.g. AoA) provide high accuracy but require complex hardware and extensive synchronization. Whereas RSSI- and Fingerprinting-based methods are fairly easy to use but with lower accuracy or, in the case of Fingerprinting, with a dependence on a predefined map that is sensitive to any change in the home environment [22].

3.4 Derived Parameters

In addition to using raw sensor data for use cases like HAR or IPS, consumer devices often provide pre-calculated values and derived parameters, such as about sleep. Many device manufacturers try to draw conclusions about sleep duration, sleep quality, and sleep phases. Additionally, information such as screen time, the frequency with which the phone is picked up, or the number of calls and messages is also documented. Even though many of these values are pre-calculated and in some cases do not provide much information on their own, they can give insights when combined with each other and with data from additional devices.

4 EXEMPLARY SYSTEM OVERVIEW

This section describes example characteristics and their monitoring possibilities, and proposes a connected system architecture.

4.1 Characteristics Monitoring

The possible elements of a monitoring system presented in the previous section, offer particular value when combining them. Different systems and methods are needed to measure specific psychological characteristics. To illustrate this, we looked at some symptoms and characteristics of mental illnesses and considered how these can be measured. The following Table 1 shows a short list of mental health characteristics and possible ways of measuring them. This table represents an exemplary overview and therefore does not claim to be complete. With this table, we show that different characteristics can be measured and documented with the same sensors, wearables, and systems but also that one characteristic can be determined with more than one measurement. We focused on the three main elements for monitoring, namely a HAR system, measuring and evaluating physiological parameters (abbreviated with PP in the table), and using an IPS. Additionally, we list other parameters or devices which can support the measurement of the respective characteristic. For some characteristics, additional information might increase the accuracy and lead to a greater knowledge gain (indicated by (x) in the table). In general, it can be said that oftentimes the combination of different input signals and parameters leads to a better system quality [7]. We do not present the exact algorithms and devices, as these depend heavily on other external factors (availability of devices, overall use case, acceptance of the user, privacy aspects).

It has long been known that sleep, e.g. in form of insomnia, is an essential feature of mental disorders such as depression or anxiety [16]. Sleeping behavior can be observed across a variety of systems and devices. By means of a HAR system, for example, it is possible to document how often a person wakes up at night, how restful the sleep is, and when and whether one gets out of bed in the morning. Monitoring this behavior can help in observing depressive phases, where patients sometimes find it difficult to get out of bed at all. But beyond that, it can also make sense to include other information, such as the position in the apartment in order to get more contextual information.

The measurement of physiological parameters can help for the majority of the characteristics. By measuring skin conductance, for example stress, which plays a major role in many mental illnesses, could be detected. Furthermore, it is also known that social behavior changes in some mental disorders. For example, social interaction decreases in depressive or anxiety patients but increases in people in a manic phase.

For some characteristics, it is particularly interesting to look at changes over time because mental illnesses often have very individual expressions. For this purpose, it can be helpful to train a personalized machine learning model for a potential patient in order to observe variations from normal behavior. In general, personalized models are well suited to represent the individual aspects of everyday activities.

Table 1: Listing of exemplary mental health characteristics and possibilities of monitoring them. HAR corresponds to the detection of human movements with motion sensors, PP stands for measuring physiological parameters and IPS implies the positioning of a person in the room or home.

Characteristic	HAR	PP	IPS	Others
Sleeping Behaviour	x	x	(x)	derived smartphone and smartwatch parameters (sleep hours, sleep phases, sleep quality)
Compulsive Handwashing	x		(x)	
Compulsive Checking	x		x	
Stress Level	(x)	x		
Eating Behavior	x	x	x	interaction with IoT devices, e.g. personal scale, microwave
Social Interaction		(x)	(x)	derived smartphone and smartwatch parameters (screen time, pick up times, phone call and messages frequencies)

4.2 Connected System

In Figure 1 we demonstrate how the individual components of a system for monitoring characteristics of mental disorders can be connected. Depending on the concrete use case, data from multiple devices will be constantly collected. For energy efficiency, it makes sense to store the collected data on the respective device first, and only send it to a data hub from time to time. For this, smartphone applications like SensorHub [7] are very useful. Multiple (wearable) sensors can be connected via Bluetooth, collecting and storing the data in a central place and a unified format to provide complete control over the data. Additionally, systems like SensorHub provide the possibility to get point-in-time feedback from the user by repeatedly querying certain conditions (behavior, feelings, experiences), so-called EMAs. This is extremely valuable and these subjective sensations could be supported and enriched by objective, quantifiable sensor measurements.

A system designed to give a holistic view of a current state is not intended to make assessments or provide results at any time. That means these kinds of systems have a long-term character rather than being a snapshot. Moreover, when working with raw sensor data, this often means that it needs a lot of pre-processing and cleaning. This includes e.g. filtering and de-noising. When it comes to machine learning, domain-specific knowledge is also helpful in order to come up with meaningful features.

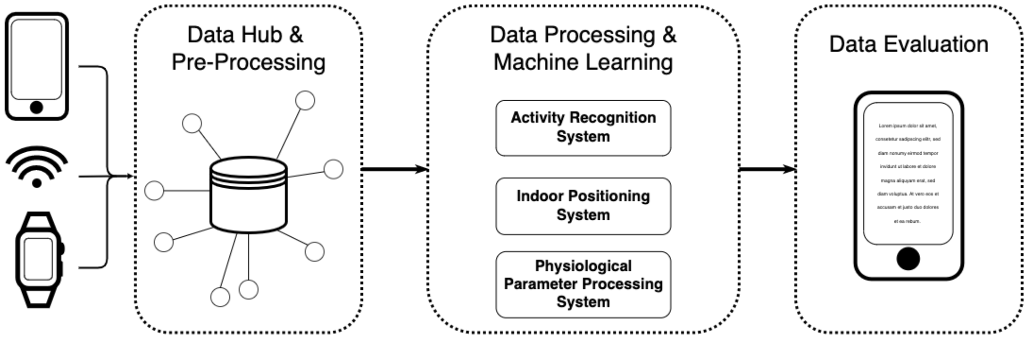

Data Hub & Pre-Processing
Data Processing & Machine Learning
Activity Recognition System
Indoor Positioning System
Physiological Parameter Processing System
Data Evaluation

Figure 1: System Overview to Monitor Mental Health Characteristics at Home

5 CHALLENGES AND LIMITATIONS

Each component of an overall system has its advantages and disadvantages. It always has to be determined which features predominate for the specific use case. It should also be noted that issues like data security, especially with such sensitive topics as mental disorders, play a tremendous role. For this, all actions should be transparent to the user. The consumer must be informed in advance about all processes, devices and measurements, and be able to stop the monitoring at any time. This complete transparency can result in the user consciously or subconsciously adapting his/her behavior when he/she feels observed. However, these effects should be negligible, as this type of monitoring happens over a longer period of time and thus integrates into everyday life over time.

Furthermore, it should be kept in mind that systems that integrate everyday user devices (smartphone, smartwatch, and activity tracker) are also always limited in their battery power, especially if they are in constant use. Here, a balance must be found between monitoring frequency and consumption. The times when the devices have to be charged (usually daily) must also be taken into account in the system design.

In general, one of the most important factors is that the monitoring system is as pleasant and unobtrusive as possible for the user. It must be installed with as little effort as necessary and be perfectly integrated into everyday life.

6 CONCLUSION

This paper presented possible ways to measure various characteristics of mental disorders. We want to emphasize that systems of this type are not diagnostic tools and are in no way equivalent to professional assessments. But they can support and help to describe a given state and to perceive and document changes. In general, it is helpful to make psychological characteristics measurable and thus to support the subjective feelings of patients by means of objective measurements. Moreover, even small changes can be detected and documented at an early stage and help to take countermeasures in time. It could provide new insights into behavioral patterns, overlaps of different diseases, and personal aspects. Furthermore, these forms of monitoring systems cannot only be used for early detection but also for relapse supervision.

In future work, an exemplary monitoring system will be built for detecting compulsive behavior as it occurs in patients suffering from OCD. We also want to determine to what extent such systems are accepted by potential patients and also what other limitations and possibilities are encountered.

REFERENCES

[1] 2018. Depression (major depressive disorder). https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007
[2] 2019. Mental illness. https://www.mayoclinic.org/diseases-conditions/mental-illness/symptoms-causes/syc-20374968
[3] 2020. Bipolar Disorder. https://www.nimh.nih.gov/health/topics/bipolar-disorder
[4] 2022. Anxiety Disorders. https://www.nimh.nih.gov/health/topics/anxiety-disorders
[5] The American Psychiatric Association (APA). [n.d.]. *What Is Obsessive-Compulsive Disorder?* https://www.psychiatry.org/patients-families/ocd/what-is-obsessive-compulsive-disorder/
[6] Elisabetta Burchi, Eric Hollander, and Stefano Pallanti. 2018. From treatment response to recovery: a realistic goal in OCD. *International Journal of Neuropsychopharmacology* 21, 11 (2018), 1007–1013.
[7] Jonas Chromik, Kristina Kirsten, Arne Herdick, Arpita Mallikarjuna Kappattanavar, and Bert Arnrich. 2022. SensorHub: multimodal sensing in real-life enables home-based studies. *Sensors* 22, 1 (2022), 408.
[8] Maria Cornacchia, Koray Ozcan, Yu Zheng, and Senem Velipasalar. 2016. A survey on activity detection and classification using wearable sensors. *IEEE Sensors Journal* 17, 2 (2016), 386–403.
[9] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition* 108 (2020), 107561.
[10] Miguel Ángel Álvarez de la Concepción, Luis Miguel Soria Morillo, Juan Antonio Álvarez García, and Luis González-Abril. 2017. Mobile activity recognition and fall detection system for elderly people using Ameva algorithm. *Pervasive and Mobile Computing* 34 (2017), 3–13.
[11] Wayne K Goodman, Dorothy E Grice, Kyle AB Lapidus, and Barbara J Coffey. 2014. Obsessive-compulsive disorder. *Psychiatric Clinics* 37, 3 (2014), 257–267.
[12] Oscar D Lara and Miguel A Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials* 15, 3 (2012), 1192–1209.
[13] Felipe Barbosa Araújo Ramos, Anne Lorayne, Antonio Alexandre Moura Costa, Reudismam Rolim de Sousa, Hyggo O Almeida, and Angelo Perkusich. 2016. Combining Smartphone and Smartwatch Sensor Data in Activity Recognition Approaches: an Experimental Evaluation.. In *SEKE*. 267–272.
[14] Mahsa Sheikh, Meha Qassem, and Panicos A Kyriacou. 2021. Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Frontiers in Digital Health* (2021), 33.
[15] Santosh Subedi and Jae-Young Pyun. 2020. A survey of smartphone-based indoor positioning system using RF-based wireless technologies. *Sensors* 20, 24 (2020), 7230.
[16] Daniel J Taylor, Kenneth L Lichstein, H Heith Durrence, Brant W Reidel, and Andrew J Bush. 2005. Epidemiology of insomnia, depression, and anxiety. *Sleep* 28, 11 (2005), 1457–1464.
[17] Timothy J Trull and Ulrich Ebner-Priemer. 2013. Ambulatory assessment. *Annual review of clinical psychology* 9 (2013), 151.
[18] Ramachandran Varatharajan, Gunasekaran Manogaran, Malarvizhi Kumar Priyan, and Revathi Sundarasekar. 2018. Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. *Cluster Computing* 21, 1 (2018), 681–690.
[19] Vini Vijayan, James P Connolly, Joan Condell, Nigel McKelvey, and Philip Gardiner. 2021. Review of wearable devices and data collection considerations for connected health. *Sensors* 21, 16 (2021), 5589.
[20] World Health Organization (WHO). 2022. Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders
[21] Ali Yassin, Youssef Nasser, Mariette Awad, Ahmed Al-Dubai, Ran Liu, Chau Yuen, Ronald Raulefs, and Elias Aboutanios. 2016. Recent advances in indoor localization: A survey on theoretical approaches and applications. *IEEE Communications Surveys & Tutorials* 19, 2 (2016), 1327–1346.
[22] Faheem Zafari, Athanasios Gkelias, and Kin K Leung. 2019. A survey of indoor localization systems and technologies. *IEEE Communications Surveys & Tutorials* 21, 3 (2019), 2568–2599.

Towards Multi-Modal Recordings in Daily Life: A Baseline Assessment of an Experimental Framework

Christoph Anders*
Sidratul Moontaha*
Bert Anrich
firstname.lastname@hpi.de
Hasso Plattner Institute
Potsdam, Germany

ABSTRACT

Background: Wearable devices can record physiological signals from humans to enable an objective assessment of their Mental State. In the future, such devices will enable researchers to work on paradigms outside, rather than only inside, of controlled laboratory environments. This transition requires a paradigm shift on how experiments are conducted, and introduces new challenges. **Method:** Here, an experimental framework for multi-modal baseline assessments is presented. The developed test battery covers stimuli and questionnaire presenters, and multi-modal data can be recorded in parallel, such as Photoplethysmography, Electroencephalography, Acceleration, and Electrodermal Activity data. The multi-modal data is extracted using a single platform, and synchronized using a shake detection tool. A baseline was recorded from eight participants in a controlled environment. Using Leave-One-Out Cross-Validation, the resampling of data, the ideal window size, and the applicability of Deep Learning for Mental Workload Classification were evaluated. In addition, participants were polled on the acceptance of using the wearable devices. **Results:** The binary classification performance declined by an average of 7.81% when using eye-blink removal, underlining the importance of data synchronization, correct artefact identification, evaluating and developing artefact removal techniques, and investigating on the robustness of the multi-modal setup. Experiments showed that the optimal window size for the acquired data is 30 seconds for Mental Workload classification, with which a Random Forest classifier and an optimized Deep Convolutional Neural Network achieved the best-balanced classification accuracy of 70.27% and 74.16%, respectively. **Conclusions:** This baseline assessment gives valuable insights on how to prototype stimulus presentation with different wearable devices and suggests future work packages, paving the way for researchers to investigate new paradigm outside of controlled environments.

1 INTRODUCTION

The concept of Mental Workload (MW) originates from the field of psychology, refers to the amount of working memory used in the brain, and is historically researched on in the context of laboratories [1]. High levels of MW experienced over an extended period of time lead to Mental Fatigue (MF). It can be assumed that the onset of MF depends on contextual factors such as level of sleep during previous nights, overall health, emotional state, and more. MF can increase the amount of mistakes an individual does, and hinder work-performance amongst others. The impact of MF on economies can be estimated from the finding that a fatigued work-force costs the US economy an approximation of 18 billion USD per year [2]. Methods that quantify the level of MW an individual experiences in and outside of laboratory environments are of interest to a broad community.

MF can be circumvented in various ways, e.g. by taking more Micro-Breaks [2]. To quantify the impact of interventions, measurement frameworks have to be developed in controlled environments and evaluated for use in uncontrolled environments. Subjective measurements of MW can be performed using questionnaires or discussions with individuals. However, these approaches take time, require active truthful participation, and are therefore not suited for every context. To overcome this hurdle, objective measurement methods are researched, amongst which EEG seems promising [3].

To-be-developed measurement frameworks for experiments mainly conducted in controlled environments, such as MW quantification, need to be combined with research on the quality and amount of sensor data needed, accurate synchronization between different modalities, and precise data labeling. Merging research on all these aspects into one skeleton would increase the overall usability of the resulting framework. This paper presents an experimental framework for baseline assessment on the use-case of objective measurements of MW conducted across university students. As data storage, compression, and transmission consume a lot of battery power [4], the length of time windows required for accurate classifications, the sampling-rate required, and the time-series classification performance were evaluated. Finally, participants of this study were surveyed about their experiences with the two well-established wearable devices used, since this framework can be customized in terms of stimulus presentation and multi-modality used for the Affective Computing research community in general. The measurement framework is presented in detail, and necessary steps towards an experimental framework for multi-modal recordings in uncontrolled environments are outlined.

2 EXPERIMENTAL FRAMEWORK

The experimental framework for this study was built using PsychoPy (v2022.2.0) [5] running under Python 3.10.4 in a controlled environment, as a preliminary step for recordings in daily life. Among the most frequently used software packages for visual stimulus presentation[1], Psychopy was preferred due to the usability, automated calibration feature, and the real-time stimulus

*Both authors contributed equally to this research.

[1] http://hans-strasburger.userweb.mwn.de/psy_soft.html#imagen

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia

© 2022 Copyright held by the owner/author(s).

presentation [6]. The setup was implemented to induce MW in line with common practice from state-of-the-art studies (e.g. [7]). As a first step, participants were asked the put all the devices into a box and shake them, to synchronize the devices. Then, high magnitude tapping onto the space bar was performed to synchronize with Psychopy. After instructing participants to minimize movement, a five minute relaxation video[2] was presented for baseline recording. An eye-closing session of one-minute duration followed, before the MW was induced. Participants had to work on the N-Back task (*n=3*) for five minutes. Afterwards, participants had to work for five minutes on the Stroop task, where four colors (yellow, green, blue, and red) were shown for a duration of 3 seconds. For every wrong answer, a *buzz* sound was played to intensify the workload and provide feedback to the participants. Both tasks were followed by the pairwise NASA Task Load Index (NASA-TLX) questionnaire [8]. By using physiological data recorded during the relaxation video and eye-closing session as *'Low-to-No-Workload'*-class, and using the data from both MW tasks as *'High-Workload'*-class, a binary classification task was formed. Physiological data recorded during answering of the questionnaires, or reading instructions for the MW tasks, was excluded. With a ratio of *4:10* for *'Low-to-No-Workload'* to *'High-Workload'*, the recorded data was imbalanced.

Two wearable devices were used: The Empatica E4 which records skin temperature (4 Hz), PPG (64 Hz), and GSR (4 Hz), alongside acceleration-readings (32 Hz) that can be used for the identification and removal of artefacts. The Muse S device was used, which records EEG (256 Hz) and accelerometer data (50 Hz). Following the 10/20-system for electrode placement [9], the EEG electrodes of the Muse S [3] device are located at TP9, AF7, AF8, TP10, with a reference electrode at FPz.

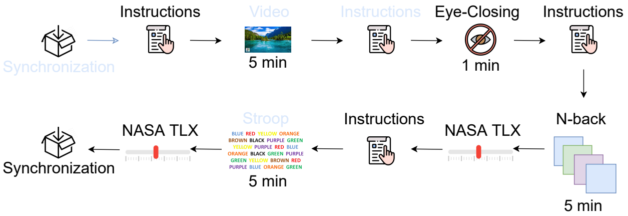


Figure 1: Study design of the experimental paradigm utilized for the multi-modal framework

3 METHODS

The Muse S data was recorded using MindMonitor [4] and loaded by *devicely* [5], whereas the Empatica E4 data was recorded using the SesnsorHub Application [10]. Synchronization was performed at simultaneous peaks in the accelerometer data, using *jointly*[6] on readings from both wearable devices. Acceleration was caused once in the beginning and once at the end of the experimental protocol: The devices were placed in the same box, and the box was shook. This procedure was repeated after the experiment. Potential offsets and time-shifts in the recordings were automatically corrected by Jointly. Labeling of the sensor data was performed using the information contained in the logs from PsychoPy. How data labeling will be performed for recordings in uncontrolled environments remains an open question.

Once the data was labeled, data cleaning needed to be performed. As time-series data is not uniform over time (e.g. due to a temporary loss of connection), missing values needed to be interpolated. Linear interpolation was performed by filling missing data with the mean value of two neighboring data points. Additionally, head-movements and eye-blinks predominantly compromised the EEG recordings, while movements of the hand predominantly compromised readings from the Empatica E4. Removal of artefacts in the data from the Empatica E4 was performed in three steps: First, both the raw values for acceleration and BVP were normalized to the range of [-1, 1]. Second, a fourth-order Butterworth band-pass filter with 0.5 Hz and 3.5 Hz cutoff-frequencies was applied. Third, a Savitzky-Golay filter was applied, using a 101-sample window and a 5th-degree polynomial. These steps removed the baseline-drift in the recorded BVP signal. Additionally, adaptive noise cancellation was performed to remove movement-artefacts from the BVP signal, by using linear recursive least-squares filtering. Removal of artefacts from the EEG signal was performed using spectral filtering with an infinite impulse response filter. Following parameter recommendations from the literature [11], a Chebyshev type 2 band-pass filter with 0.5 Hz and 48.5 Hz cutoff-frequencies and 40 dB attenuation in the pass-band was applied. Thereby, the power-line interference and other artefacts such as jaw-clenching were removed. Strong artefacts for EEG recordings, especially in the frontal channels, are eye-blinks [12]. Here, eye-blink were removed using the independent component analysis (ICA) [13].

Spatial filtering of the EEG data was investigated using the common spatial pattern (CSP) algorithm [14] implemented in the *meet*[7] repository [15]. CSP performs a generalized eigenvalue decomposition of two distinct mutlivariate sets of data, for which an additive underlying mixture of sources is assumed. CSP basically maximizes power differences between the two conditions *'Low-to-No-Workload'* and *'High-Workload'*. After derivation of filter values for each channel, the filter with the highest Eigenvalue is chosen and applied to both the *'Low-to-No-Workload'*-, and the *'High-Workload'*-, classes. The result is the sum of all the multiplications of the respective scalar-filters with the corresponding electrode-channels, resulting in one single channel which best describes the underlying phenomenon optimized for.

Temporal filtering describes the process of either rejecting recordings from the process of building trials all-together (e.g. physiological data recorded during answering of questionnaires), or of building trials from the recorded data. Two important parameters have to be taken into account: *window-size*, and *window-overlap*. Here, multiple parameters for the *window-size* were evaluated: *5 sec, 10 sec, 30 sec*. The *window-overlap* was constantly chosen to be *0.5 sec* smaller than the respective *window-size*: *4.5 sec, 9.5 sec, 29.5 sec*.

To extract different features, the cleaned BVP signal was used to extract the heart rate variability using *NeuroKit2* [8] package [16], which locates the peaks in the peak to peak (RR) interval of the hear rate variability and calculates different time-and frequency-domain features, partially mentioned below. Additionally, the mean and standard deviations (SD) from GSR and skin temperature were extracted. The different feature-sets utilized

[2] https://www.youtube.com/watch?v=S6jCd2hSVKA

[3] https://choosemuse.com/de/muse-s/

[4] https://mind-monitor.com/

[5] https://github.com/hpi-dhc/devicely

[6] https://github.com/hpi-dhc/jointly

[7] https://github.com/neurophysics/meet

[8] https://neuropsychology.github.io/NeuroKit/

were extracted from the training data only, and can be summarized as follows: **CSP features:** Gamma, Beta, Alpha, Theta and Delta band powers, mean over the band powers, mean and SD of the absolute band powers; **BVP features:** Mean and SD of the RR intervals (peak to peak of Hear Rate Variability), SD of the successive differences between RR intervals, ratio of SD and mean RR intervals, low frequency band power (0.04 - 0.15 Hz), high frequency band power (0.15 - 0.4 Hz), very high frequency band power (0.4 - 0.5 Hz), ratio of low-high band power; **GSR:** Mean and SD of absolute values, mean amplitude of Skin Conductivity Response (SCR) peaks; **Local Skin Temperature:** Mean and SD of absolute values; and **PSD features:** Power spectral density of raw EEG of TP9, TP10, AF7, AF8.

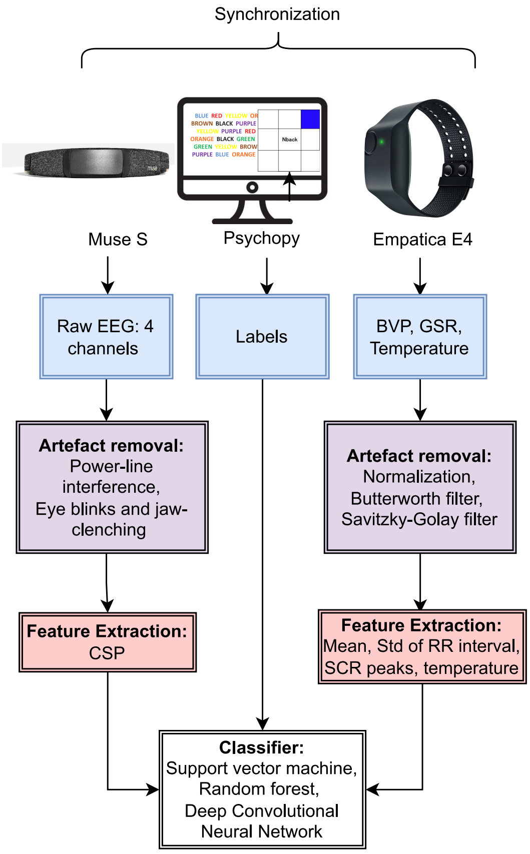

Synchronization
Nback
Muse S
Psychopy
Empatica E4
Raw EEG: 4 channels
Labels
BVP, GSR, Temperature
Artefact removal: Power-line interference, Eye blinks and jaw-clenching
Artefact removal: Normalization, Butterworth filter, Savitzky-Golay filter
Feature Extraction: CSP
Feature Extraction: Mean, Std of RR interval, SCR peaks, temperature
Classifier: Support vector machine, Random forest, Deep Convolutional Neural Network

Figure 2: The flowchart of the employed study protocol with the necessary intermediate steps.

In total, three different evaluations were performed on the data recorded in a controlled environment. First, two different feature sets were investigated for data resampled to 10 Hz, using a Random Forest (RF) classifier. This evaluation was performed to investigate on the possibility of reducing the sampling rate required per modality. Second, the optimal time window for time series classification (TSC) of MW was investigated on by comparing the performance of different feature sets utilized by RF and a Support Vector Machine (SVM). Therefore, the modalities were utilized at the respective sampling rates recorded with and simply combined. Third, the application of Deep Learning to this task was investigated using a Deep Convolutional Neural Network (DCNN) [17]. The DCNN was built of ten layers: Input (2D convolution, 5x5, ReLu), 1st Hidden (2D Max Pooling, 2x2), 2nd Hidden (2D convolution, 5x5, ReLu), 3rd Hidden (2D Max Pooling, 2x2), 4th Hidden (Flatten), 5th Hidden (Fully-Connected, ReLu), 6th Hidden (Dropout), 7th Hidden (Fully-Connected, ReLu), 8th Hidden (Dropout), Output (Fully-Connected, Single-Output, Sigmoid).

For the RF, the default hyperparameters of the *RandomForestClassifier* from *scikit-learn* were chosen. For the SVM, a radial basis function kernel was utilized, and the gamma value was calculated for each evaluation. The best hyperparameters of DCNN were identified using the sequential model based optimization (SMBO) algorithm with the tree-structured parzen estimator (TPE), which has been shown to outperform both grid search and random search [18]. The derived hyperparameters are listed in Table 1. The inputs to all classifiers were min-max normalized.

Hyperparameter	Value Range	Baseline	Optimized
Dropout	0 - 0.5 (0.1)	0.5	0.3
Epochs	1 -200 (5)	200	25
Batch Size	1 - 1000 (50)	500	350
Conv. Layer 1	10 - 100 (10)	20	70
Conv. Layer 2	25 - 250 (25)	50	125
Hidden Layer 1	100 - 1000 (50)	500	200
Hidden Layer 2	100 - 1000 (50)	250	750
Window Size	5 - 30	5	30
Input Height	20 - 130 (10)	28	110
Input Width	20 - 130 (10)	28	110

Table 1: Hyperparameters for the DCNN. Values in parenthesis indicate incremental steps. Window size in seconds.

4 RESULTS

The first experimental evaluation used two different sets of features, each resampled to 10 Hz. Averaged results of all of the Leave-One-Out Cross-Validation for the classification tasks are shown in Table 2.

Set #	Window Size	Blink Removal	Balanced Acc.
Set 1	1200 sec	no	74.06
Set 1	1200 sec	yes	65.52
Set 1	**6000 sec**	**no**	**82.21**
Set 1	6000 sec	yes	73.49
Set 2	1200 sec	no	77.31
Set 2	1200 sec	yes	72.43
Set 2	**6000 sec**	**no**	**80.94**
Set 2	6000 sec	yes	71.84

Table 2: TSC Performance for RF. Set 1: Raw TP9, TP10, AF8, AF7, Skin Temperature, BVP features. Set 2: Set 1 + GSR. The row of the best performance is printed in bolt face.

The second experiment evaluated on the optimal *window-size*. Results are visualized in Figure 3, where the PSD feature set refers to all the extracted features mentioned in 3, and the FE feature-set refers to all but the PSD features. With the FE feature-set, while RF performed best across all time-windows, the average time series classification performance increased only marginally across all TSC models when varying the *window-size*. The best performance of 70.27% balanced accuracy was achieved for RF with FE for a *window-size* of *30 sec*.

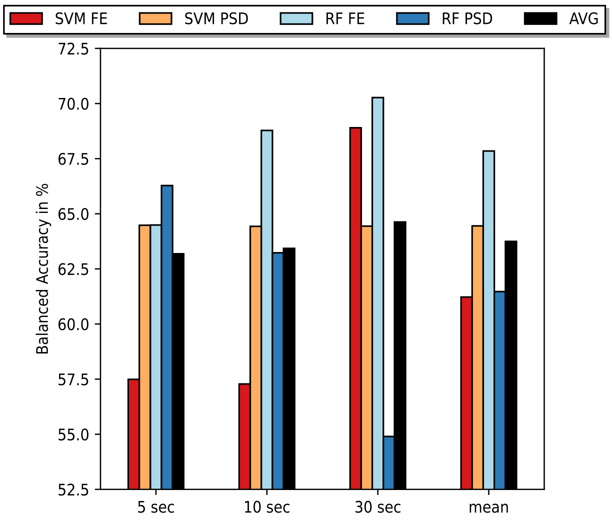


Figure 3: TSC Performance for RF and SVM. The choice of features and windows significantly impacted inter-subject TSC performance.

The third experiment investigated on the applicability of Deep Learning to this task. The *baseline*-DCNN achieved a balanced accuracy of 59.79%, whereas the *optimized*-DCNN achieved a balanced accuracy of 74.16%.

Eight participants were recruited in this baseline assessment and provided subjective feedback on their experiences with the setup: No participant complained about uncomfortable feelings due to pressure from the sensors, but sensors felt too bulky, and the utilization of three different devices—two sensors and one phone for recordings—seemed too complicated.

5 CONCLUSION

In the first experiment, it was found that eye-blink removal worsened the TSC performance. This finding was consistent across all test-runs, and the average loss in balanced classification accuracy was with 7.81% substantial. Amongst others, reasons for this circumstance are: Firstly, the existence of only one eye-blink per time window of 20 seconds duration was assumed, which proved false. Secondly, more advanced algorithms for automatic eye-blink removal and signal restoration exist, which outperformed ICA-based methods [19, 20] and should have been applied.

In the second experiment, it was found that the best accuracy was achieved for a time-window with *window-size* of *30 sec.* This finding is in line with findings in the literature on affective computing (e.g. [21]). Furthermore, the FE feature set performed better for this task than the PSD feature set, for which the TSC performance stagnated or even declined. Future work should investigate on computing PSD features from further cleaned EEG data, and on features such as power in key frequency bands.

Finally, it was found that the optimization of the DCNN also led to choosing a *window-size* of *30 sec.* This finding is in line with the results from the second experiment, where the average performance also peaked for the *window-size* of *30 sec.* However, as this was the maximum value evaluated for, it might be that the models would have performed better for longer time-windows.

The performed baseline assessment highlights future work, such as to investigate on better algorithms for artefact removal (e.g. [19, 20]); on longer *window-sizes*, different DL models, more features such as power-ratios; to recruit more participants; and to investigate on feature-importance. Also, resampling the sensor data to frequencies other than 10 Hz and investigating the effect of interventions to remove MF in controlled environments, should be performed. The presented framework needs to be extended to allow automatic randomization of the tasks, recovery from crashes, more robust data extraction, to be evaluated for applicability to uncontrolled environments, and published. Experimental paradigms for measuring MW need to be taken from controlled environments, and frameworks that are under development need to be tested and evaluated in uncontrolled settings.

6 ACKNOWLEDGMENTS

We appreciate the contribution of Ahmed Azzouz, Alina Krichevsky, Leonhard Hennicke, Nastassia Heumann, Nikita Shishelyakin and Tanja Manlik as a part of their master project. This research was (partially) funded by the HPI Research School on Data Science and Engineering.

REFERENCES

[1] Hart et al. 1988. Development of nasa-tlx (task load index): results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183. DOI: 10.1016/S0166-4115(08)62386-9.

[2] Khosro Sadeghniiat-Haghighi and Zohreh Yazdi. 2015. Fatigue management in the workplace. *Industrial Psychiatry Journal*, 24, 1, (Jan. 1, 2015), 12. DOI: 10.4103/0972-6748.160915.

[3] Hogervorst et al. 2014. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in Neuroscience*, 8. DOI: 10.3389/fnins.2014.00322.

[4] Casson. 2019. Wearable EEG and beyond. *Biomedical Engineering Letters*, 9, 1, (Jan. 2019), 53–71. DOI: 10.1007/s13534-018-00093-6.

[5] Peirce et al. 2022. *Building experiments in PsychoPy*. ISBN-13: 978-1473991392. Sage.

[6] Rolf Kötter. 2009. A primer of visual stimulus presentation software. *Frontiers in neuroscience*, 21.

[7] Giorgi et al. 2021. Wearable technologies for mental workload, stress, and emotional state assessment during working-like tasks: a comparison with laboratory technologies. *Sensors*, 21, 7, 2332. DOI: 10.3390/s21072332.

[8] Ernesto A. Bustamante and Randall D. Spain. 2008. Measurement invariance of the nasa tlx. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52, 19, (Sept. 2008), 1522–1526. DOI: 10.1177/154193120805201946.

[9] Chatrian et al. 1985. Ten percent electrode system for topographic studies of spontaneous and evoked eeg activities. *American Journal of EEG technology*, 25, 2, 83–92. DOI: 10.1080/00029238.1985.11080163.

[10] Chromik et al. 2022. Sensorhub: multimodal sensing in real-life enables home-based studies. *Sensors*, 22, 1, 408. DOI: 10.3390/s22010408.

[11] Apicella et al. 2021. High-wearable EEG-based distraction detection in motor rehabilitation. *Scientific Reports*, 11. DOI: 10.1038/s41598-021-84447-8.

[12] Ajay Kumar Maddirala and Kalyana C. Veluvolu. 2021. Eye-blink artifact removal from single channel EEG with k-means and SSA. *Scientific Reports*, 11. DOI: 10.1038/s41598-021-90437-7.

[13] A. Hyvärinen and E. Oja. 2000. Independent component analysis: algorithms and applications. *Neural Networks*, 13. DOI: 10.1016/S0893-6080(00)00026-5.

[14] Blankertz et al. 2008. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal Processing Magazine*. DOI: 10.1109/MSP.2008.4408441.

[15] Waterstraat et al. 2017. On optimal spatial filtering for the detection of phase coupling in multivariate neural recordings. *NeuroImage*, 157. DOI: 10.1016/j.neuroimage.2017.06.025.

[16] Makowski et al. 2021. Neurokit2: a python toolbox for neurophysiological signal processing. *Behavior research methods*, 53. DOI: 10.3758/s13428-020-01516-y.

[17] Sarkar et al. 2016. Wearable eeg-based activity recognition in phm-related service environment via deep learning. *international Journal of Prognostics and Health Management*, 7. DOI: 10.36001/ijphm.2016.v7i4.2459.

[18] Bergstra et al. 2013. Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*. Vol. 13. Citeseer.

[19] Nguyen et al. 2020. A deep wavelet sparse autoencoder method for online and automatic electrooculographical artifact removal. *Neural Computing and Applications*, 32. DOI: 10.1007/s00521-020-04953-0.

[20] Olaf Dimigen. 2020. Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *NeuroImage*, 207. DOI: 10.1016/j.neuroimage.2019.116117.

[21] Athavipach et al. 2019. A wearable in-ear EEG device for emotion monitoring. *Sensors*, 19, 18. DOI: 10.3390/s19184014.

Assessing Sources of Variability of Hierarchical Data in a Repeated-Measures Diary Study of Stress

Junoš Lukan
Jožef Stefan Institute
Department of Intelligent Systems
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
junos.lukan@ijs.si

Larissa Bolliger
Department of Public Health and
Primary Care
Ghent University
Ghent, Belgium
larissa.bolliger@ugent.be

Els Clays
Department of Public Health and
Primary Care
Ghent University
Ghent, Belgium
els.clays@ugent.be

Primož Šiško
Jožef Stefan Institute
Department of Intelligent Systems
Ljubljana, Slovenia
sisko.primoz@gmail.com

Mitja Luštrek
Jožef Stefan Institute
Department of Intelligent Systems
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
mitja.lustrek@ijs.si

ABSTRACT

There are different methodological approaches to stress recognition in different disciplines. In machine learning literature, a typical approach is to select a target variable and try to predict it as generally as feasible, but possibly with person-specific feature normalization or personalization of models. In medical, psychological, and social sciences, the nested nature of data is often taken into account by using multilevel models, especially with repeated measures data. In our diary study, we asked participants to assess different aspects of stress every 90 min for 15 working days. They accessed their questionnaires through an Android application which also served to passively record phone usage and sensor data. At the same time they wore Empatica E4 wristbands which collected physiological data. This study design lends itself well to hierarchical consideration. In this paper, we use variance partitioning, a technique which is also a part of multilevel modelling, to inform a machine learning pipeline. We show how consideration of different sources of variability can help us decide how to personalize normalization of data or machine learning models.

KEYWORDS

stress detection, ecological momentary assessment, variance partitioning, hierarchical data

1 INTRODUCTION

Chronic stress is a well researched medical, psychological, and sociological phenomenon which has been shown to have detrimental health consequences [8]. It is less clear, however, how daily experiences of stress translate into a long-term experience of chronic stress [13]. In the STRAW project, we have tackled this question by carrying out a longitudinal diary study [6].

In machine learning literature, this problem falls under the topic of affective computing [19]. Typical studies settle for one definition of stress and either measure it by simply asking about it or using one of the established psychological questionnaires [2]. Next, stress detection is relayed to machine learning models as a supervised problem in which objectively measured data are used as predictors of self-reports, serving as labels.

The aim of this paper is to employ statistical techniques from medical and social sciences to inform machine learning modelling. Specifically, we analyse daily aggregated data collected in our study and consider possibilities for analysis on a lower, within-day level. We do this by describing the data in terms of multilevel models and then assess how each level of measurements contributes to the overall stress variability.

2 METHODS

2.1 Data Collection

Three main data types were collected using different measuring devices. Physiological parameters were measured by Empatica E4 wristbands, while participants filled in questionnaires on their smartphones for 15 working days. These ecological momentary assessments (EMAs) were presented at random intervals throughout the working day, roughly 90 minutes apart, while an additional, longer questionnaire was offered in the evening, asking about the day as a whole. The questions in each EMA session (a set of questions) were selected from questionnaires that measure different aspects of stress and related constructs, such as stress appraisal, negative affect, job demand and job control. Smartphone sensor data and phone usage data were continuously collected by a self-developed Android application based on the AWARE framework [9]. The contents of the questionnaires and the data types collected have already been described in an extensive protocol paper [6].

We collected the data of 56 participants, recruited from academic institutions in Belgium (29 participants) and Slovenia (26 participants). Only the data pertaining to $N = 55$ participants were complete, which included 26 women and 29 men. Their mean age was 34.9 years with the range from 24 years to 63 years and they held various positions in their institutions, such as PhD students, employees in administration, and tenured professors.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

The participants adhered to the study protocol well. In their participation period, each participant responded to more than 96 EMA sessions on average. The median time difference between two subsequent workday EMA sessions was 93 minutes, just a bit over what was designed [12].

2.2 Classical Machine Learning Data Analysis

As the first step of the analysis, we followed a classical machine learning approach for detecting stress (see Figure 3 in [2]). After preprocessing, we calculated hand-crafted features. For phone sensor data, we used a modified *Reproducible Analysis Pipeline for Data Streams* (RAPIDS, [20]) library, which calculates behavioural features using R, Python, and Snakemake [16] following a well-defined set of rules (steps). For physiological data, we used our in-house developed Python library, `cr-features` [11].

The data were aggregated on a daily basis, by averaging target variables and calculating statistical physiological features that were first calculated on short segments. Next, we standardized the data *within* participants, i.e., by subtracting the daily mean and dividing by daily standard deviation. Finally, we used a leave-one-subject-out validation technique and tested various linear (e.g., linear regression), non-linear (e.g., support vector regression) and ensemble machine learning techniques (e.g., ADA boost regressor) from `scikit-learn` [17].

2.3 Variance Partitioning

Multilevel models (also known as mixed-effect, random-effect or mixed models) are methods commonly used in medical, biological, and social sciences to analyse hierarchical (nested) data [10]. Labels in our dataset are nested in at least three levels: each participant collected data on multiple days and each day included several measurements. We analysed self-perceived data from questionnaires using mixed models in other publications [4, 5], while in this paper we use the related technique of variance partitioning for exploring variability of the data at different levels. Variance partitioning (or partitioning of sums of squared deviations) can be used to ascribe the overall variability in a dataset to different sources of variability. In multilevel models, this sources can be different levels of analysis.

2.3.1 *Simple Linear Regression.*

To model daily stress, we can use linear regression in the following form:

$$y_j = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_p x_{jp} + \epsilon_j \tag{1}$$

Here, y_j represents the mean of the chosen indicator of stress on a day j, β_0 is the intercept term, $\left\{x_{j1}, \ldots, x_{jp}\right\}_{j=1}^{n}$ represent daily values of p features (or predictors), $\left\{\beta_1, \ldots, \beta_p\right\}_{j=1}^{n}$ their corresponding regression coefficients, while ϵ_j is the error term which captures all other factors related to variable y, which are not described by the available features (predictors included in the model). The index j runs from 1 to n, where $n = N \times n_d$ is the product of the number of participants (N) and the number of days each one participated in the study (n_d).

As we are interested in variance partitioning only, we can focus on the intercept and omit all the predictor terms. Equation (1) thus becomes:

$$y_j = \beta_0 + \epsilon_j \tag{2}$$

In the context of machine learning, this is known as a baseline or a dummy model, which predicts the same value for all days and participants: the mean.

2.3.2 *A Two-Level Model.*

To model the differences between participants using a linear regression model, we can include a personalized intercept term. The regression equation can be described in two parts, where the first level is given by[1]:

$$y_{ij} = \beta_{i0} + \epsilon_{ij} \tag{3}$$

Here, we are trying to predict the stress score for each day $j = 1, \ldots, n_d$ *within* each participant $i = 1, \ldots, N$.

We model the intercepts as the sum of the overall intercept, γ_{00} and person-specific intercepts, u_{i0}, also called the random error component. The second level regression equation is given by[2]:

$$\beta_{i0} = \gamma_{00} + u_{i0} \tag{4}$$

2.3.3 *A Three-Level Model.*

Since participants in our study answered the EMA prompts repeatedly throughout the day, we can add a third level of analysis, that is we consider *within-day* variability. In this case, we are trying to predict the score for each EMA session $k = 1, \ldots, n_s$ *within* each day j *within* each participant i. This is a more fine-grained level of analysis and includes many more instances, namely $n = N \times n_d \times n_s$

Joining the expressions for all three levels of intercept, the equation can be written as:

$$\begin{aligned} y_{ijk} &= \beta_{ij0} + \epsilon_{ijk} \\ &= \left(\gamma_{i00} + v_{ij0}\right) + \epsilon_{ijk} \\ &= \left(\left(\delta_{000} + u_{i00}\right) + v_{ij0}\right) + \epsilon_{ijk} \end{aligned} \tag{5}$$

Now, the top level intercept, β_{ij0} is composed of three different components. The first one, δ_{000}, is fixed for all participants and days, and it represents the overall intercept corresponding to the mean of scores aggregated per EMA session. The other two are random effects, where u_{i00} is the person-specific intercept, while v_{ij0} is the intercept specific to each day within each person.

3 RESULTS

3.1 Machine Learning on Daily Aggregated Data

As described in Section 2.2, we followed a typical machine learning approach to detect daily stress. We chose negative affect as an indicator for stress, which was measured with the Positive and Negative Affect Schedule (PANAS, [22]). This is the most commonly used questionnaire in similar diary studies looking at daily measures of stress [13]. It is composed of a list of adjectives describing emotional states, which are self-assessed on a scale from 1 to 5.

This approach did not yield good predictions as shown in Fig. 1. In fact, most of the models performed no better than the dummy model, as evaluated by the median of the R^2 metric across all participants. Even when considering the individual rounds of the leave-one-subject-out validation scheme, the best model (in this case an instance of an XGBoost regressor) achieved a maximum of $R^2 = 0.52$. This corresponds to 52 % of explained variance for that particular participant.

We considered modelling within-day stress as the natural next step. However, this gives the possibility of processing the data on the level of days, rather than only subjects. For example, standardization, feature selection, and model cross validation could

[1] In general, this equation would include predictor terms, such as $\beta_{i1} x_{ij1}$, but they are omitted for clarity as mentioned above.

[2] Similarly, we could write the equation for person specific regression coefficients as $\beta_{1i} = \gamma_{10} + u_{1i}$ and also model person-specific predictors as $\gamma_{01} W_i$.

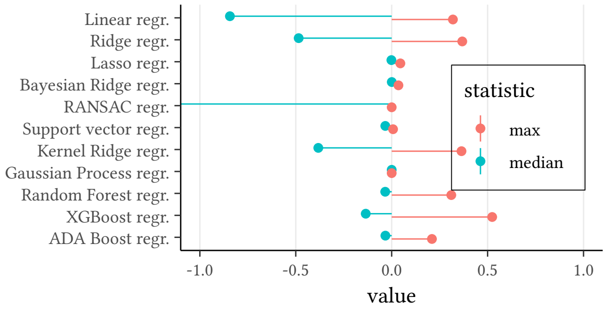


Figure 1: Median and maximum R^2 value as achieved by different regression methods in a leave-one-subject-out validation scheme.

all be done on the lowest, daily level. To get an idea of whether a more fine-grained analysis of the data might be warranted, we turned to variance partitioning.

3.2 Sources of Variability

As mentioned in Section 2.2, the data for machine learning experiments were standardized within participants, i.e., the normalization was personalized. In multilevel modelling terms, this is equivalent of introducing a participant random effect. By defining an intercept-only linear mixed model using the `lme4` library [3], it turned out that the variance explained by these person-specific intercepts was $\sigma_u^2 = 0.20$, which amounted to 57 % of the total variance.

The random effect of participants is illustrated in Figure 2. It shows that the participants differ in how they evaluated their negative affect. Their mean assessments are mostly distributed within 1 point away from the overall mean, but some differed from it by almost 2 points

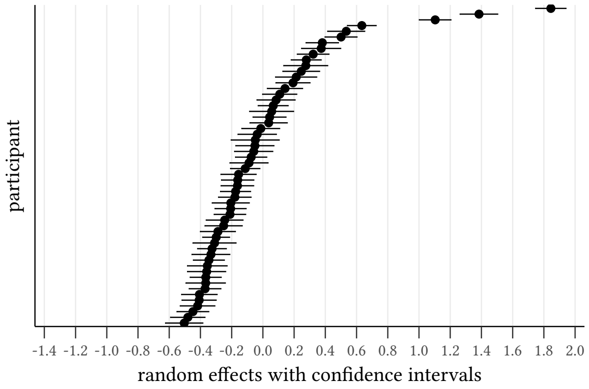


Figure 2: The offset of person random effects (roughly corresponding to person-specific means of daily stress) from the main intercept effect (roughly corresponding to the overall mean).

Next, we considered a three-level model with data aggregated on an EMA session basis. We modelled a random effect by varying the intercept among subjects *and* among days within subjects. The variance that was explained by adding the day level was $\sigma_v^2 = 0.08$ or 11 % of the total variance. This is in addition to the proportion of variance already explained at the subject-level, so the total proportion of explained variance increased to 68 %.

This is also illustrated in Figure 3 which shows that individual days differ from the overall mean by maximum of 1.5 points. On the ordinal axis, the random effects are ordered by participant, similarly to Fig. 2. Within participants, however, the data are ordered consecutively by date. This is manifested in the noisy structure of the confidence intervals as opposed to the monotonously increasing random effects shown in red points.

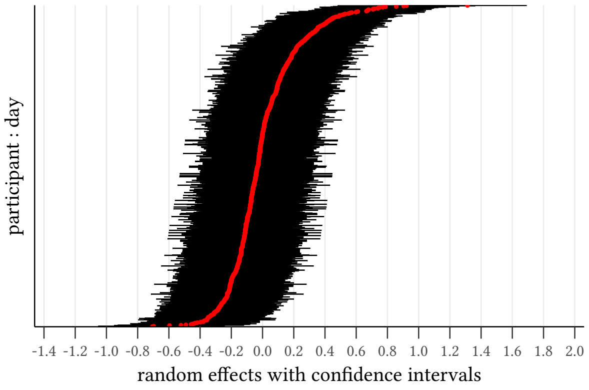


Figure 3: The offset of random effects of interaction terms of person and day (roughly corresponding to person-day-specific means of stress in one EMA session) main intercept effect (roughly corresponding to the overall mean).

By including day-specific intercepts, this model performs significantly better ($\chi^2 = 509$, $p < 0.001$). We next consider what that means in the context of machine learning.

4 DISCUSSION

When considering two sources of variability, the person and the day level, we showed that much of the total variance can be ascribed to within-person differences. This can be interpreted to confirm the merit of personalized normalization of the data, but other interpretations are also possible.

It should be noted that we only dealt with the target variable in this work. Thus, variance partitioning does not help with deciding whether to normalize independent variables. In general, it is advised to normalize physiological data since there exists inherent variability of physiological functioning in the general population [18]. Similarly, explorative analysis indicated that phone sensors vary across devices and it is also feasible to assume that people's phone usage varies significantly (independent of their stress level).

For the target variable itself, the proportion of variance explained with within-person differences can be interpreted in at least two ways. Either the participants were on average exposed to different levels of stress and this is why their assessments differ in a systematic way. Alternatively, participants can have differing thresholds of evaluating something as stressful. Since the self-reports are completely subjective, it is not possible to differentiate between these two interpretations with the self-assessments as labels. It would be possible to explore this further by taking physiological measures as ground truth for stress and use them to explain subjective measures. Treating the physiological measures as universal is problematic, however, and they might not even be related to stress deterministically. Physiological responses are not specific to different stress states, but rather a more complex relationship exists between the stimuli, physiology, and the parameters that control dynamics between them [7].

Finally, normalization is not at all the only option of removing the person-specific variation. Methods such as linear discriminant analysis offer ways that have been shown to perform better [1].

Including person-day random effects in the three-level model, the intercept model performs better than the one with only person random effects included. Following the same reasoning as for the two level model, this could be interpreted that day-specific normalization would be beneficial. There are several arguments against this interpretation, however.

First, as indicated in Section 2.2, participants responded to questionnaires 5 or 6 times a day. Standardizing with this little data is dubious, while using such small samples for feature selection or model validation is unacceptable. Second, the questionnaire data are not truly continuos, but in fact interval data (at best) that can take 5 possible values. Since each EMA session included only two items from each questionnaire, aggregating at this level brings the number of possible values to only 9. Aggregating on a daily level, however, summarises about 10 different measurements, increasing the resolution to 0.1 point. This makes daily means much closer to a continuous variable which can be modelled by regression methods.

We can therefore argue that normalizing data by considering each day as a separate unit is not appropriate. We can conclude, however, that treating each EMA session as its own instance is beneficial. As stated in Section 3.2, analysis on the EMA session level can explain at least 11 % of variance that is not captured by the variability between participants. This conclusion is also illustrated in Figs. 2 and 3: while the general pattern of random effects shown by red points in Fig. 3 can already be sensed in Fig. 2, the noisy structure of confidence intervals is noticeable and worth exploring further.

5 CONCLUSIONS

Multilevel models are a well established method in medical, biological, and social sciences for analysing nested and longitudinal data. In machine learning, research of comparable methods is in its early stages [15]. Some tree-based methods are capable of taking into account hierarchical (or clustered) nature of data, such as MixRF [21], and least squares support vector machines (LS-SVM) have been extended for handling longitudinal data, resulting in a mixed effects LS-SVM [14].

The aim of this paper was not to build multilevel models, statistical or machine learning ones, but rather use variance partitioning to explore how different levels of nested data can be leveraged. We have shown that while standardization or similar techniques do not lend well to the lowest level due to small sample size, restricting analysis to a higher level discards an important part of variance. In this way, variance partitioning can help us build better machine learning models by enabling us to systematically explore different levels of hierarchical data and decide what data transformations to apply to each level.

REFERENCES

[1] Folami Alamudun, Jongyoon Choi, Hira Khan, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2012. Removal of subject-dependent and activity-dependent variation in physiological measures of stress. In *Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE. doi: 10.4108/icst.pervasivehealth.2012.248722.

[2] Ane Alberdi, Asier Aztiria, and Adrian Basarab. 2016. Towards an automatic early stress recognition system for office environments based on multimodal measurements. A review. *Journal of Biomedical Informatics*, 59, (Feb. 2016), 49–75. doi: 10.1016/j.jbi.2015.11.007.

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1. doi: 10.18637/jss.v067.i01.

[4] Larissa Bolliger, Ellen Baele, Elena Colman, Gillian Debra, Junoš Lukan, Mitja Luštrek, Dirk De Bacquer, and Els Clays. 2022. The association between day-to-day stress experiences, recovery, and work engagement among office workers in academia. An ecological momentary assessment study. *PLOS ONE*. Submitted.

[5] Larissa Bolliger, Gillian Debra, Junoš Lukan, Rani Peeters, Mitja Luštrek, Dirk DeBacquer, and Els Clays. 2022. The association between day-to-day stress experiences and work–life interference among office workers in academia. An ecological momentary assessment study. *International Archives of Occupational and Environmental Health*. doi: 10.1007/s00420-022-01915-y. In press.

[6] Larissa Bolliger, Junoš Lukan, Mitja Luštrek, Dirk De Bacquer, and Els Clays. 2020. Protocol of the stress at work (STRAW) project: how to disentangle day-to-day occupational stress among academics based on EMA, physiological data, and smartphone sensor and usage data. *International Journal of Environmental Research and Public Health*, 17, 23, (Nov. 2020), 8835. doi: 10.3390/ijerph17238835.

[7] Justin Brooks, Joshua C. Crone, and Derek P. Spangler. 2021. A physiological and dynamical systems model of stress. *International Journal of Psychophysiology*, 166, (Aug. 2021), 83–91. doi: 10.1016/j.ijpsycho.2021.05.005.

[8] Daniel J. Brotman, Sherita H. Golden, and Ilan S. Wittstein. 2007. The cardiovascular toll of stress. *The Lancet*, 370, 9592, 1089–1100. doi: 10.1016/s0140-6736(07)61305-1.

[9] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. 2015. AWARE: Mobile context instrumentation framework. *Frontiers in ICT*, 2, 6, 1–9. doi: 10.3389/fict.2015.00006.

[10] Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 648. isbn: 9780521686891.

[11] Vito Janko, Matjaž Boštic, Junoš Lukan, and Gašper Slapničar. 2021. Library for feature calculation in the context-recognition domain. In *Proceedings of the 24nd International Multiconference INFORMATION SOCIETY – IS 2021*. Slovenian Conference on Artificial Intelligence (Ljubljana, Slovenia, Oct. 4–8, 2021). Mitja Luštrek, Rok Piltaver, and Matjaž Gams, editors. Vol. A, 23–26. https://library.ijs.si/Stacks/Proceedings/InformationSociety/2021/IS2021_Volume_A.pdf.

[12] Junoš Lukan, Larissa Bolliger, Els Clays, Oscar Mayora, Venet Osmani, and Mitja Luštrek. 2021. Participants' experience and adherence in repeated measurement studies among office-based workers. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Virtual, Sept. 21–24, 2021). ACM. doi: 10.1145/3460418.3479367.

[13] Junoš Lukan, Larissa Bolliger, Nele S. Pauwels, Mitja Luštrek, Deirk De Bacquer, and Els Clays. 2022. Work environment risk factors causing day-to-day stress in occupational settings. A systematic review. *BMC Public Health*, 22, 1. doi: 10.1186/s12889-021-12354-8.

[14] Jan Luts, Geert Molenberghs, Geert Verbeke, Sabine Van Huffel, and Johan A. K. Suykens. 2012. A mixed effects least squares support vector machine model for classification of longitudinal data. *Computational Statistics & Data Analysis*, 56, 3, 611–628. doi: 10.1016/j.csda.2011.09.008.

[15] Daniel Patrick Martin. 2015. *Efficiently Exploring Multilevel Data with Recursive Partitioning*. PhD thesis. University of Virginia, Enfield, Connecticut.

[16] Felix Mölder et al. 2021. Sustainable data analysis with snakemake. Version 2. *F1000Research*, 10, 33. doi: 10.12688/f1000research.29032.2. peer review: 2 approved.

[17] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[18] R. W. Picard, E. Vyzas, and J. Healey. 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 10, 1175–1191. doi: 10.1109/34.954607.

[19] Jianhua Tao and Tieniu Tan. 2005. Affective computing: a review. In *Affective Computing and Intelligent Interaction*. Springer Berlin Heidelberg, 981–995. doi: 10.1007/11573548_125.

[20] Julio Vega, Meng Li, Kwesi Aguillera, Nikunj Goel, Echhit Joshi, Kirtiraj Khandekar, Krina C. Durica, Abhineeth R. Kunta, and Carissa A. Low. 2021. Reproducible analysis pipeline for data streams. Open-source software to process data collected with mobile devices. *Frontiers in Digital Health*, 3. doi: 10.3389/fdgth.2021.769823.

[21] Jiebiao Wang, Eric R. Gamazon, Brandon L. Pierce, Barbara E. Stranger, Hae Kyung Im, Robert D. Gibbons, Nancy J. Cox, Dan L. Nicolae, and Lin S. Chen. 2016. Imputing gene expression in uncollected tissues within and beyond GTEx. *The American Journal of Human Genetics*, 98, 4, 697–708. doi: 10.1016/j.ajhg.2016.02.020.

[22] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect. The PANAS scales. *Journal of Personality and Social Psychology*, 54, 6, 1063–1070. doi: 10.1037/0022-3514.54.6.1063.

Academic Performance Relation with Behavioral Trends and Personal Characteristics: Wearable Device Perspective

Berrenur Saylam, Ekrem Yusuf Ekmekci, Eren Altunoğlu, Ozlem Durmaz Incel
Computer Engineering Department, Boğaziçi University
İstanbul, Turkey
{berrenur.saylam,ozlem.durmaz}@boun.edu.tr

ABSTRACT

Understanding the relevant factors related to students' academic performance can help to construct a more precise methodology for conducting successful academic life. Several studies examine the relationship between students' lives and academic performances using statistical techniques with subjective responses collected via questionnaires in the literature. In the last decade, wearable devices, such as smartwatches and smartphones, have gained popularity in the research community since they can provide objective measurements of the users' activity, sleep, and mood states with integrated sensors. It is possible to extract markers related to individuals' physiological and psychological states. This study explores the most important factors from wearables and questionnaires about students' academic grades using the NetHealth dataset. We utilize machine learning techniques, specifically Random Forest, rather than classical statistical analyzes in literature. We believe that we contribute to interpreting the underlying factors related to grade by examining objectively-measured multi-modal datasets. We also focus on classifying the grades with Random Forest and achieve overall 76% accuracy. The most important factors affecting academic performance are observed to be sleep, big five personalities, health, and mental health.

KEYWORDS

Wearable computing, machine learning, multi-modality, well-being, pervasive computing, student grades, behavioral patterns, personality traits

1 INTRODUCTION

Understanding the underlying factors of academic performance may help students to perform better throughout their academic life. Many studies have investigated these factors affecting academic performance, including family history, psychological well-being, and physical activity [**1**, **2**, **3**, **4**]. Some approached the situation from family history [**1**], and some focused on the existence of physical activity in the curriculum [**2**]. Also, some studies considered sleep based on self-reported measures [**3**]. However, they are based on one modality, focusing on one factor and trying to understand its effect on the target (i.e., students' academic performance). This approach does not provide a meta-understanding between different modalities. Thus, a multi-modal approach is necessary to obtain a more expanded view.

This study focuses on multi-modal data analysis collected from objectively measured wearable devices' sensors and several surveys corresponding to the subject's origin, sex, education level, bad habits, as well as state-of-the-art sleep, big five, mental health inventories (the details are given in Table 1). We aim explore the factors affecting students' academic performances.

We utilize the NetHealth open source data [**5**] which contains students' sleep routines, daily physical activities, communication behaviors collected with mobile phones, and a detailed survey about family history, living conditions, and personality. Data related to sleep and activity is collected from wearable devices and documented. We aim to find the relation between some of the abovementioned aspects and academic performance.

We have a large dataset from different academic periods (waves) and various survey data. However, the surveys were not filled in every period, hence, we focused on one period with the least amount of missing information. Before applying our models, we performed a preprocessing procedure by imputing the data with proper techniques to handle missing values and preparing them for the final analysis. We utilized machine learning techniques, specifically Random Forest (RF) algorithm, both for factor selection and classification. In addition, we provide essential parameters for the student's academic performance. These are related to sleep, big five personalities, health, mental health, personal information, and origin data in order. We believe that these information can be helpful in understanding affecting factors for further improvement of student life to get better performance during their academic life.

One of the essential contribution of our work is bringing different factors together and trying to produce a combination of them. In that way, we aim to find the most important predictors for students' academic performance by combining other focus areas, such as sleep, mental health, and activities, in the scope of one study.

Considering the studies utilizing NetHealth data, some are analyzing the data on different topics such as biometric-based authentication [**6**], physical activity and sleep pattern [**7**]. There are studies doing network analysis [**8**, **9**], physical activity prediction [**10**]. To the best of our knowledge, no similar study exists among the listed papers.

The rest of the paper is organized as follows: In Section 2, we explain state of the art on student grades studies and from point of wearable domain. In Section 3, we explain dataset details and the preprocessing steps for further analyses. In Section 4, we present academic grade's classification results with different balancing strategies. We give factors for best case. Finally, in Section 5, we discuss our findings with other future study ideas.

2 RELATED WORKS

Many related works exist about student's academical performance from the point of different domains such as educational, psychological and smartphone sensing [**11**, **12**, **13**, **14**].

Objectively measured signals sensed from wearables applied into the research field related to student's mental health and

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

academic performance, to the best of our knowledge, starts with StudentLife [**11**] project.

In [**12**], authors collect the day-to-day and week-by-week impact of workload on stress, sleep, activity, mood, sociability, mental well-being and academic performance via smartphone sensors. They examined strong correlations between smartphone sensors and student's mental health along with their academical scores by not counting behavioral differences.

In [**13**], authors extracted related factors to the students academical grades from academic related behaviors, personality, affect, stress, lifestyle and sensed behaviors with wearables. They modelled behavior change points to capture individual's behaviors while having the same final grade. One of the findings is study duration has positive correlation with the final grade.

In [**14**], researchers examined the relation between wearable device sensors and survey with student's grade in a similar manner. They used SVM with different kernel setups. They found social features such as negative email contacts and negative interactions are lower on students with high GPA. Also, accelerometer sensor in wearables have an impact on discriminating the higher and lower performants. This study is similar to our experiment, where there is multi-modal data from wearable sensors and surveys. We also examine the related factors on different datasets, but our study also explores class balancing scenarios.

3 METHODOLOGY

3.1 Dataset

We utilized the NetHealth dataset[1]. It is collected from undergraduate students from Notre Dame (ND) University between Fall 2015 and Spring 2019. Thus, there are 8 waves corresponding to each semester. There are approximately 700 students' data from the 2015 – 2017 period and 300 from the 2017 – 2019 period caused by the drops in participation. Data collection consists of the social network, physical activity, sleep data from Fitbit wearable device, and ground truth data from questionnaires about physical and mental health, social-psychological states, tastes, and various self-reported behaviors, demographics, and background traits. The collection procedure is approved through IRB protocols, and each participant has consented. Nevertheless, not all data collection is publicly shared due to privacy concerns.

The details of the collected dataset per modality are as follows. We performed our study with boldly-marked sub-datasets.

- Communication data: Collection of smartphone-based communication logs data.
- **Wearable data**: Collected measurements regarding activity and sleep such as the number of steps, active minutes, heart rate, sleep duration, sleep time, and awaken time using Fitbit.
- **Courses and grades data**: Administrative records from ND Registrar's Office containing course and grade information.
- Calendar: Weekly calendar showing the days about the beginning of classes, break weeks, holidays, etc.
- **Survey data**: Self-reported questionnaires related to physical and mental health, social-psychological states, tastes, and various self-reported behaviors, and demographics and background traits.
- Network survey data: Interactions' network data with the related information such as relationship type, duration, frequency of interaction, similarity, etc.

3.2 Preprocessing

As stated in Section 3.1, there are 8 waves. Each wave has different survey questions and thus responses. For instance, in waves 1, 2, 3, 7, there are no questions related to stress, while in 4, 5, 6, 8, there are. Similarly, sleep ground truth is not collected during the study waves 5, 7. Thus, we chose to work on wave 1 as it contains relatively higher responses than other waves.

Firstly, we constructed a sub-dataset from NetHealth concentrating on our purpose. The details are explained in Section 3.2.1. Then, we prepossessed our data by deleting highly correlated ones (in Section 3.2.3). Finally, we applied the Random Forest algorithm for the rest of the study.

3.2.1 Dataset Preparation. As the dataset includes many different data types, each of them has various parameters, we decided which parameters to use before starting our study. We considered all parameters from wearable devices and course-grades datasets. However, we selected some of the collected data from the survey dataset. Surveys constitute, mainly, *bad habits, big-five personality inventory, education, exercise, health, mental health, origin, personal information, sex, and sleep* related answers. We used only the summarizing parameters provided by the survey for mental health, personal information, and sleep. We select some parameters from the origin category manually. We used the parameters of parents' status, economic condition, number of siblings, and religion. Table 1 gives the final list of utilized parameters. At the end of the naming, some parameters have _1 indications, which relate to the measuring from wave1.

3.2.2 Handling Missing Values. Once the dataset was prepared for analysis, we noticed missing values over columns. We preferred to keep these columns and impute them since they are partially missed. We applied the most frequent imputation technique to the categorical ones and the mean imputation technique to the numerical ones. However, there is enough correlation for activity-related wearable data to use the KNN imputation technique. Thus, we used this technique. Finally, sleep data from wearables did not contain any missing values.

3.2.3 Correlation. We checked the correlation between parameters to reduce dimensionality. We deleted the ones which exhibit higher than %80 correlations. These are *cardiomins, fatburnmins, lowrangemins, minsasleep, minsawake, peakmins* parameters. We can deduct the information related to them from other parameters, for instance, cardiocals for cardiomins and fatburnmins. We decided on the threshold value after many experiments. When we increase it, we keep the highly correlated ones, and when we decrease the threshold, more parameters will be deleted, which causes unnecessary parameter loss. Eliminating them prevents misleading results due to highly correlated features in detecting interactions between different features. We had 93 parameters. After removal of the 6 highly correlated ones, we have 87 features.

3.2.4 Target value's distribution. In this study, we are working towards the identification of important parameters and the application of machine learning methods regarding students' grades. Thus, before starting the analysis, we examined target values, i.e., student grades distribution, to observe whether there is class imbalance. The distribution is in Figure 1. Here, it is seen that we

[1] http://sites.nd.edu/nethealth/

Table 1: Details of the features

Dataset	Measured Values
Wearable data (Activity)	complypercent (percent minutes using Fitbit), meanrate (mean heart rate), sdrate (st. dev. heart rate), steps, floors, sedentaryminutes, lightlyactiveminutes, fairlyactiveminutes, veryactiveminutes, lowrangemins (low range minutes), fatburnmins, cardiomins, peakmins, lowrangecal, fatburncal, cardiocal, peakcal
Wearable data (Sleep)	timetobed (time went to bed), timeoutofbed (time out of bed), bedtimedur (minutues in bed in minutes), minstofallasleep (minutes to fall asleep), minsafterwakeup (minutes in bed after waking), minsasleep (minutes asleep), minsawake (minutes awake during sleep period), Efficiency (minsasleep/(minsasleep + minsawake)
Courses and grades	AcademicPeriod, CourseReferenceNumber, FinalGrade
Survey data (Bad habits)	usetobacco_1 (used tobacco), usebeer_1 (drank beer), usewine_1 (drank wine or liquor), usedrugs_1 (used rec drugs like marij. or cocaine), usedrugs_prescr_1 (used presc. drugs not prescribed), usecaffine_1 (drank caffenated drinks)
Survey data (BigFive/Personal inventory)	Extraversion_1, Agreeableness_1, Conscientiousness_1, Neuroticism_1, Openness_1
Survey data (Education)	hs_1 (high school type), hssex_1 (high school sex composition), hsgrade_1 (high school average grade), apexams_1 (# of hs ap exams), degreeintent_1 (highest intended degree), hrswork_1 (paid hours senior year), ndfirst_1 (Notre Dame first choice of applied colleges?)
Survey data (Exercise)	hsclubrc_1 (club activities), exercise_1 (excersise), clubsports_1 (play club, intramural or rec sports) , varsitysports_1 (play varsity sports), swimming_1 (swim), Dieting_1 (special type of diet), PhysicalDisability_1 (physical disability)
Survey data (Health)	SelfEsteem_1 (on the whole, I am satisfied with myself), Trust_1 (most people can be trusted), SRQE_Ext_1 (external self-regulation (exercise)), SRQE_Introj_1 (introjective self-regulation (exercise)), SRQE_Ident_1(identified self-regulation (exercise), SelfEff_exercise_scale_1 (when i am feeling tired), SelfEff_diet_scale_1 (self_efficacy score (diet items)), selfreg_scale_1 (i have trouble making plans to help me reach my goals)
Survey data (Mental health)	STAITraitTotal_1 (state_trait anxiety score), CESDOverall_1 (CES depression score), BAIsum_1 (beck anxiety score), STAITraitGroup_1 (state_trait anxiety 2 category), CESDGroup_1 (CES depression - 2 categories), BAIgroup_1 (beck anxiety (3 category)), majorevent_1 (life changes)
Survey data (Origin)	momdec_1 (is your mother deceased?), momusa_1 (was mother born outside usa?), daddec_1 (is your dad deceased?), dadusa_1 (was your dad born outside usa?), parentstatus_1 (parents living together or divorced/living apart), dadage_1 (father's age), momage_1 (mom's age), numsib_1 (number of siblings), birthorder_1 (which # in birth order are you?), parentincome_1 (parent's total income last year), parenteduc_1 (combined parent education), momrace_1 (mother's race), dadrace_1 (father's race), momrelig_1 (mother's religious preference), dadrelig_1 (father's religious preference), yourelig_1 (your religious prefence)
Survey data (Personal info)	selsa_rom_1 (romantic loneliness), selsa_fam_1 (family loneliness), selsa_soc_1 (social loneliness)
Survey data (Sex)	gender_1 (gender)
Survey data (Sleep)	PSQI_duration_1 (computed time in bed), PSQIGlobal_1 (PSQI total score), PSQIGroup_1 (PSQI two categories), MEQTotal_1 (MEQ (chronotype) score - high score morning person), MEQGroup_1 (MEQ (chronotype) groups - 5 categories))

have *A* grade on the majority, and we have very few instances from the *B-, C+, C, C-* classes. More specifically, we have 41856, 19321, 10048, 7265, 2526, 1617, 1258, 354, 4346 from classes *A, A-, B+, B, B-, C+, C, C-, S (satisfactory)*, respectively. To well classify minority classes, we applied the SMOTE (synthetic minority over-sampling) technique to produce synthetic data by keeping the same class distribution [**15**]. After SMOTE, we got 41856 instances from each class.

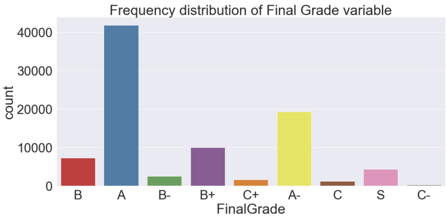


Figure 1: Target value distribution: Grade

3.3 Model details and performance metrics

As a classification method, we used RF algorithm because it is an ensemble method and performs better than the other used methods in literature in this domain [**16**]. The used parameters for RF are n estimators 1000, criterion Gini, and max features sqrt from scikit-learn toolkit[2]. %75 and %25 train and test sizes are chosen, respectively.

4 CLASSIFICATION PERFORMANCE EVALUATION

Since our target variable is already categorical, we used the dataset after preprocessing without any other change in the classification task. In Table 2, we present f1-score details of each class performance and the global average of the f1-scores with the accuracy metric. We obtained %76 average accuracy. We see that the best performances are achieved for the classes *B-, C+, C-, S*. Before SMOTE application, it was %65 average accuracy; furthermore, we had lower f1 scores for these indicated classes, but we did not present the details due to the page limit. The confused instances may be observed in Figure 2. For instance, *A* class is confused mostly with *A+* with an important ratio. It is expected since these are very close classes. The class *S* is mostly confused with others. It can be interpreted as expected since a satisfactory result corresponds to passing the course. SMOTE generates instances based on a similarity measurement rather than replicating existing ones. Thus, the bias is relatively lower compared to simple replications of instances since these are newly generated ones. Nevertheless, we also applied the under-sampling strategy and down-sampled higher class instances to be equal to the class with fewer instances. Thus, we obtained 354 instances for each class. When we applied RF to that data, we obtained even worse performance, which is 47% average accuracy. It is expected since we deleted most data points, so learning with few instances led to lower results.

In addition, in Figure 3, we provide the most critical factors to obtain this classification performance by calculating the most important 20 parameters via RF feature selection. The order is following: *MEQTotal (sleep), Trust (health), Extraversion (big five), selsa_soc (personal info), selsa_rom (personal info), Openness (big five), Neuroticism (big five), SRQE_Ext (health), dadage (origin), PSQI_duration (sleep), PSQIGlobal (sleep), BAISum (mental health), hsgrade (education), SRQE_Introj (health), CESDOverall (mental health), SelfEff_exercise_scale (health), Agreeableness (big five), momage (origin), MEQGroup (sleep).* The explanation of these parameters is presented in Table 1. We can interpret this result as the most important factors arrive from survey datasets. The important sub-surveys are sleep, big five, health, mental health, personal information, and origin.

Table 2: Classification performance details

	precision	recall	f1-score	support
A	0.53	0.56	0.55	10434
A-	0.52	0.44	0.47	10410
B	0.66	0.71	0.68	10449
B+	0.67	0.60	0.63	10507
B-	0.85	0.87	0.86	10507
C	0.89	0.92	0.91	10494
C+	0.88	0.89	0.88	10418
C-	0.98	0.98	0.98	10502
S	0.78	0.83	0.80	10455
accuracy			0.76	94176
macro avg	0.75	0.75	0.75	94176
weighted avg	0.75	0.76	0.75	94176

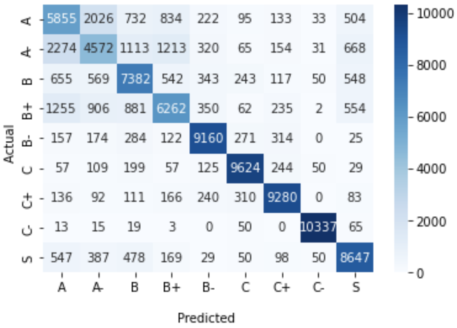


Figure 2: Confusion Matrix

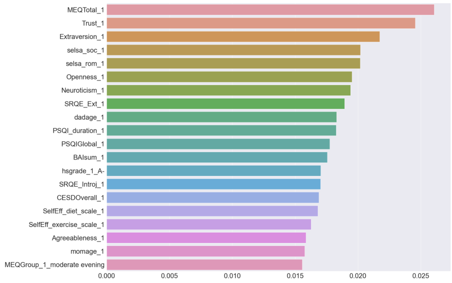


Figure 3: Feature Importance for Classification

5 DISCUSSION AND CONCLUSION

In this study, we applied a machine learning technique, RF, to see how accurately we can classify and predict students' grades using surveys and wearable data. In addition, we extract the most important factors affecting the model's performance. Results indicate *sleep, big five, health, mental health, personal information, and origin* survey parameters have higher effects on performance. We differ from state-of-the-art [**12**, **13**, **14**] by applying SMOTE.

For further research, one may examine other waves since there are 8 to obtain more instances from each class. Also, since the dataset is collected from one of the top University students, it is expected to have higher grades, i.e., A, A+. Thus, applying a similar experimental data collection setup to students with lower performances in the courses may be helpful.

ACKNOWLEDGEMENTS

Tübitak Bideb 2211-A academic reward is gratefully acknowledged. This work is supported by The Turkish Ministry of Development under the TAM Project number DPT2007K120610.

[2] https://scikit-learn.org/stable/

REFERENCES

[1] Misty, Lacour, and D. Tissington Laura. "The effects of poverty on academic achievement." Educational Research and Reviews 6.7 (2011): 522-527.

[2] Shephard, Roy J. "Curricular physical activity and academic performance." Pediatric exercise science 9.2 (1997): 113-126.

[3] Purta, Rachael, et al. "Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits." Proceedings of the 2016 ACM international symposium on wearable computers. 2016.

[4] GOMES, Maria V., Luciano Francisco Sousa ALVES, and Louelson AL COSTA. "de Azevedo." Dinâmica socioespacial urbana de Cuité–PB resultante da implantação do campus de saúde e educação da UFCG. 152f. João Pessoa (2014).

[5] Purta, Rachael, et al. "Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits." Proceedings of the 2016 ACM international symposium on wearable computers. 2016.

[6] Vhaduri, Sudip, and Christian Poellabauer. "Multi-modal biometric-based implicit authentication of wearable device users." IEEE Transactions on Information Forensics and Security 14.12 (2019): 3116-3125.

[7] Purta, Rachael, et al. "Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits." Proceedings of the 2016 ACM international symposium on wearable computers. 2016.

[8] Fridmanski, Ethan, et al. "Clustering in a newly forming social network by subjective perceptions of loneliness." Journal of American College Health (2020): 1-6.

[9] Liu, Shikang, et al. "Network analysis of the NetHealth data: exploring co-evolution of individuals' social network positions and physical activities." Applied network science 3.1 (2018): 1-26.

[10] Faust, Louis, et al. "Physical activity trend extraction: a framework for extracting moderate-vigorous physical activity trends from wearable fitness tracker data." JMIR mHealth and uHealth 7.3 (2019): e11075.

[11] StudentLife Dataset 2014. http://studentlife.cs.dartmouth.edu/.

[12] Wang, Rui, et al. "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones." Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing. 2014.

[13] Wang, Rui, et al. "SmartGPA: how smartphones can assess and predict academic performance of college students." Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing. 2015.

[14] Sano, Akane, et al. "Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones." 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, 2015.

[15] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[16] Can, Yekta Said, et al. "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey." Journal of biomedical informatics 92 (2019): 103139.

Detection of postpartum anemia using machine learning

David Susič
david.susic@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Lea Bombač Tavčar
bombac.lea@gmail.com
University Medical Centre Ljubljana, Division of Gynaecology and Obstetrics
Šlajmerjeva 3
Ljubljana, Slovenia

Hana Hrobat
hana.hrobat@icloud.com
University of Ljubljana, Faculty of Medicine
Vrazov trg 2
Ljubljana, Slovenia

Lea Gornik
lea.gornik@gmail.com
University of Ljubljana, Faculty of Medicine
Vrazov trg 2
Ljubljana, Slovenia

Miha Lučovnik
miha.lucovnik@kclj.si
University Medical Centre Ljubljana, Division of Gynaecology and Obstetrics
Šlajmerjeva 3
Ljubljana, Slovenia

Anton Gradišek
anton.gradisek@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

Postpartum anemia is seen as a health problem and should be treated. We evaluate performance of nine machine learning regression models in predicting the postpartum anemia six weeks after childbirth. We focus on tree key parameters: ferritin, haemoglobin, and transferrin saturation. Our models are compared with the baseline model, which always predicts the mean value of the training data. We found that the models for ferritin and transferrin saturation have good predictive performances, whereas this was not the case for haemoglobin prediction, as all of the implemented models were outperformed by the baseline model.

KEYWORDS

postpartum anemia, haemoglobin level, machine learning

1 INTRODUCTION

Postpartum anemia is a common maternal health problem globally and constitutes a significant health problem in women after birth, even in the developed world. Women may develop it either because of antepartum depletion of iron stores or peripartum excessive blood loss [1]. It is associated with several negative consequences, such as maternal fatigue [2, 3]. With the unacceptably high prevalence of anaemia in women after childbirth in both, up to 50% in developed and up to 80% in developing countries [4], it appears to be of great importance to treat iron deficiency effectively. Ferrum sulphate perorally is the most commonly used iron for pospartum anemia because of its low cost and simple use. Definition of postpartum anaemia rely on haemoglobin values alone, defined as Hb level <100 g/L. Postpartum haemorrhage defined as a blood loss of 500 ml or more within 24 hours after birth is one of the most frequent complications of delivery. This makes women vulnerable and frequently results in postpartum anemia. Consequently, this increases the risk for a peripartum blood transfusion, a treatment with potential severe adverse outcomes [5]. With the unacceptably high prevalence of anaemia in women after childbirth in both, up to 50% in developed and up to 80% in developing countries [4], it appears to be of great importance to treat iron deficiency effectively. In addition to the increased transfusion risk, peripartum iron deficiency anaemia can affect the wellbeing of both the mother and child. It causes cardiovascular symptoms like palpitations and dizziness, breathlessness. It increases a risk of infections as well as excessive postpartum bleeding. Furthermore, postpartum anemia adversely affects maternal mood, cognition, and behavior resulting in increased fatigue, reduced physical and mental performance [6]. This is associated with several negative consequences, such as impaired health-related quality of life [3]. Impaired health-related quality of life linked to postpartum anemia include depression, fatigue, and reduced cognitive abilities. All of these symptoms significantly interferes with mother-child interactions and impact a woman's ability to breastfeed [1].

Postpartum anemia should be treated by restoring iron stores. Although there is a number of treatment options for women with postpartum anaemia, the debate about iron supplementation and the ideal form of administration is ongoing and is not universal in all countries. Currently, common treatment includes iron supplementation administered orally or intravenouslly (IV). The traditional treatment for mild to moderate iron deficiency anaemia is oral supplementation of iron with iron sulfate perorally because of its low cost and simple use. There are advantages and disadvantages of either of the two approaches, which we will not go into detail here. Since the postpartum anaemia contributes to a major healthcare problem even in developed countries, it is important to treat it efficiently [7]. However, IV iron may be preferred because the non-compliance and absorption challenges of oral iron, but it includes increased drug costs and the need for supervised treatment in healtcare institutions. Recent robust studies have compared different iron preparations and there has been a network meta-analysis of different iron medications. However, no randomized clinical trial has directly compared intravenous derisomaltosie, intravenous carboxymaltose and peroral ferrous sulphate for treatment of postpartum anemia, including fatigue measurements.

In this paper, we address the question on predicting the postpartum anemia six weeks after childbirth. We look at three key parameters from blood tests that are related to anemia, namely

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia

© 2022 Copyright held by the owner/author(s).

Table 1: Dataset features.

Personal	Blood test
Age [years]	Haemoglobin [g/L]
Gestational age [weeks]	Serum iron [μmol/L]
Number of children born	TIBC [μgmol/L]
Number of total pregnancies	Transferrin saturation [%]
Number of total childbirths	Ferritin [μg/L]
Number of total abortions	Phosphate [mg/dL]
Type of childbirth	CRP [mg/L]
Transfusion	
Marital status	
Education	
BMI before childbirth	
BMI after childbirth	
Medication	

the ferritin, haemoglobin, and transferrin saturation. Using a database containing 296 patients that were diagnosed with anemia, we investigate the possibilities to predict these relevant blood test values using machine-learning models. We present the results of our initial studies.

2 DATA

The initial dataset included 296 patients that were diagnosed with anemia and 27 features that had some missing values. As this was our initial study, we did not perform any missing data inputation, but rather dropped the patients that had missing values in any of the columns. We were left with 224 patients that had data for all 27 features. Based on the medications that the patients were given during their treatment, they can be separated in three groups: 80 of the patients were treated with Iroprem, 75 were treated with Monofer, and 69 were treated with Tardyfer. Both Monofer and Iroprem are IV medications with iron, while Tardyfer is administered orally as tablets.

The data included personal data and blood test results. Blood tests were performed both right after the childbirth as well as six weeks after. The list of personal and blood test features is given in Table 1.

In the dataset, there are 13 personal features and $2 \cdot 7$ blood test features. Among personal features, gestational age corresponds the number of weeks since the last period. The type of childbirth is a categorical variable and can either be vaginal delivery, planned Cesarean section, or elective Cesarean section. Transfusion is a binary variable indicating whether a patient needed a blood transfusion after the childbirth or not. Marital status is a categorical variable and can either be lives alone, married, or non-marital partnership. Education is ordinal variable of 10 different values with the lowest representing elementary school education and the highest representing a doctoral degree. Lastly, BMI stands for body mass index.

In the blood test features, serum iron describes the amount of iron in the blood. TIBC stands for total iron binding capacity, which is a good indicator of the amount of iron in blood. If the iron level in blood is low, the TIBC is higher as the free capacity for binding of the iron is higher. Transferrin saturation is the value of serum iron divided by the TIBC of the available transferrin. The higher the transferrin saturation, the bigger the iron stores in the body. Lastly, CRP stands for C-reactive protein, which is high is there is inflammation in the body. Inflammation can also be cause as a consequence of an injury during childbirth or Cesarean section. Typically, CRP levels are increased after childbirth. If the high level of CRP (>8 mg/L) still persists after six weeks after the childbirth, this indicates inflammation.

3 METHODOLOGY

The aim of this initial study was to evaluate the performance of several machine learning (ML) models in predicting the values of haemoglobin, ferritin, and transferrin saturation levels in blood of the anemia patients six weeks after childbirth, as these parameters are related to anemia. The input of the models were personal features and the features of the blood test immediately after the childbirth. In each experiment, only one of the three quantities was the output. Thus, we ran three experiments with the same input and different outputs. Additionally, we ran additional separate experiments for each of the three medication groups. We compared our results with the baseline, which always predicted the mean output value of the training data.

4 RESULTS

Our dataset included 224 patients with 20 predictor features. We used mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) as the evaluation metrics, with MAE as the main metric of performance evaluation. Formulas for calculation of MAE, RMSE , and MAPE are given in equations (1), (2), and (3). Parameter y_i denotes predicted values, x_i denotes true values, and n denotes the total number of data points.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}} \tag{2}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{x_i - y_i}{x_i} \right| \tag{3}$$

We implemented nine ML regression models. Regression models predict a continuous variable(s). Linear regression (LR), Kernel Ridge (KR), and elastic net regression (EN) find linear correlations between the predictor features and the output. Bayesian ridge regression (BR) formulates linear regression using probability distributions rather than point estimates. Support vector regression (SVR) finds a hyper-plane in the feature space that has maximum number of data points. Gradient boosting regressor (GB), Light gradient boosting machine (LGBM), extreme grading boosting regressor (XGB), and CatBoost regressor (CB) are ensemble methods that combine the predictions of multiple decision tree regressors. A decision tree regressor uses a tree diagram for decision making, where each branch is partitioned based on a threshold for a predictor feature.

The models trained on the whole dataset were compared in a 10-fold cross validation with the folds stratified with respect to the medication. The models trained for separate medication only were compared in a 5-fold cross validation due to the smaller dataset size. For each of the output variables, we also show a histogram of values distribution along with the mean and standard deviation (SD).

The models' training and performance evaluation was done using Python 3.7 and libraries Numpy 1.18.5 [8], Scikit 0.24.2 [9], LightGBM 3.2.1 [10], XGBoost 1.4.2 [11], and CatBoost 0.26 [12].

4.1 Ferritin

Distribution of ferritin blood levels six weeks after childbirth is given in Figure 1. We see that the patients that were given medication Tardyfer had significantly lower levels than those that were given medications Iroprem or Monofer. The mean and SD values of the distribution are 185.88 μg/L and 141.31 μg/L, respectively. Results of the regression models are given in Table 2.

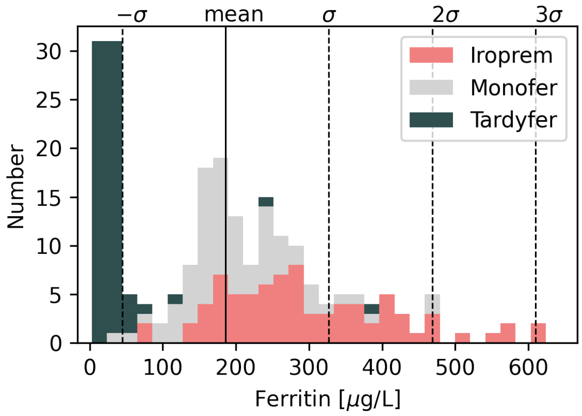


Figure 1: Distribution of ferritin blood levels in patients six weeks after childbirth.

Table 2: Results for the prediction of ferritin.

Model	MAE	RMSE	MAPE $[10^{-2}]$
CB	**61.96**	**87.44**	80.11
XGB	62.76	93.97	**61.23**
LGBM	63.07	88.31	65.88
GB	64.14	91.32	83.86
LR	68.42	89.45	86.26
KR	69.3	90.62	80.2
EN	79.64	99.56	158.81
BR	80.43	101.93	135.88
Baseline	111.81	138.88	272.51
SVR	112.91	140.25	260.76

We see that the best performing model according to both metrics was the CB. Except for the SVR, other models have had similar performances to that of CB. Additionally, we see that most of the models significantly outperform the baseline.

The results of the models performance of predictions for separate medications only are shown in Table 3. The models within each medication have similar performances. In the case of Monofer, all of the models' performances are worse than that of the baseline.

4.2 Haemoglobin

Distribution of haemoglobin blood levels six weeks after childbirth is given in Figure 2. We see that the distributions are very similar between all three medication groups. The mean and SD values of the distribution are 133.87 g/L and 8.10 g/L, respectively. Results are given in Tables 4 and 5.

Table 3: Results for the prediction of ferritin for each medication separately.

Model	Iroprem MAE	Monofer MAE	Tardyfer MAE
LR	93.65	70.85	41.33
LGBM	86.19	57.03	21.74
XGB	95.48	62.44	19.43
CB	**81.26**	58.78	20.48
KR	98.90	69.93	33.11
EN	92.76	63.77	31.81
BR	96.24	61.47	21.99
GB	88.37	70.00	25.69
SVR	97.41	58.20	**19.27**
Baseline	94.61	**55.87**	23.42

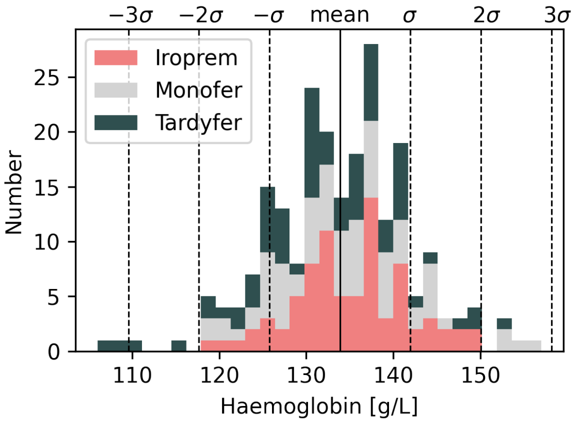


Figure 2: Distribution of haemoglobin blood levels in patients six weeks after childbirth.

Table 4: Results for the prediction of haemoglobin.

Model	MAE	RMSE	MAPE $[10^{-2}]$
Baseline	**6.11**	**8**	**4.62**
BR	6.31	8	4.77
SVR	6.33	8.01	4.80
EN	6.56	8.22	4.96
LR	6.67	8.41	5.03
CB	6.74	8.34	5.10
LGBM	7.16	8.93	5.41
XGB	7.2	9.19	5.44
GB	7.28	9.03	5.52
KR	7.43	9.45	5.59

We see that the models do not perform well in predicting haemoglobin, as they perform worse than the baseline for both the general case and the separate medication cases.

4.3 Transferrin saturation

Distribution of transferrin saturation in blood six weeks after childbirth is given in Figure 3. We see that the distributions are very similar between all three medication groups. The mean and SD values of the distribution are 33.56 % and 11.53 %, respectively. Results of the regression models are given in Table 6.

Table 5: Results for the prediction of haemoglobin for each medication separately.

Model	Iroprem MAE	Monofer MAE	Tardyfer MAE
LR	6.99	7.04	8.17
LGBM	5.46	6.41	8.21
XGB	6.38	6.58	9.20
CB	5.75	6.58	8.03
KR	7.65	7.13	8.63
EN	5.69	6.43	7.36
BR	5.31	6.45	7.33
GB	5.79	7.23	9.41
SVR	5.42	6.62	7.28
Baseline	**5.17**	**5.85**	**7.22**

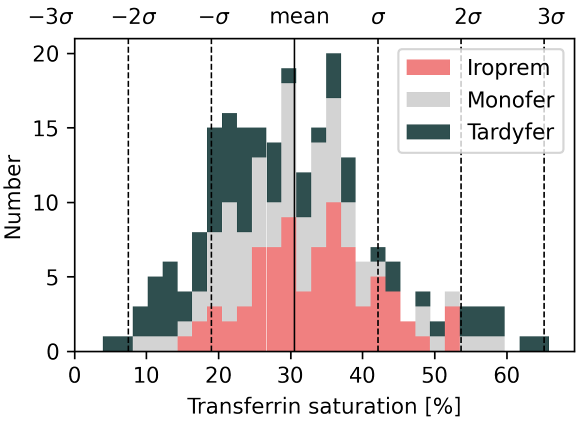


Figure 3: Distribution of transferrin saturation in blood of patients six weeks after childbirth.

Table 6: Results for the prediction of transferrin saturation.

Model	MAE	RMSE	MAPE $[10^{-2}]$
KR	**8.74**	**10.93**	**36.74**
LR	8.78	10.97	36.80
EN	8.82	11.14	38.45
Baseline	8.88	11.12	39.16
SVR	9.11	11.38	39.51
CB	9.11	11.31	38.81
BR	9.22	11.41	40.49
GB	9.51	11.89	40.20
LGBM	9.55	12.10	39.58
XGB	9.62	12.11	39.64

We see that the top three performing models outperform the baseline, with the best model being the KR. The results of the models performance of predictions for separate medications only are shown in Table 7. Unlike Monofer and Tardyfer, the models do not perform well in the case of Iroprem.

5 DISCUSSION AND CONCLUSION

We evaluated nine classic machine learning regression models for the prediction of three key parameters associated with anaemia collected from blood tests six weeks after childbirth: ferritin, haemoglobin, and transferrin saturation. We compared the results with the baseline model, which always predicted the output mean of the training data. We found that the models for ferritin and transferrin saturation had good predictive performance, whereas this was not the case for haemoglobin prediction, as all models were outperformed by the baseline model.

Table 7: Results for the prediction of transferrin saturation for each medication separately.

Model	Iroprem MAE	Monofer MAE	Tardyfer MAE
LR	7.68	9.4	11.84
LGBM	7.16	**7.59**	11.39
XGB	8.36	9.12	12.86
CB	7.2	7.87	11.44
KR	7.73	8.94	12.01
EN	7.16	8.54	**11.24**
BR	6.8	7.83	12.02
GB	7.88	8.78	11.61
SVR	6.94	7.62	11.62
Baseline	**6.49**	7.75	11.82

ACKNOWLEDGMENTS

The authors acknowledge the funding from the Slovenian Research Agency (ARRS), Grant (PR-10495) and Basic core funding P2-0209. The dataset was collected as a part of the study Clinicaltrials.gov registration number NCT03957057.

REFERENCES

[1] Nils Milman. 2011. Postpartum anemia i: definition, prevalence, causes, and consequences. *Annals of hematology*, 90, 11, 1247–1253.
[2] Kathryn A. Lee and Mary Ellen Zaffke. 1999. Longitudinal changes in fatigue and energy during pregnancy and the postpartum period. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, 28, 2, 183–191.
[3] Kiyoshi Ando et al. 2006. Health-related quality of life among japanese women with iron-deficiency anemia. *Quality of life research*, 15, 10, 1559–1563.
[4] Nils Milman. 2012. Postpartum anemia ii: prevention and treatment. *Annals of hematology*, 91, 2, 143–154.
[5] Andreas Greinacher, Konstanze Fendrich, Ralf Brzenska, Volker Kiefel, and Wolfgang Hoffmann. 2011. Implications of demographics on future blood supply: a population-based cross-sectional study. *Transfusion*, 51, 4, 702–709.
[6] Christian Breymann. 2005. Iron deficiency and anaemia in pregnancy: modern aspects of diagnosis and therapy. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 123, S3–S13.
[7] Lisa M. Bodnar, Anna Maria Siega-Riz, William C Miller, Mary E Cogswell, and Thad McDonald. 2002. Who should be screened for postpartum anemia? an evaluation of current recommendations. *American journal of epidemiology*, 156, 10, 903–912.
[8] Charles R. Harris, Jarrod K. Millman, Stefan J. van der Walt, Ralf Gommers, Pauli Virtanen, and David Caurnapeau. 2020. Array programming with numpy. *Nature*, 585, 357–362. doi: https://doi.org/10.1038/s41586-020-2649-2.
[9] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS'17). Curran Associates Inc., Long Beach, California, USA, 3149–3157. isbn: 9781510860964.
[11] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). ACM, San Francisco, California, USA, 785–794. isbn: 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.
[12] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (NIPS'18). Curran Associates Inc., Montréal, Canada, 6639–6649.

Covid symptoms home questionnaire classification and outcome verification by patients

Goran Jakimovski
Faculty of Electrical Engineering and Information Technology
University of "Ss. Cyril and Methodus"
Skopje, Macedonia
goranj@feit.ukim.edu.mk

Dragana Nikolova
Faculty of Electrical Engineering and Information Technology
University of "Ss. Cyril and Methodus"
Skopje, Macedonia
nikolova.dragana98@gmail.com

ABSTRACT

Testing for Covid, in a time of pandemic, can put a lot of overhead on the medical and testing facilities. Moreso, in a pandemic crisis, people become more hypochondriacs and get tested even if a slightest symptom of Covid is detected. This leads to many people, infected and not infected, to gather at the medical facilities, thus increasing the possibility for not infected people to get Covid infection. Our application registers patients and, by using a medical survey, determines if the patient is supposed to get tested for Covid or even more severe measure are to be taken. Additionally, our application uses medical tests results from patients to determine the success rate of the prediction. The case study has shown that the application has 89% success rate of classification. Using this application, only people with the right symptoms will be advised to get tested, thus lowering the overload placed on the medical facilities and minimizing the virus spread.

KEYWORDS

analysis, classification, Covid, survey, symptoms, test cases

1 Introduction

Although world-wide pandemics are not that often, yet Covid pandemic hit the world fast, with many patients dying and doctors not being able to understand the cause in time. The aftereffect of the pandemic has left many people with health issues with more and more people becoming hypochondriacs. Technology was and is still used to alleviate the hit from the virus and help prevent the spread of the corona virus and maintain the current lifestyle as much as possible. On the other side, technology was used to help fight against the virus and return life to its original form.

A lot of research has been done on the Covid virus and Corona outbreak, including image processing, machine learning and so on. In [1], they give a summary of the different machine learning techniques to predict and classify covid-19 cases. They are using mathematical models and machine learning to predict Covid-19 cases. The authors in [2], have further used machine learning and image processing to determine the cause of pneumonia in covid-19 infected patients. They are using X-rays and CT images to create a software to determine how to classify patients based on pneumonia and Covid-19 images.

The machine learning approach is also used in most papers but in [3], authors are trying to investigate the best possible options and weight distribution in the ML techniques to get the best results when working with Covid-19 data. They are using different approaches to get the best results when combining ML and Covid. Furthermore, [4] is again using CT scans and ML to classify patients as infectious or not, which would be useful to decrease infection spread amongst the population.

Much like in [1], authors in [5] are helping other authors with an overview of the ML techniques. Additionally, they are offering data sets to help with the further investigation. The research done in [1] and [5], coupled with the research in [6], gives authors the means, the knowledge, the data set and the information on how to proceed with the research for covid and ML. The research in [6] evaluates all the data and the publishing process of papers regarding Covid and ML and how the publication process changes the initial paper submission.

Further analysis is done in [7] about covid detection and CT images using a pre-trained data set that can help classify the new data set before training and testing using deep learning and multi-layered convolution algorithms. This way, the data set can be increased and overcome the persisting problem of ML with not having enough data to perform the training and testing. The overall analysis of all the research in ML and data set is concluded by the authors in [8], where they give a detailed analysis of the functions and usage of ML and Covid.

There is a lot of research of Machine Learning/Deep Learning techniques to detect Covid using medical images. Our approach is simpler and uses medical questionnaires and human input to improve the detection of Covid 19 in patients. The architecture of our Covid Medical App system is described in Section 2, whereas the behavior and case scenarios are described in Section 3. Section 4 concludes the paper and gives information about further development.

2 Architecture of the system

The Covid Medial Application is designed to help patients and the health system by classifying patients into six categories. These categories range from the patient not having Covid (or the least suspect of a Covid infection), to an almost certain Covid infection (requires isolation and medical treatment). The users of the applications are taking a short survey (questionnaire) about their wellbeing and symptoms, and the result of the questionnaire is the classification of the user into one of categories [9]. The application accesses the survey from an API that is standardized and provided by the InferMedica Medical Platform, implemented and approved by World Health Organization (WHO). The API contains all sorts of Covid data that can be retrieved and many surveys the users can take, our application utilizes only the API for classification of Covid, which is done based on symptoms and patient's wellbeing.

Besides all the Covid recommendations and information that is displayed in the application, the users can take the survey and find out, based on their symptoms, in which category they belong. The categories are:

- No risk – the patient is the least likely to have Covid
- Self-monitoring – the patient should continue to monitor the symptoms but is not likely to have Covid
- Call a doctor – there is an infection, but it is not Covid-related
- Quarantine – the patient is advised to quarantine himself from the environment and perform Covid tests
- Isolation call – the person should isolate themselves from the environment with high probability for Covid infection
- Isolation ambulance – the person has high probability for Covid infection and should call for ambulance since the symptoms are severe.

The architecture and the organization of the application is presented on Figure 1.

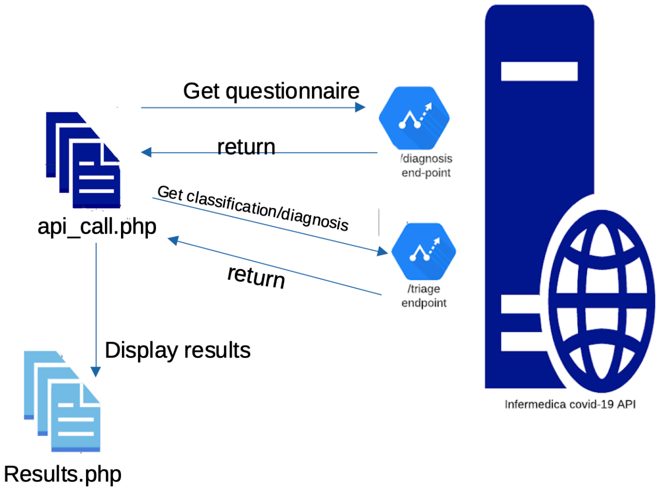


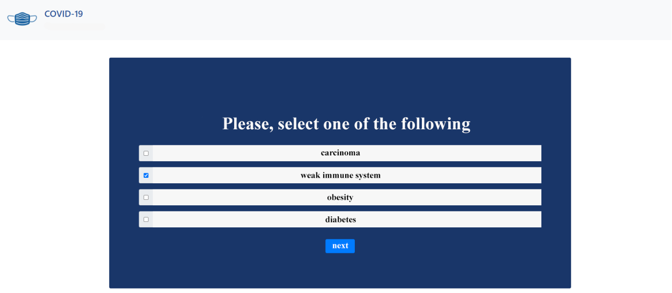
Figure 1 Organization of the application

On Figure 1 we can see that our application is a wrapper around the API provided from InferMedica, which first and foremost, provides a human readable survey that patients can take and classify their symptoms into a category. The questionnaire helps patients with symptoms of Covid to determine the best possible action to take, in case they are suspecting Covid infection. Users of the application access it via web link, where users can get Covid-related information, access their profile and take the questionnaire. The questionnaire taken from a patient is packed, formatted and sent to the API, the API returns the result, which is displayed back to the patient.

As presented on Figure 1, we can see that the application uses two APIs from InferMedica. The first API is diagnosis endpoint, that we use to obtain the questions to form the questionnaire. These questions are predetermined, can easily be translated into any language, and be adapted if the questionnaire changes from the endpoint. The second API is the triage endpoint that is used to perform the diagnosis and classification of the patient. Also, the result returns a short info status that is presented to the patient with information about how to proceed with the diagnosis and recommendations. This information can also be easily translated and wrapped.

Patients that might have higher risk of Covid infection (placed in that category by the API) can isolate themselves in time to prevent others to be infected. Furthermore, the entire pandemic made many patients hypochondriacs and suspect Covid symptoms even for a small cough. Thus, by using this application, if they get classified in no Covid infection categories, uninfected patients can avoid going to the hospitals for unnecessary Covid tests, and reducing the possibility to get infected in the testing areas.

On the figure below (Figure 2), we can see a part of the survey interface and the questions that the users have to answer to be classified in the categories.



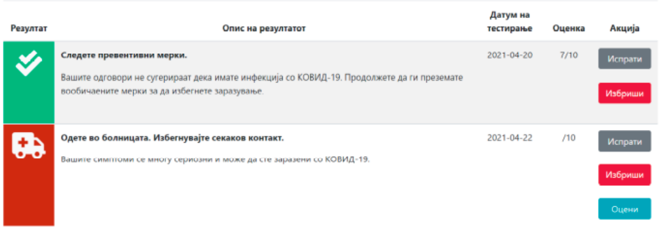
Please, select one of the following

☐ carcinoma
☑ weak immune system
☐ obesity
☐ diabetes

Figure 2 Questions from the survey (multipart)

The series of questions can vary from input fields for body temperature measured or blood pressure, to multiple choice questions and Yes/No questions. The requirements from the questionnaire are simple and easily understandable that every patient can answer even if with severe health issues. The interface is adjusted and simplified as to not impose any incorrect information that could lead to a faulty classification.

On Figure 3 we can see a list of results that the patient received, as a result of the survey. From Figure 3, we can see that the information is presented in different color based on the severity of the classification, followed by a short information summary intended for the classification. The patient can take the questionnaire multiple times, and each result is marked and presented to the user with the date and time of the questionnaire taken and the result.



Резултат	Опис на резултатот	Датум на тестирање	Оценка	Акција
✓	**Следете превентивни мерки.** Вашите одговори не сугерираат дека имате инфекција со КОВИД-19. Продолжете да ги преземате вообичаените мерки за да избегнете заразување.	2021-04-20	7/10	Испрати Избриши
🚑	**Одете во болницата. Избегнувајте секаков контакт.** Вашите симптоми се многу сериозни и може да сте заразени со КОВИД-19.	2021-04-22	/10	Испрати Избриши Оцени

Figure 3 Result of the classification

On the other hand, medical personnel also have access to these classifications, but only to patients that they have been assigned to. Based on the outcome of the classification, the medical personnel can schedule an appointment for testing or send an ambulance to the appointed address. The panel of the medical personnel is similar to the one of the patient's, except it additionally displays the information of the patient that took the survey and contact information.

3 Evaluation of the system

Each medical classification system cannot guarantee a faultless classification method, so there is always a chance that the classification might not be correct. If there are numerous of medical tests and findings, a different doctor might give a different diagnosis and classification of a patient's condition. Even more so in our case, where we are using a questionnaire to classify a patient in a six different Covid categories, it gives a rough classification as a basic step of the diagnosis. The questionnaire, as stated before, is taken from Infermedica, which was previously issued by the WHO, but it is not something that can be used with absolute certainty and fully depended upon. That is why, in this section of the paper, we are also making an evaluation of the results of the questionnaire.

Our medical application allows users to take a Covid survey based on their symptoms and be classified into categories of high to low Covid infection. Alongside with the classification, a short information is presented on how to proceed with their result and how to minimize further infection on other patients. The survey, as stated before, is intended to keep patients with low risk of infection to visit Covid testing places in order to avoid getting infected. Also, by advising patients with low possibility of Covid infection to not get tested, reduces crowding the medical facilities and Covid test centers, thus reducing overhead of the medical system. However, patients can still ignore the results from our application and get tested to make sure if they have Covid or not.

The case study of the API and our application was conducted with 20 patients who already have been tested with Polymerase Chain Reaction (PCR) test for Covid in the past. More than half of the patients (15 of them) have been tested twice for Covid, thus the total number of test cases is 35. The patients already had the diagnosis for Covid from their PCR test before the survey was taken on our application. After which, we have compared the results from the survey with the results from the PCR tests of the patients. The results from the case study are presented on Figure 4.

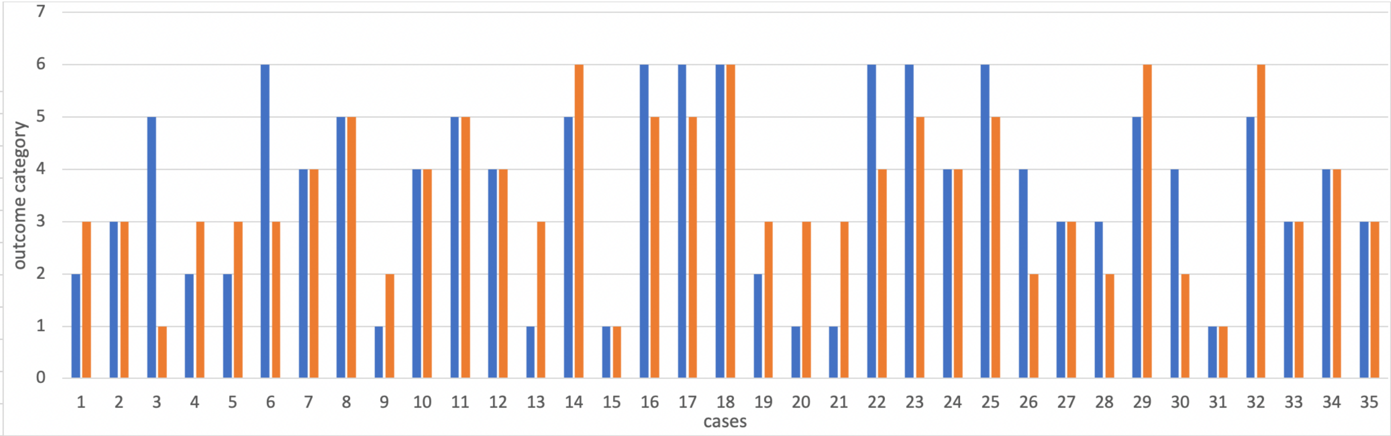


Figure 4 Results from the case study of our application with 35 tests

On Figure 4 we can see the results from our application (shown with blue bars) and the results from the PCR tests (shown with orange bars). As we can see from the results, the PCR and the application bars are mostly the same. The deviation in the PCR and the application results are mostly in categories one, two and three. The most common error is when the API suggests category one, but the PCR shows category three. This error is minimal since the first three categories are linked with low to no infection. The next frequent error is in the last two categories, when the API suggests category five, but the PCR suggests category six and vice versa. If we put the results of the questionnaire in binary form (the patient has Covid or the patient doesn't have Covid), the first three categories will form the result that the patient doesn't have Covid, whereas the last three categories will form the result that the patient has Covid. If the categories are binary, the error between the API and the PCR is close to zero. The minimal diversion is detected in the subcategories presented by the questionnaire. Also, the PCR gives information as to whether the patient has or hasn't got Covid, the subcategorizing is done based on hospitalization of the patient and the recommendations received from their doctor.

If we consider the six categories offered by the API, the overall success rate of the API, compared with the PCR tests is at 85% of accurate prediction and classification. If we consider the binary classification, the success rate of the API is increased to 89%.

4 Conclusion

Our application tends to use a simplified system for online diagnosis of Covid patients that uses questionnaire designed to give initial diagnosis of the patient. This initial diagnosis is used to give patients information as to whether they have Covid or not and to suggest testing and medical care, only if necessary, thus reducing the overhead on the testing places and the medical facilities from patients that are with low risk or no infection at all. The case study in section III shows that the questionnaire is accurate enough to give initial diagnosis and sufficient enough to determine if the patient has Covid or not with 89% accuracy.

For future work we propose testing the system with patients before they go to the hospital or testing facilities for Covid. The user can update the results of the API with the results from the medical/test facilities. This can be done by result category, and the system can present the accuracy of the API result next to the result. Thus, users can get classified into the categories, but also receive accuracy information provided by users of the application that have been classified and afterwards tested.

References

[1] Swapnarekha, H, Behera, S., Nayak, J., Naik, B., Role of intelligent computing in COVID-19 prognosis: A state-of-the-art review, Chaos, Solitons & Fractals, Volume 138, 2020, ISSN 0960-0779

[2] Bharati, S., Podder, P., Mondal, R. H, Prasath, S., Medical Imaging with Deep Learning for COVID- 19 Diagnosis: A Comprehensive Review, arXiv:2107.09602

[3] Mohammed, M., Abdulkareem, K., Al-Waisy, A., Benchmarking Methodology for Selection of Optimal COVID-19 Diagnostic Model Based on Entropy and TOPSIS Methods, IEEE Access, May 2020, 10.1109/ACCESS.2020.2995597

[4] Subhalakshmi, R.T., Appavu, S, Sasikala, S., Deep learning based fusion model for COVID-19 diagnosis and classification using computed tomography images, oncurrent Engineering: Research and Applications 2022, Vol. 30(1) 116–127

[5] Chadaga, K., Prabhu, S., Vivekananda, B., Battling COVID-19 using machine learning: A review, Cogent Engineering, 8:1, 1958666, DOI: 10.1080/23311916.2021.1958666

[6] Jemioło, P.; Storman, D.; Orzechowski, P. Artificial Intelligence for COVID-19 Detection in Medical Imaging—Diagnostic Measures and Wasting—A Systematic Umbrella Review. *J. Clin. Med.* **2022**, *11*, 2054. https://doi.org/10.3390/ jcm11072054

[7] Mehboob, F., Rauf, A., Jiang, R. *et al.* Towards robust diagnosis of COVID-19 using vision self-attention transformer. *Sci Rep* **12**, 8922 (2022). https://doi.org/10.1038/s41598-022-13039-x

[8] Swapnarekha, H., Behera,H., Nayak, J., Naik, B., Role of intelligent computing in CʜOVID-19 prognosis: A state-of-the-art review, Chaos, Solitons & Fractals,Volume 138,2020,ISSN 0960-0779,https://doi.org/10.1016/j.chaos.2020.109947.

[9] Infermedica Medical Platform, Covid-19 survey API, https://developer.infermedica.com/docs/api

Piloting ICT Solutions for Integrated Care

Mitja Luštrek
Department of Intelligent Systems
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
mitja.lustrek@ijs.si

Samo Drobne
Faculty of Civil and Geodetic
Engineering
University of Ljubljana
Ljubljana, Slovenia
samo.drobne@fgg.uni-lj.si

Sokratis G Papageorgiou
Neurology Department
Aiginition Hospital, National and
Kapodistrian Univ. of Athens
Athens, Greece
sokpapa@med.uoa.gr

Efthalia Angelopoulou
Neurology Department
Aiginition Hospital, National and
Kapodistrian Univ. of Athens
Athens, Greece
angelthal@med.uoa.gr

Roberta Matković
Teaching Institute
for Public Health
of Split and Dalmatian County
Split, Croatia
roberta.matkovic@nzjz-split.hr

Bojan Blažica
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
bojan.blazica@ijs.si

Pietro Hiram Guzzi
Municipality of Miglierina
Miglierina, Italy
sindaco@comunemiglierina.it

Miodrag Miljkovic
Marketing Department
Special hospital Merkur
Vrnjacka banja, Serbia
miljkovicdzoni@gmail.com

ABSTRACT

The SI4CARE project is aiming to develop a strategy and action plans to improve health and social care in the Adriatic-Ionian region. It started with surveying the state of affairs in the region, identifying needs and challenges, as well as best practices that can answer them. Based on these, wishes for improvement were formulated. The paper describes the methodology of this process and the key findings. Some of the best practices are being piloted to support the development and monitoring of the policy actions. In the paper, we describe nine pilots that involve pervasive health technology and otherwise strongly leverage ICT to benefit senior users. Most employ wearables and other sensing devices to monitor the users and provide health and care services, or provide telehealth and care through web and mobile technology.

KEYWORDS

Social innovation, integrated care, telehealth, telecare, transnational strategy, action plan

1 INTRODUCTION

The population of Europe and the rest of the developed world is rapidly aging. In the last 20 years, the old-age dependency ratio of working-age population vs. seniors in Europe decreased from 4 : 1 to 3 : 1, and it is projected to further decrease to 1.75 : 1 by 2050 [1]. This will result in a range of problems, including a lack of people who can support the seniors once they can no longer live independently. These problems will have to be tackled from multiple angles: with demographic policies, increases in retirement age, and social and technological innovations that can improve the care for the seniors and their quality of life.

The SI4CARE project [2] aims to create a transnational ecosystem for social innovation in integrated care with a focus on ICT technology. It started with surveying the status quo of health and social care in the Adriatic-Ionian region, identifying needs and challenges, as well as best practices that can answer them. It then formulated wishes and actions for improved health and social care, which will eventually result in a transnational strategy and national/regional action plans.

To gain a deeper insight into the benefits of the identified best practices and ways of implementing them, the project started 13 pilots in seven countries. We describe the nine that involve pervasive health technology and otherwise strongly leverage ICT to benefit senior users. Most employ wearables and other devices to monitor the users and provide health and care services, or provide telehealth and care through web and mobile technology.

2 SI4CARE PROJECT: FROM STATUS QUO TO ACTION

The SI4CARE project used a systematic and evidence-based approach for devising a strategy and actions to improve integrated care via social and technological innovation, with the aim of presenting solid arguments to decision makers.

2.1 Status Quo of Health and Social Care

The first step was to survey the status quo (the state of affairs) in health and social care in the Adriatic-Ionian region, comprising Slovenia, Croatia, Bosnia and Herzegovina, Serbia, Montenegro, Greece and Italy. Four key activities were done:

- We surveyed the literature, such as statistical reports, national and regional policy documents and legislation.
- We conducted semi-structured interviews with high-level stakeholders such as highly placed employees at relevant

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

ministries, non-governmental organizations and educational institutions. The interviews included 26 questions on the healthcare system, financial and physical accessibility of healthcare services, future challenges and other topics. 31 stakeholders were interviewed in total. Qualitative analysis of the answers was performed, focusing on the main points raised among the participants.

- We administered a questionnaire to various people employed in health and social care services. The questionnaire included 29 items on the use of healthcare services by seniors, their accessibility, the ability to obtain information on healthcare, and the status of seniors in the society and their social care. A subset of these questions was asked specifically about people with memory impairment or dementia. We received responses from 222 health and social care staff.
- We administered the same questionnaire to seniors. We received responses from 619 people.

Our finding was that in general, the provision of healthcare services is moderately good, with a lack of human resources cited as a key problem. Rehabilitation was noted to be less available than other services, and people with dementia face more problems than the general elderly population. A significant problem is that seniors are poorly informed about healthcare.

Even though healthcare is mostly covered by insurance, many seniors face significant financial problems, mainly due to low pensions. In part, this appears to be because, despite the insurance, they sometimes still need to resort to private services. Waiting times are a common issue, which may explain the use of private services. Physical accessibility is also a major issue – the seniors have significant difficulties using public transport. Secondary healthcare for people living in rural areas was also found to be difficult to access.

Seniors have a low digital literacy and find anything involving the internet (e.g., booking an appointment) a major problem. High-level stakeholders feel that new technologies have not been successfully integrated in the healthcare system, and this is even more true for the questionnaire respondents. Most stakeholders believe such technologies are important, though, validating the objectives of the SI4CARE project.

2.2 Best Practices

The SI4CARE project identified and documented 115 best practices in social and technological innovation to improve the care and quality of life of seniors, selected based on their effectiveness as demonstrated by experience.

Since SI4CARE emphasizes the use of ICT technologies in integrated care, most of the identified best practices are technology-based. The largest group involve pervasive health technology, such as wearables to monitor users, either to help them manage their health or to provide functions such as fall detection. Some also use sensing integrated in fitness devices or 3D cameras to support rehabilitation. There are also web and mobile platforms that support various activities interesting to seniors (e.g., gardening, cognitive training), facilitate communication and social inclusion. A few best practices are intended for hospitals and other care organizations (e.g., for management of health records).

Some of the best practices – less relevant to this paper but otherwise just as important – are non-technological. Examples include providing information and training to seniors about health(care), particularly dementia, and digital technology; promotion of social inclusion; and organizing provision of (health)care (e.g., via mobile medical units).

2.3 Wishes to Improve Care and Quality of Life

Based on the analysis of status quo (Section 2.1) and inspired by the best practices (Section 2.2), the SI4CARE project formulated a number of wishes that – if fulfilled – would leverage social and technological innovation to improve the care and quality of life of seniors. These were developed for each of the involved countries, and validated in a focus group involving stakeholders.

Since the analysis of the status quo found a strong need for the introduction of new technologies, and many technology-based best practices were identified, it is not surprising that various initiatives aimed at increasing the use of telehealth and telecare comprise the largest group of wishes. They had different focus: rehabilitation (where the current availability is particularly poor), cost-effectiveness (which is a prerequisite for institutional funding), non-pharmacological interventions (that tend to be neglected), applications that do not require institutional support (which are typically inexpensive and non-pharmacological) … Activities to improve digital skills of seniors were also wished for, as well as better digital infrastructure.

Unlike best practices, most wishes were not technological. This is perhaps because wishes are about goals, whereas technology in health and social care is often a means of achieving these goals. The non-technological wishes include increases in human resources (which were found to be a key reason for the inadequacies of healthcare provision), improved overview of the state of care and solutions for improvement (essentially activities similar to SI4CARE's but put on a more sustainable basis), improvements in home care, training and better policies.

2.4 Transnational Strategy and Action Plans

The preparation of the transnational strategy and national/regional action plans – one for each country involved – is still in progress. The strategy is organized in five pillars:

- Digital transitions are concerned with pervasive health technologies and other ICT-based innovations exemplified by the pilots presented in this paper.
- Digitalization process will support digital transitions by providing the required infrastructure and knowledge.
- Economic and financial implications deal with appropriate funding for healthcare and other aspects of long-term care.
- Governance and policies address sustainable and geographically appropriately distributed provision of care, ensuring its quality and properly trained staff.
- The SI4CARE community will ensure the sustainability of the project via organizations that will exist after the end of the funding period.

The national/regional action plans aim at implementing this strategy in individual countries. Their main components are specific actions, which essentially fulfill the wishes discussed in Section 2.3. These wishes are being validated by stakeholders in events organized in each country, one of which is also taking place at the Information Society 2022 multiconference. Afterwards, the action plans will be presented to high-level decision makers.

3 PILOTS OF ICT SOLUTIONS FOR INTEGRATED CARE

3.1 Mobile Application for Self-management of Heart Failure

Heart failure is a common and debilitating disease among seniors, and a leading cause of hospitalizations. It requires complex management difficult for many seniors. Healthcare institutions provide only periodic checkups and cardiac rehabilitation, the latter not to all who would benefit. Resources to provide more support are hard to come by, so a mobile application to assist self-management is an attractive solution.

The HeartMan application [3] provides a personalized exercise program and nutrition advice, support for measurement of vital parameters, medication reminders, mindfulness exercises intended to improve the patients' mental health and wellbeing, and cognitive behavioral techniques to improve the adherence to the application's advice. The first step of the pilot was to make the application easier to deploy and to remove physiological monitoring as input for its decisions, as this is a barrier from the usability and regulatory perspective. The user experience was also improved. The ongoing second step is a feasibility study with 20 patients using the application and 10 controls.

The lesson learned so far is that designing an application for heart-failure patients is difficult due to the complex topic and poor digital and health literacy of this group. Our solution was to guide less advanced users by simple automatic prompts, and not require them to do much on their own initiative.

3.2 ICT Solution for Monitoring the Health of Patients after Returning Home

Special Hospital Merkur is a secondary health institution in Serbia specializing in diabetes. Upon discharge, patients often return to bad habits, and diabetes complications occur. In addition, they face problems when they need to see a doctor.

The main aim of the pilot was to investigate the integration of modern communication technology in diabetes treatment to facilitate better coordination between stakeholders. The patients were trained to use the SmartCare mobile application, and to input the necessary data (insulin, sugar, mass, blood pressure, temperature, etc.). Merkur's medical team had insight into the patient's condition and intervened as needed. In addition, patients were trained to contact doctors for consultations from home.

The combined effect of the involvement of patients in their health condition, and the remote intervention of doctors, proved to reduce the risk of diabetes complications. The pilot demonstrated the feasibility of remote treatment in Serbia, which can also lead to significant financial savings. It should be repeated on a larger sample on a national level to provide a basis for the introduction of telemedicine in the health system.

3.3 ICT to Enable Accessibility to Health Systems by the Elderly

In the Italian healthcare system, regional governments are responsible for ensuring the delivery of a health benefits package through a network of health management organizations. There is a remarkable difference among regions, with northern regions providing better services, resulting in migration of patients from south to north. In 2018, healthcare mobility in Calabria amounted to approx. € 310 million. This is particularly relevant for small towns and villages where people suffer from a lack of general medicine and efficient public transportation to regional hubs.

Due to some recent programs, many rural areas in Calabria have good internet connections. In this pilot, with the help of UCCP del Reventino (a team of physicians), we are evaluating the use of tele-assistance and remote monitoring of chronic patients (elderly people and people affected by dementia).

The developed services are particularly useful for patients who require a re-evaluation of an already known clinical picture, people suffering from rare diseases, and frail people who require constant contact with health facilities. Teleadvice also proved of great utility in the context of COVID-19.

3.4 Specialized Outpatient Clinic for Memory, Dementia and Parkinson's Disease

Approximately 20% of the population above the age of 65 are affected by mild cognitive impairment or dementia. As the status quo analysis indicated, these people have limited access to specialized healthcare. This is more pronounced in remote areas. Greece has many small and isolated islands with a high percentage of elderly inhabitants and understaffed health centers.

The Aeginition Hospital of the National and Kapodistrian University of Athens developed an outpatient clinic pilot through the National Telemedicine Network, in collaboration with the 2nd Regional Healthcare Administration of Piraeus and the Aegean Islands. Through this clinic, patients with cognitive or movement disorders living in remote Aegean islands are examined by a specialized healthcare team (neurologist, psychiatrist and neuropsychologist) through video-conferencing.

Based on the questionnaires from 58 telemedicine visits, all stakeholders are highly satisfied with this telemedicine service, mentioning improved care, better health, and convenience, reduced transportation and cost. The low number of cases compared to the available capacity points to the need to better disseminate the information about the availability of telemedicine in the area by involving local health professionals and other telemedicine services in Greece.

3.5 Tele-exercise for the Elderly and Patients with Cognitive Disorders/Dementia

Physical activity is a well-established non-pharmaceutical intervention for health improvement in the elderly. It improves mobility, fitness, and cognitive function, prevents falls, improves functionality and quality of life as well as increases socialization.

The Aeginition Hospital of the National and Kapodistrian University of Athens in collaboration with the Medical School of Athens developed a tele-exercise pilot to provide specialized online physical activity programs for the elderly. Small groups of about 10 individuals receive aerobic and resistance training with a frequency of 2–5 times/week and duration of 40 min per intervention, guided in real-time via video-conference by specialized healthcare professionals. The elderly involved were trained to use the tablets though which they are participating.

All participants report high satisfaction rates and improved functionality in everyday life. Key lessons learned are that tele-exercise is feasible and effective non-pharmacological treatment that enhances social interaction, and that effective collaboration between healthcare providers is necessary. The elderly face difficulties in the use of new technologies and training is needed.

3.6 Individualized Training Based on Biomechanical Measurements

The importance of physical activity was already discussed in the previous pilot description. The status quo analysis in Slovenia showed that the availability of physical exercising and rehabilitation services is not adequate. Resorting to the private sector may result in lower quality of services as they might be provided by people without the necessary knowledge and skills.

We prepared a pilot in which training was based on initial screening of the participants by an orthopedist and experienced coaches, followed by biomechanical measurements of lower extremities. Isometric measurement of peak torque and tensiomyography were used along with a body composition measurement. 24 participants performed training 2 times per week for 3 months under two conditions: half of the participants exercised in a gym, while the other half online. In the in-person scenario, participants were divided in small groups. The focus was proper posture and exercise execution. Only after absorbing proper technique, the training increased intensity.

Both conditions were warmly accepted by participants, with the in-person one slightly preferred. Working in small groups not only enabled individual training, but also group cohesion, resulting in socialization after exercising in the nearby café.

3.7 Nursing by Monitoring

The pilot carried out in Split, Croatia, was motivated by the well-established issue of inadequate resources to provide quality care to seniors who cannot live independently.

The pilot used monitoring technology that requires minimal interactions with senior users, since they are not familiar with digital technology. 10 medically non-certified wristbands, equipped with LoRaWAN radio, ensure data delivery to large distances without using mobile phones as a gateway. The wristbands enable 10-minute acquisition of heart rate, GPS location, steps, calories, and wrist temperature, as well as having alarms for low heart rate and falls, and a help button. The data is received by a system called IoT Wallet, which allows future expansion since it supports adding add more wristbands.

LoRaWAN technology turned out to provide broad coverage with a relatively low power consumption.

3.8 Access to Public Social Services by Telemedical Monitoring (Click for Life)

Seniors represent a high percentage of the population of Region of Central Macedonia (RCM) in Greece (22% are over 65), with a significant proportion of them living alone (approx. 100,000). They face difficulties in access to public social services, especially in high-density urban places and remote rural areas.

The RCM regional authority launched the pilot project 'Click for Life', offering telemedicine/homecare assistance to seniors with a low income living alone. Approx. 3000 users participate so far. They are provided: (1) 24-hour monitoring via devices with fall detection and a panic button. The panic button enables communication with a call center 24 hours/day. (2) Medical history is accessible to relatives and health professionals, and the users can receive notifications from the relatives. (3) Behavioral assessment service interprets movement and activity data from devices in the user's home. The aim is to automatically detect abnormal behaviors that may indicate an emerging disease.

The lesson learned so far is that there is a need for a more systematic coordination of the call center with public health care units, doctors, social care workers and emergency units.

3.9 Accessibility to Integrated Long-term Care

In the pilot project we analyzed both spatial accessibility and accessibility of information. Slovenia is rural country. Older people in rural Slovenia face poor access to public services and especially to health facilities. In terms of spatial accessibility, we identified the locations of buildings where seniors live alone. In 2021, there were 42,344 seniors living alone in houses (27,136 aged 65–79 and 15,208 aged 80 and older) in Slovenia.

There are a number of elderly care services advertised online, but the offer is scattered and searching for such information is time-consuming. To avoid these obstacles, we set up a web platform where different providers (formal and informal) are presented in one place. We included all formal providers in Slovenia in the database. We enabled self-registration of service providers and spatial representation of providers via the web.

We highlighted areas with poor accessibility to health and social care services, and will present them to local decision-makers and caregivers to improve integrated long-term care and transport for them. We will also present them our web app.

4 CONCLUSION

The paper presented the SI4Care project and its methodology to bring social innovation to integrated care. The focus was on the presentation of the pilots that address the identified needs and wishes in the region. The fact that most, nine out of thirteen, of the piloting activities within the SI4Care project involve some sort of pervasive health technology testifies to the importance of such technologies also for integrated care. Preliminary results from most pilots show benefits for stakeholders and good acceptance. However, digital literacy is a significant barrier, and in some cases also infrastructure, organizational readiness and legislation. Pervasive technology clearly cannot be introduced in isolation, which is why our strategy consists of five pillars, only one of which is concerned with pervasive technology.

ACKNOWLEDGMENTS

This paper has been produced with the financial assistance of the European Union. The content of the paper is the sole responsibility of project partners and can under no circumstances be regarded as reflecting the position of the European Union and/or ADRION programme authorities. The SI4CARE project is supported by the Interreg ADRION Programme funded under the European Regional Development Fund and IPA II fund.

REFERENCES

[1] Eurostat, 2021. *Eurostat Regional Yearbook* (2021 edition). DOI: 10.2785/894358

[2] SI4CARE – Social Innovation for integrated health CARE of ageing population in ADRION Regions. https://si4care.adrioninterreg.eu/

[3] M. Luštrek, M. Bohanec, C. Cavero Barca, M. C. Ciancarelli, E. Clays et al., 2021. A personal health system for self-management of congestive heart failure (HeartMan): Development, technical evaluation, and proof-of-concept randomized controlled trial. JMIR Med. Inform. 9, 3, e24501. DOI: 10.2196/24501

Network Anomaly Detection using Federated Learning for the Internet of Things

Ana Cholakoska
Ss. Cyril and Methodius University in Skopje
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia
acholak@feit.ukim.edu.mk

Bojan Jakimovski
Ss. Cyril and Methodius University in Skopje
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia
kti1562018@feit.ukim.edu.mk

Bjarne Pfitzner
Hasso Plattner Institute
Digital Health — Connected Healthcare
Potsdam, Germany
bjarne.pfitzner@hpi.de

Hristijan Gjoreski
Ss. Cyril and Methodius University in Skopje
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia
hristijang@feit.ukim.edu.mk

Bert Arnrich
Hasso Plattner Institute
Digital Health — Connected Healthcare
Potsdam, Germany
bert.arnrich@hpi.de

Marija Kalendar
Ss. Cyril and Methodius University in Skopje
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia
marijaka@feit.ukim.edu.mk

Danijela Efnusheva
Ss. Cyril and Methodius University in Skopje
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia
danijela@feit.ukim.edu.mk

ABSTRACT

The widespread use of IoT devices has contributed greatly to the continuous digitisation and modernisation of areas such as healthcare, facility management, transportation, and household. These devices allow for real-time mobile sensing, use input and then simplify and automate everyday tasks. However, like all other devices connected to a network, IoT devices are also subject to anomalous behaviour primarily due to security vulnerabilities or malfunction. Apart from this, they have limited resources and can hardly cope with such anomalies and attacks. Therefore, early detection of anomalies is of great importance for the proper functioning of the network and the protection of users' personal data above all. In this paper, deep learning and federated learning algorithms are applied in order to detect anomalies in IoT network traffic. The results obtained show that all the models achieve high accuracy, with the FL models providing slight worse results compared to the DL models. However, with the increase in the amount of user data, the model based on federated learning is expected to have better results over time.

KEYWORDS

federated learning; deep learning; malware; internet of things; anomaly detection

1 INTRODUCTION

In the last decade, a significant increase in the usage of Internet of Things (IoT) devices has been observed. The ability to connect various kinds of devices from different manufacturers to a network wirelessly and share data has proven beneficial to nearly every domain where this technology is involved, including household, industry, infrastructure, transportation, and healthcare[3]. Additionally, the actions that end users can take are increasing everyday and vary from changing ambient parameters of a home or car setting easily and on-the-go to remotely and securely controlling a manufacturing process inside a smart factory setting. Implementing these devices into an ambient assisted living (AAL) setting has proven to be beneficial both for the patients and for the medical staff, as it can improve monitoring and medical assistance (if needed), as well as medication dose adjustment[7].

However, the diversity of IoT devices, accompanied by wireless networking and a slow standardisation process, have led to many issues regarding the privacy and security of data and also the processes based on that data. The occurrence of various cyber attacks on networks composed of IoT devices, but also on individual IoT devices performing specific tasks, is becoming more common [8]. By disabling, reconfiguring or reprogramming such devices, attackers can manipulate the network, obtain private data illegally and maybe even induce a life-threatening situation, especially in the e-health domain. Therefore, it is significantly important to detect potential attacks and anomalies that occur in an IoT setting.

This paper examines the detection of anomalies in IoT network traffic by using deep learning and federated learning algorithms. The remainder of this paper is structured as follows. Section

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

2 gives an overview of the approaches tackling IoT network anomaly detection using deep and federated learning algorithms. Section 3 describes the used dataset and gives an insight into the importance of the features. The experiments done in this research and the discussion of the results obtained are presented in Section 4, while Section 5 gives a brief summary and provides further research directions.

2 RELATED WORK

One of the most popular approaches when tackling network anomaly detection is the usage of network intrusion detection systems (NIDS). By examining network data flow patterns (signatures), the NIDS can track inconsistencies (also called anomalies) and resolve them in a timely manner. However, directly analysing the behaviour of the IoT devices has proven to be more beneficial in detecting newer and unknown types of attacks, in spite of the overall lower detection accuracy and higher computational cost [6].

Using machine learning (ML) techniques has had a big impact on the development of NIDS and malware anomaly detection systems in general. Lin et al. [9] propose a combination of Support Vector Machines (SVMs) and Artificial Fish Swarm algorithms for IoT botnet detection. A combination [5] using different ML algorithms, also including an SVM has been done to evaluate the accuracy in detecting Mirai DDoS attacks. The authors in [16] used Convolutional Neural Networks (CNN) with binary visualisation to provide fast zero-day malware detection. However, some of the datasets used in these research papers provide only network traffic flow from conventional networks and have little to do with the attacks which target IoT networks. A further issue is that using traditional ML techniques increases the security risk, as data has to be moved away from the network and the data source to a powerful system performing the ML training.

Federated learning (FL) has emerged as a new decentralised way of training models on privately held datasets that can or should not be shared for security and privacy reasons. The training process consists of a central server and several clients, where the former facilitates the training and the latter possess the private data. In each round of federated training, the server randomly selects a subset of clients who receive the current model parameters. Then, local training is performed by each of the clients, keeping the local data on-site. The updated model parameters are then sent back to the server, where the global server model is updated. Opposed to centralised ML or classical decentralised techniques, FL can work with both independent and identically distributed (IID) and non-IID datasets. [10]

Several approaches have been using this decentralised technique in order to detect anomalies in IoT networks. The DIoT approach [2] uses federated learning to aggregate profiles of IoT network behaviour. It was evaluated in real-world conditions and reported no false alarms. Saharkhizan et al. [14] used a recurrent neural network with ensemble learning to detect cyberattacks on IoT devices. The evaluation of the model was performed on a Modbus dataset of network traffic. Some of the approaches even used a combination of FL and a distributed ledger (blockchain) [12, 17] in order to detect anomalies in networks. In [13], the federated deep learning model created for zero-day botnet attacks on IoT devices outperformed traditional decentralised approaches, as well as both localised deep learning (DL) and distributed DL methods. In [15], a novel privacy-by-design FL model using a stacked long short-time memory (LSTM) model is introduced for tackling anomaly detection in smart buildings. The results showed twice as fast convergence during training, compared to the centralised LSTM.

3 DATASET AND EXPLORATORY DATA ANALYSIS

For the purpose of this research we used the publicly available dataset N-BaIoT [11]. It is a dataset created by a group of researchers from the University of California, Irvine, School of Information and Computer Sciences in the USA. The dataset addresses the lack of public botnet datasets, especially for the IoT domain. It is composed of real-time network traffic data gathered from nine commercial IoT devices, including a baby monitor, security cameras, a webcam, doorbells, and a thermostat, which have been infected by the most common families of botnet attacks: Mirai and Bashlite [1].

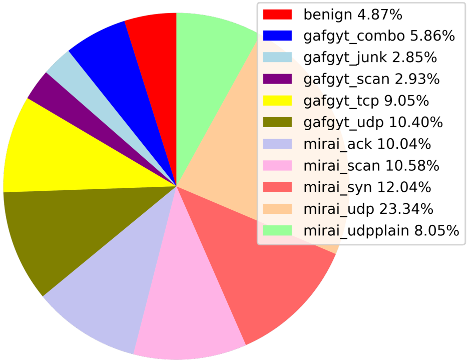


Figure 1: N-BaIoT dataset distribution by class

The N-BaIoT dataset consists of 7,062,606 entries with 115 different features, which are further divided into 10 attack categories: gafgyt_combo, gafgyt_junk, gafgyt_scan, gafgyt_tcp, gafgyt_udp, mirai_ack, mirai_scan, mirai_syn, mirai_udp, mirai_udpplain and one benign category, which contains the normal traffic flow of the observed devices. As it can be seen from Figure 1, which shows the distribution of the dataset used in the upcoming experiments, only a portion (509,149 entries) is considered for the model training in both DL and FL experiments. For the DL experiments, the dataset is further divided into a train and test partition including 80% and 20% of the data, while maintaining the distribution intact. As for the FL experiments, the data is divided into 50 IID datasets which include a train and test subsets. They represent the 50 clients which will take part in the FL process.

Table 1: Most important dataset features

Number	Feature
1	H L0.01_mean
2	Ml_dir_L0.01_mean
3	Ml_dir_L0.01_variance
4	H_L0.01_variance
5	H_L0.1_mean

After preprocessing the data, an exploratory analysis was done in order to obtain the features which have the greatest

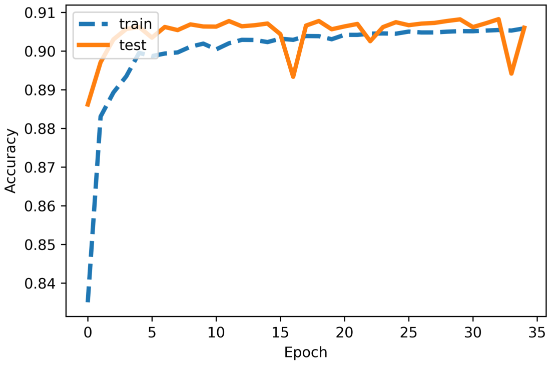


Figure 2: DL model using the five layer NN - accuracy

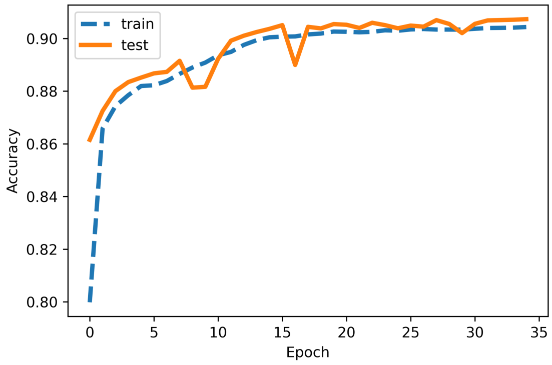


Figure 3: DL model using the three layer NN - accuracy

influence. The mutual dependence between the features and the class was determined with the help of Mutual Information Gain. From Table 1, it can be noticed that the five features with the greatest importance are H L0.01_mean, Ml_dir_L0.01_mean, Ml_dir_L0.01_variance, H_L0.01_variance and H_L0.1_mean.

4 EXPERIMENTS AND DISCUSSION

This paper compares two DL and two FL models for network anomaly detection, which are able to distinguish anomalous behaviour or a deviation from the normal traffic flow of IoT devices. After performing the training, all models were evaluated in order to see their accuracy in detecting anomalies. In the first experiment, a feed-forward neural network with 5 layers, an input layer, 3 hidden layers and an output layer was used. In the second experiment, a simple feed-forward neural network with one hidden layer was used. In both cases, the output layer has 11 neurons, which represent all the classes in the dataset.

Both models have the same hyperparameters. We used the Adam optimiser with a learning rate of 0.001, which works well for many use cases and models. Since the model performs a multi-class prediction task, we minimised the categorical cross entropy loss during training. The DL experiments were performed using the TensorFlow framework and the FL experiments were performed using the Flower [4] framework and TensorFlow Federated, applying the FedAvg aggregation strategy [10] on the server. In the FL experiments 35 rounds were performed, which corresponds to approximately 35 epochs in the DL experiments.

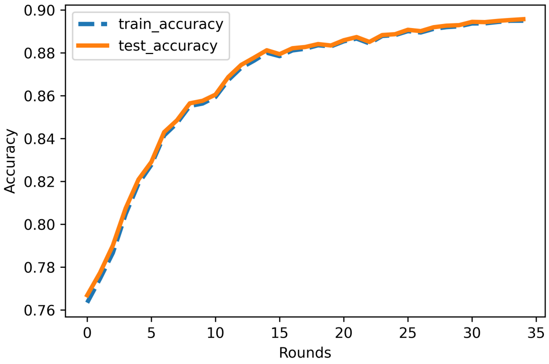


Figure 4: FL model using the five layer NN - accuracy

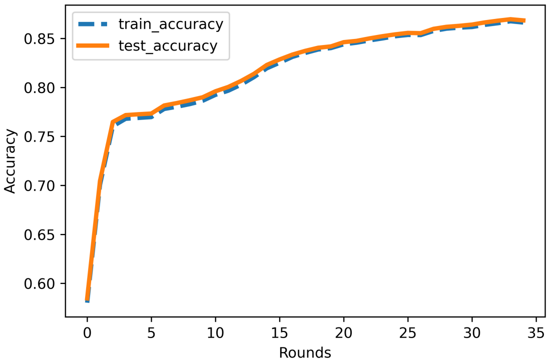


Figure 5: FL model using the three layer NN - accuracy

As previously mentioned, two DL models, the first one using a NN with multiple layers and the second one using a simple NN were trained and tested. From Figures 2 and 3 we can notice that the accuracy between the two models is very similar - the first model obtained an accuracy of 90.75% on the test data, while the second model obtained an accuracy of 90.18%. Furthermore, if the confusion matrices of both DL models are analysed, it can be noted that both models make the same mistake - predicting class 4 (gafgyt_scan) as class 5 (gafgyt_tcp).

When it comes to the results obtained from the FL process after 35 rounds it can be seen that the first model obtained an accuracy of 88% (Figure 4). As for the second simplified model, the accuracy is 86% (Figure 5). This means that even though a simpler NN was used, the second model actually performed similarly in terms of FL. We can also observe the minor differences in accuracy (1-5%) between the DL and FL models, which means that although the DL models performed slightly better, the FL models can also accurately predict anomalies.

From Figures 6 and 7 we can analyse the SHAP (SHapley Additive exPlanations) force plot, which shows the contribution of each feature in making a prediction. We can see that the features 69, 25, 75, 87, 56 and 101 (HH_jit_L3_mean, H_L0.1_mean, HH_jit_L0.1_mean, HpHp_L3_weight, HH_L0._covariance and HpHp_L0.1_weight) have the greatest influence in making the prediction. The features 69, 25 and 75 have a positive impact on

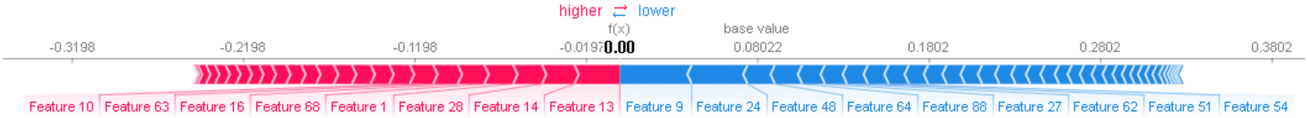


Figure 6: SHAP force plot for DL model using the five layer NN.

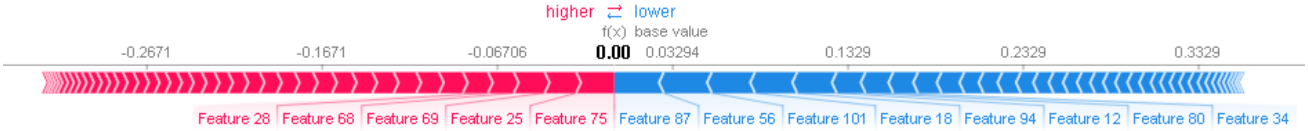


Figure 7: SHAP force plot for DL model using the three layer NN.

decision-making, i.e. prediction, while the features 87, 56 and 101 affect negatively on the performance. When we compare Figures 6 & 7 and Table 1, we can see that the most important features are different. This is because the SHAP method deals with the model and its output, while Mutual Information Gain deals with the preprocessed data.

5 CONCLUSION AND FUTURE WORK

This paper compares two models of DL and FL for accurate anomaly detection purposes in IoT networks. The FL model distributes the learning process to several clients, thus preserving data privacy and security. Both models achieve high accuracy, with the FL models providing similar results to the DL models.

Future work will include implementing some security mechanisms to the FL models and evaluating the trade-off between privacy and accuracy. Also, these models can be further tested and improved by being provided with new substantial datasets which may combine similar categories of attacks and/or include novel attacks on IoT networks. New federated learning algorithms can also be tested and evaluated on the same and new datasets, which can lead to a novel federated learning algorithm for anomaly detection purposes.

ACKNOWLEDGMENTS

The authors would like to thank Daniel Denkovski and Valentin Rakovic for the useful discussions on the research topic. This work has been supported by the WideHealth project - European Union's Horizon 2020 research and innovation programme under grant agreement No. 952279.

REFERENCES

[1] Abdelmuttlib Ibrahim Abdalla Ahmed. 2020. Systematic Literature Review on IoT-Based Botnet Attack. *IEEE Access* 8 (12 2020). https://doi.org/10.1109/ACCESS.2020.3039985

[2] Ulrich Matchi Aïvodji, Sébastien Gambs, and Alexandre Martin. 2019. IOTFLA : A Secured and Privacy-Preserving Smart Home Architecture Implementing Federated Learning. In *2019 IEEE Security and Privacy Workshops (SPW)*. 175–180. https://doi.org/10.1109/SPW.2019.00041

[3] Saurabh Bagchi, Tarek F. Abdelzaher, Ramesh Govindan, Prashant Shenoy, Akanksha Atrey, Pradipta Ghosh, and Ran Xu. 2020. New Frontiers in IoT: Networking, Systems, Reliability, and Security Challenges. *IEEE Internet of Things Journal* 7, 12 (2020), 11330–11346. https://doi.org/10.1109/JIOT.2020.3007690

[4] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. 2020. Flower: A Friendly Federated Learning Research Framework. https://doi.org/10.48550/ARXIV.2007.14390

[5] Rohan Doshi, Noah Apthorpe, and Nick Feamster. 2018. Machine Learning DDoS Detection for Consumer Internet of Things Devices. In *2018 IEEE Security and Privacy Workshops (SPW)*. 29–35. https://doi.org/10.1109/SPW.2018.00013

[6] Satish Kumar, Sunanda Gupta, and Sakshi Arora. 2021. Research Trends in Network-Based Intrusion Detection Systems: A Review. *IEEE Access* 9 (2021), 157761–157779. https://doi.org/10.1109/ACCESS.2021.3129775

[7] Isabel Laranjo, Joaquim Macedo, and Alexandre Santos. 2012. Internet of Things for Medication Control: Service Implementation and Testing. *Elsevier Procedia Technology* 5 (10 2012), 777–786. https://doi.org/10.1016/j.protcy.2012.09.086

[8] In Lee. 2020. Internet of Things (IoT) Cybersecurity: Literature Review and IoT Cyber Risk Management. *Future Internet* 12 (09 2020), 157. https://doi.org/10.3390/fi12090157

[9] Kuan-Cheng Lin, Sih-Yang Chen, and Jason Hung. 2014. Botnet Detection Using Support Vector Machines with Artificial Fish Swarm Algorithm. *Journal of Applied Mathematics* 2014 (04 2014), 1–9. https://doi.org/10.1155/2014/986428

[10] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2017. Learning Differentially Private Language Models Without Losing Accuracy. *CoRR* abs/1710.06963 (2017). arXiv:1710.06963 http://arxiv.org/abs/1710.06963

[11] Yair Meidan, Michael Bohadana, Yael Mathov, Yisroel Mirsky, Asaf Shabtai, Dominik Breitenbacher, and Yuval Elovici. 2018. N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders. *IEEE Pervasive Computing* 17, 3 (2018), 12–22. https://doi.org/10.1109/MPRV.2018.03367731

[12] Yisroel Mirsky, Tomer Golomb, and Yuval Elovici. 2020. Lightweight collaborative anomaly detection for the IoT using blockchain. *J. Parallel and Distrib. Comput.* 145 (06 2020). https://doi.org/10.1016/j.jpdc.2020.06.008

[13] Segun I. Popoola, Ruth Ande, Bamidele Adebisi, Guan Gui, Mohammad Hammoudeh, and Olamide Jogunola. 2022. Federated Deep Learning for Zero-Day Botnet Attack Detection in IoT-Edge Devices. *IEEE Internet of Things Journal* 9, 5 (2022), 3930–3944. https://doi.org/10.1109/JIOT.2021.3100755

[14] Mahdis Saharkhizan, Amin Azmoodeh, Ali Dehghantanha, Kim-Kwang Raymond Choo, and Reza M. Parizi. 2020. An Ensemble of Deep Recurrent Neural Networks for Detecting IoT Cyber Attacks Using Network Traffic. *IEEE Internet of Things Journal* 7, 9 (2020), 8852–8859. https://doi.org/10.1109/JIOT.2020.2996425

[15] Raed Abdel Sater and A. Ben Hamza. 2021. A Federated Learning Approach to Anomaly Detection in Smart Buildings. *ACM Trans. Internet Things* 2, 4, Article 28 (aug 2021), 23 pages. https://doi.org/10.1145/3467981

[16] Robert Shire, Stavros Shiaeles, Keltoum Bendiab, Bogdan Ghita, and Nicholas Kolokotronis. 2019. Malware Squid: A Novel IoT Malware Traffic Analysis Framework Using Convolutional Neural Network and Binary Visualisation. In *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*, Olga Galinina, Sergey Andreev, Sergey Balandin, and Yevgeni Koucheryavy (Eds.). Springer International Publishing, Cham, 65–76.

[17] Devrim Unal, Mohammad Hammoudeh, Muhammad Asif Khan, Abdelrahman Abuarqoub, Gregory Epiphaniou, and Ridha Hamila. 2021. Integration of Federated Machine Learning and Blockchain for the Provision of Secure Big Data Analytics for Internet of Things. *Comput. Secur.* 109, C (oct 2021), 14. https://doi.org/10.1016/j.cose.2021.102393

Indeks avtorjev / Author index