NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
"Translational Engineering in Health and Medicine"

Unsupervised Machine Learning approach on Metabolic Syndrome patient identification

Postgraduate Diploma Thesis

Sardis Antonios

Supervisor*: Georgios Matsopoulos, Professor, NTUA*

Athens, October 2024

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

“Translational Engineering in Health and Medicine”

Unsupervised Machine Learning approach on Metabolic Syndrome patient identification

Postgraduate Diploma Thesis

Sardis Antonios

Supervisor*: Georgios Matsopoulos, Professor, NTUA*

The postgraduate diploma thesis has been approved by the examination committee on 27/09/2024

1st member	*2nd member*	*3rd member*
Georgios Matsopoulos, Professor, NTUA	*Konstantina Nikita, Professor, NTUA*	*Panayiotis Tsanakas, Professor, NTUA*

Athens, October 2024

.....

Sardis Antonios

Graduate of the Interdisciplinary Postgraduate Programme,

“Translational Engineering in Health and Medicine”,

Master of Science,

School of Electrical and Computer Engineering,

National Technical University of Athens

Abstract

Metabolic syndrome (MetS) represents a complex constellation of metabolic abnormalities, including abdominal obesity, dyslipidemia, hypertension, and impaired glucose metabolism, that significantly elevate the risk of cardiovascular diseases and type 2 diabetes mellitus. Despite its public health relevance, the clinical definition of MetS remains ambiguous due to variations in diagnostic criteria used by a plethora of organizations such as the World Health Organization definition (WHO), the European Group for the Study of Insulin Resistance definition (EGIR), the National Cholesterol Education Program (NCEP) and the International Diabetes Federation (IDF). These vague definitions create discrepancies that complicate patient clustering and segmentation, making it challenging to correctly identify them and develop consistent therapeutic approaches. Additionally, real-world datasets frequently exhibit missing values, further complicating the accurate identification of patient subgroups and the quantification of their associated risks.

This thesis addresses and is driven by two major issues in the study of metabolic syndrome: (1) the variability in diagnostic definitions and its impact on patient clustering, and (2) the challenges posed by missing data in real-world clinical datasets. Initially, the four most popular MetS definitions are translated into lines of code to identify and label patients with real data. Afterwards, clustering algorithms and combination of methods are utilized to group patients based on their clinical profiles. To address the issue of missing data, cutoffs are set and data interpolation is applied.

Using a dataset of 850 patients, this analysis reveals that differing definitions of MetS result in significant variation in MetS patient identification and that they have to be combined in order to get a robust view, with some individuals probably being still misclassified or excluded entirely depending on the criteria used. Furthermore, the presence of missing values, particularly in variables crucial to the MetS definitions like glucose and waist circumference, disrupts clustering algorithms, leading to biased assessment and non-valid clustering separation.

To mitigate these challenges, hybrid methodologies are tested combining unsupervised machine learning techniques, such as K-means and Spectral clustering with principal component analysis (PCA) to handle the dimensionality, with robust missing data handling and exclusion methods. The results demonstrate that the application of different clustering techniques can improve the clustering outcomes and patient classification, providing a clearer understanding of MetS risk stratification. On the other hand, while the data remain unreliable and the MetS definitions vary, no robust classification results are enabled to occur. This work underscores the need for standardizing MetS definitions and addressing missing data to improve clinical outcomes and the precision of metabolic risk management strategies.

Keywords: Metabolic Syndrome, patient clustering, missing data, unsupervised learning, diagnostic criteria, K-means clustering, PCA, risk stratification

Contents

1. Introduction

1.1. Motivation

Metabolic syndrome (MetS) represents a complex constellation of metabolic abnormalities, including abdominal obesity, dyslipidemia, hypertension, and impaired glucose metabolism, that significantly elevate the risk of cardiovascular diseases and type 2 diabetes mellitus. While being a serious public health factor, still a single definition of the syndrome doesn't exist. This lack of a single definition and the nature of MetS that consists of coexisting diseases create discrepancies that complicate patient clustering and segmentation, making it a challenge to identify them and develop medical approaches. Additionally, real-world datasets frequently exhibit missing values, further complicating the accurate identification of patient subgroups and the monitoring of their associated risks.

This thesis addresses and is driven by the two above-mentioned major issues in the study of Metabolic Syndrome, namely:

1. The variability in diagnostic definitions and its impact on patient clustering, and
2. The challenges posed by missing data in real-world clinical datasets.

1.2. Research Structure

Initially, the real-patient data are explored and divided into meaningful groups with different characteristics. Then, the missing values are properly handled. At the next step, the four major definitions of MetS are translated into code and with their application on the data, each patient instance is marked as positive or negative with respect to MetS. Afterwards, clustering algorithms and combination of methods are utilized to group patients based on their clinical profiles. Testing variations of these processes on the data provides an overview of their performance and raises the questions and issues that real-world datasets pose.

2. Metabolic Syndrome

2.1. Metabolic Syndrome Foundations

When talking about Metabolic Syndrome (MetS from now on), a referral to a cluster of interrelated metabolic risk factors is mentioned. These factors are found to significantly increase the risk of cardiovascular diseases (CVD) and Type 2 Diabetes (T2DM) development. [1]

MetS, also called insulin resistance syndrome, is a cluster of various physiological and metabolic abnormalities including, among others, hyperinsulinemia, hyperglycemia, hypertension, a decreased plasma concentration of high-density lipoprotein (HDL) cholesterol, an increased plasma concentration of very low density lipoprotein (VLDL) triglyceride, glucose intolerance, and abdominal obesity. Insulin resistance is the primary metabolic defect of this syndrome, with compensatory hyperinsulinemia being the common denominator ultimately responsible for other changes of this constellation. In addition to insulin resistance, abdominal obesity is a key contributor to the development of MetS (Figure 1). [2]

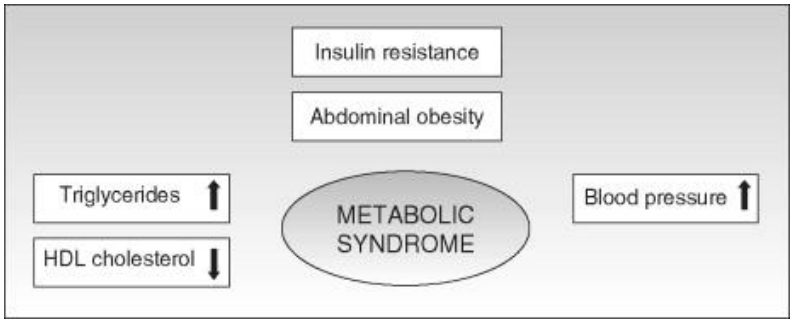

Figure 1 Metabolic Syndrome schematic approach according to (Keltikangas-Järvinen, 2007)

2.2. Metabolic Syndrome Epidemiology

The prevalence of metabolic syndrome has reached epidemic proportions globally, with variations observed across different populations and age groups, making it a significant driver of a current global cardiovascular crisis. [3] It drastically increases the risk of T2DM, CVD and premature death as well. The rising rates of obesity and sedentary lifestyle lead to metabolic derangements and the development of MetS among others. [4]

Originally emerging in the Western world, metabolic syndrome has now become a global issue due to the worldwide adoption of Western lifestyles. [5] In fact, the prevalence of metabolic syndrome is often higher among urban populations in some developing countries compared to those in the West. This widespread issue is primarily driven by two factors: the increased consumption of high-calorie, low-fiber fast food and the decline in physical activity due to mechanized transportation and sedentary leisure activities. Metabolic syndrome contributes significantly to the rising incidence of diseases such as T2DM, coronary artery disease, stroke, and other disabilities. The overall economic impact, including healthcare costs and loss of potential economic productivity, amounts to trillions of dollars.

The specific number of people with MetS can't be defined, since its nature and definition are not completely evident and straight forward, thus leaving a big space for interpretation and making it impossible to capture the whole picture. Moreover, many patients with one or more conditions that make up the MetS might not be diagnosed for the rest ones and not considered as MetS patients.

Some attempts have been made to get estimations about the prevalence of MetS, usually by utilizing T2DM data that are available, since this is one of the MetS outcomes. According to NHNES (National Health and Nutrition Examination Survey) data, during 1988–2010, average BMI in USA increased by 0.37% per year in both men and women and waist circumference (WC) increased by 0.37 and 0.27% per year in women, respectively. According to CDC (Centers for Disease Control and Prevention) data published in 2017, about 30.2 million adults aged 18 years or older or 12.2% of USA adults had T2DM. It is really interesting that one quarter of these people (23.8%) were not aware of having diabetes. Incidence of T2DM increased with age, reaching a high of 25.2% among US seniors (65 years or older). Prevalence of prediabetes or MetS was about three times more. Based on these data, about one third of US adults have metabolic syndrome. [6] Moreover, there is research done that places this percentage around 34% for the US population in 2012, leading to thoughts of further increase of it up to now. [7] When it comes to Europe, based on 2014 data, more than 4 million Europeans died of CVD with many more being hospitalized after acute episodes or treated for chronic cardiovascular ill health. [8] The MetS

estimation for Europe adult population ranges from 20% to 30% and is translated to approximately 100 to 130 million people possibly affected by MetS.

2.3. Metabolic Syndrome Definitions

Given the vague definition of MetS, many approaches exist that attempt to set specific and well-established criteria under which MetS is defined. Some of them are the following ones exhibited by the organization that set them and the year that this happened. [6]

- WHO (World Health Organization) – 1999:

Presence of insulin resistance or glucose > 6.1 mmol/L (110 mg/dl), 2h glucose > 7.8 mmol (140 mg/dl) (required) along with any two or more of the following:

1. HDL cholesterol < 0.9 mmol/L (35 mg/dl) in men, < 1.0 mmol/L (40 mg/dl) in women
2. Triglycerides > 1.7 mmol/L (150 mg/dl)
3. Waist/hip ratio > 0.9 (men) or > 0.85 (women) or BMI > 30 kg/m^2
4. Blood pressure > 140/90 mmHg

- EGIR (European Group for the Study of Insulin Resistance) – 1999:

Presence of insulin resistance (required) along with any two or more of the following:

1. Waist ≥ 94 cm (male), ≥ 80 cm (female)
2. Triglycerides ≥ 2.0 mmol/L (177 mg/dL) and/or HDL-C < 1.0 mmol/L (38.61 mg/dL) or treated for dyslipidemia
3. Blood pressure ≥ 140/90 mmHg or antihypertensive medication
4. Fasting plasma glucose ≥ 6.1 mmol/L (110 mg/dL)

- NCEP (National Cholesterol Education Program)-ATP III – 2005:

Presence of any three or more of the following:

1. Blood glucose greater than 5.6 mmol/L (100 mg/dl) or drug treatment for elevated blood glucose
2. HDL cholesterol < 1.0 mmol/L (40 mg/dl) in men, < 1.3 mmol/L (50 mg/dl) in women or drug treatment for low HDL-C
3. Blood triglycerides > 1.7 mmol/L (150 mg/dl) or drug treatment for elevated triglycerides
4. Waist > 102 cm (men) or > 88 cm (women)
5. Blood pressure > 130/85 mmHg or drug treatment for hypertension

- IDF (International Diabetes Federation) – 2006:

Waist > 94 cm (men) or > 80 cm (women) along with the presence of two or more of the following:

1. Blood glucose greater than 5.6 mmol/L (100 mg/dl) or diagnosed diabetes
2. HDL cholesterol < 1.0 mmol/L (40 mg/dl) in men, < 1.3 mmol/L (50 mg/dl) in women or drug treatment for low HDL-C

3. Blood triglycerides > 1.7 mmol/L (150 mg/dl) or drug treatment for elevated triglycerides
4. Blood pressure > 130/85 mmHg or drug treatment for hypertension

While there are many factors and limits stemming from the definitions, one can note that the NCEP and IDF definition are very similar except in the waist parameter of 102 vs. 94 cm in men and 88 vs. 80 cm in women.

Other organizations like the American Association of Clinical Endocrinologist (AACE) and the European Group for the Study of Insulin Resistance (EGIR) used slightly different definitions but they are not as commonly used.

Below, a tabular representation of the above definitions is exhibited. [1]

	NCEP ATP III (2005 revision)	WHO (1998)	EGIR (1999)	IDF (2005)
Absolutely required	None	Insulin resistance* (IGT, IFG, T2D or other evidence of IR)	Hyperinsulinemia‡ (plasma insulin >75th percentile)	Central obesity (waist circumference§): ≥94 cm (M), ≥80 cm (F)
Criteria	Any three of the five criteria below	Insulin resistance or diabetes, plus two of the five criteria below	Hyperinsulinemia, plus two of the four criteria below	Obesity, plus two of the four criteria below
Obesity	Waist circumference: >40 inches (M), >35 inches (F)	Waist/hip ratio: >0.90 (M), >0.85 (F); or BMI >30 kg/m^2	Waist circumference: ≥94 cm (M), ≥80cm (F)	Central obesity already required
Hyperglycemia	Fasting glucose ≥100 mg/dl or Rx	Insulin resistance already required	Insulin resistance already required	Fasting glucose ≥100 mg/dl
Dyslipidemia	TG ≥150 mg/dl or Rx	TG ≥150 mg/dl or HDL-C: <35 mg/dl (M), <39 mg/dl (F)	TG ≥177 mg/dl or HDL-C <39 mg/dl	TG ≥150 mg/dl or Rx
Dyslipidemia (second, separate criteria)	HDL cholesterol: <40 mg/dl (M), <50 mg/dl (F); or Rx			HDL cholesterol: <40 mg/dl (M), <50 mg/dl (F); or Rx
Hypertension	>130 mmHg systolic or >85 mmHg diastolic or Rx	≥140/90 mmHg	≥140/90 mmHg or Rx	>130 mmHg systolic or >85 mmHg diastolic or Rx
Other criteria		Microalbuminuria†		

*IGT, impaired glucose tolerance; IFG, impaired fasting glucose; T2D, type 2 diabetes; IR, insulin resistance; other evidence includes euglycemic clamp studies.
†Urinary albumin excretion of ≥20 μg/min or albumin-to-creatinine ratio of ≥30 mg/g.
‡Reliable only in patients without T2D.
§Criteria for central obesity (waist circumference) are specific for each population; values given are for European men and women.
Rx, pharmacologic treatment.

Table 1 Metabolic Syndrome definitions according to (Huang, 2009)

Each definition has its specific cutoff points and measures taken into consideration. In general, the NCEP ATP III definition is widely used due to its simplicity and inclusion of key features of metabolic syndrome, while the IDF definition has been criticized for its emphasis on obesity rather than insulin resistance. It is expected that when referring to various populations, the standards of the body weight and waist distributions will differ and this is a fact that is recognized by the IDF definition that takes different cutoff points based on the population in scope. Overall, the definition of MetS exhibit variations in the requirements for insulin resistance, obesity, and other criteria for diagnosing MetS, thus not providing a single framework when it comes to defining the exact conditions of MetS. [9]

2.4. Metabolic Syndrome Risk Factors

The development of MetS, regardless of the definition used to identify it, is in general based on a plenty of underlying risk factors that affect the patients' health and might be interconnected. According to existing research carried out on MetS risk factors, there are plenty categories that lead to its emergence. [10], [11], [12] A summary of the risk factors linked to MetS is the following.

- **Hypertension**
    - Elevated blood pressure is a significant component of MetS, contributing to the increased risk of cardiovascular diseases. It is often associated with other MetS components such as obesity and insulin resistance. [13]
- **Central Obesity**
    - Central obesity, particularly abdominal fat accumulation, is a critical risk factor for MetS. It is typically measured by waist circumference and is strongly linked to insulin resistance and dyslipidemia. The distribution of fat rather than the total amount is a crucial determinant of MetS.
- **Impaired Glucose Metabolism and Insulin Resistance**
    - Impaired fasting glucose or elevated blood glucose levels, along with insulin resistance, are central to the pathophysiology of MetS. These factors are precursors to type 2 diabetes mellitus (T2DM) and contribute significantly to the overall risk of CVD.
- **Dyslipidemia**
    - Dyslipidemia in MetS is characterized by elevated triglycerides and reduced high-density lipoprotein cholesterol (HDL-C). This lipid profile increases the risk of atherosclerosis and related cardiovascular conditions.
- **Sedentary Lifestyle and Poor Dietary Habits**
    - A sedentary lifestyle combined with poor dietary habits, including high intake of refined carbohydrates and saturated fats, exacerbates the risk factors for MetS. These lifestyle factors contribute to obesity, insulin resistance, and dyslipidemia. [14]
- **Age and Life Expectancy**
    - The prevalence of MetS increases with age, partly due to a natural decline in metabolic function and changes in body composition. As life expectancy increases globally, the burden of MetS is expected to rise, making age a non-modifiable but significant risk factor. [15]
- **Genetic Predisposition**
    - Genetic factors play a role in determining an individual's susceptibility to MetS. However, the interaction between genetics and environmental factors (such as diet and physical activity) is complex and not fully understood.
- **8. Socioeconomic Status**
    - Lower socioeconomic status is associated with a higher prevalence of MetS, possibly due to limited access to healthcare, healthy foods, and opportunities for physical activity. Socioeconomic disparities can lead to differences in MetS prevalence and outcomes.

Based on the above-mentioned risk factors, it is evident that MetS is mainly linked with indications regarding CVD problems, obesity (with the distribution of it being the most important factor, not just its existence) and insulin resistance. Many of the factors are related, with one being the indication of another's occurrence and vice versa. Additionally, the majority of the risk factors lead to three main diseases, namely cardiovascular diseases (CVD), type 2 diabetes mellitus (T2DM) and obesity. Another disease that is mentioned as linked with the risk factors is arteriosclerosis and the metabolic disorder of dyslipidemia.

2.4.1. Risk Factors Identification

In order to prevent and efficiently cope with any disease, timely identification of its occurrence or of the signs that are linked to it has to take place. In cases of diseases such as the MetS, where a plethora of possible risk factors exist, this process is significantly harder compared to other diseases and disorders. In fact, as already mentioned, MetS is a group of conditions and diseases. To identify someone as a MetS patient, many risk factors will have to be present according to the definition that is followed each time. An identification approach that takes into account the clustering of these risk factors seems to be the best option. More on these approaches will be mentioned and elaborated through this work. But even though the development of some risk factors may help in the early detection of MetS, the regular screening of health indicators is essential. Blood pressure, waist circumference, blood glucose levels, and lipid profile should be regularly monitored and recorded in order to enable an early detection of MetS.

2.4.2. Risk Factors Prevention and Management

Addressing the risk factors is key to preventing MetS and the diseases connected with it. At the core of this approach lifestyle interventions lie. Lifestyle tunings such as dietary changes, increased physical activity and weight management are crucial for preventing and even managing MetS. The everyday life plays a considerable big part in MetS and can play a decisive role in keeping the individual's levels of each risk factor in healthy levels. Starting from the working conditions that are nowadays mainly sedentary and lead to health problems and afterwards considering one's diet and physical exercise is important to maintain the MetS indicators in healthy levels. By adding more exercise in the daily routine and simple diet habits, meaningful results can be achieved in terms of overall health and MetS factors management. While the personal lifestyle choices of a person are crucial for its wellbeing, public health policies can also be an enabler towards a healthier population. The policies should aim on reducing the prevalence of MetS through community-based interventions that could address dietary habits, physical activity and socioeconomic factors. Such initiatives could make the difference when it comes to the general population's health. Paying attention and targeting the MetS risk factors should be considered in the context of the broader metabolic and cardiovascular health strategies, which will ensure that interventions are specific and effective in reducing the burden of MetS-related morbidity and mortality. [16]

2.5. Advancements in MetS research

The current development of research fields such as artificial intelligence (AI), drug development and biological systems modelling have empowered new approaches of many diseases and syndromes through innovative tools. MetS couldn't have been left out of this trend, with scientific and technological advancements being integrated in the identification and the management of it as well. [17]

2.6. MetS identification Methods

The identification of MetS has seen significant advancements in recent years, driven by the development of new biomarkers, genetic profiling, and the application of AI. These innovations have improved the accuracy and timeliness of MetS diagnosis, which is crucial for preventing its associated comorbidities, such as CVD and T2DM. There are two major categories in which the identification advancements could be divided.

1. Advanced Biomarkers and Genetic Profiling

Recent research has expanded the range of biomarkers used to identify MetS. Traditionally, diagnosis has relied on clinical criteria such as waist circumference, blood pressure, fasting glucose, triglycerides, and high-density lipoprotein cholesterol (HDL-C) levels. However, novel biomarkers like C-reactive protein (CRP), adiponectin, and advanced lipid profiles are now being utilized to enhance the precision of MetS diagnosis. These biomarkers provide insights into the underlying pathophysiological processes, such as inflammation and insulin resistance, that contribute to the syndrome. [18]

In addition to biomarkers, genetic profiling has emerged as a powerful tool in identifying individuals at risk for MetS. Specific genetic variants associated with MetS have been identified, offering the potential for personalized risk assessment and early intervention. The use of genome-wide association studies (GWAS) has been instrumental in uncovering these genetic predispositions, which, when combined with environmental factors, can significantly influence MetS development. [19]

When looking for specific examples of such research developments, the utilization of serum NMR metabolomics and the exosome analysis can be mentioned. The utilization of serum NMR metabolomics is a method that quantifies a broad spectrum of metabolites in serum, facilitating the identification of MetS by analyzing specific metabolic profiles. Through a detailed evaluation of metabolites, such as amino acids, lipids, and sugars, researchers can classify individuals based on their risk of developing MetS with high accuracy. The introduction of tools like MetSCORE, which combines metabolomics data with statistical modeling, has improved the predictive power for MetS risk, allowing for earlier and more personalized interventions. [20] Additionally, exosome analysis has emerged as a promising tool in the identification of MetS. Exosomes, small extracellular vesicles, are involved in intercellular communication and carry various biomolecules that reflect the metabolic state of cells. Recent studies have shown that exosomes can be used to detect early metabolic disturbances associated with MetS, offering a non-invasive method to monitor disease progression and response to treatment. This method represents a leap forward in understanding the pathogenesis of MetS and offers a novel biomarker for early diagnosis. [21]

2. Artificial Intelligence and Machine Learning (ML)

The integration of AI and ML into MetS identification represents a significant leap forward. These technologies are capable of analyzing vast datasets to detect patterns that may be indicative of MetS risk. AI-driven models can predict the likelihood of developing MetS based on a combination of traditional risk factors, biomarkers, and genetic data. Such approaches not only improve early detection but also allow for the stratification of patients based on their risk levels, enabling more targeted and efficient interventions. [22]

Moreover, AI tools are being used to refine diagnostic criteria and develop personalized diagnostic algorithms that account for individual variability in MetS manifestations. These innovations are critical for moving beyond the one-size-fits-all diagnostic approach, allowing healthcare providers to tailor screening and monitoring strategies to each patient's unique risk profile.

2.7. Management of MetS

The management of MetS has evolved with the introduction of new pharmacological treatments, lifestyle intervention strategies, and digital health technologies. These advancements aim to address the multifaceted nature of MetS, providing more comprehensive and effective approaches to reduce the burden of this syndrome.

1. Pharmacological Advances

Pharmacological treatment options for MetS have expanded beyond the traditional focus on individual components, such as hypertension or dyslipidemia, to address the syndrome as a whole. New drug classes, including GLP-1 (glucagon-like peptide-1) receptor agonists and SGLT2 (Sodium-glucose co-transporter-2) inhibitors, have shown promise in managing insulin resistance and obesity, two key components of MetS. These drugs not only improve glycemic control but also contribute to weight loss and cardiovascular protection, making them valuable in the comprehensive management of MetS. [23]

Furthermore, the use of combination therapies, where multiple pharmacological agents are used simultaneously to target different aspects of MetS, has gained traction. This approach is particularly effective in patients with advanced MetS, where single-agent therapies may be insufficient to control the complex interplay of metabolic abnormalities. Combination therapies have been shown to reduce the overall cardiovascular risk more effectively than monotherapies, thereby improving patient outcomes.

Moreover, lifestyle interventions, enhanced by digital health tools, have seen renewed emphasis in MetS management. Digital platforms now offer personalized diet and exercise plans based on individual metabolic profiles, tracked continuously through wearable technology. These interventions not only promote weight loss but also help maintain long-term metabolic health by providing real-time feedback and motivation to patients. This approach represents a significant shift towards personalized medicine, allowing for more effective and sustainable management of MetS. [24]

2. Integrated Lifestyle Interventions and Digital Health

Lifestyle modification remains a cornerstone in the management of MetS. Recent advancements in this area focus on integrating multiple lifestyle interventions into comprehensive programs that address diet, physical activity, and behavioral factors simultaneously. These programs often utilize digital health platforms to provide continuous support and monitoring, enhancing adherence and effectiveness. For example, digital tools can track dietary intake, physical activity levels, and medication adherence in real-time, offering personalized feedback and recommendations to patients. [25]

Nutrigenomics, the study of how individual genetic makeup affects response to diet, has also emerged as an innovative approach to lifestyle management of MetS. Personalized nutrition plans based on genetic profiling can optimize dietary interventions, ensuring they are tailored to the individual's metabolic response. This personalized approach can lead to more significant improvements in MetS parameters, such as weight loss and insulin sensitivity, compared to generic dietary guidelines. Based on this rationale, nutraceuticals have also gained attention for their role in managing MetS. Compounds like

omega-3 fatty acids, polyphenols, and probiotics have been shown to modulate metabolic pathways and reduce inflammation, providing an adjunctive benefit to traditional therapies. These developments highlight the growing trend towards incorporating natural compounds into MetS management protocols, offering a holistic approach to treatment. [26]

Lastly, the rise of telemedicine and remote patient monitoring has revolutionized the management of MetS. Telemedicine platforms allow for regular follow-ups and timely adjustments to treatment plans, reducing the need for frequent in-person visits. This is particularly beneficial for patients with mobility issues or those living in remote areas, ensuring they receive continuous care despite logistical challenges. [27]

3. Data

3.1. Data acquisition & information

Metabolic Syndrome is a widespread syndrome and, on this basis, one would expect that should be loads of available datasets to study it. Unfortunately, this is not the case with MetS. Finding a reliable and well-structured database to work on proved to be a tough task, with many research-oriented databases being discontinued, defective or outdated. This task was solved with the aid of a Greek university research team. Thanks to professor Manios Yannis, after the signing of an official Non-Disclosure Agreement between professor Matsopoulos George and him, we were able to acquire a real-patient dataset. The dataset mentioned is the GATEKEEPER Reference Use Case (RUC) 1 Greek Pilot Study dataset. This dataset was formed under the GATEKEEPER project [28] and the relevant research studies connected with it. [29]

GATEKEEPER's object is to build a European-led decentralized digital ecosystem, aiming to enable collaboration and provide mutually beneficial results to a multi-stakeholder ecosystem in Europe. The platform will provide evidence in real life, generate a set of first adopters and develop sustainability activities to maintain the project. Altogether, the project will empower the ageing citizens to keep themselves healthy with respect to optimal functional ability over time. GATEKEEPER will directly contribute to the United Nations Sustainable Development Goal which aims to "ensure healthy lives and promote well-being for all at all ages".

Regarding the RUC1 purpose, as stated in the relevant informative paperwork, it is targeted to the promotion of healthy lifestyles among elderly people to prevent and/or delay the onset and/or worsening of chronic conditions. RUC1 will be based on timely interventions provided by AI-based, digital coaches using Natural Language Processing techniques, structured conversations, and personalized feedback and education. Big Data Analytics techniques will be exploited to address risk stratification and early detection, based on lifestyles analysis including: pattern recognition for the improvement of public health surveillance and for the early detection of cognitive decline and frailty; data mining for inductive reasoning and exploratory data analysis; and, Cluster Analysis for identifying high-risk groups among elder citizens.

In order to achieve the above targets, data are collected from various countries including Greece according to some practices that will be explained further.

Each subject included in the study has provided health data, acquired by measurements carried out from hospital staff and wearable devices, usually with one-month intervals. These people that took part in the research were marked as being at risk of developing Metabolic Syndrome.

The IDs given to the people present in the study are 1112 according to the data received from the research team. They come with an annotation regarding groups they are divided to. These are:

- Test (20 instances)
- Control (318 instances)
- Dropout (3 instances)
- Intervention 1 (285 instances)
- Intervention 2 (268 instances)
- Not used (56 instances)

These groups add up to 950 instances, meaning that the rest IDs do not contain any information on the dataset provided and will be named as Missing, with a count of 162 instances. From this information, the following figure comes up regarding the study's subject groups.

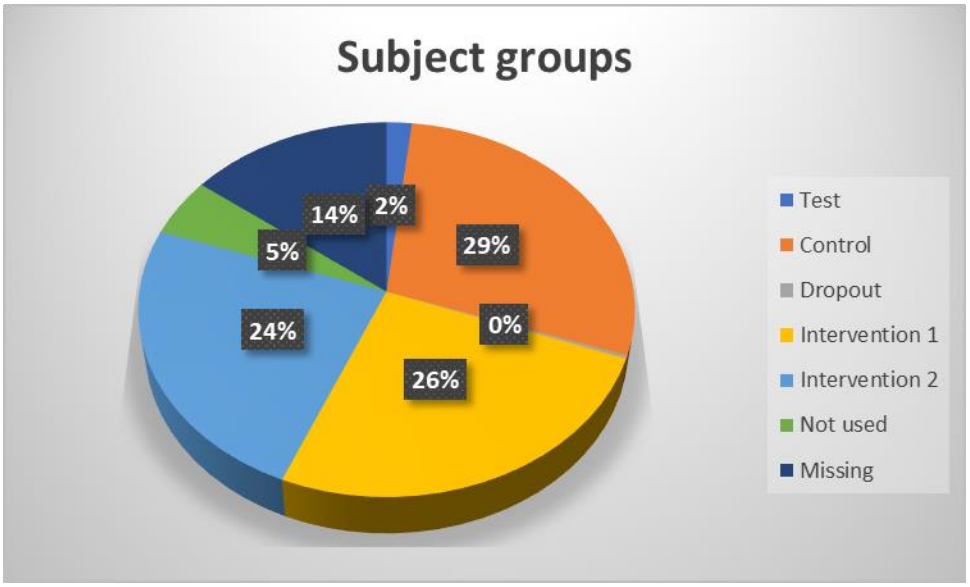

Figure 2 Study's subject groups

From the total number of subjects, only 871 individuals will be used in the following research work. These subjects are divided into three groups, namely the Control (C), the Intervention 1 (IN1) and the Intervention 2 (IN2) groups. The first, and a little larger than the other two groups, consists of the Control subjects, with a population of 318. The IN1 group is made up of 285 subjects and the IN2 group of 268 subjects.

3.2. Data Exploration

The three groups in scope (C, IN1, IN2) contain various information about each patient. These are not consistent in all three of them and a mapping of the information that each group contains should be done in order to define the next steps. With a first look, the measurement categories of each group can be easily spotted. The available health indicators for the whole dataset are 62. Some of them are

encoded according to the Logical Observation Identifiers Names and Codes (LOINC) standards, such as 3141-9 for example. LOINC is clinical terminology that is important for laboratory test orders and results, and is one of a suite of designated standards for use in U.S. Federal Government systems for the electronic exchange of clinical health information. The list of available health indicators in the dataset is presented below, in the form that they exist in the initial dataset provided by RUC1, including LOINC encodings and abbreviations that will be explained further.

Available Health Indicators			
15074-8-manual	bfmi	ggt	sgpt
3141-9	bmi	hba1c	skeletal_muscles
3141-9-manual	bmr	hct	sodium
39156-5	body_fat	hdl-cholesterol	steps
8462-4-manual	body_fat_percent	heart-rate-avg-manual	tbw
8480-6-manual	body_muscle	heart-rate-levels-cardio	tbw_percent
93829-0	body_muscle_percent	heart-rate-levels-fat-burn	temperature
93830-8	bone_mass	heart-rate-levels-out-of-range	total_cholesterol
93831-6	calories-resting-manual	heart-rate-levels-peak	total_hemoglobin_concentration
93832-4	cpk	heart-rate-resting	triglycerides
LP35925-4	creatinine	height	urea_level
album	creatinine_renal_clearance	ldl_cholesterol	uric-acid
albumin	crp	plt	visceral_fat
alkaline-phosphatase	ferritin	potassium	waiste_circumference
awake	ferrum	rbc	wbc
	ffmi	sgot	

Table 2 Available health indicators

In order to get a better understanding of the available data, the LOINC encodings had to be matched to the respective health metrics and understand the abbreviations the also exist. Moreover, the units of measurement should be identified for each category. After the relevant research and according to the documentation available from the GATEKEEPER project, the meaning and the units of measurement for each category are found and presented below. These measurements will be processed and assessed in the following steps of the present work to handle the issues they exhibit.

The methods through which the measurements are collected belong to two categories. The majority of measurements are taken with the aid of medical staff with one baseline value for each patient and then monthly follow-ups, that are recorded with the respective date that they were taken. The second category of measurements are acquired through continuous patient indicators monitoring using a fitness tracker. These patients belong to the IN2 group.

Health indicator dataset name	Health indicator	Unit of measurement
15074-8-manual	Blood Glucose (Manual)	mg/dL
3141-9	Body Weight	Kg
3141-9-manual	Body weight (Manual)	Kg
39156-5	Body Mass Index (BMI)	Kg/m^2
8462-4-manual	Diastolic blood pressure	mmHg
8480-6-manual	Systolic blood pressure	mmHg
93829-0	REM Sleep duration	minutes
93830-8	Light sleep duration	minutes
93831-6	Deep sleep duration	minutes
93832-4	Sleep duration	minutes
LP35925-4	BMI	Kg/m^2
album	Albumin protein concentration	g/L
albumin	Albumin protein concentration	g/L
alkaline-phosphatase	Alkaline phosphatase (ALP) enzyme concentration	µkat/L
awake	Awake period (mins)	minutes
bfmi	Body Fat Mass Index	Kg/m^2
bmi	BMI	Kg/m^2
bmr	Basic Metabolic Rate	kcal
body_fat	Body Fat	Kg
body_fat_percent	Body Fat Percentage	%
body_muscle	Body Muscle	Kg
body_muscle_percent	Body Muscle Percent	%
bone_mass	Bone Mass	Kg
calories-resting-manual	Calories Resting	kcal
cpk	Creatine phosphokinase (CPK) concentration	units/L
creatinine	Creatinine concentration	mg/Dl
creatinine_renal_clearance	Creatinine renal clearance	mmol\mol
crp	C-reactive protein (CRP) concentration	µg/mL
ferritin	Ferritin concentration	ng/mL
ferrum	Ferrum concentration	ng/mL
ffmi	Fat-Free Mass Index	kg
ggt	Gamma glutamyl transferase (GGT) concentration	µkat/L
hba1c	Glycated haemoglobin (HbA1c)	%
hct	Haematocrit	mg/dL
hdl-cholesterol	HDL cholesterol concentration	mg/dL
heart-rate-avg-manual	Average heart rate	bpm
heart-rate-levels-cardio	Heart rate levels cardio (duration)	minutes
heart-rate-levels-fat-burn	Heart rate levels fat burn (duration)	minutes
heart-rate-levels-out-of-range	Heart rate levels out of range (duration)	minutes
heart-rate-levels-peak	Heart rate leves peak (duration)	minutes
heart-rate-resting	Heart rate resting (duration)	minutes
height	Height	m
ldl_cholesterol	LDL cholesterol concentration	mg/dL
plt	Thrombocyte/Platelet count (PLT) concentration	x1000/µL
potassium	Potassium (K) concentration	mg/dL
rbc	Red blood cell count	number of cells
sgot	Serum Glutamic Oxaloacetic Transaminase (SGOT)	µkat/L
sgpt	Serum Glutamic Pyruvic Transaminase (SGPT)	µkat/L
skeletal_muscles	Skeletal muscles	Kg
sodium	Sodium (Na)	mg/dL
steps	Steps count	number of steps
tbw	Total Body Water	Kg
tbw_percent	Total Body Water percentage	%
temperature	Temperature	Celsius
total_cholesterol	Total cholesterol	mg/dL
total_hemoglobin_concentration	Total haemoglobin concentration	mmol/mol
triglycerides	Triglycerides	mg/dL
urea_level	Urea level	mg/dL
uric-acid	Uric acid level	mg/dL
visceral_fat	Visceral fat	no unit
waiste_circumference	Waist circumference	cm
wbc	White blood cell count	number of cells

Table 3 Available health indicators explanation

4. Data Preprocessing

4.1. General Preprocessing

In this work, Python version 3.11.9 is used to exploit the data and build the models that will be presented.

Up to this point, the categories of available data and the population in scope are established. In order to get a manageable form of the dataset and fetch the measurement categories, the initial raw data were processed. The raw form of the dataset consists of a CSV file with every unique measurement of a patient makes up a row along with the patient ID, the measurement's value and its timestamp (i.e. the exact time and date that the measurement was carried out). A view of the initial dataset follows.

	Unnamed: 0	patient_id	variable	value	timestamp
0	0	1	heart-rate-resting	78.0	2022/02/07 08:00
1	1	1	93831-6	45.0	2022/02/05 09:38
2	2	1	heart-rate-levels-peak	0.0	2022/02/05 22:14
3	3	1	heart-rate-levels-cardio	1.0	2022/02/05 22:14
4	4	1	heart-rate-levels-fat-burn	13.0	2022/02/05 22:14

Figure 3 Initial dataset view

First of all, since the measurement interval in scope is monthly, the timestamp variable has to be transformed to include month and year in order to be useful.

	patient_id	variable	value	timestamp
0	1	heart-rate-resting	78.0	2022-02-01
1	1	93831-6	45.0	2022-02-01
2	1	heart-rate-levels-peak	0.0	2022-02-01
3	1	heart-rate-levels-cardio	1.0	2022-02-01
4	1	heart-rate-levels-fat-burn	13.0	2022-02-01

Figure 4 Timestamp transformation

Once this is done, it is obvious that the dates of the measurements begin in year 2020 and continue afterwards, with patients being measured through other years as well. The minimum date available is spotted with it being 2020-02 (i.e. February 2020) and then all of the timestamps are calculated as the difference from this date. For example, if a timestamp is 2022-04 (April 2022), it will be transformed as 26, since it is 26 months after the minimum timestamp of February 2020. The maximum available timestamp is 35, translated to January 2023.

Next up is the exploration of the single patient data. As mentioned above, based on the documents that accompanied the raw dataset, there were a total of 1112 patient IDs. In contrast to that, when the raw data are explored, 1014 available patient IDs exist.

Following this finding, a view of the timestamps of each patient measured health indicators is needed to get a better understanding of the data. This will help in knowing for how much time were the patients monitored and enable a better selection of data.

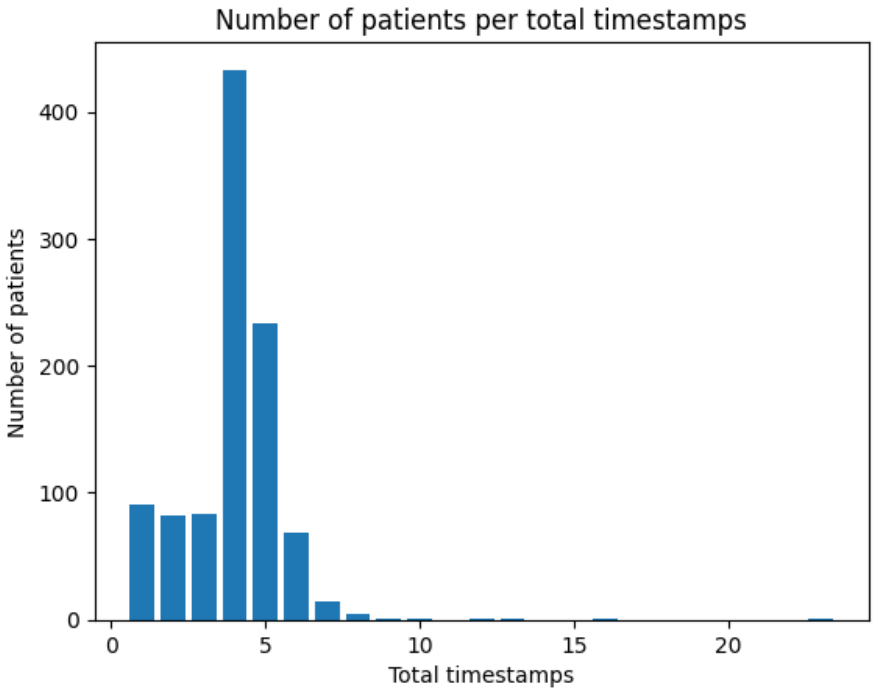

Figure 5 Number of patients per total timestamps

For example, if a patient's data were acquired in February 2020, March 2020 and April 2020, then the respective total timestamps value would be three (3), since there are 3 total available timestamps for this subject. From this process is calculated the exact number of patients that provide a specific number of measurements timestamps.

Total timestamps	Number of patients
1	91
2	82
3	83
4	433
5	233
6	68
7	14
8	4
9	1
10	1
12	1
13	1
16	1
23	1

Table 4 Number of patients per total timestamps

According to this finding, it makes sense to set a cutoff point at 4 timestamps. This selection will leave out patients that exhibit one, two and three timestamps (total of 256 individuals) but will include all the rest that have four or more timestamps (758 individuals).

The next step involves the exploration of the initial visit timestamp point of all the subjects, since they exhibit various timestamps. To get a view, the total subjects per initial visit timestamp are collected in a plot.

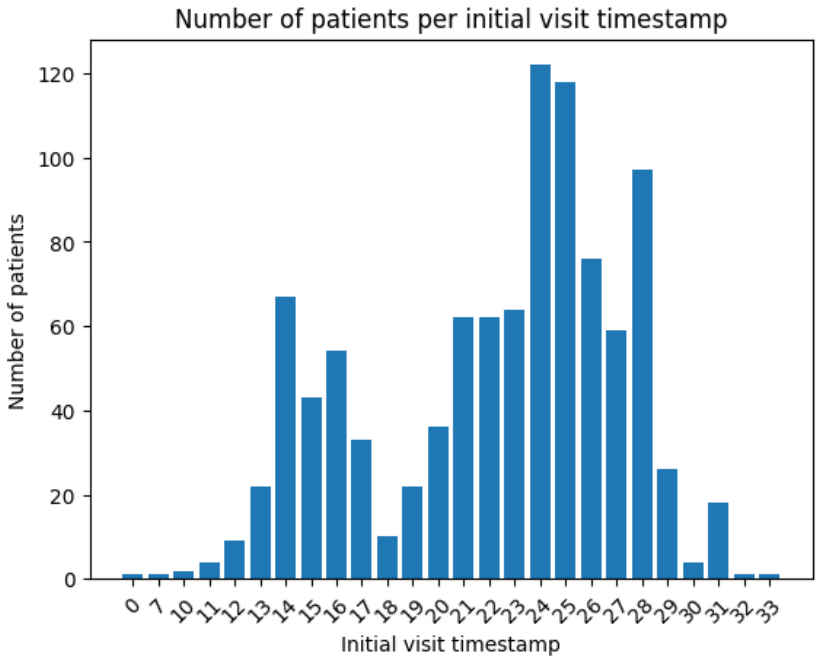

Figure 6 Number of patients per their initial visit timestamp

According to the results, there are plenty of initial visit points through the dataset. To handle the measurements in a universal and efficient way, all of the visits have to be initialized. Meaning that the starting point of all available measurements will be starting from the same timestamp. If for example a patient's visits are timestamps (16,17,18,19), they will be transformed to (0,1,2,3) based on the initialization process. Carrying out this initialization process for the timestamps that the measurements are acquired, a better view of the subjects' distribution can be gained to help with the timestamp selection that will be made afterwards.

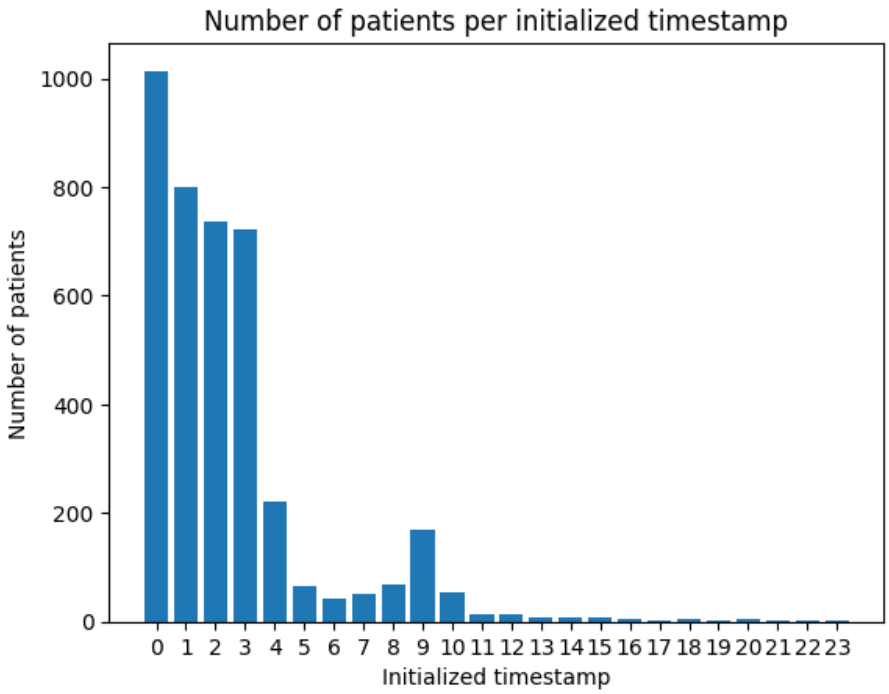

Figure 7 Number of patients per initialized timestamp

The data that make up the above figure are presented, since they play an important role in the timestamp inclusion for the whole approach.

Initialized timestamps	Number of patients	Initialized timestamps	Number of patients
0	1014	12	14
1	801	13	7
2	738	14	9
3	721	15	9
4	221	16	6
5	65	17	3
6	42	18	4
7	50	19	3
8	69	20	5
9	170	21	2
10	53	22	1
11	14	23	1

Table 5 Number of patients per initialized timestamp

The initialized timestamp calculated as zero is considered as the first measurement of every subject. As the numbers suggest, there should be a maximum of 721 subjects that are present through the 4 first visits. The maximum is mentioned, since it is not clear from this process whether all the 721 subjects that are present in the 4th timestamp have consecutive presence through the rest three timestamps that exist before of it. For example, one patient may have measurements for timestamps (0,1,3). At this case, this subject will not have data for all the intervals from timestamp 0 to 3 and if they are needed, it should be excluded. More on the handling of the subjects' data will follow.

Going back on the graph's outcomes, it is seen that after the 4th measurement timestamp (timestamp 4 and on according to the numbering), a drastic drop of the patients is exhibited, from 721 to 221. This characteristic leads to the establishment of a cutoff point at the 4th timestamp (timestamp number 3) to be able to include as much information as possible in this work.

To enable a meaningful use of the available data, their shape has to be redefined. The best way of utilizing them is a central matrix (dataframe in Python notation) in which all the available information is grouped per patient and timestamp. In this way, a comprehensive and easy to process form of the data is built. This matrix will have as columns all the available health indicators that appear through the dataset and the rows will consist of a subject's data acquired at a specific timestamp. Consequently, an issue arises. When a patient's health indicator measurement is not available at a specific timestamp, a NaN (Not a Number - indicating a missing value in Python) will appear at the respective matrix cell. For example, the data of patient 15 for the indicator 3141-9 might exist in timestamp 0,2,3 but not in timestamp 1 and a NaN will be shown in this cell. This missing data can be caused by a measurement actually not having been done at that point or being left out during the dataset formation. These measurements can't be found and will be handled as missing values. Almost every real-patient dataset exhibits missing values that can be exploited in various ways through its processing and have to be taken into consideration when they are in large numbers.

To explore the data availability of the health indicators present in the dataset, these indicators are plotted against their non-NaN percentages, i.e. the percentage of central matrix cells that contain values versus the overall available cells of each indicator. This percentage provides the available information in the dataset of each health indicator. The missing values are linked with information shortage and the indicators that do not include information should be left out, so that the models and results are not biased or misleading because of the missing values. The following figure displays the information percentages for the health indicators across the whole dataset.

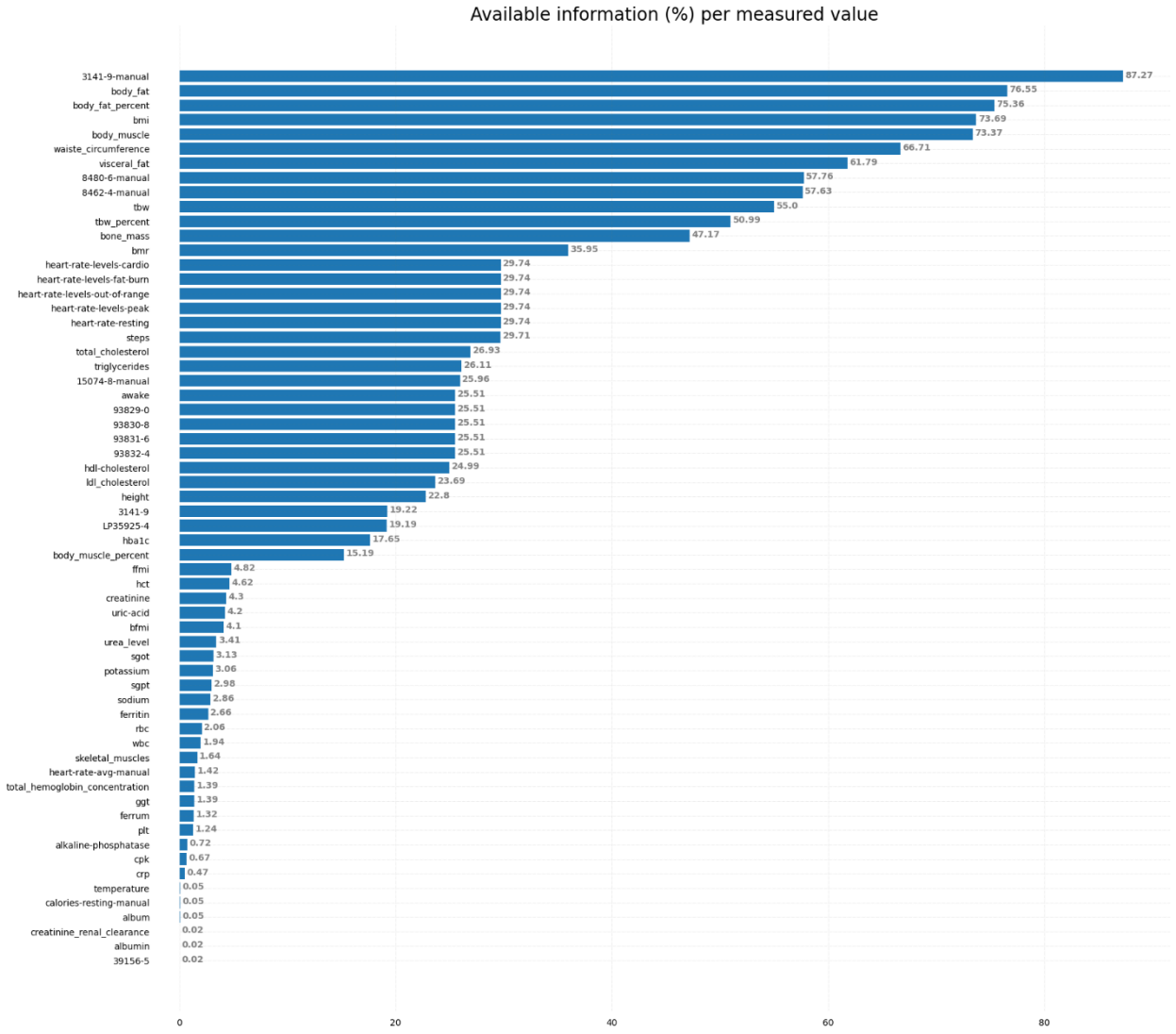

Figure 8 Available information (%) per measured value

It is obvious that there is a vast lack of information when the dataset is considered as a whole, meaning that Control, IN1 and IN2 are all taken into account for each and every health indicator. While some indicators are clearly near-zero filled with information, others' percentages are possibly marginal in this plot. The underlying reason can be found at the foundations of the dataset creation. As previously mentioned, (see Data Exploration section) IN2 subjects are the only ones that include fitness tracker monitoring indications. Subsequently, some indications are available only at specific subjects and when compared to the whole population they seem to have a big information loss that doesn't reflect the reality of the dataset.

4.2. Subject subgroups definition

Since the three patient groups that will be used have to be defined, the available data will be separated in them. An important parameter that should be also taken into consideration is that the subjects that will be used must have four consecutive months measurements and more specifically through the timestamps 0 to 3, aligned to the initialization that was carried out earlier.

After the calculations, the number of patients that present 3 consecutive monthly visits is 774 and these with 4 consecutive visits are 594. The subgroups population of interest for each category are shown in the following table.

Group	Subjects per Group	
	3 consecutive timestamps	4 consecutive timestamps
Control (C)	231	162
Intervention 1 (IN1)	204	133
Intervention 2 (IN2)	231	210

Table 6 Subject per group and consecutive timestamps

4.3. Four consecutive timestamps datasets

The three subgroups that are defined will be used for the cases that have four consecutive timestamps available. Each subgroup exhibits different capacity when it comes to the data included in it and the health indicators that are present. The respective figures that represent these differences are shown below. Each figure refers to a subgroup (Control – Intervention 1 – Intervention 2). Each figure is shown in a separate page for better visualization.

4.4. Control Group data availability

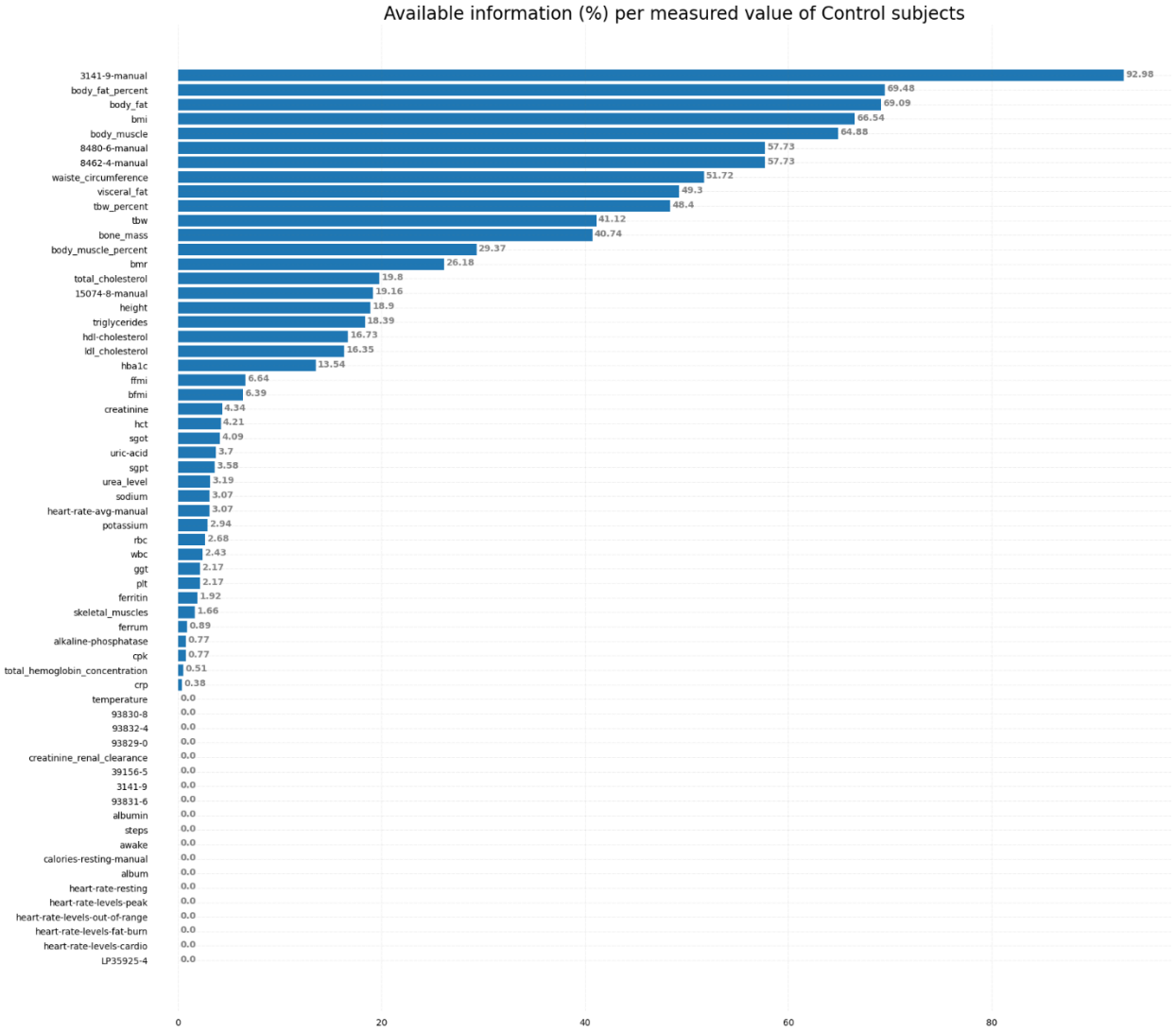

Figure 9 Available information (%) per measured value of Control group

4.5. Intervention 1 data availability

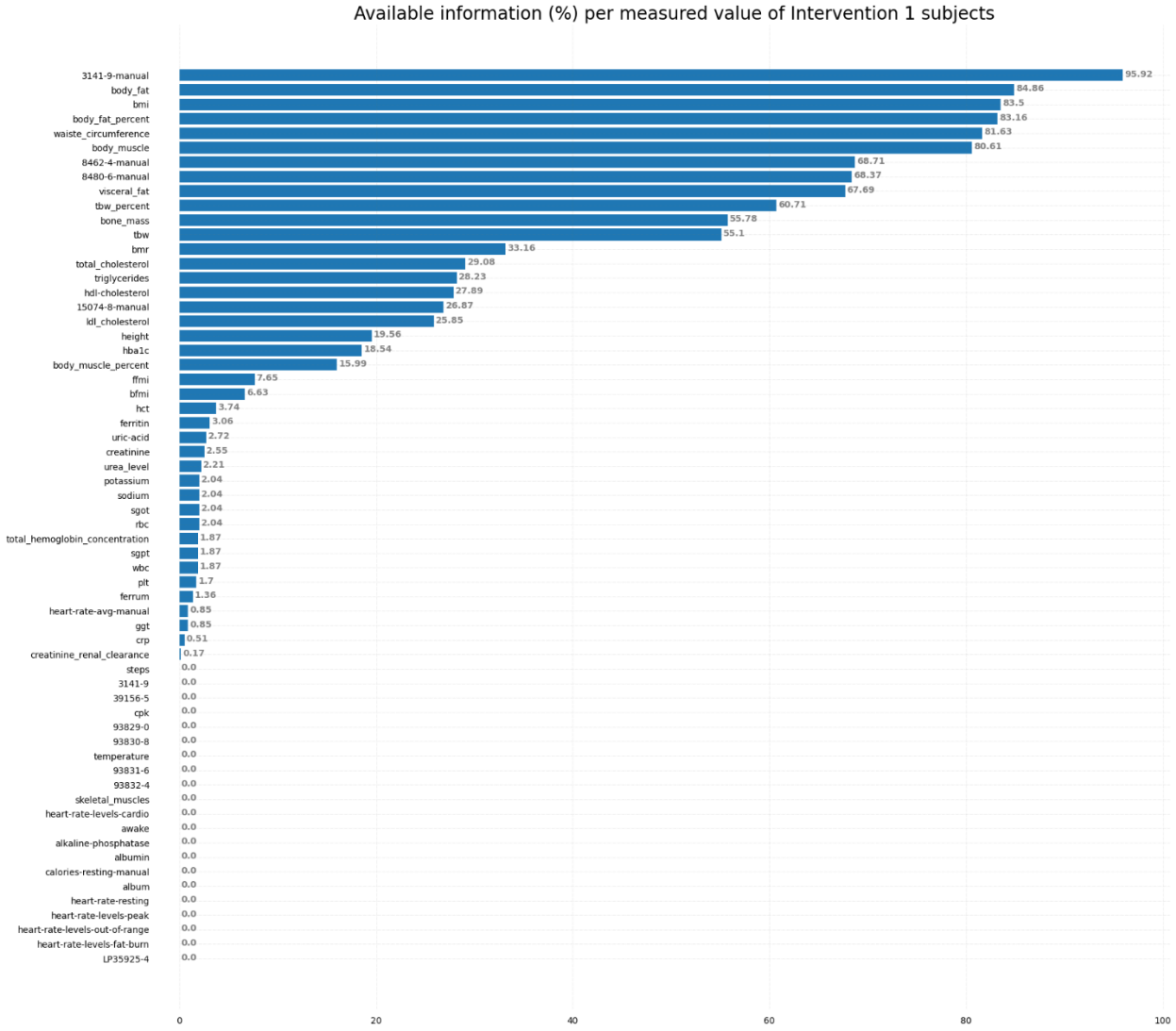

Figure 10 Available information (%) per measured value of Intervention 1 group

4.6. Intervention 2 data availability

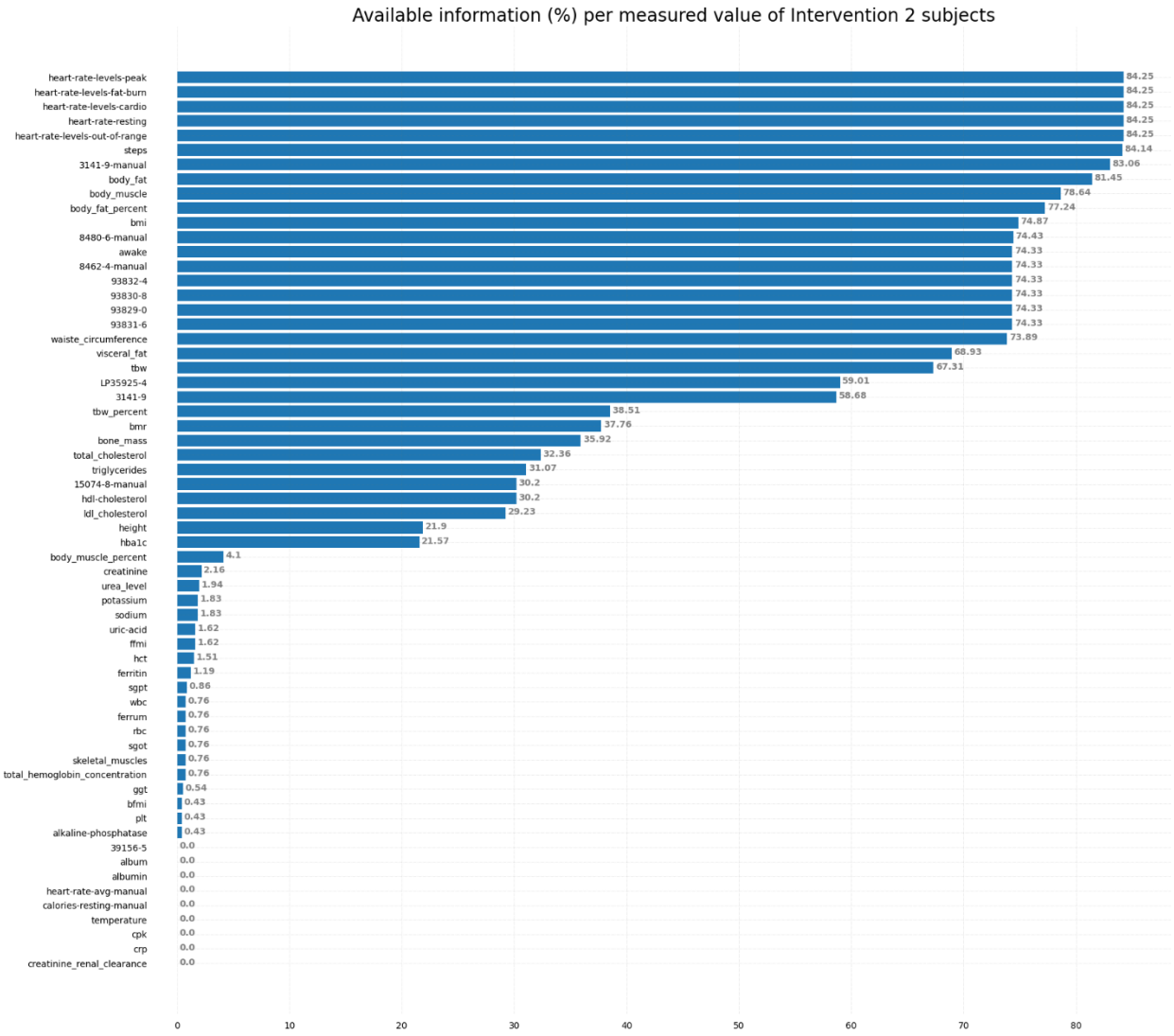

Figure 11 Available information (%) per measured value of Intervention 2 group

As displayed, the Control and Intervention 1 groups have similar distributions of the data availability. This can be explained by the fact that the fitness tracker is not utilized in these two groups and the health indicators that are acquired by the tracker are not present. On the other hand, Intervention 2 group comes with a significantly different distribution, especially on the measurements that are related to the fitness tracker, namely various heart rate levels, steps, awake time and also the sleep related measurements (93829-0, 93830-8, 93831-6, 93832-4). The existence of extended measurements in the Intervention 2 group makes it more preferable for being used as the main group for the current project. The methods used will be applied firstly on the Intervention 2 data and afterwards tested on the other groups as well.

4.7. Missing values

Missing values are always an issue with any dataset that includes real-life measurements. It couldn't be different in this case with real patients' health measurements. As shown above, the majority of the 62 health indicators existing in the dataset unfortunately prove to contain significant number of missing values.

If we were to set a missing value threshold at the initial data at 50%, it is impressive that only a small portion of the available data measurement categories satisfies this condition. This portion is the smallest at Control group, increases in IN1 group and reaches its highest value at IN2 group.

Health Indicators with above 50% data					
Control		**Intervention 1**		**Intervention 2**	
Indicator	**Available values**	**Indicator**	**Available values**	**Indicator**	**Available values**
3141-9-manual	93%	3141-9-manual	96%	heart-rate-resting	84%
body_fat_percent	69%	body_fat	85%	heart-rate-levels-peak	84%
body_fat	69%	bmi	84%	heart-rate-levels-out-of-range	84%
bmi	67%	body_fat_percent	83%	heart-rate-levels-fat-burn	84%
body_muscle	65%	waiste_circumference	82%	heart-rate-levels-cardio	84%
8462-4-manual	58%	body_muscle	81%	steps	84%
8480-6-manual	58%	8462-4-manual	69%	3141-9-manual	83%
waiste_circumference	52%	8480-6-manual	68%	body_fat	81%
		visceral_fat	68%	body_muscle	79%
		tbw_percent	61%	body_fat_percent	77%
		bone_mass	56%	bmi	75%
		tbw	55%	8480-6-manual	74%
				awake	74%
				93829-0	74%
				93830-8	74%
				93831-6	74%
				93832-4	74%
				8462-4-manual	74%
				waiste_circumference	74%
				visceral_fat	69%
				tbw	67%
				LP35925-4	59%
				3141-9	59%

Table 7 Health Indicators with above 50% data in the dataset

According to the matrix, it is proven that the data availability increases through the groups with Control having 8 indicators above 50%, IN1 having 12 and IN2 23.

4.7.1. Linear Interpolation

In order to achieve higher information existence in the dataset, a method to handle the missing values has to be utilized. The method selected based on the dataset characteristics is the linear interpolation. The method is the one suggested when it comes to timeseries with missing values. The dataset in scope satisfies the characteristics of a timeseries since the values are directly linked with consecutive health measurements through a specific time period (timestamp) for each patient.

In mathematics, linear interpolation is a method of curve fitting using linear polynomials to construct new data points within the range of a discrete set of known data points. If two known points are given

with their coordinates (x_1, y_1) and (x_2, y_2), then for a value x in the interval of (x_1, x_2) , the value y is given from the linear interpolation equation.

The linear interpolation form is the following:

$$Linear\ interpolation\ (y) = y_1 + (x - x_1)\frac{(y_2 - y_1)}{(x_2 - x_1)}$$

where,

- x_1 and y_1 are the first coordinates
- x_2 and y_2 are the second coordinates
- x is the point to perform the interpolation
- y is the interpolated value

In the use case of linear interpolation at the dataset, x is matched with the respective timestamps of the previous and next datapoints, while y is the health indicator value that is subject to the interpolation and has to be calculated.

When applied to the dataset, interpolation is used for each subject's data through the available timestamps. The Python command used is **DataFrame.interpolate(*method*='linear', *limit_direction*='both', *axis*=0)**.

In this way, the interpolation is applied both forward and backwards, meaning that a missing value can be calculated whether it is before or after existing real measurements.

After applying this method to the IN1 and IN2 datasets, they are filtered to two categories from timestamps 0 to 3 (4 timestamps in total). The one includes measurements that exhibit more than 70% of information (not NaNs) after the interpolation and the other includes measurements that exhibit more than 50% of information (not NaNs) and saved in order to be used in the following steps.

4.8. Baseline establishment

As mentioned already, the data that will be processed and used as a baseline for this project are the Intervention 2 data, since they provide a greater amount of information and health indicators than the rest of the data groups.

The IN2 group, after the linear interpolation and the filtering according to the 70% information existence threshold is defined by a total of 210 patients, with 29 available health indicator measurements for each one during their 4 consecutive measurements timestamps. The categories available and their translation to health indicators are shown below.

Intervention 2 available data over 70%		
Health indicator dataset name	**Health indicator**	**Unit of measurement**
15074-8-manual	Blood Glucose (Manual)	mg/dL
3141-9	Body Weight	Kg
3141-9-manual	Body weight (Manual)	Kg
8462-4-manual	Diastolic blood pressure	mmHg
8480-6-manual	Systolic blood pressure	mmHg
93829-0	REM Sleep duration	minutes
93830-8	Light sleep duration	minutes
93831-6	Deep sleep duration	minutes
93832-4	Sleep duration	minutes
LP35925-4	BMI	Kg/m^2
awake	Awake period (mins)	minutes
bmi	BMI	Kg/m^2
body_fat	Body Fat	Kg
body_fat_percent	Body Fat Percentage	%
body_muscle	Body Muscle	Kg
hdl-cholesterol	HDL cholesterol concentration	mg/dL
heart-rate-levels-cardio	Heart rate levels cardio (duration)	minutes
heart-rate-levels-fat-burn	Heart rate levels fat burn (duration)	minutes
heart-rate-levels-out-of-range	Heart rate levels out of range (duration)	minutes
heart-rate-levels-peak	Heart rate leves peak (duration)	minutes
heart-rate-resting	Heart rate resting (duration)	minutes
height	Height	m
ldl_cholesterol	LDL cholesterol concentration	mg/dL
steps	Steps count	number of steps
tbw	Total Body Water	Kg
total_cholesterol	Total cholesterol	mg/dL
triglycerides	Triglycerides	mg/dL
visceral_fat	Visceral fat	no unit
waiste_circumference	Waist circumference	cm

Table 8 Available health indicators in Intervention 2 group above 70%

In order to establish a baseline that could be used a reference point, the various definitions of MetS presented in the beginning are used (see Metabolic Syndrome Definitions paragraph). Each definition utilizes some of the health indicators. There is overlapping between three of the definitions, but none of them is identical to the others since each one sets different absolutely required conditions and also different limits for the values in scope. The four definitions that are applied to the available data are the:

1. World Health Organization definition (WHO)
2. European Group for the Study of Insulin Resistance definition (EGIR)
3. National Cholesterol Education Program - ATP III definition (NCEP)
4. International Diabetes Federation definition (IDF)

The data needed for each definition (see Table 1) exist in the Intervention 2 form under investigation after the thresholding and health indicators selection. Their existence enables the test of all four definitions on the data. Still, plenty of missing values exist in the dataset, even after the interpolation carried out to fill them and this is going to have an impact on the application of the definitions on the patients' data.

In this step, the available information about the gender of each patient that was provided along with the initial data is incorporated. This is important since the MetS definitions have different limits on male and female populations.

The four definitions are going to be applied to each timestamp available in the dataset. In this way, a patient can be marked as having the MetS or not at each point his or her measurements were taken. So, there might be cases of patients developing the syndrome while being monitored, or others that might have it at the beginning and improve their health ending up not being labeled as MetS patients during the monitoring period of four measurements.

Based on the methodology followed, the definitions are applied according to their conditions and each one provides an image of whether the timestamps data fulfill its criteria or not. In case there are missing data that should be included in the criteria, it is considered that the patient does not fulfill the MetS conditions.

MetS Definitions Indicators			
WHO	**EGIR**	**NCEP**	**IDF**
Blood Glucose (Manual)	Blood Glucose (Manual)	Blood Glucose (Manual)	Blood Glucose (Manual)
HDL cholesterol concentration	HDL cholesterol concentration	HDL cholesterol concentration	HDL cholesterol concentration
Triglycerides	Triglycerides	Triglycerides	Triglycerides
Visceral fat	Diastolic blood pressure	Diastolic blood pressure	Diastolic blood pressure
BMI	Systolic blood pressure	Systolic blood pressure	Systolic blood pressure
Diastolic blood pressure	Waist circumference	Waist circumference	Waist circumference
Systolic blood pressure			

Table 9 Metabolic Syndrome Definitions Indicators

As mentioned already, the EGIR, NCEP and IDF definitions utilize the same metrics but their way of providing required levels and limits of each indicator vary. When these definitions are applied to the data of four consecutive measurements (timestamps 0 to 3) for the 210 subjects of the Intervention 2 dataset, the following identification of timestamp measurements that should be considered as MetS occurs. The definitions are applied per measurement group, so they are 840 in total (4 for every single subject).

	MetS instances identification per Definition			
	WHO	**EGIR**	**NCEP**	**IDF**
MetS instances	44	207	164	173
Non-MetS instances	796	633	676	667

Table 10 Metabolic Syndrome instances per definition

The results show that WHO definition is probably stricter, with only 44 instances identified as MetS. On the other hand, EGIR definition identifies the most MetS instances adding up to 207. Since there is no

universally accepted definition and the missing data makes it hard to get the whole view of the dataset, a combination of the four definitions is done. Each identified MetS instance is taken into account as correctly identified and they are calculated across the whole Intervention 2 dataset's instances. The total amount of unique instances identified at MetS based on one of the definitions given is 252. Thus, leaving us with 588 non-MetS incidents.

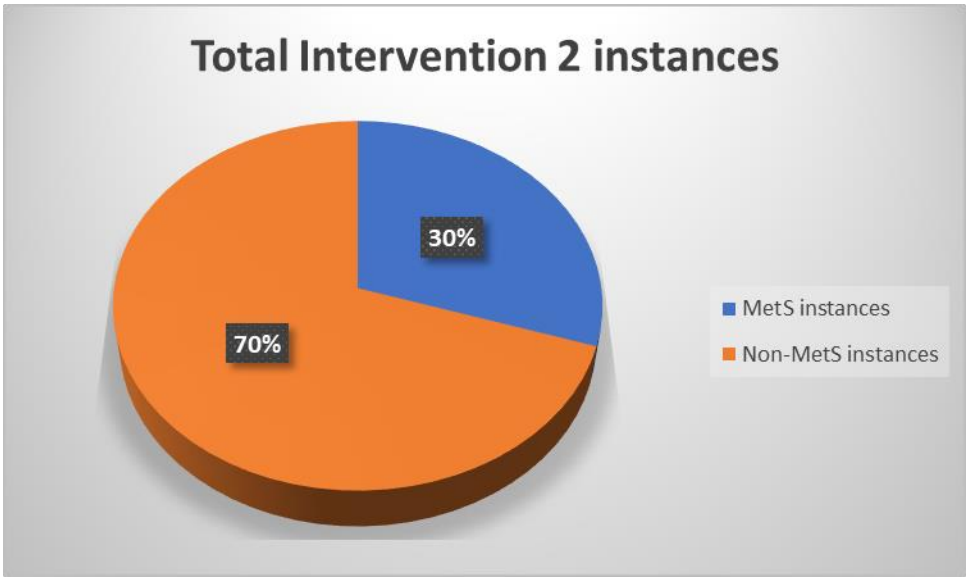

Figure 12 Total identified instances of MetS in Intervention 2

5. Unsupervised Machine Learning Methods

5.1. Clustering

Clustering is an unsupervised machine learning technique designed to group unlabeled examples based on their similarity to each other. In case the examples are labeled, this kind of grouping is called classification. Clustering is used to define groups of data points that exhibit similarities within each group and are distinct from the others.

This method is the most appropriate to apply on the data available at this stage. Each patient's timestamp is up to this point marked as MetS instance or non-MetS instance based on the various definitions. The clustering method is going to be applied to all of the data points and will be the means towards a distinguishment of two groups. One will be the MetS instances and the other the non-MetS. After clustering the data, a comparison will be made between the MetS instances based on the definitions and the clusters that will be formed. This approach resembles a classification problem.

The dataset in hand consists of multifactorial data points. This means that each instance subject to clustering consists of many measurements. These measurements are the factors that define the data to be used in the clustering technique. To assess and get an understandable representation of the clustering results, Pair Plot Representation figures will be used. Pair Plot Representation is the ideal way of illustrating multifactorial clustering results.

A pair plot is utilized to display the clustering structure across multiple variables by plotting each variable against every other variable in a grid of scatter plots. Each cell in the grid represents the relationship between two features, revealing patterns, correlations, and separability of clusters. Diagonal elements display the distribution of the individual features, while off-diagonal plots depict pairwise interactions between variables.

When used to represent multifactorial clustering, the pair plot highlights how clusters, made up of multiple factors, are distributed in the feature space. By coloring data points according to cluster assignments, it becomes possible to observe:

- Cluster separability: how distinct the clusters are across different pairs of features.
- Overlaps and boundaries: areas where clusters overlap or diverge.
- Feature relevance: which variables or combinations are most influential in distinguishing clusters.

Thus, the pair plot provides a comprehensive overview of the relationships between features and the structure of the clustering results, aiding in the interpretation of the multidimensional data and their underlying relationships.

5.1.1. K-means Clustering Algorithm

In the situation under investigation, K-means algorithm will be used to apply the clustering process. K-means clustering is an unsupervised learning algorithm used for data clustering, which groups unlabeled data points into groups or clusters. K-means clustering assigns data points to one of the K clusters depending on their distance from the center of the clusters. It starts by randomly assigning the clusters centroid in the space. Then each data point assign to one of the clusters based on its distance from the centroid of the cluster. After assigning each point to one of the clusters, new cluster centroids are assigned. This process runs iteratively until it finds good cluster.

In case K is not clearly defined, the optimal number of K should be defined. K-Means clustering performs best with data that are well separated. K-Means is faster as compare to other clustering techniques. It provides strong coupling between the data points. Different initial assignment of cluster centroid may lead to different clusters. K-means is applied on the Intervention 2 (IN2) dataset in order to uncover possible clustering of the instances according to the measurements in scope each time that could be similar to the MetS definitions results.

5.1.2. Spectral Clustering Algorithm

K-means clustering is one the most known and used clustering algorithms. It is used in many cases and did really indicate that some of the data selections carried out above exhibit better results than others. On this rationale, the most consistent data of the previous clustering applications will be used as input in a more sophisticated algorithm, which is Spectral Clustering.

Spectral Clustering is a graph-based clustering method that transforms data into a lower-dimensional space using the eigenvectors of the Laplacian matrix. This technique is useful for complex or non-linear data that are not well-separated in the original high-dimensional space. The dataset under investigation fits the latter data description.

Spectral clustering and K-means clustering are both clustering methods, but they differ in their approach. K-means clustering assigns data points to the nearest centroid in a low-dimensional space, while Spectral Clustering first embeds the data points into a lower-dimensional space using the spectrum of an affinity matrix and then applies a clustering algorithm to the embedded data points. In the application presented, the affinity matrix is built using a radial base function kernel.

Application of Spectral Clustering:

1. First, an undirected graph G = (V, E) with vertex set V = $\{v_1, v_2, \dots, v_n\}$ = 1, 2, …, n observations in the data is created. Going forward, a parameter epsilon is fixed beforehand.
2. Then, each point is connected to all the points which lie in its epsilon-radius. If all the distances between any two points are similar in scale, then typically the weights of the edges i.e. the distance between the two points are not stored since they do not provide any additional information. Thus, in this case, the graph built is an undirected and unweighted graph.
3. An adjacency matrix for this graph is constructed with $A_{ij} \rightarrow 1$ when the points are close and $A_{ij} \rightarrow 0$ if the points are far apart.

Close data points are in the same cluster. Data points in different clusters are far away. But data points in the same cluster may also be far away - even farther away than points in different clusters. The goal then is to transform the space so that when 2 points x_i, x_j are close, they are always in same cluster, and when they are far apart, they are in different clusters. The Gaussian Kernel K is directly used for this purpose through the following equation.

$$A_{ij} = \exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right)$$

4. The next step is the construction of the Graph Laplacian. This is another matrix representation of a graph, but it comes with some advantages. In particular, it can be used to construct low-dimensional embeddings (which is why is needed in this application). There are many Laplacians that could be constructed. In constructing all of these, a diagonal matrix D is inevitably built (a matrix where only the principal diagonal elements are nonzero).

$$D_{i,i} = \sum_{j=1}^{n} A_{ij}$$

Therefore, each diagonal element is simply the sum of the corresponding row of the affinity matrix. Some Laplacians that could be constructed using this matrix are:

- Simple Laplacian: $L = D - A$
- Normalized Laplacian: $L_N = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$
- Generalized Laplacian: $L_G = D^{-1} L$

- Relaxed Laplacian: $L_\rho = L - \rho D$
- Ng, Jordan, and Weiss Laplacian: $L_{NJW} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$

The whole purpose of computing the Graph Laplacian L is to find eigenvalues and eigenvectors for it, in order to embed the data points into a low-dimensional space. To identify good clusters, Laplacian L should be approximately a block-diagonal, with each block defining a cluster.

5. At the final step, K-means clustering is used. For K clusters, the first K eigen vectors are computed $(v_1, v_2, \dots, v_k)$. The vectors are vertically stacked to form the matrix with eigen vectors as columns. Every node is represented as the corresponding row of this new matrix and these rows form the feature vector of the nodes. K-means is utilized to cluster these points into k clusters $C_1, C_2, \dots, C_k$.

The described method will be used on the two datasets that exhibited robust clustering before and after normalization. These are the whole of IN2 dataset, without the fitness tracker data and the Sleep data from the fitness tracker. Spectral Clustering will be applied on both of them, at their raw format but at their L1-Normalized format as well.

5.2. Normalization

5.2.1. L1 Normalization

L1 normalization, also known as Least Absolute Deviations (LAD) or Manhattan Norm, is a technique used to normalize data. It involves transforming the data such that the sum of the absolute values of the vector (like a column in a dataset) is equal to 1.

L1 Normalization is used and preferred in comparison to L2 Normalization since:

1. L1 normalization is beneficial when dealing with sparse data (data with many zeros). It is able to help in preserving the sparsity of the data, which is often desirable in high-dimensional data scenarios like the one in scope.
2. Due to its nature of taking the absolute values, L1 normalization is less sensitive to outliers compared to L2 normalization. This makes it a suitable choice in datasets where outliers are present and should not dominate the feature's importance. [30]

The mathematical formula of L1 Normalization is the following:

$$x_{L1\ Normalized} = \frac{x}{\sum abs(x)}$$

These characteristics of L1 Normalization technique make it the most appropriate normalization method to apply on the available dataset. All of the data groups will be clustered after L1 Normalization as well in order to observe the results and identify groups that perform better than the others when it comes to correctly clustering the available dataset (i.e. classifying the data the closest possible to what the MetS definitions indicate).

5.2.2. MinMax Scaler Normalization

This scaler is used to transform features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range provided by the user. The range that will be used in the data in scope is [0,1]. MinMax scaler doesn't reduce the effect of outliers, but it linearly scales them down into a fixed range, where the largest occurring data point corresponds to the maximum value and the smallest one corresponds to the minimum value.

5.3. Dimensionality Reduction

5.3.1. Principal Component Analysis

Principal Component Analysis is one of the easiest, most intuitive, and most frequently used methods for dimensionality reduction, projecting data onto its orthogonal feature subspace. Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Reducing the number of variables of a dataset naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Smaller datasets are easier to explore and visualize, and thus make analyzing data points much easier, faster and efficient without extraneous variables to process. The idea of PCA can be summed up to reduction of the number of variables of a dataset, while preserving as much information as possible.

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. PCA can be explained through five steps. [31]

1. Standardization

   The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

   More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges and it will eventually lead to biased results. So, transforming the data to comparable scales can prevent this problem.

   Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable through the following equation.

$$z = \frac{value - mean}{standard\ deviation}$$

   After the standardization, all the variables will be transformed to the same scale.

2. Covariance Matrix computation

   This step aims to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because

sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, the covariance matrix is built.

The covariance matrix is a $p \times p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables x, y, and z, the covariance matrix is a 3×3 data matrix of the form:

$$\begin{matrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{matrix}$$

Since the covariance of a variable with itself is its variance, $Cov(a,a) = Var(a)$, in the main diagonal (top left to bottom right) the values are actually the variances of each initial variable. Moreover, since the covariance is commutative, $Cov(a,b) = Cov(b,a)$, the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal. The important characteristic in such matrices is the sign of the covariance. In case the sign is positive, it means that the two variables are correlated (i.e. when the one increases or decreases, the other follows the same). In case it is negative, then the two variables are negatively correlated (i.e. when the one increases, the other decreases and vice versa).

3. Computation of the eigenvectors and eigenvalues of the covariance matrix to identify the principal components

Eigenvectors and eigenvalues need to be computed from the covariance matrix in order to determine the principal components of the data. Their number is equal to the number of dimensions of the data. For example, for a 3-dimensional dataset, there are 3 variables, therefore there are 3 eigenvectors with 3 corresponding eigenvalues.

Eigenvectors and eigenvalues are behind all the principal component effectiveness because the eigenvectors of the covariance matrix are actually the directions of the axes where there is the most variance (most information) and that we call Principal Components. And eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component.

By ranking the eigenvectors in order of their eigenvalues, highest to lowest, the Principal Components arise in order of significance.

4. Formation of a Feature Vector

In this step the choice of keeping all these components or discard those of lesser significance (of low eigenvalues) is carried out. With the remaining ones, a matrix of vectors called Feature Vector is built.

The Feature Vector is simply a matrix that has as columns the eigenvectors of the components that is decided to be kept. This makes it the first step towards dimensionality reduction, because since only p eigenvectors (components) are selected to be kept out of n, the final data set will have only p dimensions. It is up to the problem setting whether all the components are kept or the ones of lesser significance are discard, depending on the needs.

5. Recasting the data along the Principal Components axes

Through the previous steps, apart from standardization, no changes were done on the data. The calculation and selection of the Principal Components and the formation of the Feature Vector are carried out, but the input dataset remains intact and follows the original axes (i.e. in terms of the initial variables).

In the last step, the aim is to use the Feature Vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the Principal Components (hence the name Principal Components Analysis). This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

$$PCA\ Dataset = FeatureVector^T * StandardizedOriginalDataset^T$$

While PCA is proven to be a powerful tool for dimension reduction, in order to be useful in investigating the MetS dataset available, it will have to be combined with a clustering method that will follow the PCA application on the dataset. In this way, the multiple measurement categories available will be decreased to the respective number of Principal Components that will be defined through the method.

PCA is meaningful when many measurements are present since it is all about dimensionality reduction. For this reason, it will be applied in three cases where multiple data factors are present. These will be:

a. All IN2 data
b. IN2 data excluding the fitness tracker data
c. IN2 fitness tracker data

For each case, the Explained Variance by Components is calculated in order to select the suitable number of components that will be used in the PCA. Afterwards, these components will be clustered testing the K-means and the Spectral Clustering algorithms.

The Cumulative Explained Variance plot is a graphical representation that shows the proportion of the dataset's variance that is cumulatively explained by each component. When PCA is performed, the data are transformed into a new coordinate system with axes ranked by how well they capture the variance in the data. Each axis (Principal Component) can explain a certain amount of the variance.

The plot usually starts with the variance explained by the first principal component on the left. Each subsequent component adds to this cumulative value. Ideally, it is desired to choose a number of components such that you can capture a high percentage of the total variance with as few components as possible, which means a simpler model and less computational expense. [32]

Visually, the Cumulative Explained Variance plot often shows a sharp turn or "elbow," indicating the point at which adding more components has diminishing returns in terms of explained variance. A rule of thumb is to select the number of Principal Components that reach over 80% of Cumulative Explained Variance and this is the way that PCA will be performed through the following cases.

Additionally, the representation of the clustering of the two most important Principal Components and the points respective clustering is presented for each PCA result. In the plots, K-means is presented, while Spectral clustering results will also be shown at the final results section.

5.4. Classification

5.4.1. Correctly classified (clustered) data

As far as correctly clustered is considered, this is a metric that refers to the proportion of data that are separated according to the MetS definitions outcomes on the dataset. The positive outcome is when the data are correctly separated. The following example provides the way the correct clustering is defined.

Ten datapoints are assumed with their MetS definitions results and their clustering results. MetS definitions results indicate whether they are indicated as MetS patients, while clustering results refer to the datapoints belonging either to Cluster 0 or Cluster 1 according to the algorithm results. Consequently, the following information is available:

Datapoints	MetS definitions results	Clustering method results
1	0	1
2	0	1
3	1	0
4	0	0
5	1	0
6	0	0
7	1	1
8	0	1
9	0	1
10	0	0

Table 11 Correctly clustered data - example

According to it, if the MetS definitions are considered as the ground truth and clustering goal, there are two groups of data. The one consists of the data that are clustered accordingly to the MetS definition, e.g. Datapoint 4, is 0 at the MetS definition and 0 at the Clustering as well and the other group that is made up from the data that are in the opposite cluster than the MetS definition, e.g. Datapoint 1 that has a MetS definition of 0 and belongs to the Cluster 1. Each of this groups can be considered correctly clustered, since in the first case, Cluster 1 and 0 are matched with the MetS definition results, while in the second one, the Clusters formed have just the opposite annotation of the MetS definitions results. So, in this case group 1 includes four datapoints (datapoint 4, 6, 7, 10), while group 2 includes six datapoints (datapoint 1, 2, 3, 5, 8, 9). The group with the most correctly separated datapoints makes up the “Correctly clustered” metric and indicates the percentage of the correctly separated datapoints over the overall available datapoints of each clustering attempt.

6. Results

6.1. Intervention 2 K-Means Clustering (without Normalization)

A number of dataset alternatives and normalization methods will be used based on the Intervention 2 data to address the clustering problem. As the technique is applied and assessed, specific data might pose problems to the cluster’s formation and data separation. These data might have to be left out from

the process. On the other hand, normalization methods could be proven really helpful when handling data with different scales and will be also tested.

6.1.1. All IN2 dataset categories above 70% – filling the NaN with zero

At the beginning, all of the available data points of Intervention 2, with above 70% information availability, are used to get an idea whether the separation in two clusters is possible. To handle the remaining missing values, even after the interpolation is applied, these measurements that are still missing are filled with zeros. So, if for example a heart rate measurement was not available, it will be set to zero.

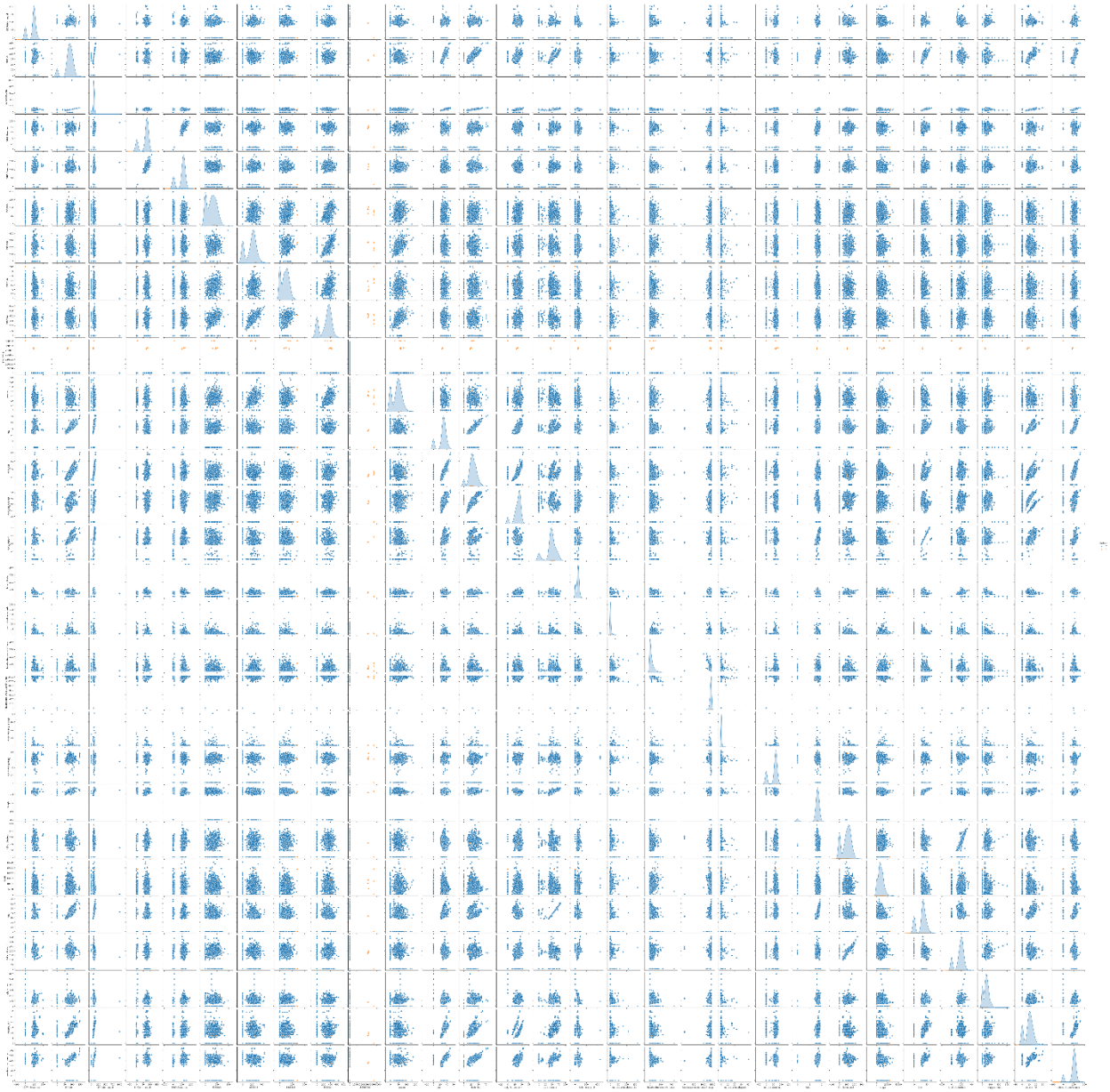

Figure 13 Pair plot of IN2 all categories – filling the NaN with zero

It is obvious that this approach doesn't provide meaningful results due to the existence of outliers that form a group and make it impossible to get a distinction of two bigger groups.

6.1.2. All IN2 dataset categories above 70% – dropping the NaN instances

The next approach that involves all of the available data points of Intervention 2 with above 70% information availability is to remove the instances that include NaN in their data. In this way, the zeros will be decreased compared to the initial clustering attempt. In this case, if for example a heart rate measurement was not available, the whole instance will be omitted from the dataset subject to clustering.

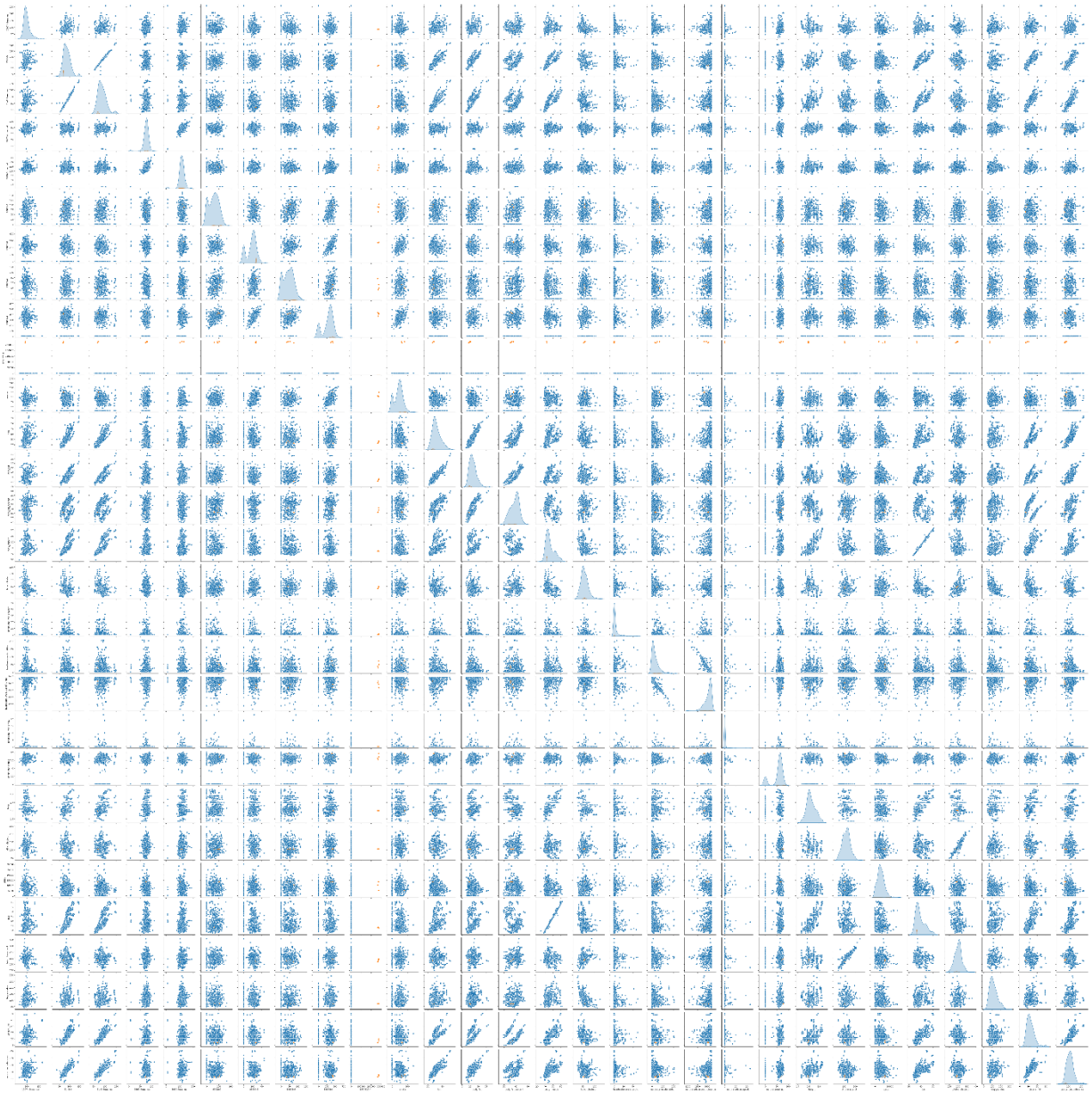

Figure 14 Pair plot of IN2 all categories – dropping the instances with NaN

Dropping the NaN results to a dataset with significantly less zeros present but still doesn't make a difference because outliers exist and affect a lot the clustering results, that practically don't show distinct groups. An evident variable that presents outliers through both initial clustering attempts and is affecting the outcome is the 'LP35925-4'. Based on the LOINC standardization, this variable is matched to the Body Mass Index, for which another variable with a better distribution is available. Having this in mind, the 'LP35925-4' is left out and clustering carried out once more with everything else staying the same.

6.1.3. All IN2 dataset categories above 70% without LP35925-4 – dropping the NaN instances

Figure 15 Pair plot of IN2 categories without LP359925-4– dropping the instances with NaN

Leaving out the LP359925-4 category leads to a better clustering that seems to provide more information in a first look. Out of the 456 instances left after dropping the NaN, the 233 are correctly classified based

on the definitions of MetS results. This is translated to 51% of the total instances being correctly grouped, which resembles an almost random selection.

6.1.4. IN2 dataset categories above 70% excluding the fitness tracker data – dropping the NaN instances

To acquire a view of the capability of the measured values to be used in order to study MetS, an initial case is the exploration of the measurements without the fitness tracker ones that are unique to the IN2 dataset compared to the IN1 and Control. LP359925-4 is again left out.

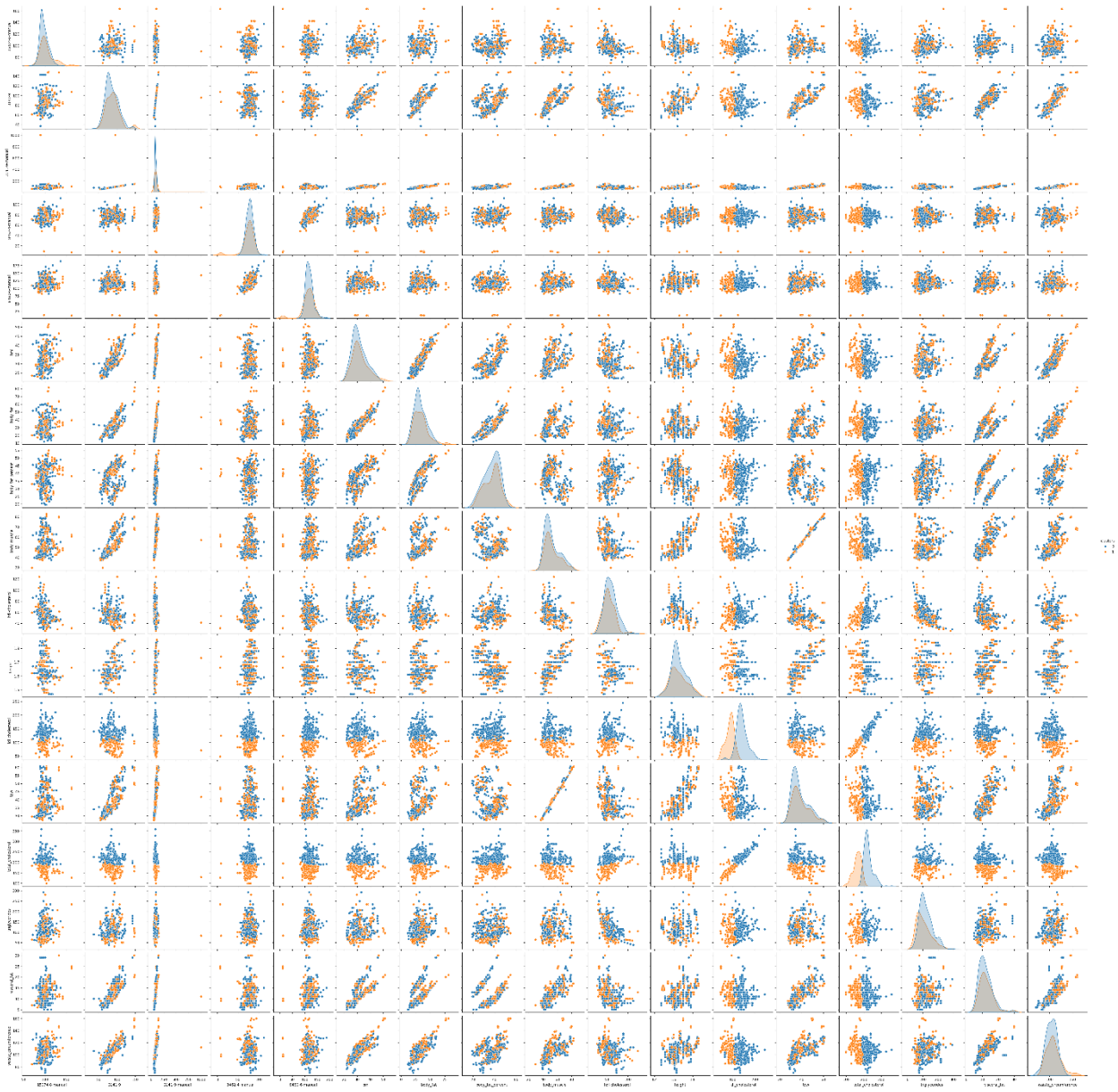

Figure 16 Pair plot of IN2 categories without tracker data – dropping the instances with NaN

6.1.5. IN2 dataset heart-related categories – dropping the NaN instances

Since the heart-related data are the ones with the most information across the IN2 dataset with around 84% information availability before interpolation, this clustering attempt takes only these variables into consideration. Again, the instances with missing values (NaN) are completely removed. This preprocessing makes the available instances to be clustered from 840 to 836, exhibiting the great amount of information that the heart-related categories show, leaving only 4 instances out due to missing data.

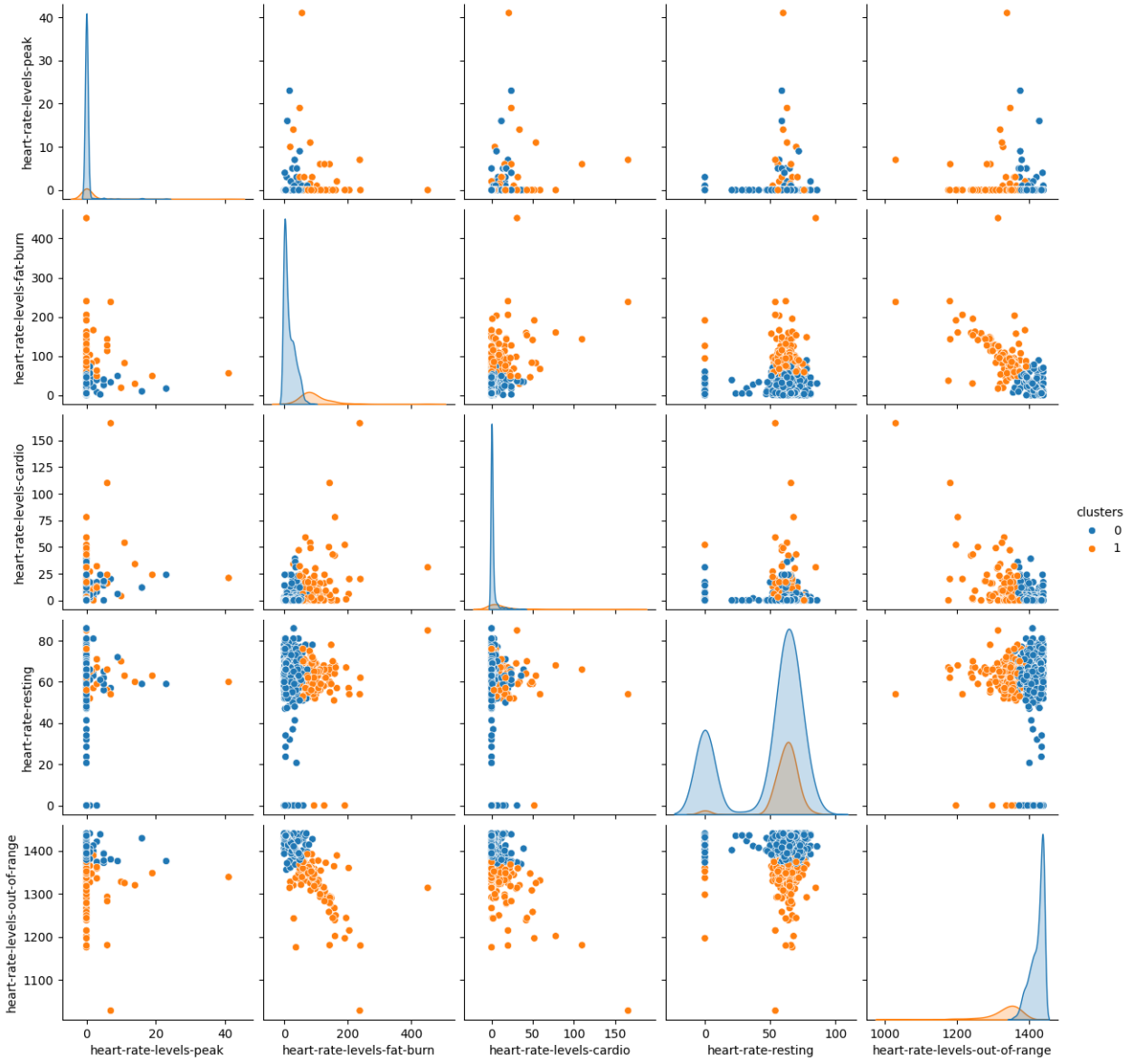

Figure 17 Pair plot of IN2 heart-related categories

This clustering approach seems to be more meaningful, with the data points of the pair plot showing clear distinction between clusters. To decide whether these data are capable of separating the data close to the MetS definitions, the clusters formed have to be compared to the results of the definitions of MetS. When done so, it is calculated that out of the 836 instances the 511 are correctly classified based

on the definitions of MetS results. This is translated to 61% of the total instances being correctly grouped and provides a promising result of the heart-related measurements utilization on clustering.

6.1.6. IN2 dataset fitness tracker categories – dropping the NaN instances

Heart-related data seem to come up with a promising result and this leads to another option. This is the selection of all 11 categories that are related to the fitness tracker and contain significant amount of information across the IN2 dataset with over 73% information availability before interpolation. Again, the instances with missing values (NaN) are completely removed. This preprocessing makes the available instances to be clustered from 840 to 792, around 6% reduction of the whole available instances at this point.

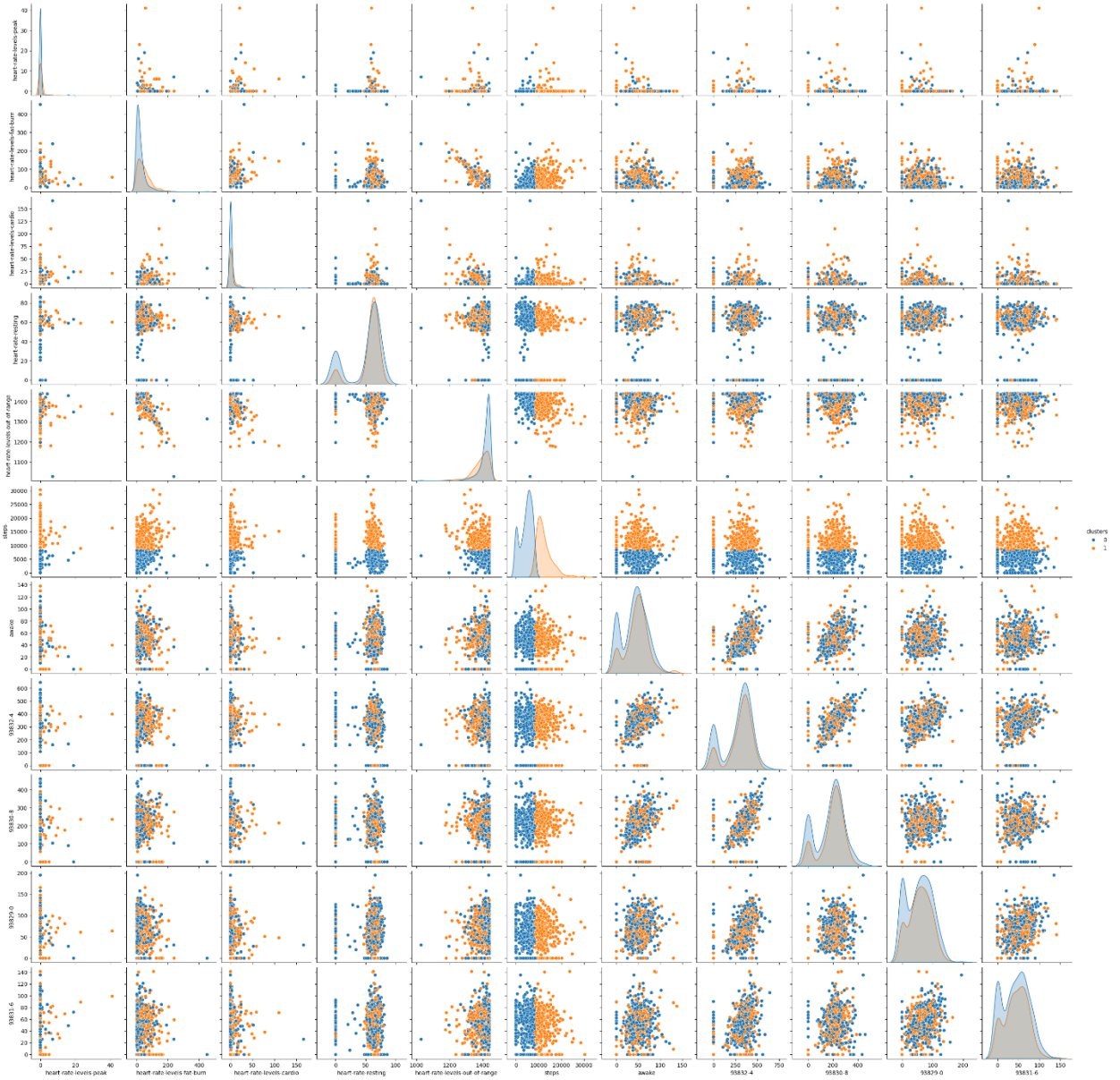

Figure 18 Pair plot of IN2 fitness tracker-related categories

Though this clustering approach seems to achieve a great separation, especially at the steps category, in fact the results are close to random. The correct clustered values are only 406 out of 792. This percentage is translated to just 51% of the data in scope, leading to a random clustering selection based on the MetS definitions characterization.

6.1.7. IN2 dataset fitness tracker categories except steps – dropping the NaN instances

In the above case, the steps values seem to be well clustered around the middle of the values. The issue is that steps are considerably bigger values than the rest of the data categories (even two orders of magnitude at some cases) and can be manipulating the clustering due to this fact. So, in the following approach steps are left out of the clustering variables and instances with NaN are removed.

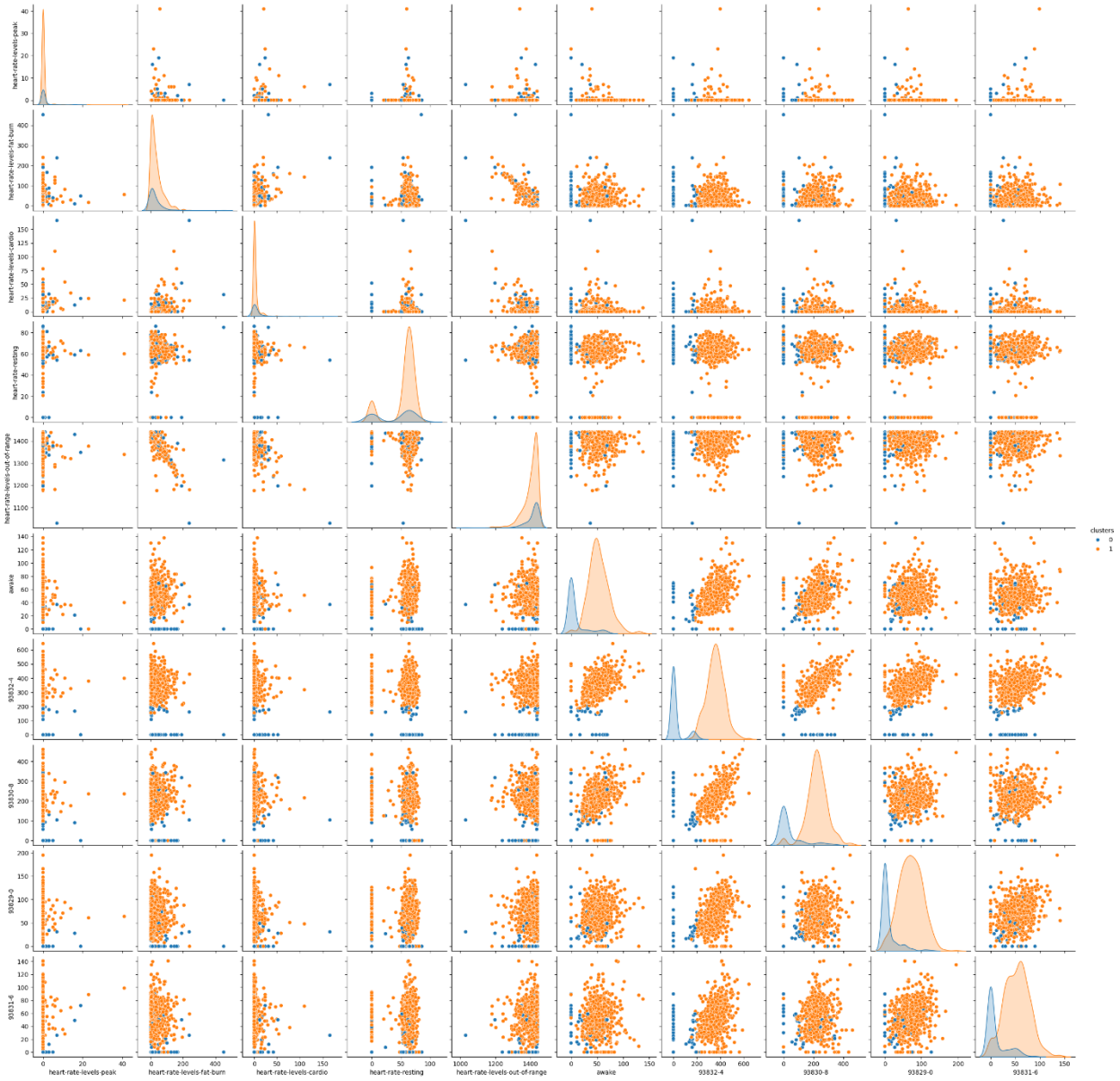

Figure 19 Pair plot of IN2 fitness tracker-related categories without steps

This approach separates correct 502 out of the 792 available instances, which means that 63% of the data are right clustered. This is one of the most promising results towards the clustering of the patients.

6.1.8. IN2 dataset sleep-related categories – dropping the NaN instances

The data categories related with sleep exhibit distributions that could be easier to be distinguished based on the diagonal elements of the above pair plots. The next approach relies only on sleep-related data acquired from the fitness tracker to cluster the data.

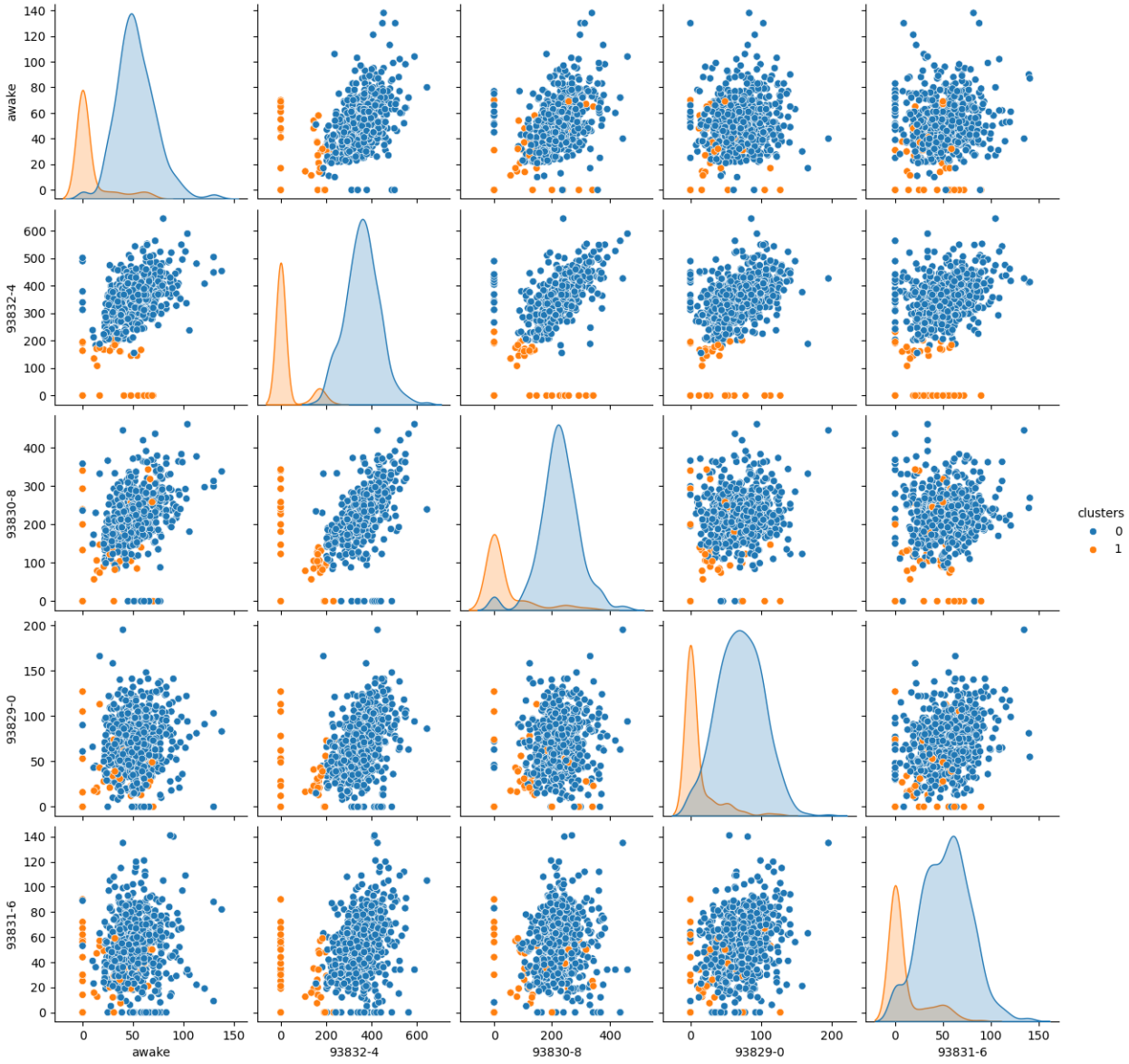

Figure 20 Pair plot of IN2 sleep-related data

Surprisingly, the results of this approach are exactly the same with the previous one, clustering right 502 samples, meaning the 63% of the dataset.

6.2. Intervention 2 K-Means Clustering with L1 Normalization

6.2.1. All IN2 dataset categories above 70% – dropping the NaN instances

L1 normalization doesn't prove to be helpful in the case of the whole dataset clustering, providing almost the same indistinguishable clusters affected by the outliers as without L1 normalization in place.

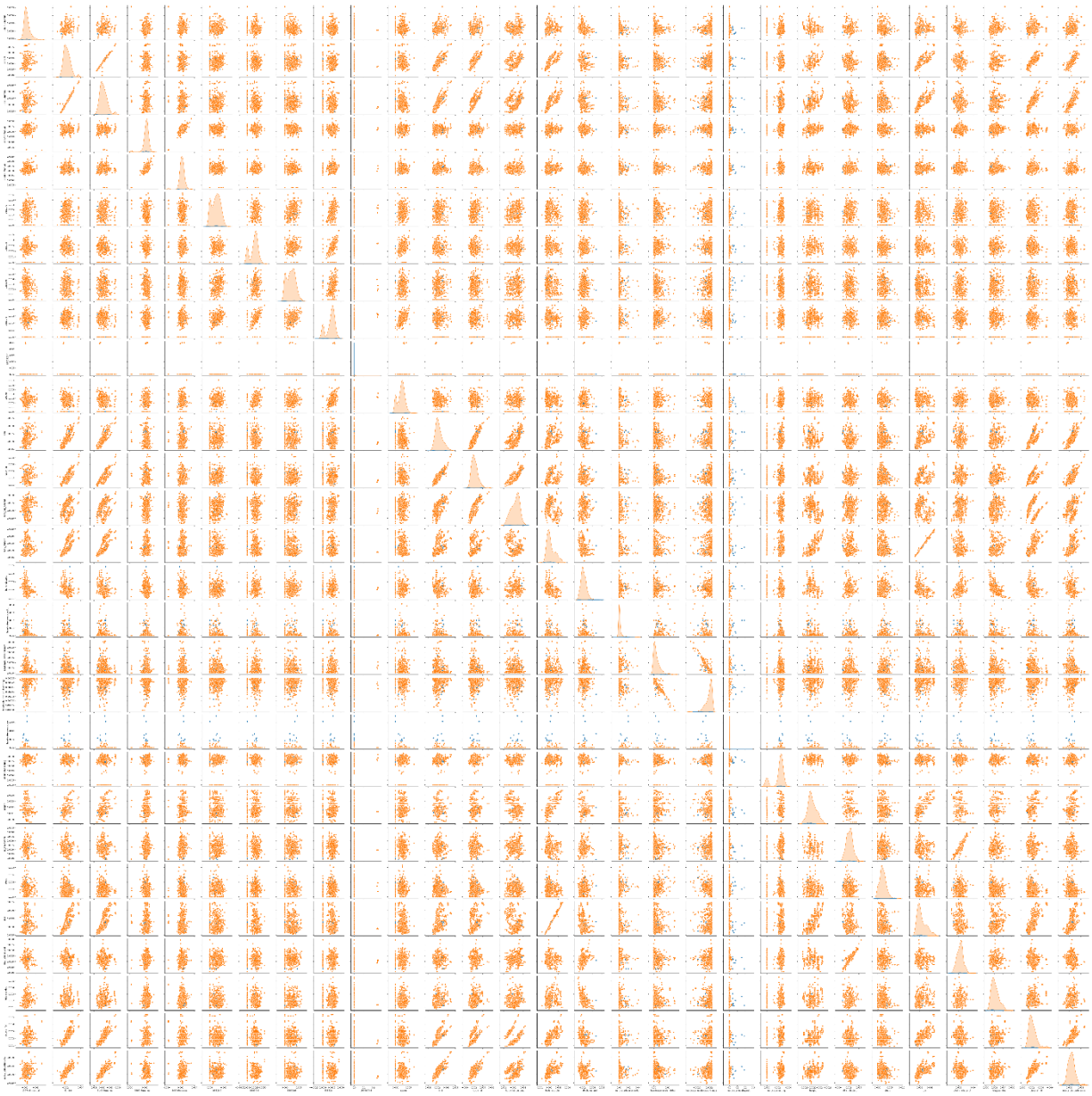

Figure 21 Pair plot of IN2 all categories after L1 Normalization

6.2.2. All IN2 dataset categories above 70% without LP35925-4 – dropping the NaN instances

Even though the LP35925-4 variable is dropped, no clustering is seen to be resulting after L1 Normalization. While for the same data, without normalization there seemed to be a better separation into two groups. In this case, almost all of the data points belong to one cluster. It is obvious that L1-Normalization doesn't provide a better solution regarding the data clustering.

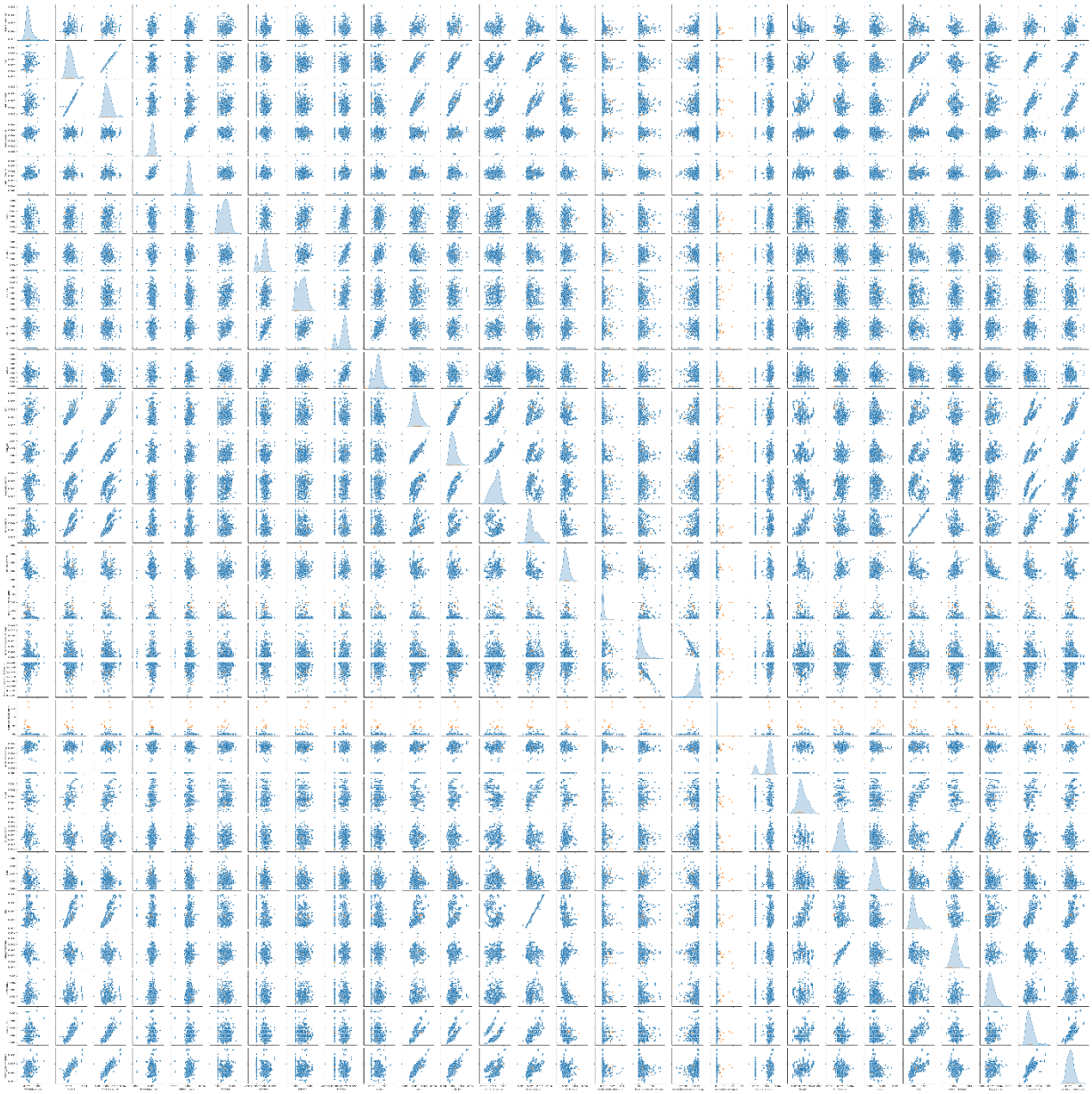

Figure 22 Pair plot of IN2 categories without LP359925-4 after the L1 Normalization – dropping the instances with NaN

6.2.3. IN2 dataset categories above 70% excluding the fitness tracker data – dropping the NaN instances

On the other hand, using all the data apart from the fitness tracker related ones and the LP35925-4 that is removed from the dataset, seems to come up with a better clustering after the L1-Normalization is applied on the data. Clustering after L1-Normalization provides 68% correct separation of the data points, improved from the 55% that was reached with the respective non-normalized data.

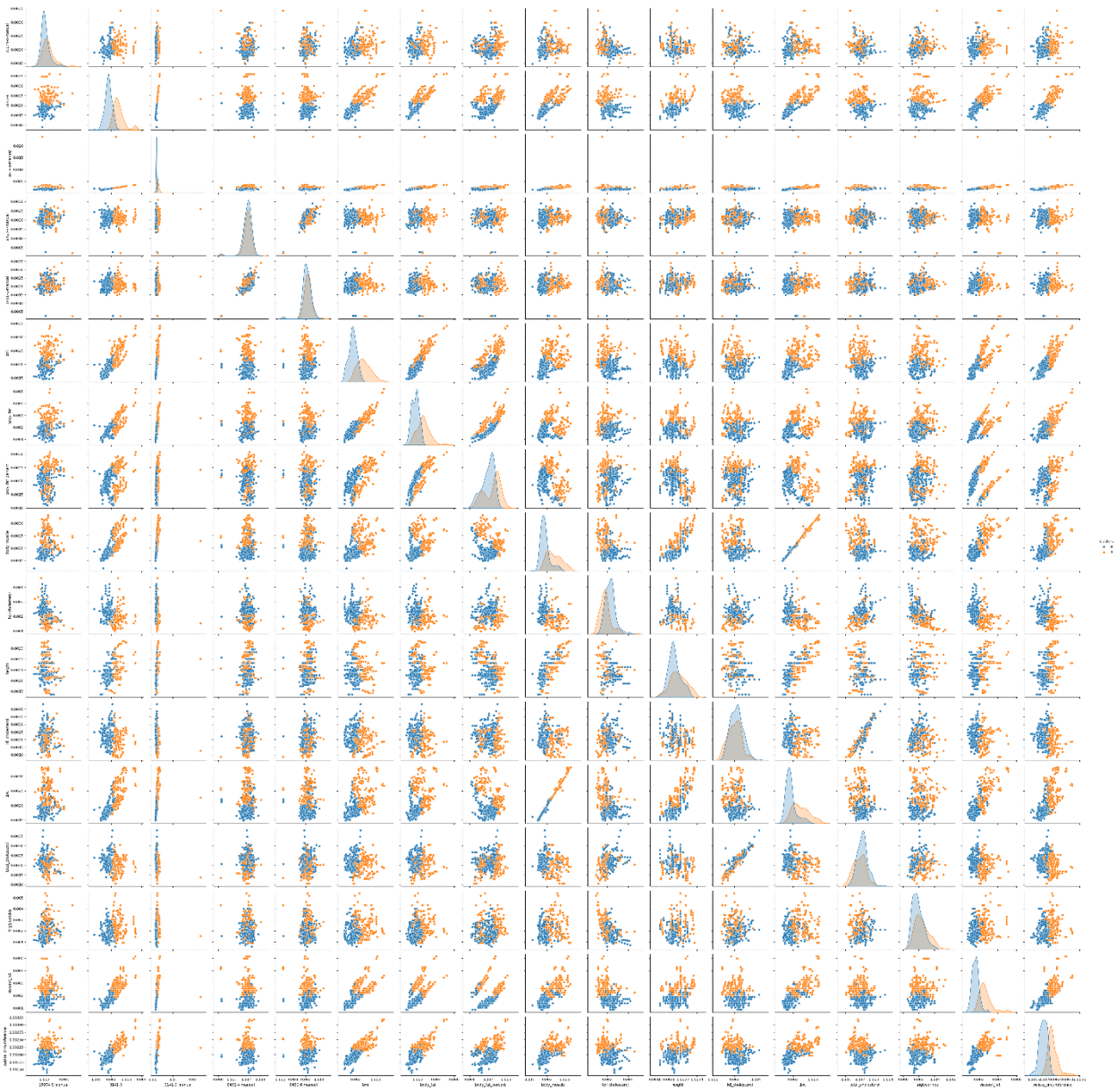

Figure 23 Pair plot of IN2 categories without tracker data after L1 Normalization – dropping the instances with NaN

6.2.4. IN2 dataset heart-related categories – dropping the NaN instances

When the heart-related data are used after the L1 normalization, the forming of two groups seems to be worse than the one without the data being normalized, with most of the data belonging to one cluster. This clustering comes with 68% of the data being grouped correctly.

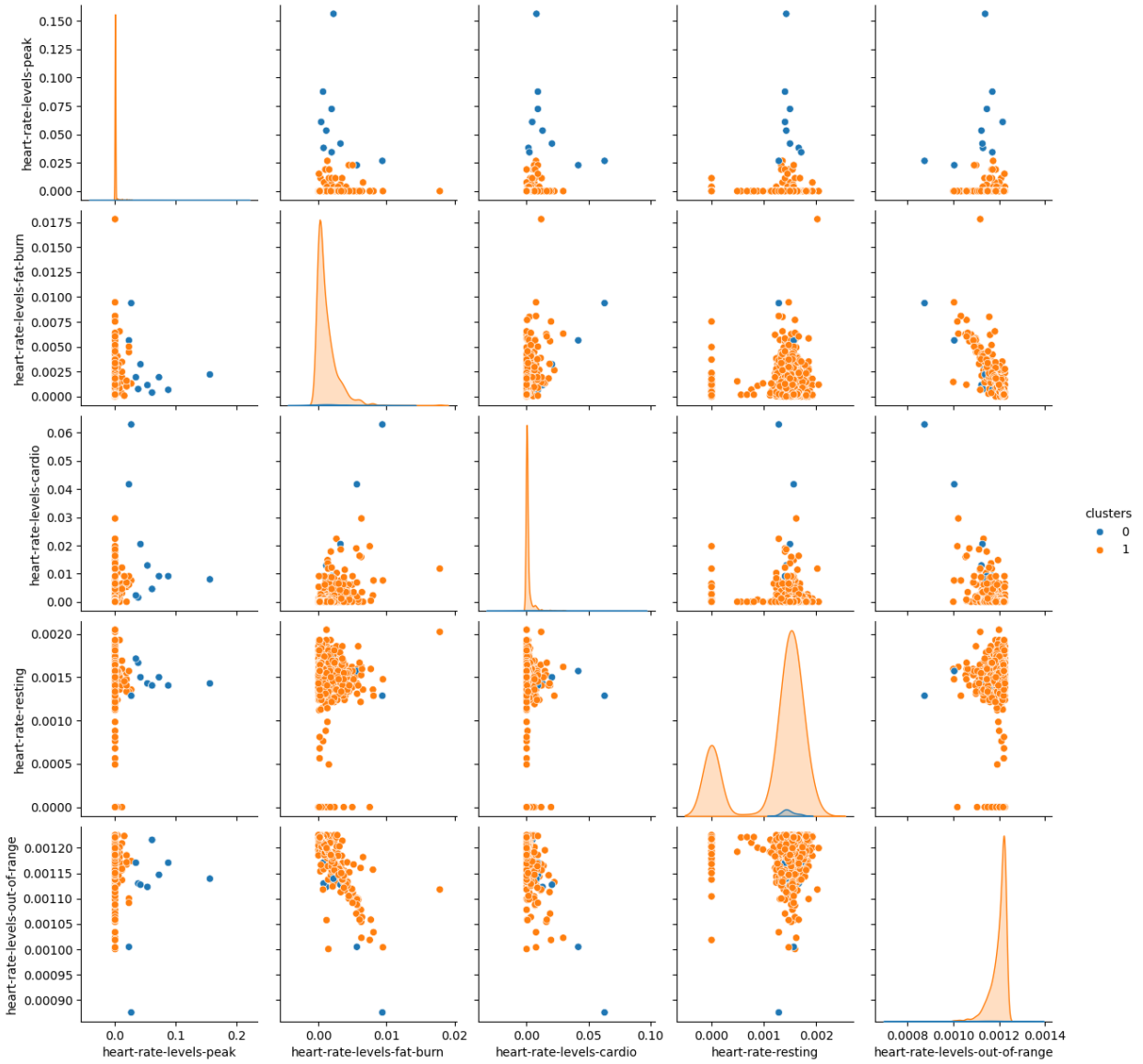

Figure 24 Pair plot of IN2 heart-related categories after L1 Normalization

6.2.5. IN2 dataset fitness tracker categories – dropping the NaN instances

Using the fitness tracker data that include heart rate, sleep and steps measurements after the L1 Normalization, the results do not exhibit better results. Unfortunately, the data are not clustered into two groups, with only little of them making up one group and almost all the rest the other group.

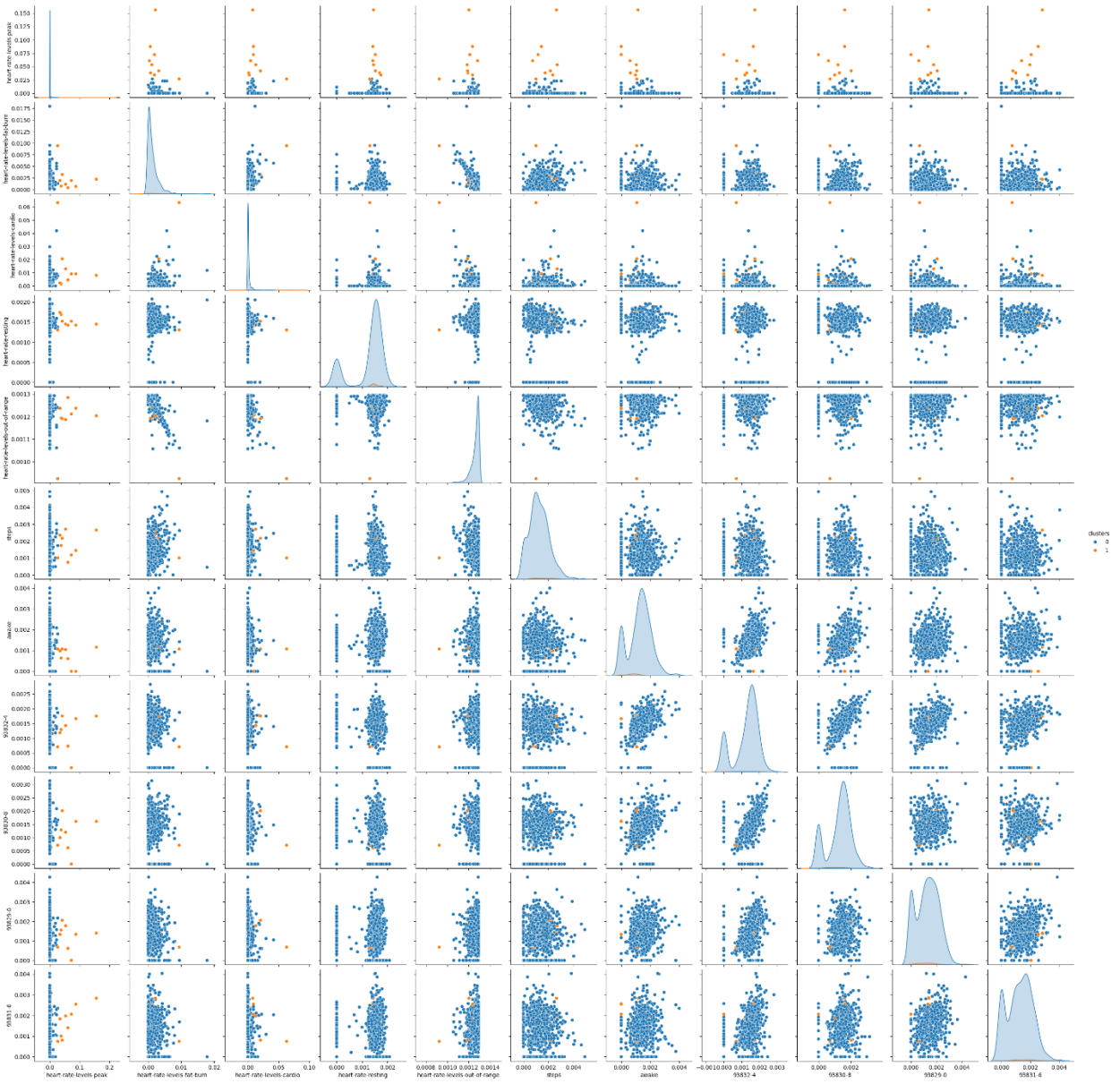

Figure 25 Pair plot of IN2 fitness tracker-related data after L1 Normalization

6.2.6. IN2 dataset fitness tracker categories except steps – dropping the NaN instances

Next up is the clustering attempt of the fitness tracker data without including the steps measurement data that were in a significantly bigger scale than the other measurements. This shouldn't be the case with L1 Normalization since the difference magnitude is handled by the normalization process, but it still should be tried as well. Unfortunately, again there is no evident clustering of the available data.

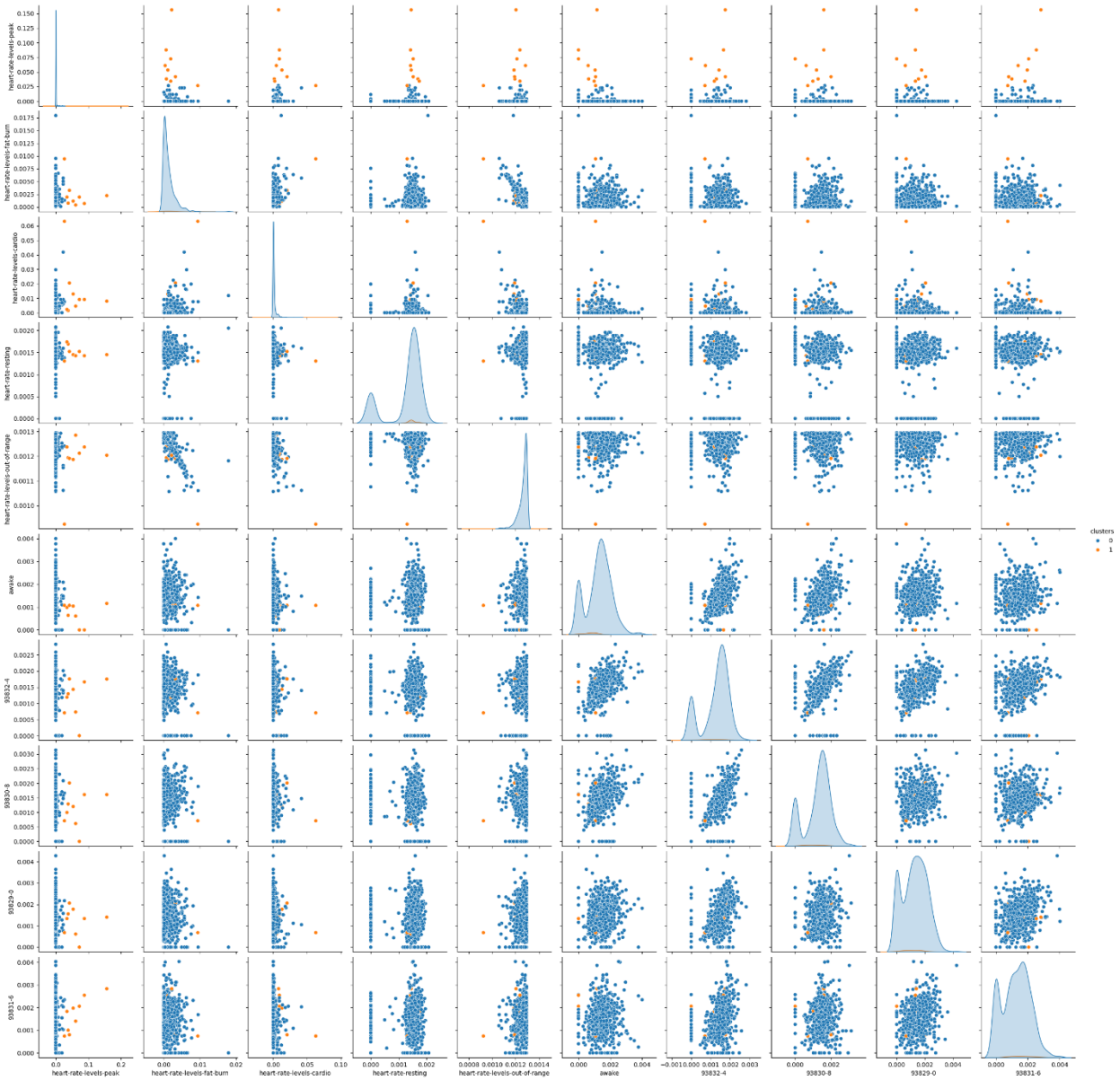

Figure 26 Pair plot of IN2 fitness tracker-related categories without steps after L1 Normalization

6.2.7. IN2 dataset sleep-related categories – dropping the NaN instances

The final application of the L1 Normalization will be done on the sleep-related data to uncover any relations that could enable a better clustering. The clusters seem to have more elements and 504 out of the 792 instances are correctly clustered, making up the 64% of the available dataset.

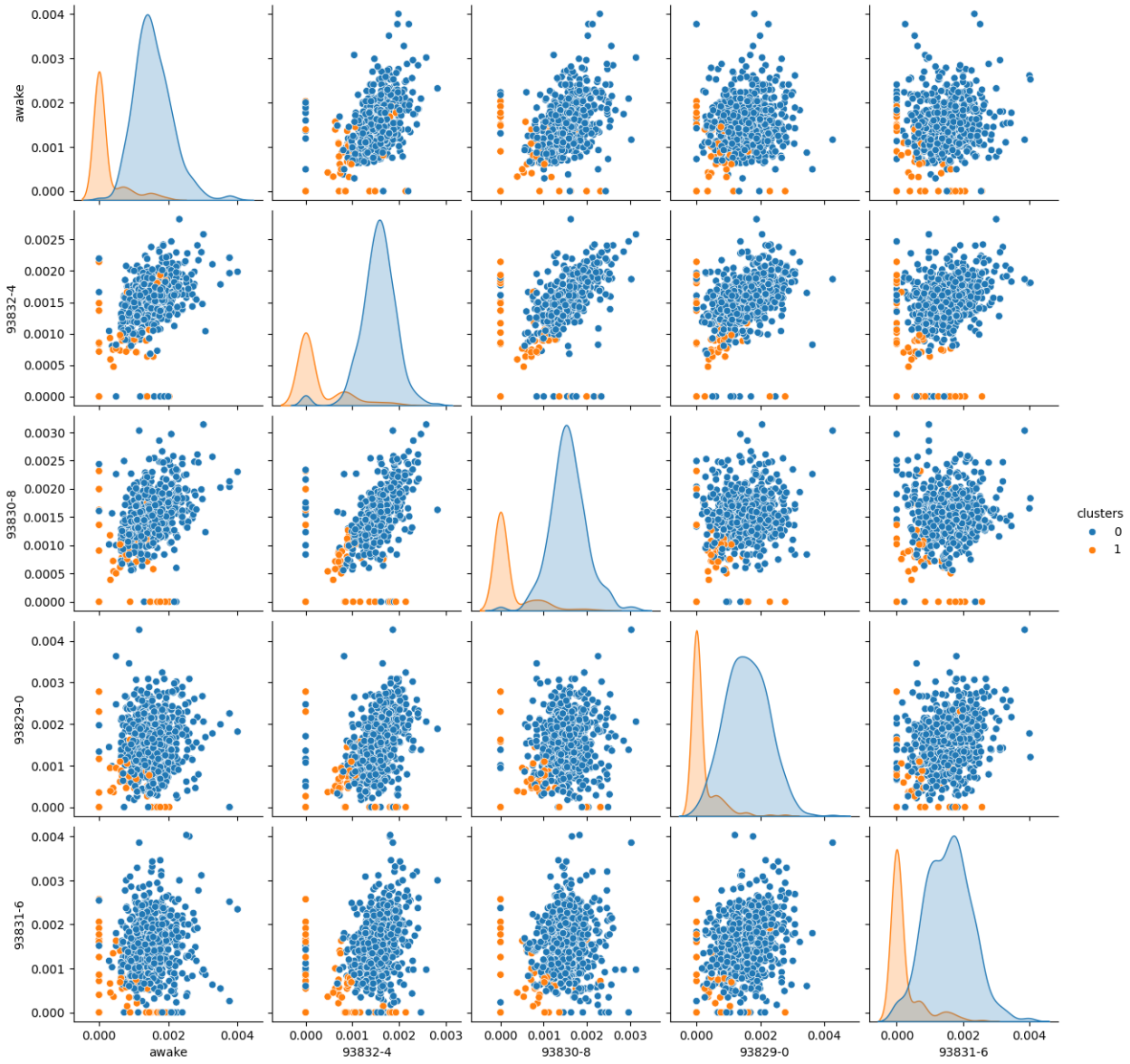

Figure 27 Pair plot of IN2 sleep-related data after L1 Normalization

6.3. Intervention 2 Overall K-means Clustering Results

When checking the instances that are part of one cluster, only a part of the results is taken in mind. A significant factor should also be the number (or percentage) of instances that a cluster includes. If, for example, most of the data available are included in one of the two clusters made up by the algorithm, and the MetS definitions indicate that almost 70% of the subjects of our datasets are not MetS patients, then most of them will be included in the one large group that clustering has defined, thus providing a

false understanding of its ability to capture and define the two clusters. For this reason, the table below is constructed in order to get a better understanding of the results, apart from the pair plot images that provide a schematic representation of them.

	All IN2 data		All IN2 data excluding LP35925-4		IN2 data excluding fitness tracker data		Heart-related data		Fitness tracker data		Fitness tracker data without steps (heart & sleep data)		Sleep data	
	Raw	L1-norm	Raw	L1-norm	Raw	L1-norm	Raw	L1-norm	Raw	L1-norm	Raw	L1-norm	Raw	L1-norm
Correctly clustered	64%	64%	51%	65%	55%	68%	61%	68%	51%	67%	63%	67%	63%	64%
Cluster 0 instances percentage	79%	1%	60%	98%	58%	57%	82%	1%	59%	99%	21%	99%	21%	78%
Cluster 1 instances percentage	21%	99%	40%	2%	42%	43%	18%	99%	41%	1%	79%	1%	79%	22%

Table 12 K-means clustering results and cluster data

As the results indicate, when moving to L1 normalization (L1-norm in the table), the clustering seems to worsen in all of the situations except from two cases, namely the IN2 data without the fitness tracker ones and the Sleep data. In these two situations there seems to be consistency and improvement compare to the non-normalized data clustering, achieving almost the same (in Sleep data) or much better (in all data excluding fitness tracker) correctly clustered portions of the available data. All of the rest clustering attempts seem to make one big cluster when using L1 normalization and in this way, they get significantly better results in terms of correct separation, but the clustering is not valid since they use 99%-98% of the data in one cluster and the remaining 1%-2% in the other. These outcomes highlight the fact that in cases with extreme outliers, L1-Normalization or even K-means clustering algorithm can be extensively affected.

6.4. Intervention 2 Spectral Clustering

The results of Spectral Clustering on the two selected dataset subgroups are presented both without normalization and with L1 normalization.

6.4.1. IN2 dataset categories above 70% excluding the fitness tracker data – dropping the NaN instances

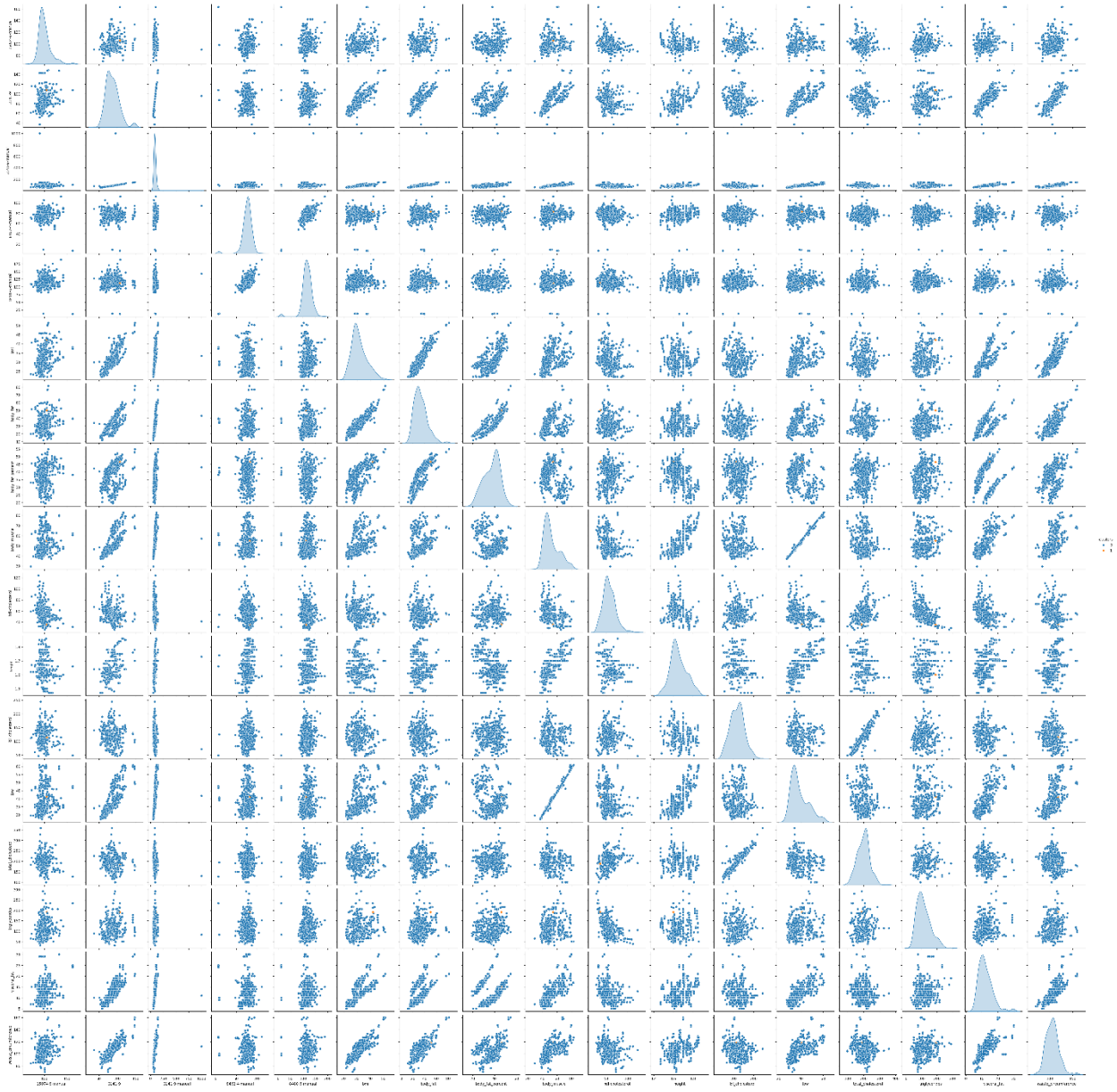

Figure 28 Pair plot of IN2 categories without tracker data – dropping the instances with NaN (Spectral Clustering)

Unfortunately, no established distinction is evident through the pair plot, indicating that probably almost all of the data are included in the one cluster and only a really small proportion in the other.

6.4.2. IN2 dataset categories above 70% excluding the fitness tracker data with L1 Normalization – dropping the NaN instances

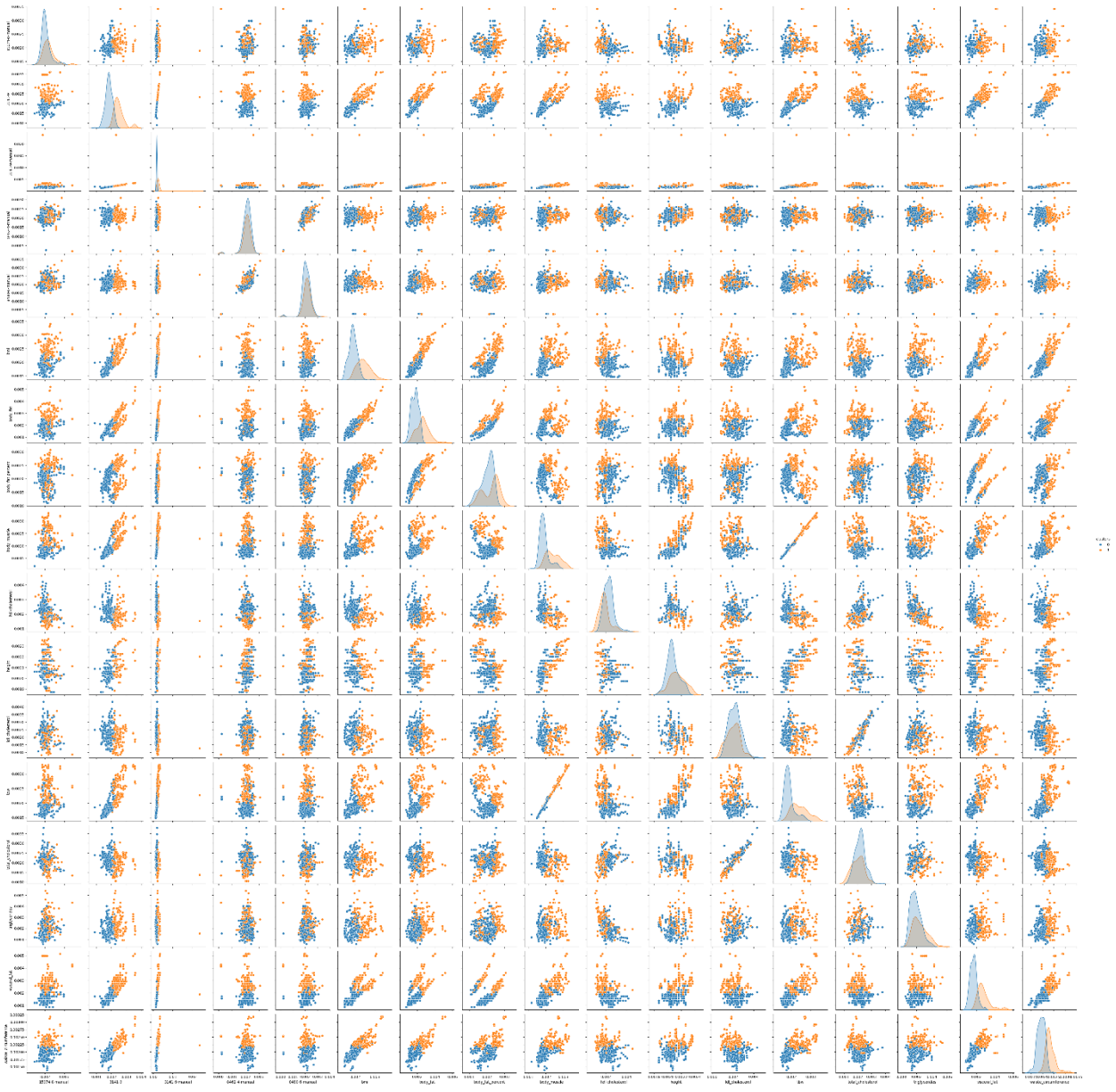

Figure 29 Pair plot of IN2 categories without tracker data after L1 Normalization – dropping the instances with NaN (Spectral Clustering)

After using L1 Normalization, the results exhibit better clustering and the clusters are obviously more balanced than without normalization.

6.4.3. IN2 dataset sleep-related categories – dropping the NaN instances

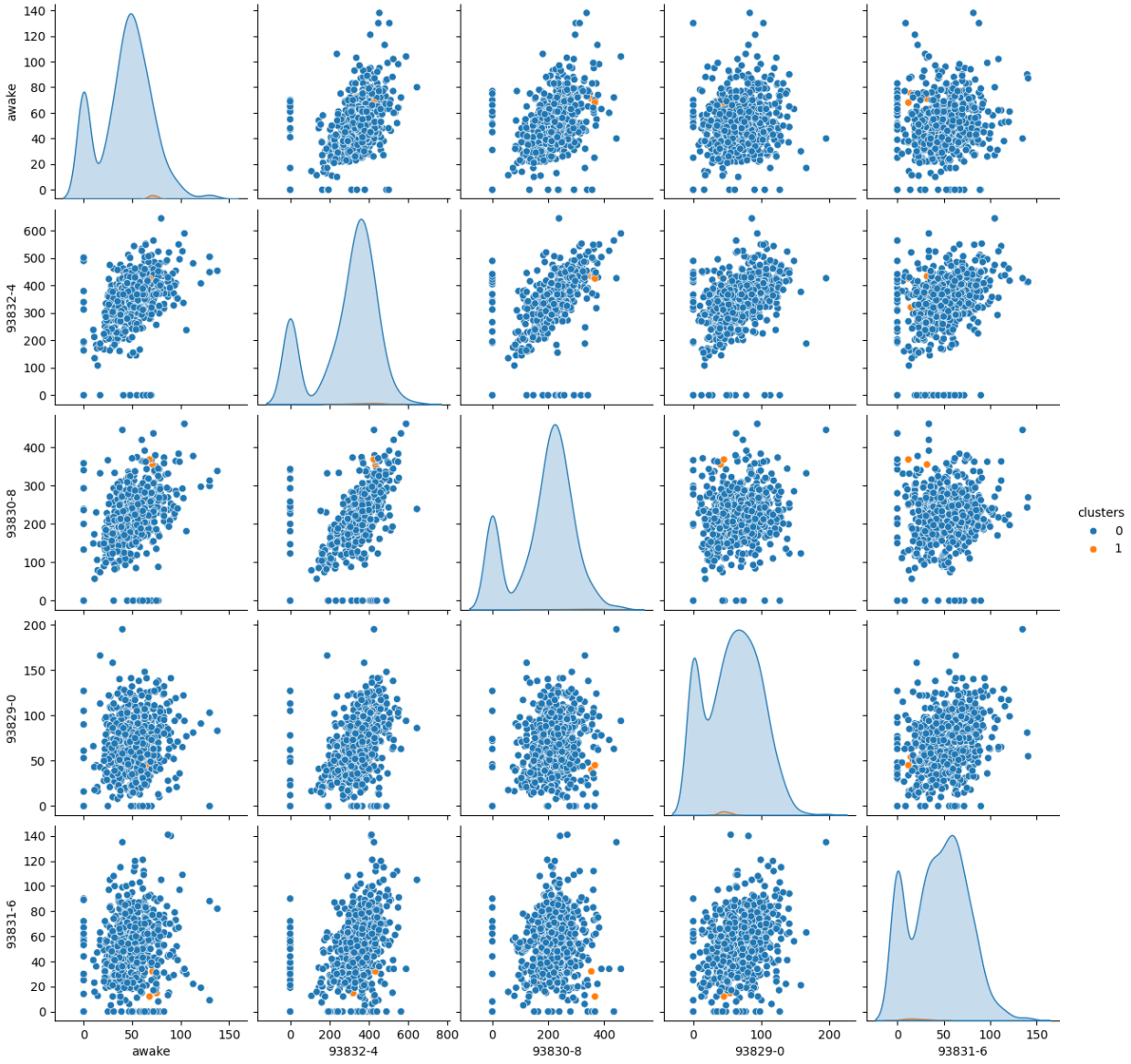

Figure 30 Pair plot of IN2 sleep-related data (Spectral Clustering)

Raw sleep data fail as well to achieve a meaningful clustering, with the majority of the points belonging to one cluster and practically making up almost the whole dataset.

6.4.4. IN2 dataset sleep-related categories with L1 Normalization – dropping the NaN instances

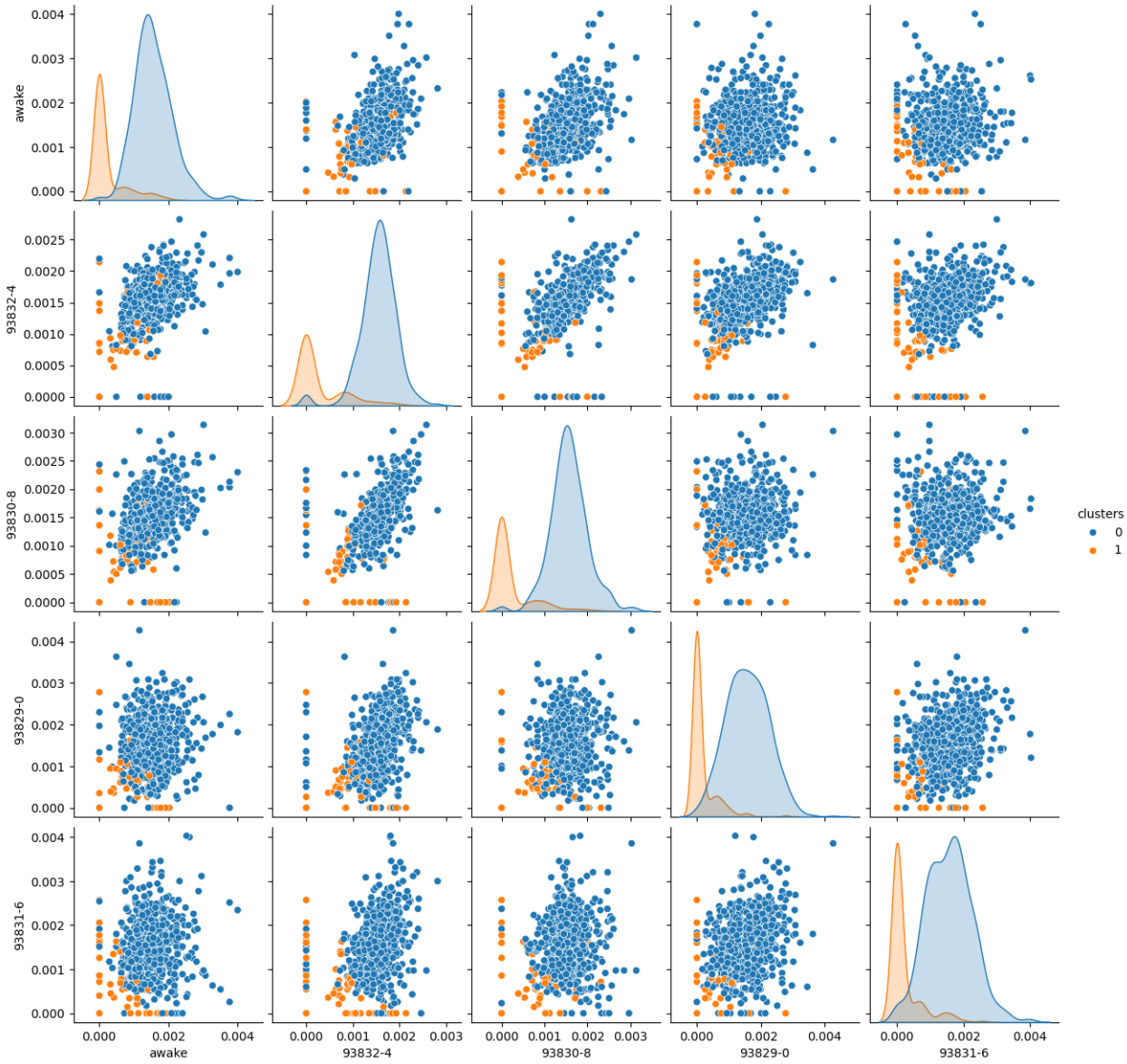

Figure 31 Pair plot of IN2 sleep-related data after L1 Normalization (Spectral Clustering)

On the other hand, and following the first application's motif, the sleep related data improve significantly when L1 Normalization is applied before the Spectral Clustering.

6.5. Intervention 2 Overall Spectral Clustering Results

From the pair plots, it is made clear that the data in their raw format shape a big cluster with their majority included there and just a really small number at the other cluster. Interestingly, after L1 normalization, the results are much better with distinct clusters that include more data points than the non-normalized ones. The respective results are shown in Table 12.

	IN2 data excluding fitness tracker data		**Sleep data**	
	Raw	L1-norm	Raw	L1-norm
Correctly clustered	66%	69%	69%	64%
Cluster 0 instances percentage	99%	60%	99%	78%
Cluster 1 instances percentage	1%	40%	1%	22%

Table 13 Spectral clustering results and cluster data

The table shows that L1 Normalization is a prerequisite to get two clusters with sufficient data in each one so that they are distinguishable and not a single cluster that contains all the data and another with just some outliers. The selected data subgroups provide promising clustering and classification (when it comes to correctly clustered data) outcomes. Nevertheless, there is possibility that combinations of other available techniques lead to even better outcomes. Some of them will be tested through the following pages.

6.6. Alternative Clustering and Normalization Techniques

In this section, alternative techniques are applied on the clustering process steps to investigate whether they enhance the clustering capabilities.

6.7. MinMax Scaler clustering results

6.7.1. IN2 dataset categories above 70% excluding the fitness tracker data with MinMax Scaling and Spectral Clustering – dropping the NaN instances

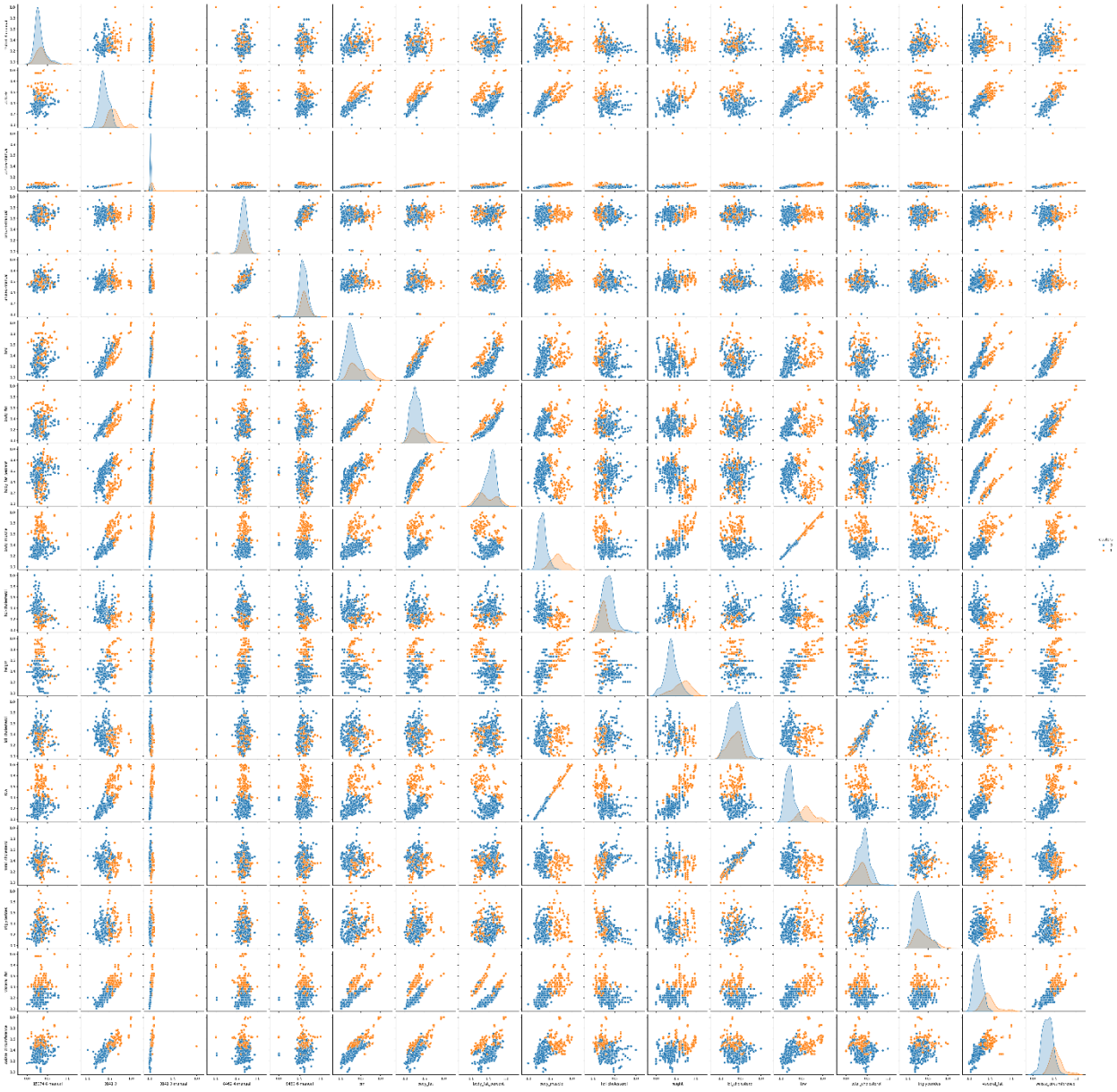

Figure 32 Pair plot of IN2 categories without tracker data after MinMax Scaling – dropping the instances with NaN (Spectral Clustering)

6.7.2. IN2 dataset sleep-related categories with MinMax Scaling and Spectral Clustering – dropping the NaN instances

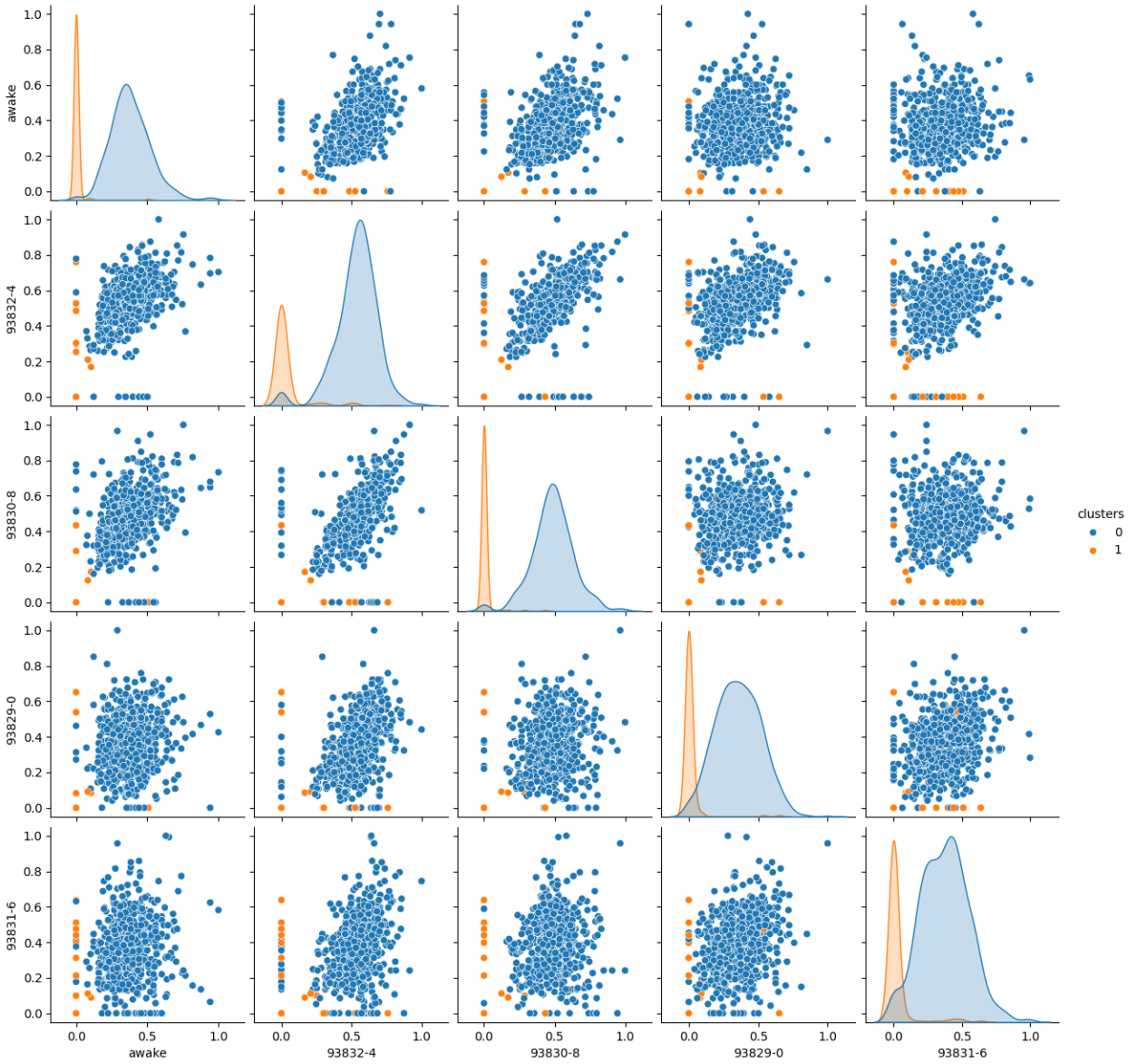

Figure 33 Figure 33 Pair plot of IN2 sleep-related data after MinMax Scaling (Spectral Clustering)

Both of the pair plots indicate that distinct clusters are formed, with the sleep data having a small and a bigger cluster, while the first dataset showing better distribution of the data in the two clusters. The results and metrics of the two clustering applications are presented in the following table.

	IN2 data excluding fitness tracker data	Sleep data
	MinMax	MinMax
Correctly clustered	69%	66%
Cluster 0 instances percentage	70%	83%
Cluster 1 instances percentage	30%	17%

Table 14 Spectral clustering results and cluster data after MinMax scaling

There seems to be almost the same outcome compared to the results of the previous applications where L1 Normalization was applied. So, MinMax could be considered as a similar technique to L1 Normalization when the dataset in scope is studied.

6.8. Principal Component Analysis (PCA)

6.8.1. All IN2 dataset categories above 70% – dropping the NaN instances

At first, PCA is applied on the whole IN2 dataset, excluding the LP35925-4 variable that is left out of the whole approach as mentioned before. The Explained Variance by Components leads to selecting 7 Principal Components out of the total 28 variable categories. K-means clustering is presented.

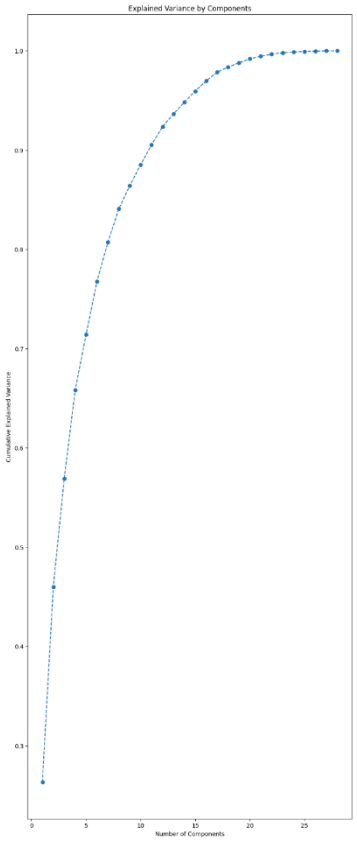

Figure 34 Cumulative Explained Variance plot for all IN2 data categories

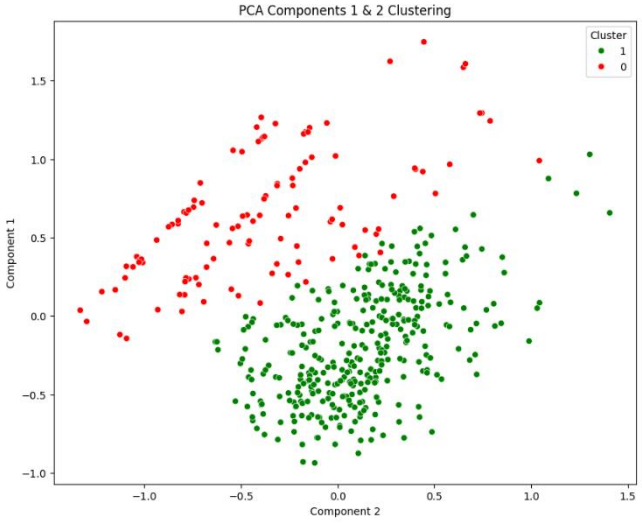

Figure 35 PCA Components 1 & 2 K-means Clustering for all IN2 data categories

6.8.2. IN2 dataset categories above 70% excluding the fitness tracker data – dropping the NaN instances

Next up, PCA is applied on the IN2 dataset excluding the fitness tracker related data. In this case, the Explained Variance by Components leads to selecting 6 Principal Components out of the total 17 variable categories.

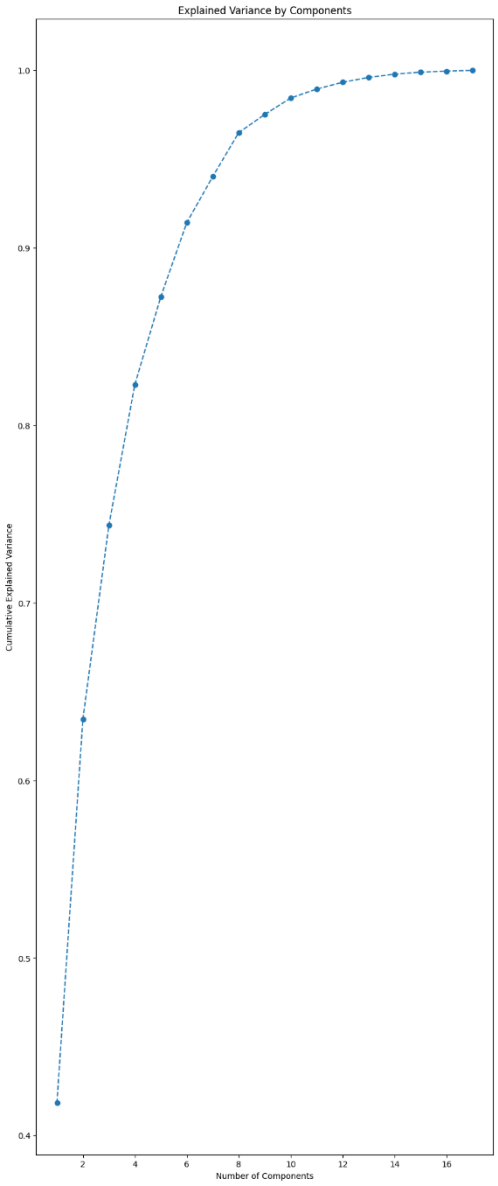

Figure 36 Cumulative Explained Variance plot for IN2 data excluding fitness tracker categories

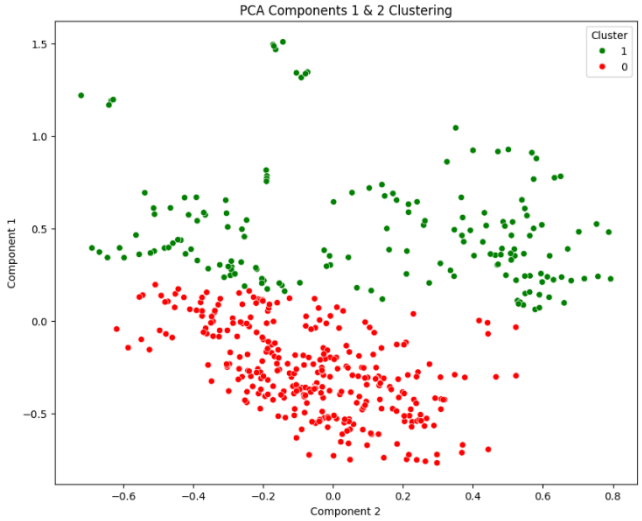

Figure 37 PCA Components 1 & 2 K-means Clustering for IN2 data excluding fitness tracker categories

6.8.3. IN2 fitness tracker categories above 70% – dropping the NaN instances

Finally, PCA is applied on the IN2 dataset fitness tracker related data. The Explained Variance by Components leads to selecting 4 Principal Components out of the total 11 variable categories.

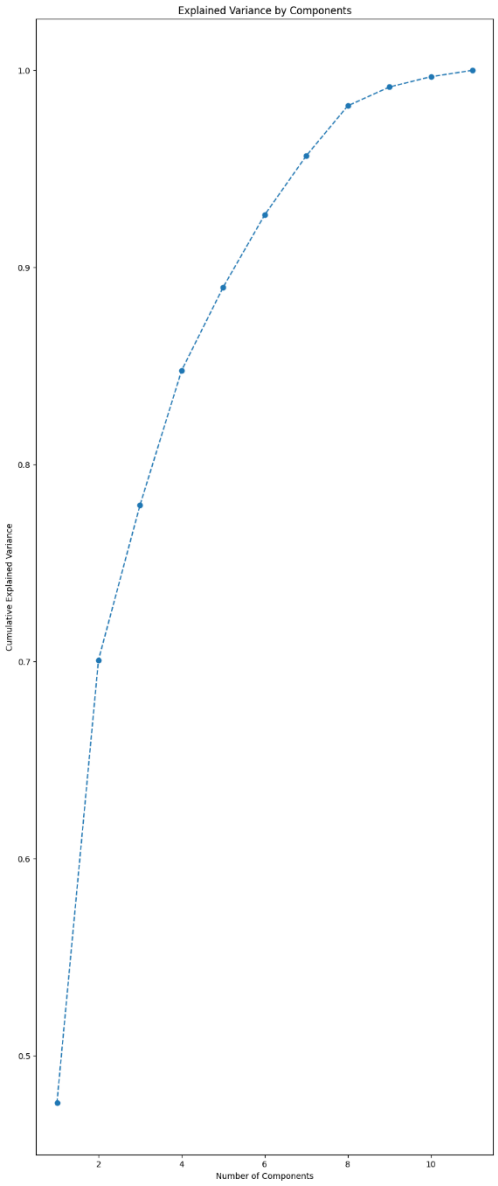

Figure 38 Cumulative Explained Variance plot for IN2 fitness tracker data

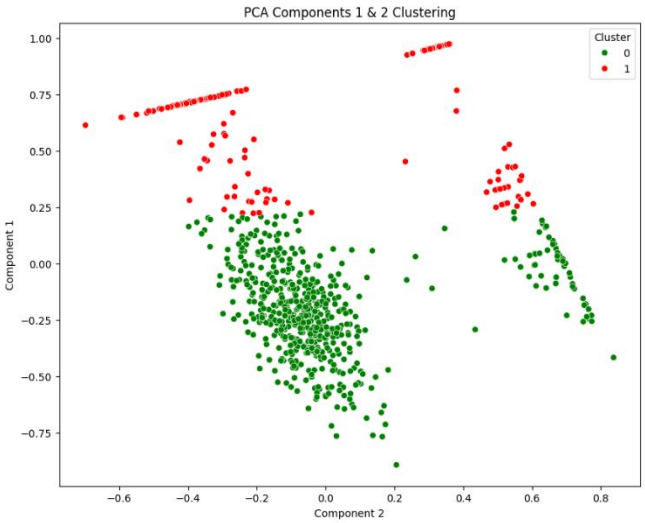

Figure 39 PCA Components 1 & 2 K-means Clustering for IN2 fitness tracker data

6.9. Principal Component Analysis clustering results

After applying PCA and then clustering of the three datasets, the following results occur.

	All IN2 data		IN2 data excluding fitness tracker data		IN2 fitness tracker data	
Clustering method	**K-means**	**Spectral**	**K-means**	**Spectral**	**K-means**	**Spectral**
Correctly clustered	63%	65%	66%	69%	59%	65%
Cluster 0 instances percentage	24%	80%	64%	30%	76%	81%
Cluster 1 instances percentage	76%	20%	36%	70%	24%	19%

Table 15 IN2 PCA Clustering results with cluster metrics

It seems that all of the dataset subgroups manage to achieve a remarkable two-cluster separation with both clusters having a share of the data and not one of them taking over the whole dataset, which was the case with many previous attempts of other approaches. The correctly separated data are still around 60%-70%, which is the maximum range that is observed through all of the clustering methods applied. Clustering the Principal Components with Spectral method provides better results in all of the groups and this is a possible indication for its superiority to K-means in such applications.

7. Overall IN2 Clustering results comparison

In order to compare the results of each method, it is necessary to put them all together in a comparable way. For this reason, the best results of each dataset category in the alternative clustering methods presented above are gathered in a chart. On the y-axis, the correctly clustered data percentage is shown, indicating how many instances were correctly clustered, while on the x-axis, the total instances of each dataset are presented. Ideally, the best clustering cases would be located in the top right quadrant of the chart.

Figure 40 Clustering methods results comparison

The points are separated in two sides. On the left side, the points refer to clustering with less instances, while the right ones are about clustering with more instances. These two groups occur due to the process of dropping the NaN values before clustering. In the case of the whole IN2 categories dataset, many NaNs are present in the data since more variable categories exist, thus missing a lot of instances. These are the left-sided points. In contrast, when less data categories are available, less NaN values exist and this leads to more instances existing after dropping the NaNs. The results of these clustering applications are shown on the right side of the figure.

Looking closely to the results, Spectral clustering seems to be doing better than the other two techniques. When moving to higher number of instances all of the techniques exhibit worse results in terms of correctly clustered data. With more data, they miss around 5% of the total correctly distinguished instances. Interestingly, PCA, which is considered the most complicated and extensive method fails to overcome the Spectral clustering results, being really close to them and above K-means.

Having the overall results in mind, Spectral clustering technique will be also applied to the two other initial datasets, Intervention 1 and Control to investigate its efficiency on these as well.

8. Clustering on IN1 & Control Datasets

8.1. IN1 Spectral Clustering

Intervention 1 dataset includes less patients than Intervention 2 (IN2), without any fitness tracker data. The available data categories are as a result significantly less than IN2 as presented at the data exploration parts of this work.

8.1.1. IN1 Available data

Due to this lack of measurement categories an issue is presented in the IN1 dataset. The blood glucose value doesn't exist as a variable when the datapoints with over 70% of information existence are selected. To handle this issue, the MetS definitions are changed for IN1 instances characterization and exclude the blood glucose variable. Below the available variables for IN1 dataset are shown.

Intervention 1 available data over 70%		
Health indicator dataset name	**Health indicator**	**Unit of measurement**
3141-9-manual	Body weight (Manual)	Kg
8462-4-manual	Diastolic blood pressure	mmHg
8480-6-manual	Systolic blood pressure	mmHg
bmi	BMI	Kg/m^2
body_fat	Body Fat	Kg
body_fat_percent	Body Fat Percentage	%
body_muscle	Body Muscle	Kg
hdl-cholesterol	HDL cholesterol concentration	mg/dL
height	Height	m
total_cholesterol	Total cholesterol	mg/dL
triglycerides	Triglycerides	mg/dL
visceral_fat	Visceral fat	no unit
waiste_circumference	Waist circumference	cm

Table 16 Intervention 1 available data over 70%

It goes without saying that leaving out blood glucose from the MetS definitions, the baseline information about the instances that are characterized as MetS patients or not are not completely right. On the other hand, in order to try the methods used and assessed in the previous parts with a meaningful amount of data, this is the only workaround available. So, this is the way going forward as long as IN1 data are concerned.

8.1.2. IN1 Spectral clustering application

Before clustering, the NaN instances are remover, leaving the IN1 dataset with just 276 instances to participate in clustering. These are scaled using the MinMax scaler and afterwards clustered, providing the following pair plot.

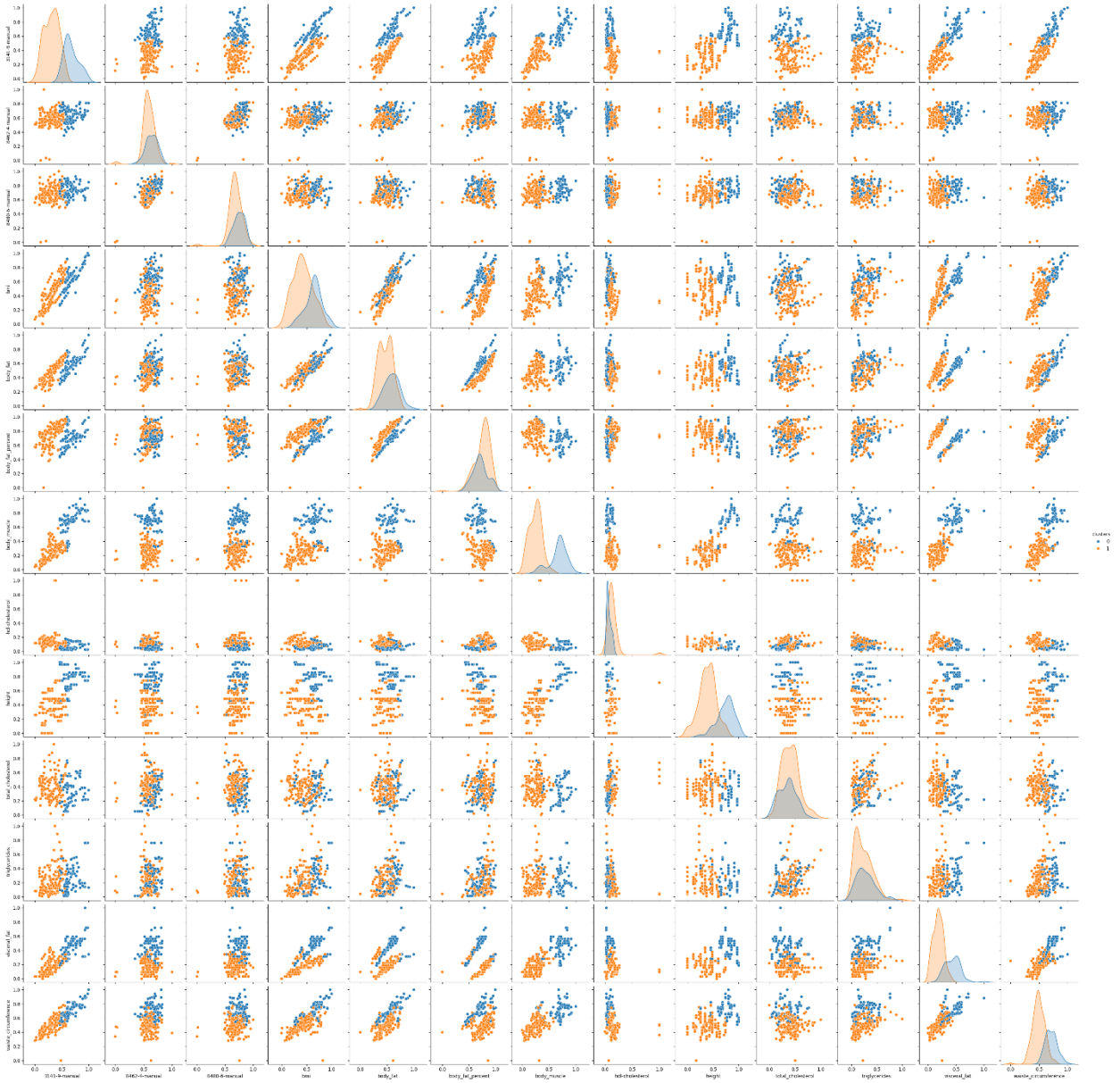

Figure 41 Pair plot of IN1 all categories (Spectral Clustering)

According to the results, 34% of the instances belong to one cluster and the rest 66% to the other. The correctly clustered instances are 171 out of the total 276, which is translated to the 62% of the data.

8.2. Control group Spectral Clustering

Unfortunately, the latest group, namely Control, that had early indications of great data loss confirms it when the data are prepared for clustering. The data categories when selecting over 70% of information existence are only 7, with many crucial variables for the MetS definitions missing. In this case, no MetS patient characterization can be carried out for the available instances, since only the below variables are present, with waist circumference, triglycerides and HDL cholesterol that are used in the definitions missing.

Control group available data over 70%		
Health indicator dataset name	**Health indicator**	**Unit of measurement**
3141-9-manual	Body weight (Manual)	Kg
8462-4-manual	Diastolic blood pressure	mmHg
8480-6-manual	Systolic blood pressure	mmHg
bmi	BMI	Kg/m^2
body_fat	Body Fat	Kg
body_fat_percent	Body Fat Percentage	%
body_muscle	Body Muscle	Kg

Table 17 Control group available data over 70%

8.2.1. Control data Spectral clustering application

Spectral Clustering is applied to the data available, with 304 total instances and results in two clusters, with 76% and 24% of the data in each one. Whether this clustering is valid based on MetS definitions cannot be tested in the Control data case, due to the absence of crucial variables.

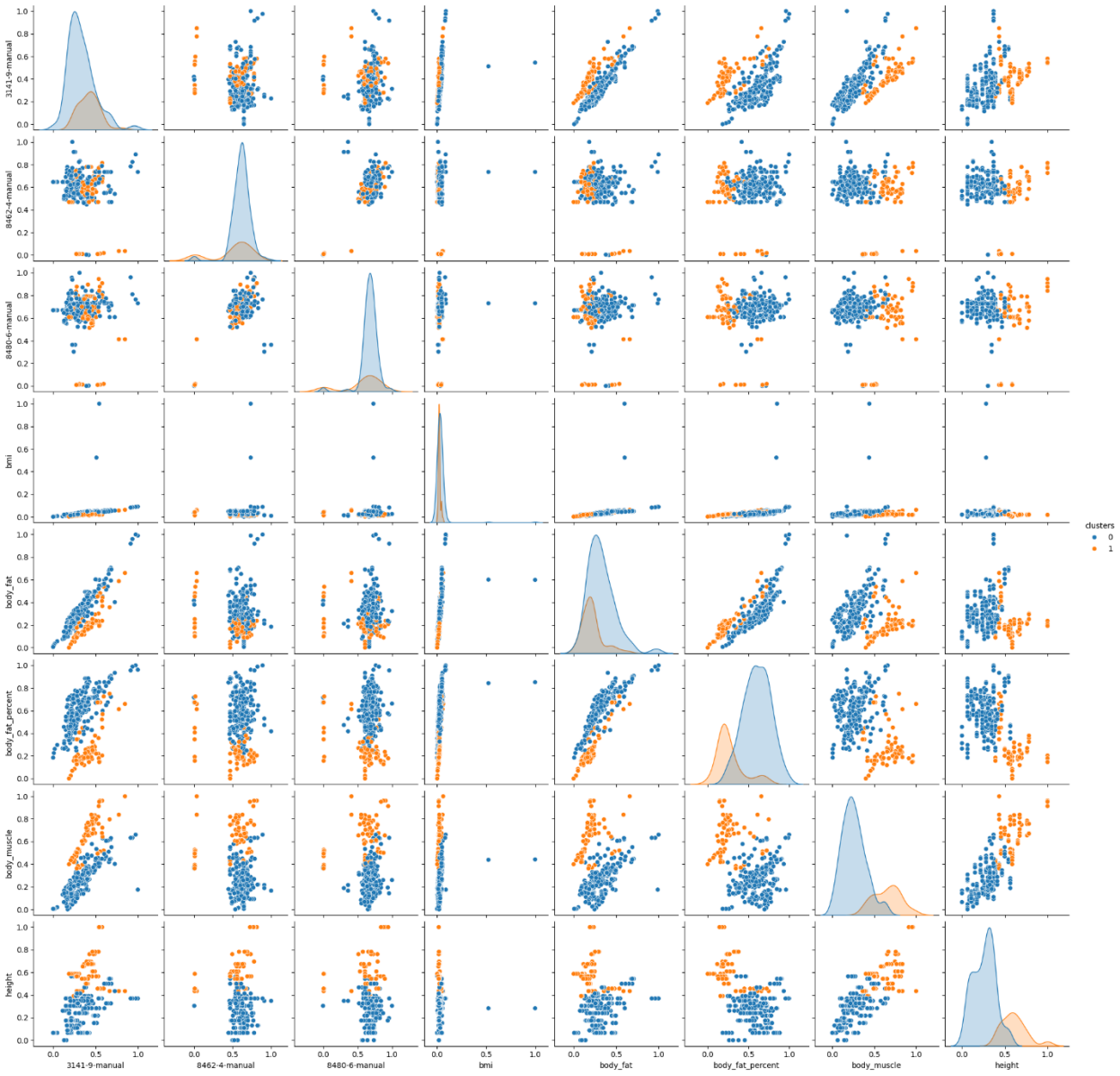

Figure 42 Pair plot of Control data all categories (Spectral Clustering)

9. Conclusions

Defining a patient positive to Metabolic Syndrome is a task which is practically hard in real life. The different definitions that exist make the MetS framework vague and definitely not straightforward. Through this work, MetS definitions are translated to code lines in order to get a view of what the MetS characterization of the patient instances would be in the real world. Subsequently, unsupervised learning and more specifically clustering methods are utilized to enable the separation of patients that are positive to MetS from others that are not. These methods are applied on real-world data with a big number of missing values.

The most trustworthy subgroup of the whole dataset is consisted of health indicators measurements along with fitness tracker data as well. This combination is used for all the different approaches and methods combinations tests regarding clustering, since it is the one that provides the most reliable data, with the least missing values.

On this dataset, linear interpolation is applied to construct data that are missing up to an extent and avoid losing information based on a cutoff at 70% of data information availability. Then, two clustering techniques and normalization methods are used in an attempt to compare their results with the MetS definitions outcomes. Clustering algorithms K-means and Spectral clustering are exploited, along with L1 normalization and MinMax scaling processes to normalize the data before clustering. Moreover, Principal Components Analysis (PCA) is implemented on the data in an attempt to reduce their variable number and is afterwards combined with K-means and Spectral clustering for the identification of patient groups.

All of the approaches' results in terms of classification based on the MetS definitions combination reaches a range of 60%-70%, with the best of them being the Spectral Clustering, but with a slight correct classification percentage difference from the others. At the final step, Spectral Clustering is used on two other subgroups of the dataset with less information and available data and the results are presented.

10. Future Work

This work brings forward the issue of the multiple MetS definitions that present a variation and differentiations between each other making an unsupervised approach almost impossible and the issue of the real-world data that usually prove to be unreliable due to the missing values. The first issue makes a strict labelling of the patients not possible while the latter significantly decreases the power and robustness of classification methods.

Working towards a better Metabolic Syndrome identification and patients' classification, the datasets ought to be more reliable and carefully built in the context that real-worlds allows, while also the definitions of diseases and syndromes based on clinical data have to straightforward and strict for correct data labelling to occur. With these two prerequisites in place, the enablement of advanced Machine Learning and Artificial Intelligence approaches will be realized towards an improved patient risk assessment and precise Metabolic Syndrome risk factors management strategies.

11. References

[1] P. L. Huang, “A comprehensive definition for metabolic syndrome,” May 2009. doi: 10.1242/dmm.001180.

[2] L. Keltikangas-Järvinen, “Metabolic Syndrome,” *Encyclopedia of Stress*, pp. 717–721, Jan. 2007, doi: 10.1016/B978-012373947-6.00230-0.

[3] M. Rus *et al.*, “Prevalence and Risk Factors of Metabolic Syndrome: A Prospective Study on Cardiovascular Health,” *Medicina (B Aires)*, vol. 59, no. 10, Oct. 2023, doi: 10.3390/MEDICINA59101711.

[4] J. J. Noubiap *et al.*, “Geographic distribution of metabolic syndrome and its components in the general adult population: A meta-analysis of global data from 28 million individuals,” *Diabetes Res Clin Pract*, vol. 188, p. 109924, Jun. 2022, doi: 10.1016/J.DIABRES.2022.109924.

[5] K. Denys, M. Cankurtaran, W. Janssens, and M. Petrovic, “Metabolic syndrome in the elderly: An overview of the evidence,” *Acta Clin Belg*, vol. 64, no. 1, pp. 23–34, 2009, doi: 10.1179/ACB.2009.006.

[6] M. G. Saklayen, “The Global Epidemic of the Metabolic Syndrome,” *Curr Hypertens Rep*, vol. 20, no. 2, Feb. 2018, doi: 10.1007/S11906-018-0812-Z.

[7] J. X. Moore, N. Chaudhary, and T. Akinyemiju, “Metabolic Syndrome Prevalence by Race/Ethnicity and Sex in the United States, National Health and Nutrition Examination Survey, 1988-2012,” *Prev Chronic Dis*, vol. 14, no. 3, 2017, doi: 10.5888/PCD14.160287.

[8] M. Nichols, N. Townsend, P. Scarborough, and M. Rayner, “Cardiovascular disease in Europe 2014: epidemiological update,” *Eur Heart J*, vol. 35, no. 42, pp. 2950–2959, Nov. 2014, doi: 10.1093/EURHEARTJ/EHU299.

[9] W. S. Hui, Z. Liu, and S. C. Ho, “Metabolic syndrome and all-cause mortality: A meta-analysis of prospective cohort studies,” *Eur J Epidemiol*, vol. 25, no. 6, pp. 375–384, 2010, doi: 10.1007/S10654-010-9459-Z.

[10] M. Eyvazlou *et al.*, “Prediction of metabolic syndrome based on sleep and work-related risk factors using an artificial neural network,” *BMC Endocr Disord*, vol. 20, no. 1, Dec. 2020, doi: 10.1186/s12902-020-00645-x.

[11] M. S. Ibrahim, D. Pang, G. Randhawa, and Y. Pappas, “Risk models and scores for metabolic syndrome: Systematic review protocol,” Sep. 01, 2019, *BMJ Publishing Group*. doi: 10.1136/bmjopen-2018-027326.

[12] J. Kim, S. Mun, S. Lee, K. Jeong, and Y. Baek, “Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea,” *BMC Public Health*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12889-022-13131-x.

[13] S. Mottillo *et al.*, "The metabolic syndrome and cardiovascular risk: A systematic review and meta-analysis," *J Am Coll Cardiol*, vol. 56, no. 14, pp. 1113–1132, Sep. 2010, doi: 10.1016/J.JACC.2010.05.034.

[14] A. Bankoski *et al.*, "Sedentary Activity Associated With Metabolic Syndrome Independent of Physical Activity," *Diabetes Care*, vol. 34, no. 2, p. 497, Feb. 2011, doi: 10.2337/DC10-0987.

[15] A. Laudisio, S. Bandinelli, A. Gemma, L. Ferrucci, and R. A. Incalzi, "Metabolic syndrome and functional ability in older age: TheInCHIANTI study," *Clinical Nutrition*, vol. 33, no. 4, pp. 626–633, 2014, doi: 10.1016/j.clnu.2013.08.005.

[16] S. M. Mohamed, M. A. Shalaby, R. A. El-Shiekh, H. A. El-Banna, S. R. Emam, and A. F. Bakr, "Metabolic syndrome: risk factors, diagnosis, pathogenesis, and management with natural approaches," *Food Chemistry Advances*, vol. 3, p. 100335, Dec. 2023, doi: 10.1016/J.FOCHA.2023.100335.

[17] H. Yang *et al.*, "Machine learning-aided risk prediction for metabolic syndrome based on 3 years study," *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–11, Feb. 2022, doi: 10.1038/s41598-022-06235-2.

[18] U. A. Tahir and R. E. Gerszten, "Molecular Biomarkers for Cardiometabolic Disease: Risk Assessment in Young Individuals," *Circ Res*, vol. 132, no. 12, p. 1663, Jun. 2023, doi: 10.1161/CIRCRESAHA.123.322000.

[19] C. M. Povel, J. M. A. Boer, E. Reiling, and E. J. M. Feskens, "Genetic variants and the metabolic syndrome: a systematic review," *Obes Rev*, vol. 12, no. 11, pp. 952–967, Nov. 2011, doi: 10.1111/J.1467-789X.2011.00907.X.

[20] R. Gil-Redondo *et al.*, "MetSCORE: a molecular metric to evaluate the risk of metabolic syndrome based on serum NMR metabolomics," *Cardiovasc Diabetol*, vol. 23, no. 1, p. 272, Jul. 2024, doi: 10.1186/S12933-024-02363-3.

[21] N. Wang, J. Li, Z. Hu, E. E. Ngowi, B. Yan, and A. Qiao, "Exosomes: New Insights into the Pathogenesis of Metabolic Syndrome," *Biology (Basel)*, vol. 12, no. 12, Dec. 2023, doi: 10.3390/BIOLOGY12121480.

[22] C. Liu, J. Liu, Z. Liu, and Y. Yang, "Machine Learning-Based Metabolic Syndrome Identification," *Communications in Computer and Information Science*, vol. 2019 CCIS, pp. 94–101, 2024, doi: 10.1007/978-3-031-52216-1_8.

[23] E. M. Muzurović *et al.*, "Glucagon-Like Peptide-1 Receptor Agonists and Dual Glucose-Dependent Insulinotropic Polypeptide/Glucagon-Like Peptide-1 Receptor Agonists in the Treatment of Obesity/Metabolic Syndrome, Prediabetes/Diabetes and Non-Alcoholic Fatty Liver Disease—Current Evidence," *J Cardiovasc Pharmacol Ther*, vol. 27, Jan. 2022, doi: 10.1177/10742484221146371/ASSET/IMAGES/LARGE/10.1177_10742484221146371-FIG2.JPEG.

[24] J. H. Lee, K. H. Lee, H. J. Kim, H. Youk, and H. Y. Lee, "Effective Prevention and Management Tools for Metabolic Syndrome Based on Digital Health-Based Lifestyle Interventions Using Healthcare Devices," *Diagnostics (Basel)*, vol. 12, no. 7, Jul. 2022, doi: 10.3390/DIAGNOSTICS12071730.

[25] J. H. Lee, K. H. Lee, H. J. Kim, H. Youk, and H. Y. Lee, “Effective Prevention and Management Tools for Metabolic Syndrome Based on Digital Health-Based Lifestyle Interventions Using Healthcare Devices,” *Diagnostics (Basel)*, vol. 12, no. 7, Jul. 2022, doi: 10.3390/DIAGNOSTICS12071730.

[26] A. Assemie, “The value of nutraceuticals in the management of metabolic syndrome,” *~ 46 ~ Journal of Advances in Microbiology Research*, vol. 4, no. 1, pp. 46–52, 2023, Accessed: Aug. 20, 2024. [Online]. Available: www.microbiojournal.com

[27] J. López-Torres, J. Rabanales, and M. J. Simarro, “Effectiveness of a telemedicine programme for patients with metabolic syndrome,” *Technol Health Care*, vol. 23, no. 2, pp. 161–169, 2015, doi: 10.3233/THC-140888.

[28] “Gatekeeper Project | GATEKEEPER PROJECT.” Accessed: Aug. 22, 2024. [Online]. Available: https://www.gatekeeper-project.eu/

[29] J. de Batlle *et al.*, “GATEKEEPER’s Strategy for the Multinational Large-Scale Piloting of an eHealth Platform: Tutorial on How to Identify Relevant Settings and Use Cases,” *J Med Internet Res*, vol. 25, 2023, doi: 10.2196/42187.

[30] “16 Data Normalization Methods Using Python (With Examples) — Part 5 of 6 | by Reina | Medium.” Accessed: Sep. 16, 2024. [Online]. Available: https://medium.com/@reinapeh/16-data-normalization-methods-using-python-with-examples-part-5-of-6-8744cb2b2e15

[31] “Principal Component Analysis (PCA) Explained | Built In.” Accessed: Sep. 24, 2024. [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[32] “Understanding Cumulative Explained Variance in PCA with Python | by Megha Natarajan | Medium.” Accessed: Sep. 25, 2024. [Online]. Available: https://medium.com/@megha.natarajan/understanding-cumulative-explained-variance-in-pca-with-python-653e3592a77c