Proceedings of the Federated Conference on Computer Science and Information Systems pp. 119–128 DOI: 10.15439/2020F202 ISSN 2300-5963 ACSIS, Vol. 21

Generating Fuzzy Linguistic Summaries for Menstrual Cycles

Łukasz Sosnowski
Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
sosnowsl@ibspan.waw.pl

Tomasz Penza
OvuFriend Sp. z o.o.,
Złota 61/100, 00-819 Warsaw, Poland
tomasz.penza@ovufriend.com

***Abstract*—This paper presents a method of generating linguistic summaries of women's menstrual cycles based on the set of concepts describing various aspects of the cycles. These concepts enable description of menstrual cycles that are readable for humans, but they also provide high-level information that can be used as control input for other data processing actions such as e.g. anomaly detection. The labels signifying these concepts are assigned to cycles by means of multivariate time series analysis. The corresponding algorithm is a subsystem of a bigger solution created as a part of an R&D project.**

I. Introduction

INFERTILITY is becoming a civilization disease. Statistics say that every fifth couple that is trying to conceive (TTC) has a problem to achieve pregnancy in the first 12 months of efforts, and this tendency is increasing [1]. In addition, the age of women trying for the first child statistically shifts towards 35. This is a problem because with age the risk of pregnancy problems increases, including the birth of a child with defects, and according to official terminology, pregnancies of women aged over 35 are referred to as ”geriatric pregnancy”.

OvuFriend[1] is a platform for women trying to conceive that allows them to document their menstrual cycle and receive feedback aimed at helping them successfully conceive. Using a mobile app, users provide declarative data about specific parameters of their body and subjective feelings recorded at specific times of the day. By providing this data, they gain access to the algorithms designed to help them conceive, a supportive community and other such tools.

The platform has collected data of over 400,000 menstrual cycles, e.g. symptoms felt in various stages of the cycle and measurements of basic factors used to determine the phase of the cycle and its fertility on a given day [4]. This data include, among others, measurements of baseline body temperature (BBT), type of cervical mucus occurring on particular days of the cycle, parameters of the cervix as well as the results of ovulation tests which measure the concentration of the LH hormone in a woman's body.

The company is conducting a research and development project co-financed from the National Center for Research and Development in Poland, aimed at eliminating barriers related to pregnancy and facilitating effective family planning at home environment. One of the elements prepared under the project is an AI algorithm dedicated to the prediction and confirmation of ovulation [15]. Its approach is to use a set of independent detectors that analyze time series of different parameters of the menstrual cycle to detect parameter-specific information about ovulation. The results of their analyses are aggregated with a set of weights that depends on the phase of the cycle to facilitate differentiation of the ovulation designation into two phases: prognostic and retrospective. Another issue that the project deals with is discovering the vulnerability of medical anomalies from data and sustaining intelligent communication between the system and the women using the OvuFriend's platform.

Automatic description of the menstrual cycle is an additional goal of the project that facilitates understanding of the processes occurring in a woman's body. The description is to describe the parameters of one's own menstrual cycle in such a way that they are understandable to the average woman of childbearing age without medical experience.

This last issue is the subject of this article. However, for a better understanding of the context, Fig. 1 presents a general diagram of the entire solution covered by the R&D project.

A. Overview of the OvuFriend's AI platform

The central element of the architecture is the AI module, which integrates the developed algorithms into a coherent interface that exchanges data between individual elements at the level of processing. This module is powered by data from a data warehouse built to store the cycle data provided by women. The most commonly used functionality is the ovulation algorithm [15]. In the part where insightful comparisons are made with available historical cycles both at the level of a single woman as well as a group of women characterized by similar cycles in terms of selected features, the algorithm uses networks of compound object comparators (NoC) described in detail in [14]. Determining the state of ovulation and the date of its eventual occurrence is the key information for further processing. The descriptions of the cycle that are generated help not only to make the decision regarding ovulation, but they are also used by other parts of the system, e.g. by the detection algorithm for the most common medical anomalies related to menstrual cycles (e.g. endometriosis [3]). However, this work will describe the part of the solution that generates

[1] www.ovufriend.pl

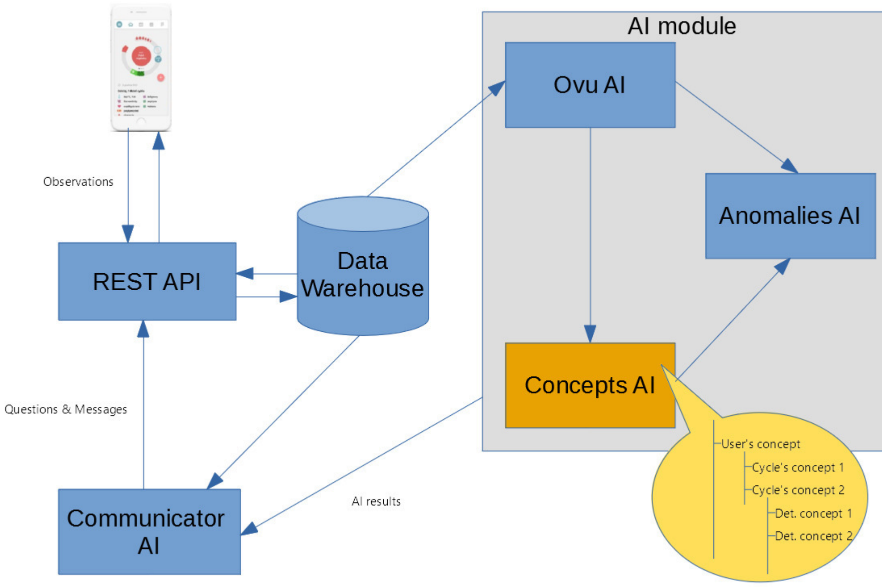

Fig. 1. General scheme of the architecture of the AI module linked with other interfaces inside the OvuFriend's platform

descriptions understandable to women who are the users of the system.

B. Goal description

The issue of cycle description applies to both open (ongoing) menstrual cycles as well as those that are already closed (past cycles). The algorithm provides, for both types of cycles, an accessible and automatically generated description of the given menstrual cycle, that pays attention to a number of features that may be medically important. This description should include local aspects of the cycle, as well as a slightly broader perspective of the whole cycle. The description is to contain the initial classification of the values of the features, compared with the applicable standards. The last aspect is the analysis in a wider time window that captures trends and tendencies arising from the repetition of certain phenomena in a defined historical window. Calculating approximated results based on meta-descriptions and summaries are widely applied in many areas of interest, such as analytical databases [13], large relational data sets [12], redesigning and accelerating machine learning algorithms [2] or systems for monitoring health conditions for members of nursing homes [5]. A slightly different approach is to use Japanese candles as summaries [9], and then compute and process that data. One example is the annual AI competition, which this year is based on such summaries [6].

C. The algorithm for automatic cycle description

The cycle description algorithm analyzes and automatically labels the menstrual cycles of women who are users of the OvuFriend's platform. The analysis takes into account various aspects important from the point of view of confirming the correctness of the entered data or probability of the occurrence of certain specific symptoms associated with medical anomalies. Finally, linguistic summaries describing the given cycle are generated and, optionally, descriptions concerning the user herself are generated (inter-cycle analysis). Generated labels are processed into natural language (messages in a language understood by the end user). This description is intended to facilitate the understanding and interpretation of the user cycle as well as to improve the quality of the data entered. In addition, it is a source of input data in the Anomalies AI module included in Fig. 1, dealing with the analysis of the possibility of medical anomalies. Generation of descriptions is a fuzzy linguistic summary by using a number of quantifiers in accordance with the techniques described in the works [8]. The basis for generating linguistic summaries are label collections, but in order to correctly present the final text, various summary generation techniques are used, ranging from simple static ones, to dynamic ones and ones that aggregate other variables.

D. Layout of the paper

This article consists of five sections. The second section presents the processed data that together constitute the compound object. This description will allow the reader to understand the complexity and relationship of the individual elements making up the representation of the menstrual cycle in the system. The third section presents the methods used to build the solution, the formal foundations and the definitions of individual components of the solution. The fourth section describes the method of evaluating the correctness of the solution and the results achieved. The last section presents the discussion of the results and the plan of further work on the issue.

II. Menstrual cycle's representation

The central object of interest is the menstrual cycle described by the ensembles of time series inter-correlated with each other, constructed from observations taken by women. The individual component time series are indexed with the same time quanta representing particular days of the cycle. Depending on the cycle and the woman's behavior, there are many possible combinations of data types that constitute this multivariate time series [4]. Time series are typically associated with financial applications [11] but in this case we are dealing with a specific case of a cyclic time series, where a single cycle in multivariate version is considered. The set of features contains: BBT, cervical mucus, cervix parameters, LH urinary tests, pregnancy tests, statistics and occurrence of user-specific symptoms that may signify approaching ovulation.

BBT data consists of temperature values. The measurements are compared to the mean temperature of the previous 6 days. At the same time other factors are computed (eg. mean, relative difference, etc.) and stored together with the BBT time series for later processing.

Cervical mucus is defined by one of five possible values taken from the enumerative scale: dry, sticky, creamy, watery, stretchy. Each value describes different state of the mucus. Making use of this parameter requires detection of patterns in its variability. Therefore it is not enough to get a single measurement. The data should be collected day by day in a certain range.

Cervix has three parameters that can be tracked: opening, position and texture. Each of them has three values respectively: {open, medium, closed}, {high, medium, low}, {soft, medium, hard}. The observations are collected independently, but the interpretation of the whole state depends on all these values combined (at least two of them). These data create an additional nested three dimensional time series that describes one feature.

Ovulation test has a binary value: positive or negative. However there are some difficulties with interpreting its result which sometimes leads to wrong classification as one of these two states on part of the user. In this type of data a series of measurements is also required, in particular one containing a transition from negative to positive values. A single positive measurement is often not enough to accurately determine ovulation day.

Pregnancy test also has binary positive and negative values. If a woman got pregnant during the cycle, the pregnancy test will come out positive, but only if it was taken an appropriate amount of time after the ovulation. Thus both positive and negative values of pregnancy test in such a cycle may convey some information on the date of the ovulation.

Statistics are useful, because the length of the luteal phase is expected to be constant across a given woman's menstrual cycles. Simple statistical data particular to the user are used: average cycle and luteal phase lengths, as well as typical values based on clustering concerning luteal phase, cycle length, ovulation days, etc.

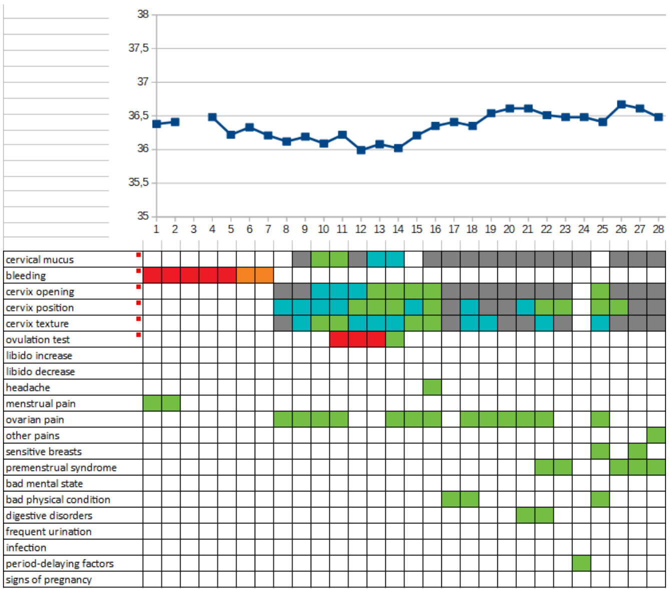

Fig. 2. Multivariate time series of menstrual cycle

Symptoms are the most complex feature in terms of stored information. It consists of more than 80 elements which describe symptoms (e.g. various pains, mental states, infections, libido, etc.) on a single day of the cycle. Most of them are binary, but together they create a complex structure. Elementary symptoms are granulated and combined into groups of similar elements.

A representation of a single menstrual cycle can be any combination of these data. Moreover, each of the time series independently may require handling of missing values and of imprecision of processed values [8]. An important element is a correlation of the particular sub-time-series. Thus the need arises to create a representation whose values are determined by the mutual influence of the individual parts of the multivariate time series.

An example of a graphical interpretation of this multivariate time series representing the menstrual cycle is shown in Fig. 2.

Such combined time series for each cycle constitutes a *compound object* described by various features and consisting of many sub-objects in the sense of the definition in [14].

III. Methods used

A. Ontology of concepts

The cycles are described through the lenses of concepts defined at three levels of the hierarchy. The lowest level of the hierarchy concerns concepts assigned at the level of a single dimension of the particular time series of the input object, e.g. the menstrual cycle described by the given data type (mucus, cervix, BBT, etc.). The second level aggregates the previous one and concerns observations for the whole cycle. On the third - highest - level of the hierarchy are the concepts obtained as a result of the analysis of user's historical cycles in a fixed time window. The concepts are connected by relationships among themselves, which generally falls under the definition of ontology.

The name of the *ontology* comes from philosophy, but now it is also frequently found in the field of artificial intelligence (AI). The formal definition (one of many) was introduced in 2001, described in [16]. Its meaning is as follows: ontology is a system marked as $O = \{C, R, H_c, rel, A, L\}$, which specifies the structure of concepts, relationships between them as well as theory defined on a model, where: C is the set of all concepts of the model and the concept is called the idea of representing a group of objects with common characteristics. R is a set of non-taxonomic relations defined as named connections between concepts [16], H_c - a collection of taxonomic relationships between concepts, rel - defined non-taxonomic relationships between the concepts, A - a set of axioms, L - lexicon defining the meaning of concepts (including relations). L is a set of the form $\{L_c,\ L_r,\ F,\ G\}$, where L_c - lexicon of definitions for concepts, L_r - lexicon of definitions of a set of relationships, F - references to concepts, G - references to relationship.

There are many interesting applications of ontology described in the literature covering many fields, e.g. pattern recognition, image analysis or modelling situational awareness by AI systems [17]. In all these cases ontology is a tool for modelling the structure of concepts and relationships describing a selected part of the local context in which the system is described [18].

In the simplest sense, ontology is a set of concepts connected one with another through named relationships. Ontological concepts can create hierarchies by grouping more specific concepts into more general entities. This form is used, for example, to model mereologic relations, which describe dependencies between parts of objects [10].

In the context of this work, ontology is used as a set of concepts describing menstrual cycles with its structure and relations. It is used for preparing meta-representation of the object, ready to further processing, e.g. comparing each other, clustering or generating human readable linguistic descriptions.

A fragment of the ontology is presented on Fig. 3 as an example. It shows in particular how concepts of higher levels are obtained from the concepts of lower levels using the algebra of labels.

B. Overview of the designed solution

First, the designed algorithm designates labels for particular detectors, the cycle and the user with terms corresponding to various ontological concepts, and then it generates the cycle description in natural language. The concepts are organized in a three-level hierarchy and processing consists of three steps. At the first level there are simple atomic concepts regarding directly the aspects of the cycle related to the parameters analyzed at the level of a given data type (one dimension of the cycle time series). Higher levels of the ontological concepts are built on the basis of atomic concepts. First, the concepts of a single cycle are built - this is the second level of the hierarchy. These concepts constitute knowledge of the cycle using calculations based on labels from the previous level. The comprehensive description of a single cycle obtained in this way allows defining the concepts of the third level of the hierarchy regarding the behavior and condition of a woman in time. These concepts are defined based on the occurrence of individual concepts related to the cycle in user cycles in a historical time window. They allow detecting the persistent features of cycles that characterize a woman (her condition and the functioning of the biological mechanisms).

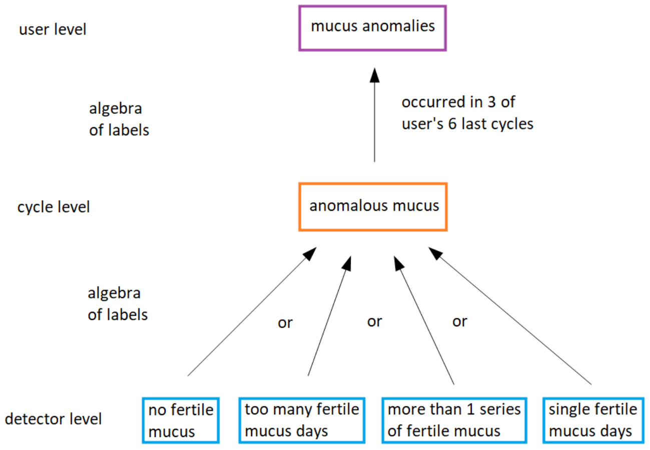
user level
mucus anomalies
algebra of labels
occurred in 3 of user's 6 last cycles
cycle level
anomalous mucus
algebra of labels
or
or
or
detector level
no fertile mucus
too many fertile mucus days
more than 1 series of fertile mucus
single fertile mucus days

Fig. 3. A fragment of used ontology together with an example of using the algebra of labels to obtain higher level concepts.

Concepts are defined in natural language as various features that cycles may possess. This set was developed by medical experts who identified interesting aspects in the cycle that should be monitored. Then these concepts were defined in the form of predicates that are verified during processing of the data. If the predicate is satisfied, the object (cycle, detector representing the given data type or woman herself) is assigned a label. If condition is not met, the label does not appear in the context. Due to the fact that labels address both positive (normative) features and anomalies (non-normative), in both cases certain subset of labels will be allocated.

The rest of the section provides ontological definitions of concepts at individual levels of the hierarchy and some details of the associated conditions that have to be met in order for a concept label to be assigned. The descriptions given for these conditions are general and short, they don't go into the details. Many of the conditions are fuzzy - they can be satisfied to a degree. The label then is assigned to that same degree. As an example consider the label *temperature jump detected*. The occurrence of the temperature jump is decided in a fuzzy way - a clear-cut jump is recorded if the temperature raises at least 0.2°C over the mean of the last 6 days. But the definition is extended in the fuzzy way to accommodate raises of even 0.18°C. If the temperature jump is satisfied to a certain degree, then the label *temperature jump detected* is considered to have such degree of being assigned.

Labels corresponding to the lowest level concepts are set by the detectors from the ovulation detection algorithm, because the same set of input data is used to determine them. Some labels are common to all detectors, and some dedicated to specific detectors. Below, the lowest level concepts are listed, divided by the type of data for which the label is issued.

1) Concepts of level 1 - data types: Common concepts (labels) appearing in all types of analyzed data at the first level of the hierarchy:

- *No data* - the analyzed data type has not been entered at all or the boundary conditions for the occurrence of a given data type have not been met (for different types, the boundary conditions specify the minimum amount of data necessary to perform calculations).
- *Few measurements* - the measurements analysis showed that the occurrence of data in critical areas of the cycle is low. The label is set if the amount of data entered is relatively low according to the specifications of the given detector (unit to process data of a given type in ovulation detection).
- *Cycle irregular* - the cycle shows a deviation from the standard pattern in terms of the parameter being processed (type of data). Label is set if the data entered deviate from the standard according to the specifications of the given detector.
- *No ovulation detected* - measurement analysis did not lead to detection of ovulation. The label is set if the detector has not determined ovulation by exceeding the activation threshold. Assigning this label does not mean that the cycle is anovulatory.

Concepts dedicated to temperature analysis:

- *Not double phased* - the cycle is not divided into a lower temperature phase and a higher temperature phase. Label is set if the temperature detector does not detect a temperature jump that has been confirmed.
- *Temperature fluctuation* - temperature fluctuations occur in the cycle. Label is set if the temperature detector has discovered at least two temperature jumps that didn't persist.
- *Imprecise thermometer* - the thermometer used for measurements has low precision - it measures the temperature only to one decimal place. Label is set, if no temperatures entered by the user have a non-zero value in second place after the decimal point, while the minimum requirement for the number of is attained.
- *Jump detected* - A temperature jump has been detected in the cycle. Label is set if the temperature jump is detected and there is no confirmed temperature jump yet.
- *Confirmed jump detected* - there is a confirmed temperature jump in the cycle.

Concepts dedicated to cervical mucus analysis:

- *No fertile mucus* - there are no days in the cycle when the mucus is fertile (stretchy or watery). Label is set if there isn't any fertile mucus, even though the cycle is closed or its length has exceeded the upper predicted limit of fertile days.
- *Too many fertile mucus* - there are too many fertile mucus days in the cycle. The label is set if the period between the first and last day of fertile mucus exceeds the specified threshold value.
- *Menstruation long* - the length of menstruation exceeds the established norm, but it falls within the extended norm (acceptable from medical point).
- *Menstruation too long* - the length of menstruation exceeds the extended norm.
- *Single fertile mucus* - single fertile mucus days occur in the cycle. Label is issued if there are single days of fertile mucus surrounded by days of infertile mucus.
- *Fertile mucus series occurred* - a series of fertile mucus appeared in the cycle. The label appears if in the cycle there were two days with fertile mucus next to each other or separated at most by one infertile day.
- *Fertile mucus series finished* - label is set, if after a fertile series there were at least two infertile days (or no data).
- *More than 1 mucus series* - label is set, if more than one series occurred in mucus data. The label means the irregularity in the cervical mucus data.
- *Vaginal infection* - vaginal infection appeared in the cycle. The label is set if at least one day has appeared in the cycle with vaginal infection.

Concepts dedicated to cervix analysis:

- *No fertile cervix* - there are no days in the cycle when the cervix is in a fertile phase, even though the cycle is closed or its length has exceeded the upper forecasted limit of fertile days.
- *Too many fertile cervix days* - there are too many days in the cycle when the cervix is fertile.
- *Single fertile cervix after series* - there are single days in the cycle when the cervix is in the fertile phase surrounded by days when the cervix is in the infertile phase.
- *Fertile cervix series occurred* - a series of fertile cervix appeared in the cycle. Label is issued if two consecutive days appeared in the cycle indicating fertile cervix parameters.
- *Fertile cervix series finished* - the series of fertile cervix indications ended. Label is issued if after a fertile series there were at least two days of infertile cervix (or no data).

Concepts dedicated to ovulation test analysis:

- *No positive ovulation test* - there are no days in the cycle when the ovulation test is positive, even though the cycle is closed or its length has exceeded the predicted limit of fertile days.
- *Too many days of positive tests* - there are too many days in the cycle when the test is positive. Label is set if the period between the first and last day when the test is positive exceeds the specified threshold.

- *First positive ovulation test* - a positive ovulation test appeared in the cycle.
- *Series of positive tests finished* - the first negative test appeared after positive tests.
- *LH hormone irregular* - generated pattern from ovulation tests results indicates irregularity.

Concepts dedicated to the ovulation monitor analysis:

- *Measurements started too late* - measurements were performed in a given cycle, but contrary to the instructions of the ovulation monitor, they were started too late (after the 10'th day of the cycle).

Concepts dedicated to symptoms analysis:

- *A lot of pain* - there are a large number of days with marked pain symptoms in the cycle. The label is issued if pain symptoms occur in the percentage of days in the cycle exceeding a certain threshold.
- *Positive pregnancy test* - a reliable positive pregnancy test appeared in the cycle.
- *Menstruation phase* - the cycle is currently in the menstrual phase. Label is issued while current cycle day is in the range of the bleeding series which started at the beginning of the cycle.
- *Follicular phase* - phase of the cycle after the end of menstruation, but before ovulation, determined for ongoing cycles. The label is issued when the menstruation is over and ovulation symptoms have not yet occurred.
- *Ovulation phase* - concept assigned usually for one day. Label is issued if the current day of the cycle coincides with the ovulation forecast.
- *Luteal phase* - phase after ovulation in the cycle. Label is issued for open cycles with designated ovulation.

Concepts dedicated to the user's history analysis:

- *No personal reference set* - there are no historical cycles of the user that are completed, ovulatory and the credibility of ovulation is at least on a certain level defined with a parameter.
- *Small personal reference set* - there are only few historical cycles of the user that meet the criteria of being closed, ovulatory and of an appropriate level of reliability of the determined ovulation.

Concepts dedicated to the history of the user's profile analysis:

- *No profile reference set* - there are no historical cycles of users from the user's profile that meet the criteria of closure, ovulatory and the required value of reliability for determining ovulation.
- *Small profile reference set* - there are only few historical cycles from user's profile that meet the condition of being closed, ovulatory and of an appropriate level of reliability of the determined ovulation.

2) Concepts of level 2 - cycles: Based on level 1 concepts (section III-B1), more general concepts are built at the level of the entire cycle. The analysis confronts the occurrence of premises in various types of data. In this way, more and more general knowledge is obtained based on the processing of individual levels of concepts. The construction of generalized concepts is performed using the so-called *algebra of labels*, e.g. the mechanism of constructing higher-level concepts based on the presence of specific lower-level concepts using logical operations (alternative, conjunction, etc). Concepts of cycles are assigning only for already closed cycle. The definitions are given below along with some details required for calculating conditions.

- *No data* - no measurements of any type have been entered in the cycle. Label is issued if each detector has assigned a label *No data* in the first level.
- *Few measurements* - few measurements have been entered to indicate ovulation reliably. Label is issued if each detector has assigned *Few measurements* or *No data* labels, and at least one has assigned the label *Few measurements*.
- *Anovulatory* - the multivariate analysis of the input time series did not show behavior characteristic for particular types of given data, or the indications were completely divergent. On this basis, it is assumed that such a cycle is anovulatory. The label is issued if the aggregation of the responses of individual detectors did not determine the day of ovulation with a certainty exceeding the learned threshold and the labels *No data* and *Few measurements* are not set.
- *Luteal phase too long* - the luteal phase exceeds the standard length and falls outside the extended norm. The label is issued if ovulation with certainty exceeding the learned threshold has been determined and the obtained luteal phase is longer than the specified value.
- *Luteal phase long* - the luteal phase exceeds the standard length but is within the enlarged norm and the ovulation has been determined with credibility exceeding the learned threshold.
- *Luteal phase ok* - the luteal phase is normal. The label is issued if ovulation credibility exceeds the learned threshold and the obtained luteal phase is 12-16 days long.
- *Luteal phase short* - the luteal phase is shorter than the specified norm but falls within the enlarged norm and ovulation was determined with credibility exceeding the learned threshold.
- *Luteal phase too short* - the luteal phase is shorter than the specified norm and does not fall within the increased norm, at the same time the ovulation was determined with a credibility exceeding the learned threshold.
- *Biochemical pregnancy* - the possibility of biochemical pregnancy occurred. The label is issued if ovulation has been designated without the luteal filter and at an appropriate interval from ovulation a pregnancy test was performed with a positive result, but menstruation occurred no later than 46 days after the start of the cycle.
- *Intermenstrual bleeding* - bleeding occurs within the cycle. Label is set if bleeding occurs after the menstruation, but not in the vicinity of ovulation (which is normal).

- *Anomalous temperature* - time series analysis of temperature shows deviations from the defined double phased pattern. Label is issued if the temperature detector has assigned the labels *Irregular cycle* or *No double phase* or *Temperature fluctuation*.
- *Anomalous mucus* - mucus time series analysis shows deviations from the defined pattern of changes in cervical mucus.
- *Anomalous cervix* - analysis of the time series of cervix parameters shows deviations from the defined pattern of changes in cervix parameters.
- *Anomalous hormones* - time series analysis of ovulation test results or ovulation monitor measurements shows deviations from the defined pattern of changes in the test results.
- *Anomalous symptoms* - prolonged persistence of individual symptoms.
- *Sufficient variety of data* - the variety of entered measurements is sufficient - one can try determine ovulation in a reliable way.
- *Good variety of data* - the variety of entered measurements is good - one can try determine ovulation in a reliable way.
- *Intercourse on ovulation day* - there was an intercourse on the ovulation day.
- *No intercourse in ovulation day* - there was no intercourse on the ovulation day.
- *Intercourse in the area of ovulation* - the woman had an intercourse close enough to ovulation to have a chance of pregnancy.
- *No intercourse in the area of ovulation* - the woman didn't have an intercourse close enough to ovulation for a chance of pregnancy.
- *The X parameter and the Y parameter do not match* - There is a mismatch between the given parameters in terms of ovulation indications.
- *The X parameter does not match the rest of the parameters* - a given parameter deviates from the rest of the parameters in terms of ovulation indications.
- *PCOS symptoms* - symptoms characteristic of PCOS were detected in the cycle.
- *Endometriosis symptoms* - symptoms characteristic of endometriosis were detected in the cycle.
- *Thyroid disease symptoms* - symptoms characteristic of thyroid disease were detected in the cycle.
- *Hyperprolactinaemia symptoms* - symptoms characteristic of hyperprolactinaemia were detected in the cycle.

3) Concepts of level 3 - woman's health: The third level of the labels hierarchy is based on the previous two levels. Top-level labels - for a woman (user of the platform) - are assigned based on the labels of her latest cycles history. The history is considered in the fixed length window, controlled by the parameter e.g. 3 months, 6 months, etc.

The most generalized concepts (labels) are defined as follows:

- *Short cycles* - in the history of the user analyzed in a fixed time window, most of the cycles have length below the established norm.
- *Long cycles* - in the history of the user analyzed in a fixed time window, most of the cycles have length above the established norm.
- *Short luteal phases* - in the history of the user analyzed in a fixed time window, most cycles have a luteal phase length below the established norm.
- *Long luteal phases* - in the history of the user analyzed in a fixed time window, most cycles have a luteal phase length above the established norm.
- *Chronic anovulation* - in the history of the user analyzed in a fixed time window, most of the cycles have the characteristics of an anovulatory cycle.
- *Temperature anomalies* - in the history of the user analyzed in a fixed time window, most of the cycles have temperature anomalies.
- *Mucus anomalies* - in the user's history analyzed in a set time window, most cycles have mucus anomalies.
- *Cervix anomalies* - in the user's history analyzed in a set time window, most cycles have cervix anomalies.
- *Hormonal anomalies* - in the user's history analyzed in a set time window, most cycles have hormonal anomalies.
- *Symptom anomalies* - in the history of the user analyzed in a fixed time window, most cycles have symptom anomalies.
- *Menstruations long* - in the user's history analyzed in a fixed time window, most cycles have menstruation length exceeding the norm.
- *Chronic pain* - in the user's history analyzed in a fixed time window, most cycles have prolonged periods of pain symptoms.
- *Does not enter parameter X* - in the history of the user analyzed in a fixed time window, in most cycles, the user did not enter data on a given parameter.
- *Insufficiently enters parameter X* - in the history of the user analyzed in a fixed time window, in most cycles, the user did not enter enough data of a given parameter to be able to reliably indicate ovulation.
- *No measurements* - in the history of the user analyzed in the fixed time window, in most cycles, the user did not enter any relevant data that would allow calculation of ovulation detection.
- *Insufficient measurements* - in the history of the user analyzed in a fixed time window, in most cycles, the user did not enter enough data to be able to reliably indicate ovulation.
- *Incorrectly measures the ovulation monitor* - in the history of the user analyzed in the fixed time window, in most cycles the user performed ovulation monitor measurements contrary to the instructions in the manual.
- *Possible PCOS* - in the history of the user analyzed in a fixed time window, most cycles have symptoms typical for PCOS.
- *Possible endometriosis* - in the user history analyzed in

a fixed time window, most cycles have symptoms typical for endometriosis.

- *Possible thyroid disease* - in the user history analyzed in a fixed time window, most cycles have symptoms typical for thyroid disease.
- *Possible hyperprolactinaemia* - in the user history analyzed in a fixed time window, most cycles have symptoms typical for hyperprolactinaemia.

The set of labels obtained in this way is a linguistic summary. Operators (quantifiers) analyzing a certain window of woman's historical cycles and counting occurrences are used to designate individual labels. This summary is fuzzy, because the predicates used to assign the labels are based on fuzzy rules. It is a very important that higher levels perform operations on labels (linguistic summaries) of lower levels, so the higher in the hierarchy of concepts, the higher level of generalization is obtained. A very important feature is the decomposability of labels, which provides the functionality of selecting subsets of cycles to be searched using more advanced methods, having a designated top-level label. The user's level label covers a certain group of cycles, and these cycles consist of an even larger set of decomposed 1-dimensional time series corresponding to a given data type (processed by specific ovulation detectors). In other cases, such an aggregated linguistic description is sufficient to perform some calculations. An example can be the summary *Short cycles*, from which you can easily conclude that most cycles are shorter than the assumed norm. This information can be used to predict the length for a new cycle. In the presented algorithm, the processing does not end at this point, another step is introduced to generate descriptions in natural language.

C. *Fuzzy linguistic summaries of menstrual cycles*

After determining the labels, each cycle is described by a set of concepts at individual levels of the hierarchy. These sets of information allow for generating of natural language description that can easily be understood by a woman trying to conceive. The next stage of the algorithm for generating linguistic summaries is based on dynamic templates responsible for defining the sets of possible options used during the generation. The template is responsible for the structure of the description, elements taken into account, their order and the information scope of the researched summaries. These templates can be built from several types of summaries. The simplest form is singleton summaries, which are responsible for mapping the label to a sentence in natural language. This corresponds to the *exists* quantifier for simple fuzzy linguistic summaries [7]. If the label appears in the cycle or user representation set then a simple summary related to this information will be generated. These are relatively simple summaries that do not take into account interrelationships and context. There can be any number of the singleton summaries in the template and they can be arranged in any order.

The second type of summaries used in the template are the so-called aggregated summaries. These summaries concern the occurrence of a given feature for many elements simultaneously. Using singleton summaries in this case would create an unnatural text with repetitions. So, it is better to use an aggregated summary that has combining capabilities, e.g. the main part of the summary states that specific premises are met, and the second specifies which objects or data it concerns. These summaries use the method of grouping and enumerating values. If a given label exists in multiple data types, then one summary sentence will be produced covering the different data types. This makes the text more user-friendly.

The third type of summaries used in the described solution are generalizing summaries. They use quantifiers such as:

- for all - requires fulfillment of fuzzy predicate for all elements,
- exist - requires at least one elements which fulfills fuzzy predicate,
- most - requires fulfillment of fuzzy predicate for a majority of elements,
- at least two - requires fulfillment of fuzzy predicate for two or more elements,
- at least three - requires fulfillment of fuzzy predicate for three or more elements,
- almost majority - requires fulfillment of fuzzy predicate for number of elements which is nearly a majority.

These quantifiers are able to count the occurrences of appropriate labels and evaluate them in relation to each other (depending on the number of data types present, e.g. the *most* operator refers to the data types defined in a given menstrual cycle and not all possible ones). Depending on the fulfillment of individual quantifiers (it is worth noting here that the conditions may be met for more than one), a different form of summary can be returned. Such a summary, using different quantifiers, can be defined in the template in the form of an alternative or a conjunction. In the first case, the order in which the components are set controls the order in which the conditions are checked. So, the first satisfied summary in this type of alternative is returned. If the conjunction is used, each of the conditions generated by the quantifiers must be met.

The description template is therefore any combination of the summaries of these three types. If the data for a given summary does not meet the condition given by using the appropriate quantifier, it does not return any value. Despite appearing in the template, it does not interfere with the generation of text in natural language for the cycle or the user.

By using three types of summaries and combining conditions using conjunctions or alternatives, it is possible to define very complex label-based schemes that generate various linguistic summaries that stylistically differ very much, depending on the input data, despite using the same description template.

D. *Illustrative example*

Fig. 4 presents an example of a real menstrual cycle coded as a multivariate time series. Summaries generated for the level 1 of the hierarchy of the ontology are shown in Table I.

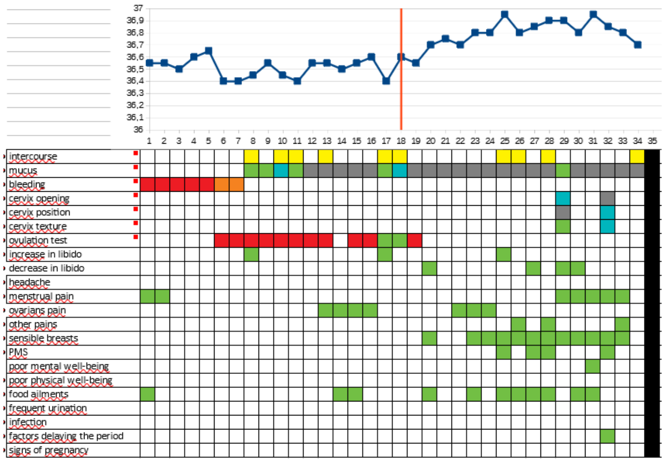

Fig. 4. Example of a real menstrual cycle retrieved from OvuFriend's platform

TABLE I
LINGUISTIC SUMMARY OF THE MENSTRUAL CYCLE BASED ON THE LEVEL 1 OF THE HIERARCHY OF THE ONTOLOGY.

Data type	Label
Ovulation test	series of positive tests finished
Ovulation monitor	no data
Cervix	few measurements
Cervix	fertile cervix series finished
Mucus	cycle irregular
Mucus	more than 1 mucus series
Mucus	fertile mucus series finished
Mucus	single fertile mucus
BBT	cycle irregular
BBT	confirmed jump detected

Next, the *algebra of labels* provided higher level of summaries, which are presented in Table II. In this simple example there is no historical cycles, so the third level of summaries cannot be designated. Even though using the pattern of linguistic description one can generate a human readable text, which for this particular cycle will be in the following form: *Cycle is ovulatory. The cycle length is normal. The length of the luteal phase is normal. The provided data is of good variety. The cervical parameters differ from the rest of the data entered. Make sure that you measure it correctly. There are anomalies in cervical mucus and temperature. Intercourse around ovulation has been observed, which gives a chance of getting pregnant.*

TABLE II
LINGUISTIC SUMMARY OF THE MENSTRUAL CYCLE BASED ON THE LEVEL 2 OF THE HIERARCHY OF THE ONTOLOGY.

ovulatory cycle
cycle length Ok
good data variety
cervix doesn't agree with the rest
anomalous mucus
cycle irregular
intercourse in fertile period
anomalous temperature
luteal phase ok

TABLE III
EVALUATION OF FUZZY LINGUISTIC SUMMARY GENERATION IN THE FORM OF LABELS FOR MENSTRUAL CYCLES. EVALUATION PERFORMED ON A SUBSET OF LEVEL 1 LABELS THAT MOST INFLUENCE PREGNANCY. EVALUATION MADE ON 100 CYCLES TAGGED BY MEDICAL EXPERTS. WHOLE SET WAS A TESTING SET.

Label	TP	TN	FP	FN	Pr	Rec	F1	Acc
Cycle too short	3	97	0	0	1	1	1	1
Cycle short	10	90	0	0	1	1	1	1
Cycle length ok	72	28	0	0	1	1	1	1
Cycle long	8	92	0	0	1	1	1	1
Cycle too long	7	93	0	0	1	1	1	1
Ovulation cycle	72	21	2	5	0.97	0.94	0.95	0.93
No double phase	20	74	4	2	0.83	0.91	0.87	0.94
Menstruation too short	5	95	0	0	1	1	1	1
Menstruation long	8	91	0	1	1	0.89	0.94	0.99
Menstruation too long	4	96	0	0	1	1	1	1
Intermenstrual bleeding	21	77	1	1	0.95	0.95	0.95	0.98
No fertile mucus	5	94	1	0	0.83	1	0.90	0.99
Too many fertile mucus	5	94	0	1	1	0.83	0.91	0.99
Mucus more than 1 series	15	83	1	1	0.94	0.94	0.94	0.98
Single fertile mucus	12	85	1	2	0.92	0.86	0.89	0.97
No fertile cervix	4	95	1	0	0.80	1	0.89	0.99
Too many fertile cervix	39	61	0	0	1	1	1	1
Cervix more than 1 series	19	78	2	1	0.90	0.95	0.92	0.97
Single cervix days	5	95	0	0	1	1	1	1
No positive o. t.	6	93	0	1	1	0.86	0.92	0.99
Too many positives o. t.	6	93	1	0	0.86	1	0.92	0.99
O. t. more than 1 series	7	93	0	0	1	1	1	1
Mucus irregularity	32	62	1	5	0.97	0.86	0.91	0.94
Cervix irregularity	48	48	2	2	0.96	0.96	0.96	0.96
LH hormone irregular	18	79	2	1	0.90	0.95	0.92	0.97
Averaged	451	2007	19	23	0.96	0.95	0.96	0.98

IV. EVALUATION AND RESULTS

As part of the project, medical experts selected a set of labels at the lowest level. The designation of these labels was carried out by the described algorithm. The development of the algorithm and its initial fine-tuning was developed on a set of 200 cycles tagged by medical experts. This set has been treated as a learning set (whole), although in this case it is not a classical learning mechanism with feedback. The algorithm has been tuned for operation on the tagged set and pre-validated in terms of content-related correctness of operation. The next step was to draw a new set of cycles having an empty intersection with the previous one and return the set for tagging. In this case, the collection of labels has been limited to those most related to pregnancy, that is, among others, labels directly or indirectly related to the anomalies. Although in the classic approach to learning, the training set is usually smaller than the testing set, here, however, due to the non-typical nature of the process, the opposite proportions were used (66% training set, 34% testing set, respectively). Therefore, the results of the experiment are given for a set of 100 menstrual cycles, which were tagged in a second attempt by medical experts to evaluate the quality of the automatic cycle description algorithm. The other higher-level labels are derivative concepts, well defined by the *algebra of labels* mentioned above, therefore they were not evaluated in this experiment. The table III presents the results of the experiment dividing into individual labels, as well as the summary efficiency calculated based on the sum of the contingency tables for individual labels. After the calculation of the summary table, the Precision, Recall, F1Score and Accuracy evaluation measures were calculated.

The obtained results are very good. All tested labels attained the minimum requirement of at least 0.8 value on both Precision and Recall, and on most labels these values are much higher.

V. Conclusions

The described solution shows how in a relatively simple way, using fuzzy quantifiers, one can create an effective algorithm that generates linguistic summaries and patterns from complex multidimensional data. The developed algorithm can be used both to generate natural language text that describes compound objects, as well as to generate high level information used to control other processes in the system. In this approach, the processed information is largely aggregated and compressed. The need to analyze decomposed data occurs sporadically, and in most cases it is enough to control the process or even make decisions based on linguistic descriptions constructed in this way.

It is worth noting that the working compliance of the mechanism with decisions of medical experts is very high. An additional difficulty in this case was the correlation with the second algorithm and the results it achieved - the algorithm of prediction and confirmation of ovulation. Linguistic summaries were generated while performing those calculations and many of them depended on the decisions made by the ovulation algorithm. Therefore, the achieved results confirm that the operation of both algorithms is compatible with the intuition and knowledge of medical experts. The average precision at the level of 0.96 and the recall at the level of 0.95 allow to treat all generated linguistic summaries and the final generation of description in natural language with enough confidence.

Further work will focus on the development of the described algorithm at the stage of higher-level labels, so that more and more information can be deduced and processed based on the generated patterns (made of linguistic summaries). In addition, in terms of implementation work, the algorithm will be implemented on the production platform and will work in fully real conditions, which will be an extension of the current state, which was developed in conditions similar to real ones (real data but supported from a backup environment).

Acknowledgement

Co-financed by the EU Smart Growth Operational Programme 2014-2020 under the project "Development of New World Scale Solutions in the Field of Machine Learning Supporting Family Planning and Overcoming the Infertility Problem", POIR.01.01.01-00-0831/17-00.

References

[1] L. Bablok, W. Dziadecki, I. Szymusik, and et al., "Patterns of infertility in Poland - multicenter study," *Neuro Endocrinol Lett.*, vol. 32, no. 6, pp. 799–804, 2011.

[2] A. Chadzynska-Krasowska, P. Betlinski, and D. Slezak, "Scalable machine learning with granulated data summaries: A case of feature selection," in *Proceedings of ISMIS 2017, Warsaw, Poland*, ser. Lecture Notes in Computer Science, vol. 10352. Springer, 2017, pp. 519–529. [Online]. Available: https://doi.org/10.1007/978-3-319-60438-1_51

[3] W. Damian, S. Iwona, W. Miroslaw, and P. Bronislawa, "The impact of endometriosis on the quality of life and the incidence of depression-a cohort study," *Int. J. Environ. Res Public Health*, vol. 17, no. 10, p. 3641, 2020.

[4] J. Fedorowicz, L. Sosnowski, D. Slezak, I. Szymusik, and et al., "Multivariate ovulation window detection at OvuFriend," in *Proceedings of IJCRS 2019, Debrecen, Hungary*, ser. Lecture Notes in Computer Science, vol. 11499. Springer, 2019, pp. 395–408. [Online]. Available: https://doi.org/10.1007/978-3-030-22815-6_31

[5] A. Jain, M. Popescu, J. M. Keller, M. Rantz, and B. Markway, "Linguistic summarization of in-home sensor data," *J. Biomed. Informatics*, vol. 96, 2019. [Online]. Available: https://doi.org/10.1016/j.jbi.2019.103247

[6] A. Janusz, M. Przyborowski, P. Biczyk, and D. Ślęzak, "Network device workload prediction: A data mining challenge at knowledge pit." in *Proceedings FedCSIS 2020, Sofia, Bulgaria*, 2020.

[7] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," *International Journal of General Systems*, vol. 30, no. 2, pp. 133–154, 2001. [Online]. Available: https://doi.org/10.1080/03081070108960702

[8] J. Kacprzyk and S. Zadrozny, "Fuzzy logic-based linguistic summaries of time series: a powerful tool for discovering knowledge on time varying processes and systems under imprecision," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 6, no. 1, pp. 37–46, 2016. [Online]. Available: https://doi.org/10.1002/widm.1175

[9] W. Kosiński and A. Chwastyk, "Ordered fuzzy numbers in financial stock and accounting problems," in *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, 2013, pp. 546–551.

[10] L. Polkowski and P. Artiemjew, *Granular Computing in Decision Approximation - An Application of Rough Mereology*, ser. Intelligent Systems Reference Library. Springer, 2015, vol. 77. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-12880-1

[11] M. Romaniuk and P. Nowak, "Monte carlo methods : Theory, algorithms and applications to selected financial problems," Warszawa, 2015.

[12] D. Slezak, J. Borkowski, and A. Chadzynska-Krasowska, "Ranking mutual information dependencies in a summary-based approximate analytics framework," in *2018 International Conference on High Performance Computing & Simulation, HPCS 2018, Orleans, France, July 16-20, 2018*. IEEE, 2018, pp. 852–859. [Online]. Available: https://doi.org/10.1109/HPCS.2018.00137

[13] D. Slezak, R. Glick, P. Betlinski, and P. Synak, "A new approximate query engine based on intelligent capture and fast transformations of granulated data summaries," *J. Intell. Inf. Syst.*, vol. 50, no. 2, pp. 385–414, 2018. [Online]. Available: https://doi.org/10.1007/s10844-017-0471-6

[14] L. Sosnowski, "Compound objects comparators in application to similarity detection and object recognition," *Trans. Rough Sets*, vol. 21, pp. 169–300, 2019. [Online]. Available: https://doi.org/10.1007/978-3-662-58768-3_6

[15] L. Sosnowski, I. Szymusik, and T. Penza, "Network of fuzzy comparators for ovulation window prediction," in *Proceedings of IPMU 2020*, ser. Communications in Computer and Information Science, vol. 1239. Springer, 2020, pp. 800–813. [Online]. Available: https://doi.org/10.1007/978-3-030-50153-2_59

[16] S. Staab and A. Maedche, "Knowledge Portals: Ontologies at Work," *AI Magazine*, vol. 22, no. 2, pp. 63–75, 2001.

[17] J. Stepaniuk and A. Skowron, "Ontological framework for approximation," in *Proceedings of RSFDGrC 2005*, ser. Lecture Notes in Computer Science, vol. 3641. Springer, 2005, pp. 718–727. [Online]. Available: https://doi.org/10.1007/11548669_74

[18] M. Świechowski and D. Ślęzak, "Introducing LogDL - Log Description Language for Insights from Complex Data," in *Proceedings FedCSIS 2020, Sofia, Bulgaria*, 2020.