Noname manuscript No.
(will be inserted by the editor)

Confidence-Enhanced Early Warning Score Based on Fuzzy Logic

Maximilian Götzinger[1,2] · **Arman Anzanpour**[1] · **Iman Azimi**[1] · **Nima TaheriNejad**[2] · **Axel Jantsch**[2] · **Amir M. Rahmani**[3,4] · **Pasi Liljeberg**[1]

Received: date / Accepted: date

Abstract Cardiovascular diseases are one of the world's major causes of loss of life. The vital signs of a patient can indicate this up to 24 hours before such an incident happens. Healthcare professionals use Early Warning Score (EWS) as a common tool in healthcare facilities to indicate the health status of a patient. However, the chance of survival of an outpatient could be increased if a mobile EWS system would monitor them during their daily activities to be able to alert in case of danger. Because of limited healthcare professional supervision of this health condition assessment, a mobile EWS system needs to have an acceptable level of reliability - even if errors occur in the monitoring setup such as noisy signals and detached sensors. In earlier works, a data reliability validation technique has been presented that gives information about the trustfulness of the calculated EWS. In this paper, we propose an EWS system enhanced with the self-aware property confidence, which is based on fuzzy logic. In our experiments, we demonstrate that - under adverse monitoring circumstances (such as noisy signals, detached sensors, and non-nominal monitoring conditions) - our proposed Self-Aware Early Warning Score (SA-EWS) system provides a more reliable EWS than an EWS system without self-aware properties.

✉ E-mail: {maxgot, armanz, imaazi, pasi.liljeberg}@utu.fi
✉ E-mail: {nima.taherinejad, axel.jantsch}@tuwien.ac.at
✉ E-mail: a.rahmani@uci.edu
[1] Department of Future Technologies, University of Turku, Finland.
[2] Institute of Computer Technology, TU Wien, Austria.
[3] School of Nursing, University of California Irvine, USA.
[4] Department of Computer Science, University of California Irvine, USA.

Keywords Early Warning Score, Self-awareness, Data Reliability, Consistency, Plausibility, Confidence, Fuzzy Logic, Hierarchical Agent-Based System

1 Introduction

Cardiovascular diseases are worldwide as one of the major causes of death [1]. The vital signs of a patient reflect the patient's health condition, and the monitoring these vital signs establishes a basis for predicting a possible deterioration of the health condition. Even up to 24 hours before a sudden health deterioration occurs, specific symptoms are visible in the vital signs of a patient [2]. The assessment of the Early Warning Score (EWS) of a patient's health condition is a common practice in hospitals and manually done by healthcare professionals. The EWS constitutes a number which indicates the level of criticality [3].

The availability of an autonomous mobile EWS system that constantly monitors patients' vital signs to calculate the EWS could increase the life expectancy of outpatients. High-risk patients could wear such a system which monitors them during their daily life activities and alert in case of an emergency. Besides a much higher survival rate, a mobile EWS system could also decrease costs related to healthcare and reduce the duration of hospitalization periods.

Internet of Things (IoT) - with its small devices and wearable technologies - is a key enabler to provide autonomous health monitoring for a mobile EWS system in a cost-efficient manner [4–7]. Such a system cannot be supervised continuously by healthcare professionals, but its reliability and the accuracy of the calculated EWS are of utter importance. The manual monitoring of a patient who is admitted and is lying in a hospital

bed, done by healthcare professionals, faces much fewer problems than automated monitoring of a patient who is at home carrying out daily tasks [8]. One of the widely acknowledged and intrinsic challenges for wearable devices is the movement artifact [9]. Moreover, incorrectly attached or detached sensors, broken sensors, and noise can affect the calculation of the EWS that could lead to a false or - even worse - a missing alarm with all its consequences [10].

Self-awareness has various properties which help to make computer systems more autonomous, smarter, and reliable [11, 12]. Therefore, it can also be an enabler to make the monitoring of patients and the calculation of EWS more robust as well as reliable. In one of our previous works [13], we already presented a data reliability assessment technique based on fuzzy logic, which gives information about the trustfulness of the calculated EWS. However, although the proposed system outputs a reliability value which correlates with the correctness of the monitored vital signs, the system can only provide an unmodified EWS, which is incorrect when the input data is corrupted. To improve the decision-making ability of a system, another self-aware property can be utilized, namely, *confidence*. In other words, data reliability and confidence are two self-aware properties that can enhance the conventional EWS system. Both reliability and confidence are metadata. Reliability is metadata of the given input data and provides information on to what degree the data is reliable; in this case, the system can trust its sensors. Besides, the system can make its decisions based on confidence, a meta-data for decisions, which have been motivated by observations of various pieces of information, and other metadata.

In this paper, we propose a self-aware EWS system which validates reliability and bases all decisions on a confidence assessment. These validations and assessments are techniques based on fuzzy logic. To show the effectiveness of these two mentioned self-aware properties, we recorded vital signs of a set of persons with high-quality and low-quality sensors. In our experiments, we demonstrate that our proposed Self-Aware Early Warning Score (SA-EWS) system calculates the EWS correctly or with a small error close the to the value it should have even if the monitoring circumstances are adverse. The results show that our proposed SA-EWS system is more reliable than an EWS system without self-awareness. In other words, we prove that self-awareness is a good foundation for a reliable EWS system that trustfully classifies the EWS even if there is some faulty sensory data. Our main contributions are:

1. We propose a fuzzy logic based confidence metric for the quality assessment of the calculated EWS,
2. we show how a fuzzy logic based reliability metric gives information about the correctness of the input data,
3. we introduce a method for combing the input data reliability and the confidence of the system to calculate output data reliability based on both factors, and
4. using extended experiments, we demonstrate that our proposed system gives equally good or better results than a similar system that does not use reliability and confidence metrics.

After reviewing relevant related work in Section 2, we explain self-awareness properties reliability and confidence in Section 3. Section 4 shows system architecture as well as the implementation of our proposed system. While Section 5 explains the experimental setup and presents the results, finally, Section 6 concludes the paper.

2 Background and Related Work

In 1997, Morgan *et al.* proposed a medical method called EWS that is currently widely used in hospitals helping to determine the degree of patients' health deterioration. The patient's vital signs, such as respiration rate, heart rate, systolic blood pressure, body temperature, blood oxygen saturation (SpO_2), and the level of consciousness are manually collected in a regular routine and classified in different scores. These scores, ranging from 0 to 3, are determined according to the observations and predefined ranges of the vital signs. Table 1 indicates an EWS chart used for obtaining the various scores. In this chart, score 0 is allocated to a vital sign that is in perfect condition; e.g., heart rate in a range between 60 and 100. If the value of a vital sign is a bit worse than this (a bit too low or too high), the corresponding score is 1[1]. If the value of a vital sign is in even a worse condition (still higher or lower), the vital sign is classified to be score 2. Any value worse (depending on the case, higher or lower in absolute value) than the above ranges is classified as score 3.

The EWS is a simple aggregate of the scores that are abstracted from the patient's vital signs. The lower the calculated EWS, the better the patient's condition. A high EWS corresponds to a high risk of death or critical medical conditions [15]. Therefore, this likelihood reveals early signs of health deterioration and can be used to trigger a rapid response team to evaluate the patient. Similarly, an approach to predict poten-

[1] As Table 1 shows, not every vital sign score have separate score for each value - a vital sign that is too high or too low.

Table 1 A conventional Early Warning Score (EWS) chart [14].

Vital sign score	**3**	**2**	**1**	**0**	**1**	**2**	**3**
Heart rate (beats per minute)	0 - 39	40 - 50	51 - 59	60 - 100	101 - 110	111 - 129	≥ 130
Systolic blood pressure (mmHg)	0 - 69	70 - 80	81 - 100	101 - 149	150 - 169	170 - 179	≥ 180
Respiratory rate (breaths per minute)		0 - 8		9 - 14	15 - 20	21 - 29	≥ 30
Body temperature (°C)		≤ 35		35.1 - 38		38.1 - 39.5	≥ 39.6
Blood oxygen saturation (%)	0 - 84	85 - 89	90 - 94	95 - 100			
AVPU score[a]				A	V	P	U

[a] AVPU (the level of consciousness): A = alert, V = reacting to voice, P = reacting to pain and U = unresponsive

tial sudden patient death have recently received FDA approval [16].

The EWS itself can be classified into three different risk levels: low (EWS: 0-3), medium (EWS: 4-6), and high (EWS: 7 or higher). A low-risk level demands a nurse to assess the patient periodically. A medium-risk level requires to inform medical team urgently. In contrast, a high-risk level should trigger an urgent clinical response as the patient's condition is critical [17–19].

There are, nevertheless, various restrictions and issues such as latency and inaccuracy in this manual data acquisition. Furthermore, this system is merely restricted to hospital settings where patients are stationary. In this regard, an IoT-based health monitoring system is proposed to monitor the vital signs autonomously and deliver the EWS score to healthcare providers [20]. Estimations suggest that the ratio between the world's population and IoT devices will be one to four [21]. These small IoT devices and wearables form a good basis for a well-structured EWS system which autonomously monitors a patient in a cost-efficient way while decreasing the mortality rate [4–7].

Despite IoT provides a potential solution for monitoring human's vital signs, the conventional EWS system is still not applicable for out-of-hospital monitoring since daily activities, and the environments influence the vital signs and subsequently the decision making. Usually, a person has a higher heart rate, blood pressure, respiratory rate, and body temperature when making physical effort (e.g., running and riding a bicycle) compared to more relaxed activities such as sitting or sleeping. Using the same score classification ranges, such as those in Table 1), would lead to a high EWS during physically demanding activities although there is no emergency. Towards this end, a modified EWS system has been proposed for everyday settings, providing a self-aware decision (i.e., the score) according to the context information and five[2] vital signs [22].

[2] The level of consciousness is excluded because it is not applicable in out-of-hospital monitoring.

Autonomous mobile EWS system still faces problems that have to be solved for being able to offer a reliable EWS calculation. Incorrectly attached or detached sensors, broken sensors, or a noisy signal affect the EWS calculation. If the calculated value is still close to the truth, it may not be a problem. In contrast, an EWS that deviates more from the truth could lead to a false or - even worse - a missing alarm with all its consequences. Self-awareness is a promising solution to tackle this problem. Self-awareness is the ability of the system to monitor itself and its environment regarding the state, behavior, performance, and goals. This is often accompanied by an adjustment of some of the components and parameters which lead to achieving or approaching to the goals of the system [23]. This process has been modeled different ways by various groups, among which some of the more well-known ones are Observe-Decide-Act (ODA) [24] and Monitor-Analyze-Plan-Execute over a shared Knowledge (MAPE-K) [25]. Several works have been done in order to implement self-awareness in various systems, and take advantage of its properties [12, 23, 24, 26–28]. However, most of these works are more focused on the smart decision-making process, while paying little attention to the observation (monitoring) part of the process. In 2016, TaheriNejad *et al.* published a paper [29] which highlighted this aspect and elaborated on different elements of observation and their potential effect on self-awareness and the overall performance of the system. Since then, several publications have appeared in the literature which demonstrated this effect in various applications [13, 26–28, 30–33].

Our previous works utilize various self-awareness properties to overcome different issues. Anzanpour *et al.* exploited the self-awareness in IoT-based EWS systems. In this work, *situation awareness* was utilized to improve the specificity of the EWS values, considering the impact of the user's physical activities in the calculation. *Attention* as another self-awareness property was also used to enable a self-organized system, dynamically adjusting the system's configuration for power

consumption reduction [26]. Such a dynamic behavior can increase system battery life, but it could decrease the reliability of the EWS in the case of low-quality signals. In another work [13], the proposed system assess the reliability of the calculated EWS. The fuzzified reliability validation tackles the fact that the knowledge about the vital signs as well as their interactions is not complete. With this technique, it was possible to recognize erroneous vital signs caused by various measurement artifacts such as detached sensors, loose sensors, and other interferences.

Our results show that self-awareness can tackle various issues that affect the reliability of a mobile EWS system. Although the proposed system of [13] provides information about the trustworthiness of the calculated EWS, the EWS itself is still incorrectly calculated if the input data is corrupted. Enhancing the decision-making mechanism of the EWS system is a way to solve this problem and improve reliability.

3 Self-Awareness Properties

In this work, we study two aspects of self-awareness, namely confidence and data reliability, and the interplay between the two as well as their effect on the overall performance of the system. Moreover, we have tried to formalize these concepts, which were initially described in [29] only conceptually, in order to establish a more uniform understanding of these concepts.

3.1 Data Reliability

Data Reliability describes the trustworthiness of a set of data at hand, which can be divided into accuracy, precision, and truthfulness. A sensor may be accurate and precise. However, if it is used outside its assumed working conditions, it does not provide reliable data; i.e., it does not provide truthful data. Moreover, even though accuracy and precision provide general measures on the overall quality of a data set (or performance of a sensor), they do not provide an explicit meta-data on each data point. A (resource constrained) self-aware system such as ours, however, sometimes needs to make decisions based on single or few data points. Therefore, accuracy and precision do not provide enough situational information for such cases, and the system needs to estimate and be aware of the overall reliability of those data points based on which it makes a decision.

3.1.1 Formal Definition

As mentioned before, data reliability can be broken to accuracy, precision, and truthfulness. Accuracy, $A(X')$, is the systematic bias of the data set at hand, i.e., $X' = \langle x'_0, ..., x'_n \rangle$, compared to the ground truth values, $X = \langle x_0, ..., x_n \rangle$. As a measure of statistical bias it can be defined as

$$A(X') = \frac{1}{n} \sum_{i=0}^{n} x_i - x'_i. \tag{1}$$

Precision presents the random errors in the data (for a measurement, it would be the random errors of repeated measurements under the same conditions). Since precision is a measure of statistical variability, it can be defined as:

$$P(X') = \sigma' = \sqrt{\frac{1}{n} \sum_{i=0}^{n} (x'_i - \mu')^2} \tag{2}$$

where $\mu' = \frac{1}{n} \sum_{i=0}^{n} x'_i$.

Truthfulness, t, is the distance of each value at hand, x'_i, with the corresponding ground truth value x_i:

$$t(x'_i) = |x'_i - x_i| \tag{3}$$

The overall truthfulness, $T(X')$, of a set of values can be defined as

$$T(X') = \frac{1}{n} \sum_{i=0}^{n} t(x'_i). \tag{4}$$

Accuracy and precision are defined on one or more data sets, X' and X, and hence are a property of a set[3], whereas truthfulness is defined on each data sample, x'_i. Therefore, even though A, P, and t (and consequently T) are correlated, a closed-form formula describing their dependency often cannot be established. Moreover, in many cases the ground truth value, x_i, is not available which makes the calculation of t impossible. In consequence, often an estimation of t, namely t', is devised which may or may not include the effect of accuracy and precision.

In summary, given a sequence of sampled data points X', the data reliability R of X' is given as (the same can be defined for each value)

$$R_f(X') = f(A(X'), P(X'), T(X')) \tag{5}$$

where f determines the role of each parameter and thus how well would R fit its purpose. For example, the reliability of $x'_i \in X'$ could be calculated as

[3] In a cyber-physical system that would be a property of a measurement device.

$$r_f(x_i') = f_{x_i'}(A, P, t) = c_1 A(X') + c_2 P(X') + c_3 t(x_i') \quad (6)$$

with constants c_1, c_2 and c_3 defining the relative weights given to the three components of the data reliability. Ideally, the reliability is defined such that the mapping domain is between one and zero:

$$R, r : X' \rightarrow [0, 1] \in \Re \quad (7)$$

In a cyber-physical system, A and P are usually provided by the producers of the sensors (even though that is not always the case), and the t and f are to be calculated or estimated by the system using the sensor. In the absence of these values, the designer needs to estimate r or R by r' and R', respectively, using custom methods. In this work, we present our proposed method to calculate r' and R', which we use as our measure of data reliability.

In the following, we present three measures which can provide an insight into the reliability of the data at hand. That is consistency, plausibility, and correlation of data. An important feature of these measures is that they could be applied to low-level data (obtained directly from sensors) or higher-level data (obtained from processes and algorithms within a system).

3.1.2 Plausibility

Data sets can often be associated with a membership function, specifically in the case of cyber-physical systems, that translates into how plausible is the existence of a data with a certain value in the data set. For example, the oxygen saturation can be only in the range of 0-100%; any other value reported is a sign of malfunction and unreliability of the data. The same could be said for a heart-rate of 300 beats per minute for an adult person. By tagging such data as less reliable or unreliable, a self-aware system could react accordingly (e.g., look for further sources of information or dismiss the data).

3.1.3 Consistency

A certain consistency is often observed within the members of a data set. This is particularly valid in the case of data sets representing natural phenomena, i.e., data collected by a sensor from the real world. Such signals often experience limited changes from one sample to the next. Therefore, the history of a signal and its consistency can provide some information on how reliable is that source of data. For example, it is established that the body temperature cannot change several degrees per minute [34]. Hence, if a larger rate of change occurs in a data set, a self-aware system should tag such an observation (which may be caused by a sensor detachment or a fault/failure in the sensor) as unreliable (regardless of its cause) and react accordingly.

3.1.4 Cross-validity

In some cases, there exists a correlation between the values of two data sets (or such correlation can be established). In such cases, this correlation can be exploited to evaluate the probability or possibility of the coexistence of two or more values. If their coexistence is not possible (e.g., a living patient with valid heart rate and respiratory rate but a negative body temperature) then one or some of those data could be tagged as an unreliable (in this example body temperature). If their coexistence is possible but not very probable (e.g., a body temperature around 30^oC with typical values for other biological signals), the reliability of the data could be reduced, signaling the system a need for further analysis. In the use-case of this work, there have been several works trying to establish such correlations between vital signals of the body [35–37]. Although they do not always provide a conclusive insight, they help us to enhance the robustness of our system by enabling additional data reliability assessments.

3.2 Confidence

Confidence is a measure of the reliability of an algorithm or a process in the system[4] [29]. Conceptually, we can say that confidence provides the system with a measure on how the results of an algorithm or a process can be relied upon. In other words, how close the output of this algorithm or process would be to the ideal output. All that with the assumption that the system has received flawless input data. Although, more often than not, the input data collected by the sensors are unideal (which we discussed in the data reliability subsection). Therefore, the reliability of the output of a system depends on both its confidence and the data reliability of its inputs.

The importance of confidence is in its ability to improve the decision-making processes [12] and allow a self-aware system to question certain abstracted data it has processed, and make more reliable decisions based on the reliability of its sub-processes and sub-algorithms.

[4] Therefore, confidence is a property of an algorithm, process or system, as opposed to Data Reliability which is a property of the data at hand.

An important application of this concept for the decision-making unit is to enable it to switch between different algorithms based on their confidence, the usefulness of which has been shown in [28].

3.2.1 Formal Definition

If I is an ideal function defined over $X = \langle x_0, \ldots, x_n \rangle$ and g is the unideal function at hand, also defined over X, then the confidence of $g(x_i)$ (defined for each member of X) can be defined as a function Δ of $g(x_i)$ and $I(x_i)$. Δ represents the "distance" between f and g based on some application specific metric for distance, normalized such that $0 \leq \Delta \leq 1$. Thus, for the confidence, c, we have:

$$c\big(g(x_i)\big) = 1 - \Delta\Big(I(x_i), g(x_i)\Big). \tag{8}$$

Overall confidence of g (as opposed to confidence at each point), represented by C, is the average confidence of g over X:

$$C(g) = \frac{1}{n} \sum_{i=0}^{n} c(g(x_i)). \tag{9}$$

We note that $0 \leq c(g), C(g) \leq 1$ and $c(I) = C(I) = 1$. How to calculate c (and consequently C), however, is case specific. Often the ground truth (I) is not available and the aforementioned distance cannot be calculated. Therefore, a Δ' function is used instead to estimate Δ, which is what we do in the rest of this work too. That is, we propose an estimation of Δ (i.e., Δ'). In other words, all the confidence (c) functions hereafter refer to Δ', which is an estimation of Δ.

3.3 Combination of Data Reliability and Confidence

In this section, we already discussed the concepts of data reliability (as a property of a data set) and confidence (as a property of a process or algorithm) independently. However, in a real-world system, these two often are tightly intertwined. Processes consume data and produce data. Assuming an ideal input, the data reliability of the output data of a process could be associated with its confidence (although not always in a straight forward or in a simple manner). However, most often, the input data are unideal and subject to a data reliability below one. Therefore, the data reliability of the output data of a process is a function (ϕ) of the input data reliability and the confidence of the process. Calculating the output data reliability of a process (which in turn could be the input data reliability of another process) is particularly more difficult when data reliability or confidence are obtained using estimation functions. In this work, we explore this realm and try to propose a method which shows a good promise in the estimation of the output data reliability of different processes in our system based on respective input data reliability and confidence of that process. More details on our practical implementation are found in Section 4.3.

3.3.1 Formal Definition

If $X' = \langle x'_0, ..., x'_n \rangle$, is the data set at hand (i.e., the unideal values), corresponding to the ground truth values $X = \langle x_0, ..., x_n \rangle$, we have:

$$R_g(x'_i) = \phi\Big(r_f(x'_i), c\big(g(x)\big)\Big). \tag{10}$$

Since, as mentioned before $\forall x; c\big(g(x)\big) \leq c\big(I(x)\big)$ and $R_f(x'_i) \leq R_f(x_i)$ we can conclude that

$$R_g(x'_i) \leq R_I(x'_i) \leq R_I(x_i). \tag{11}$$

3.4 History

History enables access to time-dependent information in a system. For example, whether the performance of a (sub)system has been improving or degrading. The historical data can provide meta-data on the current status of the system and its environment. They also help in predicting the (near) future status of the system and its environment. Given that most systems have memory limitation, choosing the type and mode of storing historical value, and a smart usage of it are important points to be considered when designing a self-aware system using history for enhancing its performance.

3.4.1 Formal Definition

There are several methods to track the past values in a sequence. Given the sequence of values or symbols $X = \langle x_0, \cdots, x_n \rangle$, $H = \langle h_0, \cdots, h_m \rangle$ is a subsequence of X, in which $m \leq n$. If $m = n$, the system is memorizing everything which is undesirable. Therefore, most often $m < n$ and preferably $m \ll n$. We note that history function is a specific form of abstraction which concerns time, i.e., the sequence length of X. As of such we can define it as

$$H = \mathfrak{H}_y(X) = \langle h_0, \cdots, h_i, \cdots, h_m \rangle \tag{12}$$

where at the sequence point of x_s,

$$h_i = y(h|_{j=0}^{r}, x|_{k=0}^{s}), \quad r \leq (i-1) \;\; \& \;\; i \leq s, \tag{13}$$

where the function y determines how exactly the history H is extracted from X. An interpretation or abstraction of X (such as the average of certain number of data points), or a direct storage of the values themselves could be some examples of y.

4 System Architecture and Implementation

A hierarchical agent-based architecture (as shown in Fig. 1) consists of independent modules which can communicate with each other and may be in different hierarchical levels. The possibility of hierarchically structuring the agents enables to process data on different levels of abstraction [38].

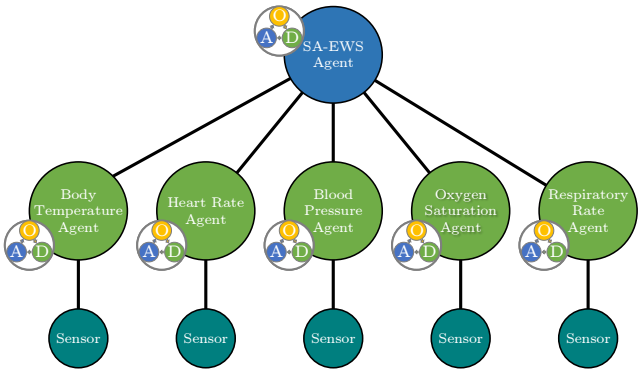


Fig. 1 Hierarchical agent-based system architecture.

The EWS is the aggregate of various scores abstracted from different vital signs. The task of abstraction is the same for each vital sign, but the ranges vary from vital sign to vital sign. The assessment of reliability and confidence-based decisions are done on different levels of abstraction. As an example, a part of the reliability assessment is based on the absolute value and the slope of the signal of a vital sign (principle of plausibility and consistency in Sections 3.1.2 and 3.1.3). To analyze whether a signal is plausible and consistent, the raw data is of interest. In contrast, for making a statement about the correlation between vital signs (principle of cross-validity in 3.1.4) already abstracted information is needed. Because of these differences horizontal direction (different vital signs) and vertical directions (different levels of abstraction), a hierarchical agent-based model (Fig. 1) constitutes an appropriate practical architecture for this purpose.

Because an ODA loop is an appropriate approach to implement self-awareness, our system is also based on this concept [23, 29, 39]. Each agent acts like an ODA loop, which means that it monitors its inputs (sensor or agent), decides what to do, and acts accordingly. Furthermore, this approach allows implementing a highly modular model easily.

While the abstraction from the raw sensor value to the vital sign score (with the help of Table 1) takes place in the lower hierarchical level, the agent on top aggregates the five scores to the overall score, the EWS. In other words, each low-level agent abstracts the actual samples obtained from its dedicated sensor and sends the result to the high-level agent, which sums up all these scores. Both, the reliability assessment, as well as the confidence-based decision-making, takes place in the lower and in the higher hierarchical level. However, the implementations of these processes are different in the two hierarchical levels. In the next two sections, we explain the reliability assessment and the confidence-based decision-making process, before Section 4.3 shows the workflow of the proposed system in detail.

4.1 Fuzzified Reliability Assessment

Due to the lack of complete knowledge of all functions of a patient's body, it is very challenging to determine whether a vital sign is monitored correctly or incorrectly. Therefore, in contrast to one of our previous works [32], we use fuzzy logic instead of simple boolean logic to assess the reliability value. The usage of fuzzy logic enables the coverage of the unsharp ranges in which a patient's vital sign is not tagged merely as correct or incorrect, but rather somewhere on the spectrum of reliability. Hence, the data reliability of a vital sign is assigned a value in the range between 0 and 1.

The reliability of a patient's vital sign, vs_i, is composed out of two different reliability assessments: the reliability of the signal's absolute value $r'_{abs,i}$ and the reliability of the signal's slope $r'_{slo,i}$. This corresponds to the plausibility and consistency of data, as described in Section 3.

The reliability for being plausible, $r'_{abs,i}$, is the output of a fuzzy membership function (Fig. 2) defined by four points and three intervals. If the absolute value is in the interval of $[p_b, p_c]$, it is certainly reliable. If it falls in one of the intervals of $[p_a, p_b]$ or $[p_c, p_d]$ - depending on the absolute value, it is more or less reliable. Otherwise, it is certainly unreliable. The reliability $r'_{abs,i}$ and its counterpart (the estimated unreliability $u'_{abs,i}$) of the actual absolute value $v_{a,i}$ are calculated by

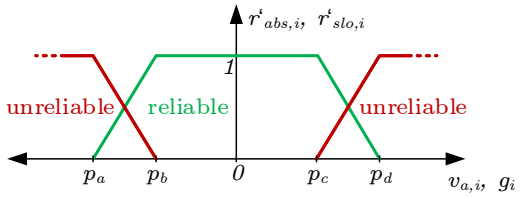


Fig. 2 Example for a fuzzy membership function to assess the reliability of the absolute value or the slope of a vital sign.

$$r'_{abs,i} = \begin{cases} \frac{v_{a,i}-p_a}{p_b-p_a} & \text{if } p_a < v_{a,i} < p_b \\ 1 & \text{if } p_b \leq v_{a,i} \leq p_c \\ \frac{v_{a,i}-p_d}{p_d-p_c} & \text{if } p_c < v_{a,i} < p_d \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

and

$$u'_{abs,i} = 1 - r'_{abs,i} \tag{15}$$

where the points p_a, p_b, p_c, and p_d respectively the intervals between them are configured in a way to match the characteristic of the assigned vital sign.

Similar to that, the reliability for being consistent, $r'_{slo,i}$, and its counterpart (the unreliability, $u'_{slo,i}$), a fuzzy membership function of the same shape exists (Fig. 2). Again, these functions are defined by for points and three intervals between them. These are as follows

$$r'_{slo,i} = \begin{cases} \frac{g_i-p_a}{p_b-p_a} & \text{if } p_a < g_i < p_b \\ 1 & \text{if } p_b \leq v_{a,i} \leq p_c \\ \frac{g_i-p_d}{p_d-p_c} & \text{if } p_c < g_i < p_d \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

and

$$u'_{slo,i} = 1 - r'_{slo,i} \tag{17}$$

where g is the gradient between the actual value, $v_{a,i}$, to the previous one, $v_{p,i}$.

This gradient is calculated by

$$g_i = \frac{v_{p,i} - v_{a,i}}{t} \tag{18}$$

where t constitutes the time between the samples.

Depending on which of the two reliabilities shall be assessed, the abscissa constitutes the absolute value or the slope of a vital sign. The ordinate of the fuzzy membership function constitutes then the reliability corresponding to it. While the abscissa gives space for all values (from $-\infty$ to $+\infty$), the reliability values on the ordinate are limited between 0 and 1.

After the assessment of $r'_{abs,i}$ and $r'_{slo,i}$, the input reliability, $r'_{in,i}$, can be calculated in many different ways such as conjunction ($\wedge$), disjunction ($\vee$), or multiplication of different inputs as well as if-then-rules and other methods. We decided to use the conjunction operator because a vital sign is reliable when its absolute value and its slope are reliable. Therefore the input reliability $r'_{in,i}$ of a vital sign is given by

$$r'_{in,i} = r'_{abs,i} \wedge r'_{slo,i} \tag{19}$$

where the fuzzy conjunction is equal to a minimum function [40].

Because the input reliability, $r'_{in,i}$, depends only on the raw sensor data (absolute value and gradient of the signal), it is calculated in the low-level agents which are also responsible for the abstraction of the vital signs. This input reliability is calculated for every vital sign and provides information on whether it is reliable or unreliable considered separately. In other words, the reliability of one vital sign omits the condition of other vital signs.

Since vital signs impact each other, and therefore, one vital sign, vs_i, usually does not have a terrible score while others have a perfect score, a cross-validation reliability value is needed. For this purpose, the cross-validation reliability, $r'_{cro,i,j}$, for the vital signs vs_i and vs_j is calculated by

$$r'_{cro,i,j} = \begin{cases} 1 & \text{if } s_i = s_j \\ \frac{1}{p_{cro,i,j}|s_i - s_j|} & \text{if } s_i \neq s_j \end{cases} \tag{20}$$

where $p_{cro,i,j} \in (0, \infty)$ denotes a coefficient of the strength of the correlation[5] between vital signs vs_i and vs_j, and s_i, as well as s_j, are the abstracted scores of these two vital signs.

Because the cross-validity reliability, $r'_{cro,i,j}$, already makes use of the abstracted information (the various vital sign scores), it is calculated in the high-level agent which is responsible for the calculation of the EWS.

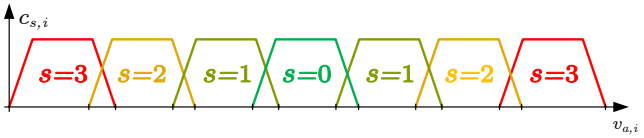
4.2 Fuzzified Confidence-Based Decisions

As already stated in Section 3.1, data reliability describes the trustworthiness of a set of data at hand, which can be divided into accuracy, precision, and truthfulness. For the case, a sample (a sensor value) is not very accurate, two different possibilities exist. If the real vital sign value (the ground truth) is somewhere in the middle of a score range of Table 1 and the sensor's inaccuracy is not very high, the abstracted score will

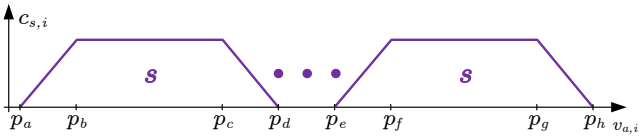
[5] The reliability module in our implementation limits the cross-validity reliability, $r'_{cro,i,j}$, to a value between 0 to 1, although theoretically, a coefficient less than 1 can lead to an $r'_{cro,i,j}$ higher than 1. The standard value of $p_{cro,i,j}$ is 1.

most likely be equal to the ground truth. In contrast, a wrong score abstraction could result out of a ground truth value very close to a boundary of such a range or a highly inaccurate sensor.

To overcome this issue, the abstraction process in the lower hierarchical level is not merely based on a simple lookup table as in Table 1, the boundaries of the different score ranges are intersecting which means that the score ranges are partly overlapping. Fig 3(a) shows an example for the vital sign abstraction, with four different fuzzy membership functions; for each score, one fuzzy membership function.



(a) All three possible scores to which the vital sign, vs_i, can be abstracted.



(b) One of the scores to which the vital sign, vs_i, can be abstracted.

Fig. 3 Example for fuzzy membership functions to assess the confidence of the abstraction of a vital sign.

In similar fashion to the reliability fuzzy functions, various intervals describe the confidence fuzzy membership functions. Because the fuzzy membership functions are an extension of Table 1, the intervals can vary between different vital signs. While heart rate and systolic blood pressure are symmetrical in a way that each score higher than 0 is available for the vital sign's value is either too low or too high. In contrast, respiratory rate, body temperature, and blood oxygen saturation are unsymmetrical; some scores are missing on one side or both sides. In the case of a symmetrical segmented vital sign (Fig. 3(b)), the confidence functions of abstracting the actual value of a vital sign to a score s_i, $c_{s,i}$ are calculated by

$$c_{s,i} = \begin{cases} \frac{v_{a,i}-p_a}{p_b-p_a} & \text{if } p_a < v_{a,i} < p_b \\ 1 & \text{if } p_b \leq v_{a,i} \leq p_c \\ \frac{v_{a,i}-p_d}{p_d-p_c} & \text{if } p_c < v_{a,i} < p_d \\ \frac{v_{a,i}-p_e}{p_f-p_e} & \text{if } p_e < v_{a,i} < p_f \\ 1 & \text{if } p_f \leq v_{a,i} \leq p_g \\ \frac{v_{a,i}-p_h}{p_h-p_g} & \text{if } p_g < v_{a,i} < p_h \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where $s \in \{1,2,3\}$ is one of three possible scores the actual value, $v_{a,i}$, of the vital sign can have.

Because score 0 of each vital sign has only one range in Table 1, the confidence function of abstracting a vital sign's actual value to score 0, $c_{0,i}$ is calculated by

$$c_{0,i} = \begin{cases} \frac{v_{a,i}-p_{0,a}}{p_{0,b}-p_{0,a}} & \text{if } p_{0,a} < v_{a,i} < p_{0,b} \\ 1 & \text{if } p_{0,b} \leq v_{a,i} \leq p_{0,c} \\ \frac{v_{a,i}-p_{0,d}}{p_{0,d}-p_{0,c}} & \text{if } p_{0,c} < v_{a,i} < p_{0,d} \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

In the configuration of the proposed system, the interval of the ramp of a fuzzy membership function is congruent with the interval of the ramp of the next fuzzy membership function (e.g., p_c and p_d of $c_{0,i}$ are equal to p_a respectively p_b of $c_{1,i}$). This approach leads to the possibility of abstracting a vital sign value to two different scores with certain confidences. As an example, let us assume that the actual value of a vital sign, $v_{a,i}$, is the interval between $p_{0,c}$ and $p_{0,d}$ (which is equal to the interval $p_{1,e}$ and $p_{1,f}$). In this case, the vital sign will be abstracted to score 0 with $c_{0,i} = \frac{v_{a,i}-p_{0,d}}{p_{0,d}-p_{0,c}}$ and to score 1 with $c_{1,i} = \frac{v_{a,i}-p_{1,e}}{p_{1,f}-p_{1,e}}$.

However, the high-level agent evaluates various confidences. Similar to Eq. 20, cross-validity confidence, $c_{cro,i,j}$ is calculated based on a patient's individual correlations of the various vital signs; e.g., Eq. 20 does not reflect the truth if a patient - in normal health condition - has tachypnea, hypertension, or another vital sign which leads to a score higher than the scores of the other vital signs. Based on the frequency of various occurring score differences, $SD_{i,j}$, between the two vital signs vs_i and vs_j, a patient profile is established which gives information about the likelihood of a score difference between two different vital signs. For this purpose, the patient (situated in normal condition) is monitored for the period T (the time of n samples). After n samples have been recorded, four different quantities, $q_{SD_{i,j}}$ for all four possible score differences $SD_{i,j} \in \{0,1,2,3\}$, are known. With the knowledge of these quantities, the cross-validity confidence between the two vitals signs sv_i and sv_j, $c_{cro,i,j}$ is calculated by

$$c_{cro,i,j} = \frac{q_{SD_{i,j}}}{n}. \tag{23}$$

4.3 Functional Description of the System

Fig. 1 shows the system architecture we propose in this work. At the bottom are five sensors which monitor the five different vital signs (Table. 1) and transmit the raw data to their dedicated agents in the lower hierarchical

level. Fig. 4 shows a simple schematic of the whole procedure for one low-level agent; the others are just faded out. The functional principle of both, the agents of the lower and the higher hierarchical, is explained in detail in the following.

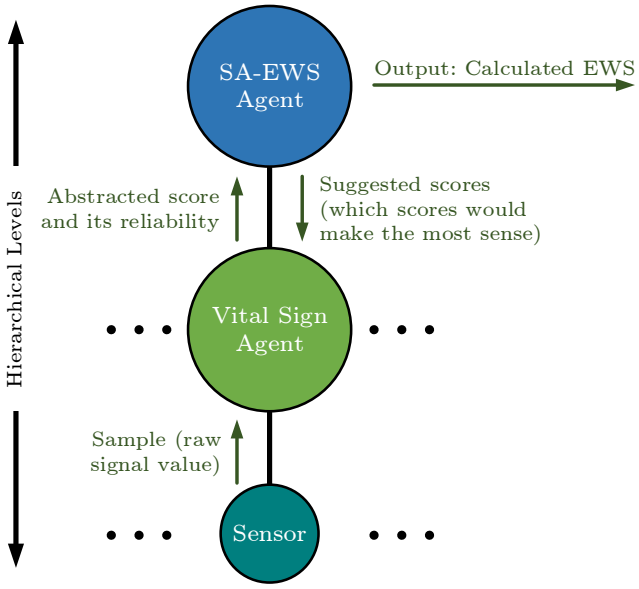


Fig. 4 Functional description explained for one vital sign.

4.3.1 Lower Hierarchical Level of Computation

Each of these five low-level agents abstracts the actual value (got from its dedicated sensor) by calculating the confidences for every possible score, $c_{s,i}$ for $s = 0, 1, 2, 3$, by Eq. 22 and Eq. 21. As shown in Fig.4, each low-level agent also receives score suggestions from the high-level agent. Each of these suggestions consists of a score and its reliability, $r'_{sug,s,i}$. Their calculation is based on Eq. 20 and Eq. 23 but Section 4.4 will show the exact procedure of generating these suggestions. Additionally, the input reliability, $r'_{in,i}$ of the corresponding vital sign is calculated by Eq. 19.

In a next step, the output reliability of each possible score, $r_{out,s,i}$ is calculated by

$$r'_{out,s,i} = r_{in,i} \wedge c_{s,i} \wedge r'_{sug,s,i} \tag{24}$$

because the score is reliable if the vital sign value is reliable, the abstraction is done with high confidence, and if it correlates with the other vital signs (based on the suggested scores).

After every possible score has been calculated, the low-level agent chooses the one with the highest output reliability and saves it in a history if the reliability is higher than a certain threshold[6]. In the next step, the low-level agent sends the last saved score and its output reliability to the high-level agent. In other words, if the reliability of the actual score is higher than the set threshold it is sent to the high-level agent; otherwise, the previous score is sent.

4.4 Higher Hierarchical Level of Computation

The high-level agent calculates the EWS and its overall reliability, r'. For this purpose, the agent reads all low-level scores and their output reliabilities, $r'_{out,i,s}$. However, these reliabilities are - from the perspective of the high-level agent - input reliabilities, and therefore, they are called $r'_{in,i}$. The EWS is just the sum of all five vital sign scores, and thus, calculated by

$$EWS = \sum_{i=1}^{5} s_i. \tag{25}$$

With all five input reliabilities, $r'_{in,i}$, the combined input reliability, r_{in} is calculated by

$$r'_{in} = r'_{in,1} \wedge \cdots \wedge r'_{in,5} = \bigwedge_{i=1}^{5} r'_{in,i}. \tag{26}$$

For two vital sign scores, the cross-validity reliability is calculated by Eq. 20, and the personalized cross-validity confidence by Eq. 23. After the calculation of both of these metrics, the personalized cross-validity reliability, $r'_{per,cro,i,j}$, can be calculated in different ways. We decided to use the disjunction ($\vee$) operator because the correlation is plausible if it is according to our general rule (Eq. 20) or matches the personalized body functions of the patient (Eq. 23). Therefore the personalized cross-validity reliability, $r'_{per,cro,i,j}$, is given by

$$r'_{per,cro,i,j} = r'_{cro,i,j} \vee c_{cro,i,j} \tag{27}$$

where the fuzzy disjunction is equal to a maximum function [40].

The overall reliability of the calculated EWS is composed of all input reliabilities and all personalized cross-validity reliabilities, $r'_{per,cro,i,j}$. All $r'_{per,cro,i,j}$ for this purpose are combined together to the combined cross-validity reliability, $r'_{per,cro}$, by

$$r'_{per,cro} = \bigwedge_{i=1}^{5} \left(\bigwedge_{j=1}^{5} r'_{per,cro,i,j} \right) \tag{28}$$

[6] If the history is empty (e.g., right after the EWS system has been started), the score and its reliability are saved in the history regardless.

where the cross-validity reliabilities for $i = j$ are not calculated because they will be 1 one for sure (Eq.20).

In further consequence, the overall reliability, r', is given by

$$r' = r'_{in} \wedge r'_{per,cro} \tag{29}$$

and constitutes, besides the EWS (Eq. 25), the output of our proposed system.

As mentioned in Section 4.3.1, the high-level agent makes also score suggestions which are sent to each low-level agent. For this purpose, theoretically personalized cross-validity reliabilities are calculated for each possible score ($s \in 0, 1, 2, 3$) with that a vital sign could be classified. In particular, the four theoretically possible scores of one agent are calculated by Eq. 27, whereas the score difference is based on the comparisons with the real scores from the other four low-level agents. The reliability of the theoretically possible score (the suggested score) is calculated by

$$r'_{sug,s,i} = \bigwedge_{j=1}^{5} (r'_{cro,i,j} \vee c_{cro,i,j}) \tag{30}$$

for each possible score $s \in 0, 1, 2, 3$. Whereas the comparison of one vital sign with itself is not performed.

This procedure is repeated for all of the five vital signs, and the results (the four possible scores and their theoretical cross-validity reliability) is sent to the dedicated low-level agent.

5 Experimental Results

In this section, we describe our experimental setup as well as the validation method of our proposed system. We also discuss the experimental results in detail.

5.1 Experimental Data

The data collection was performed on eight different participants aged from 23 to 37 (see Table 2). Half of the participants were male, and the other half were female.

As listed in Table 3, we recorded and abstracted the vital signs with different sensors respectively in different ways. A set of sensors provides a high-accuracy source, and another set of sensors provides a low-accuracy source for normal and fault-emulated signals. As the high accuracy sensor set, we use (i) a chest strap heart rate monitor for recording Electrocardiogram (ECG) signal,

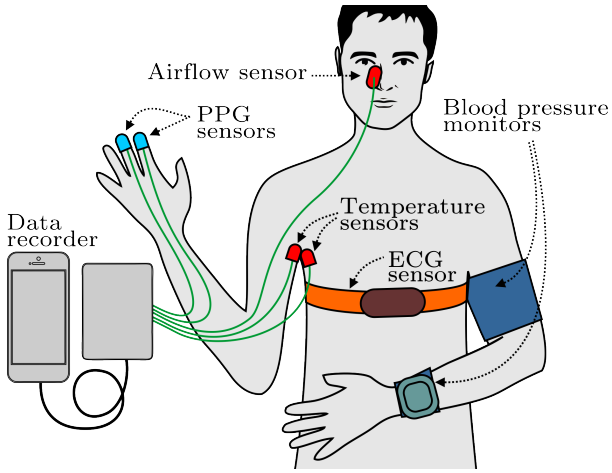


Fig. 5 Data collection sensors.

Table 2 Participants who participated in our experiments.

Person	Sex	Age	Test scenario(s)
P1	Male	37	S1, S2, S3, S4
P2	Female	23	S1 (twice)
P3	Male	29	S1, S2, S3, S4
P4	Male	25	S1 (twice)
P5	Female	23	S1 (twice)
P6	Male	30	S1 (twice)
P7	Female	23	S1 (twice)
P8	Female	28	S1 (twice)

(ii) a sensitive temperature sensor attached to the subject's nose for recording the airflow signal, (iii) an accurate temperature sensor attached to the armpit (axilla)[7], (iv) an upper arm blood pressure monitor, and (v) a high-fidelity Photoplethysmogram (PPG) sensor for recording infrared and red PPG signals[8].

The low-accuracy sensor set consists of (i) another PPG sensor which consumes less power and records PPG signal with lower Signal-to-Noise Ratio (SNR)[8], (ii) a temperature sensor with lower sensitivity is attached to armpit (axilla) measures skin temperature[7], and (iii) a wrist-type blood pressure monitor measuring an estimation of blood pressure. Table 3 shows the details of the sensors in each set. All continuously recording sensors[9] were connected to an ATMEGA328P mi-

[7] Because Table 1 shows the body core temperature, the measured skin temperature had to be converted to an estimated core temperature. This was done as Richmond *et al.* state it in [41].

[8] As shown in Table 3, the MAX30100 PPG sensor was used as accurate source for monitoring SPO_2 and as one of the inaccurate sources for monitoring heart rate and respiratory rate.

[9] The two blood pressure devices were manually operated and were not continuous.

crocontroller which reads the sensors values with a sampling frequency of 50 Hz. Finally, an Android phone, connected to this microcontroller via a USB-to-Serial converter, recorded the data.

In the next step, these recorded signals were analyzed to extract the vital signs. As listed in Table 3, we use two sets of PPG signals to obtain two sources of heart rate, respiration rate, and SpO_2 values (i.e., low-accuracy and high-accuracy values). First, a filter-based method is used to extract respiratory and heartbeat signals. In this method, the cut-off frequencies are selected based on Power Spectral Density (PSD) of the PPG signals [42–44]. Note that an acceptable SNR is needed in this method, as high noise level influences the PSD of the signal and subsequently interrupts cut-off frequency selection. Next, the respiration rate and heart rate values are determined via a peak detection method. Moreover, the SpO_2 value is calculated from the PPG signals using two light sources with different wavelengths (i.e., red has 660 nm and infrared has 880 nm) [45, 46]. In addition to the PPG signals, another high-accuracy heart rate and respiration rate values are determined by using the two other sources (i.e., ECG and airflow signals). Similarly, we use peak detection methods for the detection of these two vital signs. In total, we extracted three heart rate, three respiration rate, two SpO_2, two skin temperature[7], and two blood pressure signals.

5.2 Validation of the EWS Systems

Table 4 shows the different scenarios in which the participants were monitored. In Scenario S1, the participants were sitting without performing any physical activity. P1 and P3 were also monitored during three additional scenarios in that errors were induced in some of the low accuracy sensor setup (Scenarios S2, S3, and S4). Six participants were monitored two times, and the other two participants four times (Table 2), resulting in 20 measurements in total.

As mentioned in Section 5.1, we recorded and abstracted the vital signs of each scenario (listed in Table 4) with different sensors, respectively, in different ways (Table 3). All various combinations of the twelve vital signs (varying in quality) reveal a total number of 72 different signal setups for each measurement. The 20 measurements and the 72 different ways of monitoring/abstracting the vital signs lead to 1440 different experiments.

All these data sets were then processed with both, the conventional EWS system without any self-awareness properties and our proposed SA-EWS system. The output of these systems is the EWS signal of the same length as the experimental data sets (one EWS value for each vital sign sample set). To have a common benchmark for comparing both systems, a ground truth for each of the 20 measurement[10] is needed. Due to the lack of a real ground truth, we took the data set of each experiment, which matches the ground truth the most. These Ground Truth Datasets (GTDSs) consists of the vital signs HR_r, RR_r, $SPO_{2,r}$, ST_r, and BP_r of Table3. To ensure that the GTDSs are as close as possible to the real ground truth, all of these signals were additionally filtered[11] to remove noise. Due to corrupted measurements of the vital signs of participant P5, no valid ground truth could be established. Therefore, this participant was excluded from our analysis. This exclusion leads to a reduction of the number of measurements from 20 to 18, and in further consequence, reduced the number experiments: 1296 instead of 1440.

The EWS Ground Truth Dataset (EGTDS) was then created with the GTDSs processed by the conventional EWS system. The EWS system is used for this purpose because it does not - in contrast to the SA-EWS system - manipulate the output leveraging the self-aware properties. However, because the conventional EWS system generated the EGTDSs, it is possible that, if the vital signs of the GTDSs still contain some noise or errors, the SA-EWS system assessment is tagged as erroneous whereas, in reality, the error is in the EGTDS.

We use various metrics to compare these two systems. The Root-Mean-Square Deviation (RMSD) calculation, which indicates how close two different signals are to each other is given by:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(EWS_{GT,i} - EWS_i)^2}{n}} \tag{31}$$

where $EWS_{GT,i}$ is the i^{th} EWS value of the EGTDS and EWS_i the i^{th} outputted EWS value of the system that is compared with the EGTDS.

However, the RMSD is not the best way to compare the two systems. A signal that deviates slightly (e.g., a deviation of only one score) for a long period may have a worse RMSD than a signal that shows a much larger deviation but only for a short time. While the former signal will most likely not result in a false or missing alarm, the latter signal will raise problems. Another metric, namely the maximum absolute error (ε_{max}) which gives information about the highest deviation that occurs in signal compared to the ground truth, is more relevant in this context. It is calculated

[10] 20 measurements is the sum of all test scenarios in that the participants were monitored (Table 4).

[11] A Savitzky-Golay filter with the window size of 53 samples and a polynomial order of 3 was used.

Table 3 Details of the sensors used for data collection.

Vital sign	Reference vital sign / Source		Test vital sign 1 / Source		Test vital sign 2 / Source	
Heart rate	HR_r	Chest strap (Polar T31C)	HR_{t1}	PPG sensor (MAX30100) at 24mA	HR_{t2}	PPG sensor (MAX30102) at 3.5mA
Respiration rate	RR_r	Temperature sensor used as airflow sensor (MCP9808)	RR_{t1}	PPG sensor (MAX30100) at 24mA	RR_{t2}	PPG sensor (MAX30102) at 3.5mA
Blood Oxygen saturation	$SpO_{2,r}$	PPG sensor(MAX30100) at 24mA	$SpO_{2,t1}$	PPG sensor (MAX30102) at 3.5mA		
Skin temperature	ST_r	Temperature sensor (MCP9808)	ST_{t1}	Temperature sensor (TMP102)		
Blood pressure	BP_r	Arm-type blood pressure monitor (iHealth BP7)	BP_{t1}	wrist-type blood pressure monitor (Beurer BC32)		

Table 4 Scenarios of measurements.

Scenario	Scenario description
S1	The person was sitting and no additional error was induced during the measurement.
S2	The person was sitting and the temperature sensor was temporarily detached.
S3	The person was sitting and contracted his/her biceps for a period of the measurement.
S4	The person was sitting and the temperature sensor was temporarily detached. In addition, the person contracted his/her biceps for a period of the measurement.

by:

$$\varepsilon_{max} = max(|EWS_{GT,i} - EWS_i| : i = 1, \ldots, n) \quad (32)$$

where $EWS_{GT,i}$ is the i^{th} EWS value of the EGTDS and EWS_i the i^{th} outputted EWS value of the system that is compared with the EGTDS.

The last metric is the number of false and missing alarms. As mentioned in Section 2, the calculated EWS shows low-, medium-, or high-medical risk of a patient. If the classification of the calculated EWS deviates from the classification of the ground truth EWS, a false or missing alarm is indicated. For example, if the ground truth EWS has a value which belongs to the low or medium risk class but the EWS of the system is in one of the higher classes, a false alarm is raised. In contrast, a calculated EWS in a lower class than the ground truth EWS leads to a missing alarm, which means an alarm should be raised, but it was missed. As a third option, both, the ground truth, as well as the calculated EWS, are in the same class. In this case, there is neither a false nor a missing alarm.

5.3 Results

Table 5 shows the vital signs which are corrupted (✗) and which are uncorrupted (✓) in various experiments. To evaluate which of these signals are either correct or are containing errors, the output of the conventional EWS system processing an experiment was compared with the EGTDS of the same experiment. If a vital sign score abstracted from the vital sign (e.g., RR_{t1}) deviates, at any point during the measurement, from the value, it should have according to the EGTDS, this vital sign (in the considered experiment) is classified as erroneous.

Based on the number of different vital signs, 72 different combinations (setups) of vital sign sets are possible. Such a setup can now contain some correct and some erroneous vital signs. An important factor is how many vital signs are showing an error at the same time for an experiment. The second column of Table 6 shows this number, which ranges from 0 to 4 errors at the same time. For this purpose, all 1296 experiments (18 measurements with 72 different setups) have been processed by the conventional EWS system, and the results were compared to their dedicated EGTDS. Based on the number of simultaneous vital sign errors, the EWS and the SA-EWS system are compared for each participant. In other words, all experiments performed on each person with different vital sign setups were separated in groups regarding the number of vital sign errors that occurred at the same time. Each row in Table 6 shows the performance of the two compared systems in the form of the minimum, average, and maximum RMSD of all calculated EWS values which are in the same group of the number of vital sign errors. Additionally, and more importantly, the maximum absolute error, ε_{max}, is shown for each group.

As it can be seen, in most of the cases, our proposed system performed equally good or considerably better than a conventional EWS system without self-

Table 5 Signals with errors are marked with a ✗, while signals that are correct are marked with ✓.

Participant	Experiment	HR_r	HR_{t1}	HR_{t2}	RR_r	RR_{t1}	RR_{t2}	$SPO_{2,r}$	$SPO_{2,t1}$	ST_r	ST_{t1}	BP_r	BP_{t1}
P1	E1	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓
P1	E2	✓	✗	✓	✓	✗	✗	✓	✓	✓	✗	✓	✓
P1	E3	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗
P1	E4	✓	✗	✓	✓	✗	✗	✓	✓	✓	✗	✓	✗
P2	E1	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓
P2	E2	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗
P3	E1	✓	✗	✗	✓	✗	✗	✓	✗	✓	✓	✓	✗
P3	E2	✓	✓	✗	✓	✗	✗	✓	✓	✓	✗	✓	✓
P3	E3	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✗
P3	E4	✓	✓	✗	✓	✗	✗	✓	✓	✓	✗	✓	✗
P4	E1	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓
P4	E2	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓
P5	E1	Due to invalid ground truth these experiments were excluded.											
P5	E2	Due to invalid ground truth these experiments were excluded.											
P6	E1	✓	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓
P6	E2	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓
P7	E1	✓	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓
P7	E2	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓
P8	E1	✓	✓	✗	✓	✗	✗	✓	✗	✓	✓	✓	✗
P8	E2	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✗

awareness. For a better understanding of the table, here, we discuss the results using participant P1 as an example. In the experiments where no vital sign showed any error, both systems produced an output in that the calculated EWS did not deviate any single time ($\varepsilon_{max} = 0$). In these setups, the calculated EWS signal was exactly identical to the ground truth EWS signal ($RMSD = 0$). In these experiments, both system performances were equal.

In contrast, the conventional EWS system performed much worse than our proposed system when setups were used in which three of the vital signs contained errors at the same time. One of these experiments is shown in Fig. 6. Whereas Fig. 6(a) shows the ground truth vital signs and the corrupted signals, Fig. 6(b) presents the ground truth EWS as well as the outputs of both systems. As it can be seen, the difference between the EWS of the conventional system with the ground truth is large (up to 6-7 scores), whereas the SA-EWS shows only absolute errors of 1 or 2 in the worst case.

The RMSD values of all considered experimental results show that the output of the SA-EWS system was significantly closer to the ground truth. However, the maximum error shows the real importance of an intelligent EWS system. The EWS of the SA-EWS system only deviates in two points for all these experiments; however, the conventionally calculated EWS shows deviations up to 7 points. Participant P5 was excluded from these experiments because of corrupted measurements, which led to an invalid ground truth.

In the four cases of P3, P7, and P8 in Table 6, the conventional EWS system performed slightly better. Some of the participants were sometimes slightly uneasy, which led to temporally irregular breathing. As mentioned, we removed the majority of such noise from the GTDS. However, if there were some noise left, the EWS system may have an advantage over the SA-EWS system because the conventional EWS system generated the EGTDSs.

When comparing the RMSD and the maximum error, the SA-EWS system performed in eleven cases better than the conventional EWS system. In nine cases, the performance was equal, and only in four cases, the conventional EWS system performed slightly better. However, in the latter cases, the performance difference between the two systems was very small and did not lead to any additional false or missed alarms. As a matter of fact, the number of false alarms or missed alarms was always equal or less in the SA-EWS system. Table 7 shows how often both systems missed to raise an alarm or raised a false alarm. As mentioned, the EWS itself can be classified into three different classes, namely low-, medium-, and high risk. If the class of the system's outputted EWS deviates from the class of the ground truth EWS, it causes a false or missing alarm. In all cases, our proposed system performed better (marked in green in Table 7) or equal to the EWS system.

The wrong and missed alarms were counted based on the number of samples which deviate from the ground

Table 6 The minimum, average, and maximum RMSD as well as the maximum error of both systems compared on the base of the various participants and the number of vital sign errors occurring at the same time. Green color highlights the system with better performance.

Participants	Number of simultaneous vital sign errors	EWS System Min. RMSD	EWS System Avg. RMSD	EWS System Max. RMSD	EWS System ε_{max}	SA-EWS System Min. RMSD	SA-EWS System Avg. RMSD	SA-EWS System Max. RMSD	SA-EWS System ε_{max}
P1	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0
	1	0.10	0.94	1.66	3	0.00	0.38	0.81	2
	2	0.68	1.90	2.63	5	0.21	0.71	1.08	2
	3	1.72	2.85	3.25	7	0.37	0.87	1.08	2
P2	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0
	1	0.09	0.28	0.44	2	0.08	0.25	0.40	1
P3	0	0.00	0.00	0.00	0	0.00	0.09	0.20	1
	1	0.07	0.84	2.95	3	0.00	0.59	3.01	4
	2	0.60	1.41	3.28	5	0.23	1.05	3.01	5
	3	0.71	2.27	3.38	6	0.58	1.80	3.01	5
	4	3.35	3.35	3.35	5	2.93	2.93	2.93	5
P4	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0
	1	0.31	0.67	0.74	2	0.08	0.65	0.74	2
	2	0.68	0.70	0.72	2	0.65	0.69	0.72	1
P5	Due to invalid ground truth these experiments were excluded.								
P6	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0
	1	0.12	0.70	1.14	3	0.12	0.64	0.97	1
	2	0.92	1.07	1.20	3	0.87	0.96	0.99	2
P7	0	0.00	0.00	0.00	0	0.40	0.53	0.58	1
	1	0.38	0.57	0.70	2	0.44	0.60	0.70	2
	2	0.53	0.63	0.73	2	0.57	0.65	0.73	2
P8	0	0.00	0.00	0.00	0	0.00	0.10	0.30	1
	1	0.17	0.46	1.01	3	0.10	0.42	0.84	2
	2	0.49	0.85	1.26	5	0.52	0.71	0.88	2
	3	1.26	1.38	1.44	6	0.68	0.74	0.77	3

truth and based on the number of times (events) an alarm was incorrectly raised (false positive) or incorrectly not raised and was missed (false negative). Event-based means when two or more samples of the same event (samples in a row) deviate from the ground truth, the wrong/missed alarm is counted only once. In the example of participant P8, four and eight false alarms were raised by the SA-EWS system. However, each of these false alarms had the length of only one sample. That is why the number in both rows (samples or events) are the same. We can argue that if a doctor monitors a patient's vital signs and obtains an unrealistic result, he/she tries to redo the measurement. A logical consequence of this could be ignoring alarms of a length of only one or few sample(s), which corresponds to one second in time. However, this is out of the scope of this paper and serves only as an additional note. Therefore, we did not discount any alarms, even if they were very short.

Fig. 7 shows the occurrence frequency of absolute error in different sizes for all experiments combined. Both systems have almost the same number of absolute errors in the size of 0 and 1. Overall, except for having an error of 1 score, the proposed system is always better (including when the system has made no false recognition, i.e., 0 on Fig. 7). In particular, the SA-EWS system less often produces larger errors compared to the EWS system. We can see that the SA-EWS system never produce absolute errors larger than 5 (whereas the conventional EWS system experiences them more than a thousand times) and it produces significantly (approximately one order of magnitude) fewer errors in sizes of 4 and 5. This is particularly important since larger errors imply a deviation from the ground truth

Table 7 The number of missing and false alarms of both systems compared on the base of the various participants and the number of vital sign errors occurring at the same time. Green color highlights the system with better performance.

		EWS System				SA-EWS System			
Participants	Number of simultaneous vital sign errors	Missing alarm (samples)	Missing alarm (events)	False alarm (samples)	False alarm (events)	Missing alarm (samples)	Missing alarm (events)	False alarm (samples)	False alarm (events)
P1	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0
	2	0	0	10884	1452	0	0	0	0
	3	0	0	4124	242	0	0	0	0
P2	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0
P3	0	0	0	0	0	0	0	0	0
	1	0	0	6510	132	0	0	1676	2
	2	0	0	22034	1386	0	0	10060	16
	3	0	0	23708	924	0	0	16800	32
	4	0	0	1662	16	0	0	1676	2
P4	0	0	0	0	0	0	0	0	0
	1	0	0	144	16	0	0	0	0
	2	0	0	56	8	0	0	0	0
P5	Due to invalid ground truth these experiments were excluded.								
P6	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
P7	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
P8	0	0	0	0	0	0	0	0	0
	1	0	0	140	98	0	0	0	0
	2	0	0	380	228	0	0	4	4
	3	0	0	802	532	0	0	8	8

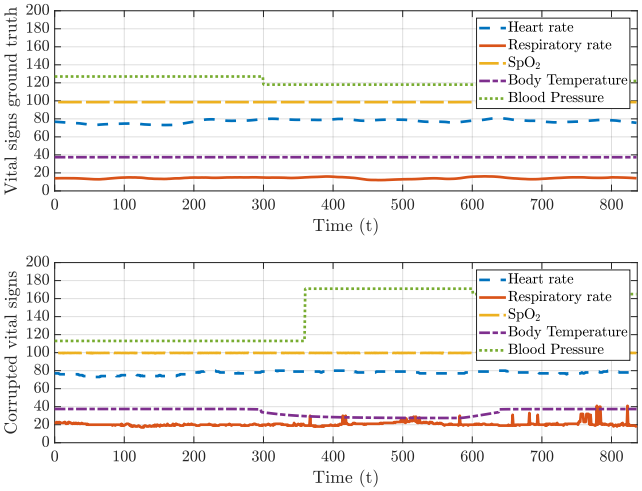
risk class, which is more important with regard to false or missing alarms. Altogether, Fig. 7 indicates that the proposed SA-EWS system is more reliable (less error-prone) than its conventional counterpart.

6 Conclusion and Future Work

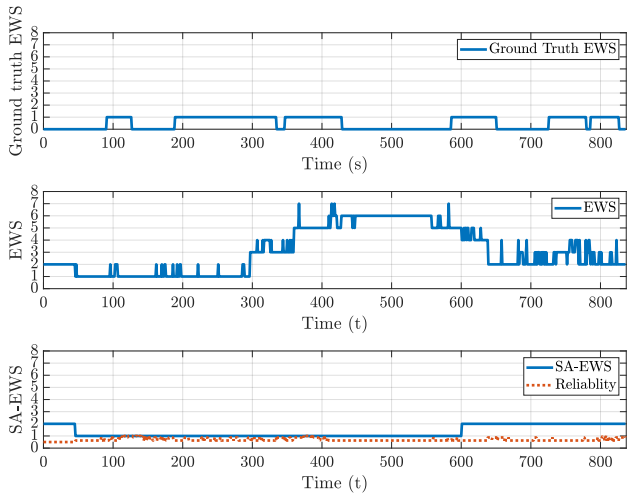
Self-awareness has proven to be advantageous in many applications, and here we show its benefits for wearable medical devices. In particular, we showed how using basic observation elements such as history, data reliability and confidence can lead to reliable results without incurring massive processing loads that conventional Artificial Intelligence (AI) algorithms impose on systems. From the application point of view, we demonstrated that - even using less reliable, low-quality sensors (which are cheaper) - our system is able to calculate the EWS properly and comparable to a system with highly reliable, high-quality sensors (which are more expensive). We also showed that our proposed system shows good resilience against intentionally introduced measurement errors.

In summary, our contributions are; (a) formalizing data reliability, confidence, and history, (b) proposing function for aforementioned self-awareness properties, in particular for combining data reliability and confidence, (c) performing extended experiments with a large number of sensors and test scenarios, and (d) improving reliability of EWS assessment using cheaper sensors and despite adversities in real life measurements.

We note that many of the proposed functions are designed heuristically. Therefore, other functions could be proposed and studied, which lead to further improved results. We leave that for future works. Moreover, in some cases, we have tried alternative parameter settings and chose the better ones; however, these studies



(a) The ground truth vital signs and the corrupted vital signs.



(b) The ground truth EWS (based on the ground truth vital signs) as well as the output of the EWS and SA-EWS system (based on the corrupted vital signs).

Fig. 6 A experiment of participant P1 in scenario S4 with a vital sign setup in that three vital sign errors simultaneously occur.

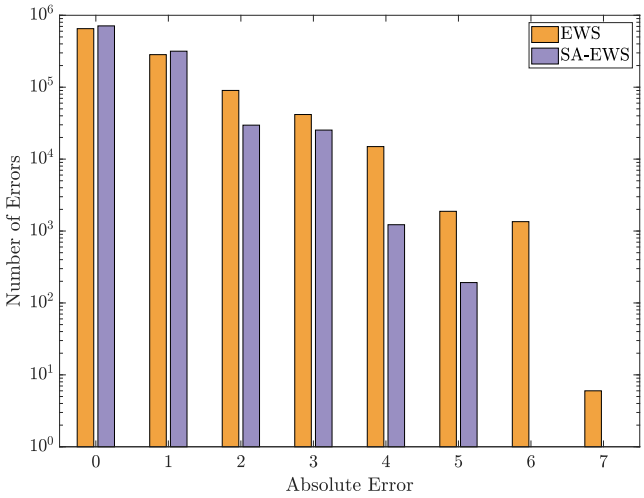


Fig. 7 Occurrence frequency of absolute errors of different sizes.

were not systematic or comprehensive. Mainly due to the extensive time that it takes to process all combination of sensors and errors using single setup values. That is, therefore, another future work.

Acknowledgements

This work was partially supported by the US National Science Foundation (NSF) grant WiFiUS CNS-1702950 and Academy of Finland grant WiFiUS 311764.

References

1. WHO. Chronic diseases and health promotion, Retr. on June 2017. `http://www.who.int/chp/en/`.
2. Jennifer McGaughey, Fiona Alderdice, Robert Fowler, Atul Kapila, Alain Mayhew, and Marianne Moutray. Outreach and early warning systems (ews) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. *The Cochrane Library*, 2007.
3. Morgan et al. An early warning scoring system for detecting developing critical illness. *Clin Intensive Care*, 8(2):100, 1997.
4. Angelika Dohr, Robert Modre-Opsrian, Mario Drobics, Dieter Hayn, and Günter Schreier. The internet of things for ambient assisted living. In *2010 seventh international conference on information technology: new generations*, pages 804–809. Ieee, 2010.
5. Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010.
6. Arman Anzanpour, Amir-Mohammad Rahmani, Pasi Liljeberg, and Hannu Tenhunen. Context-aware early warning system for in-home healthcare using internet-of-things. In *International Internet of Things Summit*, pages 517–522. Springer, 2015.
7. Daniele Miorandi, Sabrina Sicari, Francesco De Pellegrini, and Imrich Chlamtac. Internet of things: Vision, applications and research challenges. *Ad hoc networks*, 10(7):1497–1516, 2012.
8. N. TaheriNejad. Wearable medical devices: Challenges and self-aware solutions. In *IEEE Life Sciences Newsletter*, pages 5–6, 2019.
9. David Pollreisz and Nima Taherinejad. Detection and removal of motion artifacts in ppg signals. *Mobile Networks and Applications*, to appear, 2019.
10. P. Parego, A. M. Rahmani, and N. Taherinejad. *Wireless Communication and Mobile Healthcare*. ecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer International Publishing, 2017.
11. N. Dutt and N. TaheriNejad. Self-awareness in cyber-physical systems. In *2016 29th International Conference on VLSI Design and 2016 15th International Conference on Embedded Systems (VLSID)*, pages 5–6, Jan 2016.
12. N. TaheriNejad and A. Jantsch. Improved machine learning using confidence. In *The Annual IEEE Canadian*

Conference on Electrical and Computer Engineering, volume To appear, pages 1–5, May 2019.
13. Maximilian Götzinger, Arman Anzanpour, Iman Azimi, Nima TaheriNejad, and Amir M Rahmani. Enhancing the self-aware early warning score system through fuzzified data reliability assessment. In *International Conference on Wireless Mobile Communication and Healthcare*, pages 3–11. Springer, 2017.
14. R. W. Urban et al. Modified early warning system as a predictor for hospital admissions and previous visits in emergency departments. *Advanced emergency nursing journal*, 37(4):281–289, 2015.
15. Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Updated report of a working party, London: RCP, 2017.
16. Excel Medical's WAVE Clinical Platform Receives FDA Clearance, Retrieved on February 2018. `http://excel-medical.com/wp-content/uploads/RELEASE_FDA_Clearance_AchievedbyExcelMedical_1-8-18_RevMB6A.pdf`.
17. U. Kyriacos et al. Monitoring vital signs: Development of a modified early warning scoring (mews) system for general wards in a developing country. *PLoS One*, 9(1):e87073, 2014.
18. N. Holbery and P. Newcombe. *Emergency Nursing at a Glance*. Wiley, 2016.
19. National Clinical Effectiveness Committee et al. National early warning score national clinical guideline no. 1, 2013.
20. A. Anzanpour et al. Internet of Things Enabled In-Home Health Monitoring System Using Early Warning Score. In *Proc. of MobiHealth*, 2015.
21. Mark Hung. Leading the iot, gartner insights on how to lead in a connected world. *Gartner Research*, pages 1–29, 2017.
22. I. Azimi et al. Self-Aware Early Warning Score System for IoT-Based Personalized Healthcare. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, volume 181, 2017.
23. Nikil Dutt et al. Towards smart embedded systems: A self-aware system-on-chip perspective. *ACM TECS, Special Issue on Innovative Design Methods for Smart Embedded Systems*, 15(2):22–27, 2016.
24. H. Hoffmann et al. A generalized software framework for accurate and efficient management of performance goals. In *Proceedings of the International Conference on Embedded Software*, pages 1–10, Sept 2013.
25. IBM Corporation. An architectural blueprint for autonomic computing, 2006. IBM White Paper.
26. Arman Anzanpour, Iman Azimi, Maximilian Götzinger, Amir M. Rahmani, Nima TaheriNejad, Pasi Liljeberg, Axel Jantsch, and Nikil Dutt. Self-awareness in remote health monitoring systems using wearable electronics. In *Proceedings of Design and Test Europe Conference (DATE)*, Lausanne, Switzerland, March 2017.
27. N. TaheriNejad, M. A. Shami, and S. M. P. D. Self-aware sensing and attention-based data collection in multi-processor system-on-chips. In *2017 15th IEEE International New Circuits and Systems Conference (NEWCAS)*, pages 81–84, June 2017.
28. Hedyeh A. Kholerdi, Nima TaheriNejad, and Axel Jantsch. Enhancement of classification of small data sets using self-awareness - an iris flower case-study. In *To be published in the proceedings of the International Symposium on Circuit and Systems (ISCAS)*, Florence, Italy, 2018.
29. Nima TaheriNejad, Axel Jantsch, and David Pollreisz. Comprehensive observation and its role in self-awareness; an emotion recognition system example. In *FedCSIS Position Papers*, pages 117–124, 2016.
30. Maximilian Götzinger, Nima TaheriNejad, Hedyeh A Kholerdi, and Axel Jantsch. On the design of context-aware health monitoring without a priori knowledge; an ac-motor case-study. In *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–5. IEEE, 2017.
31. M Gotzinger, E Willegger, N TaheriNejad, A Jantsch, T Sauter, T Glatzl, and P Lilieberg. Applicability of context-aware health monitoring to hydraulic circuits. In *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, pages 4712–4719. IEEE, 2018.
32. Maximilian Götzinger, Nima Taherinejad, Amir M. Rahmani, Pasi Liljeberg, Axel Jantsch, and Hannu Tenhunen. *Enhancing the Early Warning Score System Using Data Confidence*, pages 91–99. Springer International Publishing, Cham, 2017.
33. M Götzinger, N TaheriNejad, HA Kholerdi, A Jantsch, E Willegger, T Glatzl, AM Rahmani, T Sauter, and P Liljeberg. Model-free condition monitoring with confidence. *International Journal of Computer Integrated Manufacturing*, 32(4-5):466–481, 2019.
34. Mathieu Pasquier, Paul-AndrÃľ Moix, Dominique Delay, and Olivier Hugli. Cooling rate of 9.4 °C in an hour in an avalanche victim . *Resuscitation*, 93:e17 – e18, 2015.
35. S. Reule and P. Drawz. Heart rate and blood pressure: Any possible implications for management of hypertension? *Curr Hypertens Rep*, 14(6):478–84, 2012.
36. P. Davies and I. Maconochie. The relationship between body temperature, heart rate and respiratory rate in children. *Emerg Med J.*, 26(9):641–3, 2009.
37. I. Zila and A. Calkovska. Effects of elevated body temperature on control of breathing. *Acta Medica Martiniana*, 2011.
38. Liang Guang, Ethiopia Nigussie, Pekka Rantala, Jouni Isoaho, and Hannu Tenhunen. Hierarchical agent monitoring design approach towards self-aware parallel systems-on-chip. *ACM Transactions on Embedded Computing Systems (TECS)*, 9(3):25, 2010.
39. Henry Hoffmann, Martina Maggio, Marco D Santambrogio, Alberto Leva, and Anant Agarwal. Seec: A framework for self-aware computing, 2010.
40. Timothy J Ross. *Fuzzy logic with engineering applications*. John Wiley & Sons, 2009.
41. VL Richmond, DM Wilkinson, SD Blacker, FE Horner, J Carter, George Havenith, and MP Rayson. Insulated skin temperature as a measure of core body temperature for individuals wearing cbrn protective clothing. *Physiological measurement*, 34(11):1531, 2013.
42. A. Garde et al. Estimating respiratory and heart rates from the correntropy spectral density of the photoplethysmogram. *PLoS One*, 9(1), 2014.
43. M. A. F. Pimentel et al. *Probabilistic Estimation of Respiratory Rate from Wearable Sensors*, pages 241–62. Springer, 2015.
44. D. Amiri et al. Edge-assisted sensor control in healthcare iot. In *IEEE Global Communications Conference: Selected Areas in Communications: E-Health*, 2018.
45. J.E. Sinex. Pulse oximetry: Principles and limitations. *The American Journal of Emergency Medicine*, 17(1):59–66, 1999.
46. Maxim Integrated. , (accessed Sept. 2018). `https://www.maximintegrated.com/en/products/sensors/MAX30102.html`.