Hindawi
Mobile Information Systems
Volume 2018, Article ID 1546210, 11 pages
https://doi.org/10.1155/2018/1546210

Research Article

Artificial Intelligence to Prevent Mobile Heart Failure Patients Decompensation in Real Time: Monitoring-Based Predictive Model

Nekane Larburu ,[1,2] Arkaitz Artetxe,[1,2] Vanessa Escolar,[3] Ainara Lozano,[3] and Jon Kerexeta[1]

[1] *Vicomtech, Paseo Mikeletegi 57, 20009 Donostia/San Sebastian, Spain*
[2] *Biodonostia Health Research Institute, P. Doctor Begiristain s/n, 20014 San Sebastian, Spain*
[3] *Hospital Universitario de Basurto (Osakidetza Health Care System), Avda. Montevideo 18, 48013 Bilbao, Spain*

Correspondence should be addressed to Nekane Larburu; nlarburu@vicomtech.org

Received 29 June 2018; Accepted 30 September 2018; Published 5 November 2018

Guest Editor: Giovanna Sannino

Rapid advances in ICT and collection of large amount of mobile health data are giving room to new ways of treating patients. Studies suggest that telemonitoring systems and predictive models for clinical support and patient empowerment may improve several pathologies, such as heart failure, which admissions rate is high. In the current medical practice, clinicians make use of simple rules that generate large number of false alerts. In order to reduce the false alerts, in this study, the predictive models to prevent decompensations that may lead into admissions are presented. They are based on mobile clinical data of 242 heart failure (HF) patients collected for a period of 44 months in the public health service of Basque Country (Osakidetza). The best predictive model obtained is a combination of alerts based on monitoring data and a questionnaire with a Naive Bayes classifier using Bernoulli distribution. This predictive model performs with an AUC = 67% and reduces the false alerts per patient per year from 28.64 to 7.8. This way, the system predicts the risk of admission of ambulatory patients with higher reliability than current alerts.

1. Introduction

Since these early days, the advances on ICT have given a huge opportunity to telemedicine applications and new e-Health services [1]. Along with this phenomenon are the large quantities of mobile data that are being collected and processed these days. The growth in these two areas are leading in advanced health-care systems that not only provide continuous support to clinicians or informal care givers (e.g., family members), but also to patients. In this context, telemedicine systems that monitor ambulatory patients and guide them in their daily routine are emerging. Nevertheless, often all the potential of the mobile-health data used to support clinical professionals and patients is not sufficiently exploited. Other times, the exploited clinical data, in the form of, for example, predictive models to identify patients at high risk, are not applied in a real setting to support clinicians and patients. Studies suggest that artificial intelligence by means of predictive models and telemonitoring systems for clinical support and patient empowerment may improve several pathologies [2], such as heart failure.

Heart failure (HF) is a clinical syndrome caused by a structural and/or functional cardiac abnormality. HF patients suffer decompensations, which is defined by Mangini et al. [3] as a clinical syndrome in which a structural or functional change in the heart leads to its inability to eject and/or accommodate blood within physiological pressure levels, thus causing a functional limitation and requiring immediate therapeutic intervention [3]. Hence, decompensations may lead in hospital admissions, which in this study are defined as emergency admissions and hospital admissions, and home interventions. As Ponikowski et al. presented in [4], the prevalence of HF depends on the definition applied, but it is approximately 1-2% of the adults in developed countries, rising to more than 10% among people >70 years of age. Hence, due to the aging population, an increase in the number of HF patients is expected in the future. Therefore, predicting the risk of a patient to suffer a decompensation may prevent admissions and readmissions, improving both patient care and hospital management, which has a high impact on costs and clinical professionals time. The first step to predict the risk of decompensation is to telemonitor ambulatory patients. Next, we need reliable systems to assess the risk. Most telemedicine systems apply alerts or rule-based systems to detect potential complications of ambulatory patients [5–8]. But these usually contain large number of false alerts, and hence, these systems are not trustworthy (Table 1).

Our hypothesis is that with the usage of artificial intelligence (AI) by means of, for instance, predictive models, it is possible to detect decompensations of ambulatory patients and reduce false alerts. In this context, this research extends the study for readmissions detection [9] and presents predictive models of a telemedicine system for heart failure patients, called INCAR. INCAR has been developed to (i) be generally applicable in HF patients, (ii) improve the clinical practice by developing an accurate system that detects the risk of decompensation and suggest actions to prevent them on time, (iii) allow professionals to maintain an efficient and personalized support and follow-up of patient, (iv) give patients support when required and guide them in risk situations, informing clinicians accordingly, and (v) reduce HF patients admission and readmission rate, which have a high economic impact.

This paper focuses on the development of predictive models to detect decompensations, and it is structured as follows: First, the *Related Work* section summarizes the state of the art on telemedicine systems for heart failure and the role of predictive models on telemedicine systems. *Materials* section presents the database used in this study and the characteristics of the dataset. *Methods* presents the applied methods to assess the risk of an ambulatory HF patient to suffer a decompensation that may lead into admission. In *Results*, the outcomes obtained for each of the developed predictive models is presented. Finally, *Discussion* presents the results and limitations of the study, and *Conclusion* gives a summary of the contributions and future work.

2. Related Work

2.1. Telemedicine Systems for HF Patients. Being HF a disease with high prevalence and high readmission rate, the usage of telemedicine systems in this area is common [7]. Chaudhry et al. [2] telemonitored patients by means of telephone-based interactive voice-response system and concluded that the simple phone-based telemonitorization does not improve the outcomes (i.e., readmission, death). Nevertheless, most of current telemonitoring systems do not simply implement telephone-based monitorization, but also the transmission of mobile health data, such as bodyweight, heart rate, and blood pressure [7]. Besides, more advanced noninvasive systems transfer electrocardiograph (ECG) tracings, oxygen saturation, and physical activity (e.g., pedometer) data. Apart from noninvasive telemedicine systems, invasive systems enable the transfer of variables measured invasively, such as transthoracic impedance and pulmonary and left atrial pressures. But literature studies do not present significantly better results when implementing invasive measurements into their telemedicine systems in terms of HF decompensation prevention. Nonetheless, some benefits have been presented when applying impedance instead of weight for detecting HF patients early decompensation, as presented by Abraham et al. [5] and Gyllensten et al. [6].

2.2. Alerts for HF Patients. Most studies implement "simple" alerts to prevent decompensations based on these data. One of the implemented techniques is *Rule of Thumb (RoT)* based on simple rules (i.e., when a measurement goes beyond or below a given threshold or when they are based on simple difference between the current value of an attribute and a previous measurement that occurs a predefined number of days in the past) [5–8]. Other studies, such as Zhang et al. [7], Gyllensten et al. [6], and Ledwidge et al. [8], make also use of more sophisticated techniques, such as the *Moving Average (MA)* or similar techniques that calculate the variations applied to usually weight. The *Cumulative Sum (CUMSUM)*, applied by Adamson et al. [10], is typically used for detecting changes and implies that when a continuous variation of a measurement is produced over time, that tendency will result in an alert. Additionally, Gilliam et al. [11], apply the *multivariate method*, which consists on the usage of several data elements that are incorporated into a multivariate logistic regression model to form the probability of an event occurring. From the studied papers, we could conclude that each type of alert may work best depending on the applied attribute. For instance, techniques related with MA work best when applied to weight. On the other hand, CUMSUM is one of the best methods when applied to transthoracic or intrathoracic impedance.

Table 1 presents the results of different studies that determine whether a monitored HF patient will have a decompensation, usually implementing alerts. Due to the large number of days that do not end in an admission, even when the computed specificity values are high, the number of false positives could remain too high for the clinical practice, so it is not an optimal testing value in this scope. Taken into account this limitation, based on the literature studies, we could consider the number of false alerts per patient per year (FA/pt-y) as de facto standard to determine the number of false positives. However, as shown in Table 1, some of the studies present the specificity value for determining how well the no admissions are detected using own techniques to compute it.

2.3. Predictive Models on Telemedicine Systems. As shown above, most telemedicine systems apply alerts or rule-based systems to detect potential complications of ambulatory patients. This is not only present in the context of HF, but also in diabetes, atrial fibrillation, and other clinical domains [12, 13]. Hence, there is a lack of the usage of collected data that could lead in more accurate solutions by means of, for instance, predictive models.

Table 1: Summary of decompensation detection studies.

Study	Data type	Dataset	Method	Results
Zhang et al. [7]	Weight	135 patients; 1964 days monitoring	RoT	Se = 58.3%, Sp = 54.1%
			MACD	Se = 20.4%, Sp = 89.4% (AUC = 0.55%)
Gyllensten et al. [6]	Weight	91 patients; 10 months	RoT	Se = 20%, Sp = 90%
			MACD	Se = 33%, Sp = 91%
			CUMSUM	Se = 13%, Sp = 91%
	Noninvasive transthoracic bioimpedance	91 patients; 10 months	RoT	Se = 13%, Sp = 91%
			MACD	Se = 13%, Sp = 91%
			CUMSUM	Se = 13%, Sp = 91%
Adamson et al. [10]	Blood pressure	274 patients	CUMSUM	Se = 83.1%, FA = 4.1/pt-y
Abraham et al. [5]	Intrathoracic impedance	156 patients; 537 ± 312 days	RoT	Se = 76.4%; FA = 1.9/pt-y
	Weight	156 patients; 537 ± 312 days	RoT	Se = 21%; FA = 4.3/pt-y
Ledwidge et al. [8]	Weight	87 patients; 23.9 ± 12 weeks	RoT	Se = 21%; Sp = 86%
			HeartPhone algorithm (based on MA)	Se = 82%; Sp = 68%
Gilliam et al. [11]	Multivariate	201 patients		Se = 41%; FA = 2/pt-y

Several studies in the context of HF develop predictive models to determine whether a patient will be readmitted within 30 days after discharge [14–20]. These predicting models make use of baseline information of patients, such as age, sex, or left ventricular injection fraction, but not daily (or weekly) telemonitored patient mobile data, such as weight, heart rate, or blood pressure, which could be crucial for detecting and preventing an ambulatory patient admission. In several telemedicine studies applied in diverse pathologies, such as chronic obstructive pulmonary disease [21, 22] and preeclampsia [23], predictive models have been successfully applied. However, in the context of HF, limited studies apply predictive models. Lafta et al. [24] is one of these studies that using several telemonitored attributes (i.e., heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, and oxygen saturation) applied basic time series prediction algorithm, regression-based time series prediction algorithm, and hybrid time series prediction algorithm. The obtained results showed that up to 75% and 98% of accuracy values could be obtained across different patients under three algorithms, but still the accuracy value is not objective enough to determine how well the system performs.

The presented study goes beyond the state of the art and applies classifiers based on alerts applied in current medical practice and state-of-the-art studies. Additionally, this study makes use of baseline information and ambulatory telemonitored information to build an integral telemedicine system that applies predictive models with double goal: assess ambulatory patients' admission risk to provide both patients and clinicians the appropriate guidance to prevent potential decompensations that may lead to hospital admissions.

3. Materials

3.1. Database. The public hospital OSI Bilbao-Basurto (Osakidetza), located in Basque Country (Spain), has been gathering HF patients' information from June 2014 until February 2018 (44 months) to closely monitor HF patients. For the present study, the dataset contained a cohort of 242 HF patients. Clinicians have collected baseline data (i.e., information collected by a clinician when the patient is diagnosed, Table 2), ambulatory patient monitored data (i.e., information collect from three to seven times per week, Tables 3 and 4), and patients admissions information (i.e., emergency admissions, hospital admissions, and home care interventions that are associated to HF associated with a patient decompensation).

Besides vital signs, a questionnaire is also included into the telemonitoring system to ask patients about their condition, with potential impact on decompensation prediction (Table 4).

3.2. Characteristics of Ambulatory Patients Dataset. In the whole study, 242 patients have been enrolled from June 2014 until February 2018. Of these 242 patients, one patient has been excluded as it is a cirrhotic man who often has interventions of evacuational paracentesis due to a liver pathology not related to HF. There is an average follow-up of 13.5 ± 9.11 months. In this time period, there have been 254 decompensations of which 202 are considered as predictable, since 52 decompensations do not have previous telemonitoring information (i.e., less than 3 times in the last week before the decompensation).

4. Methods

Following the methodology applied for the generation of the predictive models is presented: (i) training and testing dataset construction, (ii) application of alerts implemented in current clinical setting, (iii) selection of the alerts for the study, (iv) generation of the dataset to apply the machine learning classifiers, and (v) the application and comparison of different classifiers.

Table 2: Baseline characteristics of the study population.

Characteristics	Description	Median ± SD (percentage)
Age	The age of the patient (years)	78 ± 10.9
Height	The height of the patient (mm)	162.37 ± 10.34
Sex	The sex of the patient (men/women)	57% men
Smoker	If the patient smokes, did smoke, and now do not or never has smoked	15.35% do smoke, 22% did smoke (not now)
LVEF	Left ventricular ejection fraction (%)	42.4 ± 15.21
First diag	Years since first diagnosis	5.8 ± 7.04
Implanted device	If implanted device (peacemaker, implanted cardioverter defibrillator, and cardiac resynchronisation therapy)	22.7%
Need oxygen	If the patient needs oxygen	4.7%
Barthel	Barthel scale	82.98 ± 15.23
Gijón [25]	Sociofamily assessment scale in the elderly that allows the detection of risk situations or social problems.	7.47 ± 2.29
Laboratory		
Urea	Urea (mg/dl)	75.12 ± 37.8
Creatinine	Creatinine (mg/dl)	1.3 ± 0.54
Sodium	Sodium (mEq/L)	140.12 ± 4.14
Potassium	Potassium (g/dl)	4.28 ± 0.74
Haemoglobin	Haemoglobin (g/dl)	13 ± 9.6
Comorbidities		
Rhythm	If sinus rhythm, AF or atrial fluter	Sinus: 37.1%
Atrial fibrillation	If the patient has atrial fibrillation (AF)	57.4%
Pacemaker	If the patient has a pacemaker	14.5%

Table 3: Ambulatory patients monitored characteristics of the study population.

Characteristics	Description
SBP	Systolic blood pressure (mmHg)
DBP	Diastolic blood pressure (mmHg)
O2Sat	Oxygen saturation (%)
HR	Heart rate (bpm)
Weight	Body weight (kg)

4.1. Splitting Training and Testing Datasets.

To build and test a predictive model, the clinical data are divided in training and testing datasets. The training dataset is used to develop the model, and once it is finished, the resulting model is tested with the testing dataset. This way, the overfitting is prevented, and it is possible to check whether the created model will generalize well. The whole dataset is from telemonitored patients starting from June of 2014 until February 2018. The training dataset contains 132 predictable decompensations (i.e., with at least 3 monitorizations in the last week before a decompensations) out of 174 patients, with an average follow-up per patient of 13.47 ± 7.47 months. The testing set contains 70 predictable decompensations out of 162 patients, with an average follow-up per patient of 5.41 ± 3.48 months.

4.2. Applied Alerts for Ambulatory Patients Admission.

The alerts implemented in current medical practice are used as a filtering method to obtain the instances for training and building the classifiers. This way, we discard the days when there is no sign of destabilization of any attribute, leading into a more balanced dataset. Therefore, this section presents the different types of alerts that are implemented in medical practice and their performance to select the optimal ones to be applied in our study.

4.2.1. Generic Alerts.

The following tables describe the alerts that are being implemented in OSI Bilbao-Basurto Hospital and their sensitivity (Se) and false alerts per patient per year (FA/pt-y) when applied to the training dataset. They are differentiated into "yellow" and "red" alerts, being these last ones more restrictive and, therefore, more critical.

Simple Rules. Table 5 presents the rules based on each parameter individually. The alerts' thresholds presented in Table 5 are the generic ones. But based on personalized clinical cases, clinicians modified some patients' alerts thresholds. For example, if a patient's O2Sat values are always lower than 90, but the patient is stable, the O2Sat alerts are adapted. This study uses the adapted alerts.

Weight Tendency. Besides simple rules, OSI Bilbao-Basurto Hospital also checks the tendency of weight values in order to trigger an alarm (Table 6). This weight change "red" ("yellow") alert performs with a Se value of 0.52 (0.64) and a FA/pt-y of 9.55 (16.38).

Questionnaire. Additionally, OSI Bilbao-Basurto Hospital clinicians make use of the questionnaire (Table 4) and apply the following alert based on the answers from the questionnaire: if three or more answers are the wrong ones, the questionnaire alert would trigger. This alert achieves

Table 4: Ambulatory patients questionnaire.

n	Tag	Question	Possible answer
1	Well-being	Comparing with the previous 3 days, I feel:	B/W/S*
2	Medication	Is the medication affecting me well?	Yes/No
3	New medication	During the previous 3 days, did I take any medication without my clinicians' prescription?	Yes/No
4	Diet and exercise	Am I following the diet and exercise recommendations provided by my clinician and nurse?	Yes/No
5	Ankle	In the last 3 days, my ankles are:	B/W/S*
6	Walks	Can I go walking like previous days?	Yes/No
7	Shortness of breath	Do I have fatigue or shortness of breath when I lay down in the bed?	Yes/No
8	Mucus	Do I notice that I started coughing of with phlegm?	Yes/No

*B/W/S = better/worse/same.

Table 5: Simple rules implemented by Osakidetza.

Parameter to study	Threshold number	Type of alert	Se	FA/pt-y
SBP	<95 >150	Yellow	0.28	11.4
	<85 >180	Red	0.08	1.4
DBP	<60 >100	Yellow	0.23	9.1
	<50 >110	Red	0.04	0.9
HR	<55 >90	Yellow	0.30	11.2
	<50 >110	Red	0.08	1.4
O2Sat	<94	Yellow	0.15	3
	<90	Red	0.39	13.5

Table 6: Weight alerts implemented by Osakidetza.

Parameter to study	Time period	Minimum (kg)	Maximum (kg)	Type of alert
Weight change	5 days	1	2	Yellow
	3 days	1	25	Red
	5 days	2	25	Red

a Se of 0.31 and FA/pt-y of 9.55. To determine which are the questions that perform best, Table 7 presents the Se and FA/pt-y for each of them based on each possible answer.

The answers of "Worse" in the questions of n1 and n5 (Table 4) result in very good predictors of the decompensations considering Se and FA/pt-y values. Questions n3, n4, n6, and n7 also have predictive power, but not as good as n1 and n5. The other questions cannot be considered as alerts, because of their low/null prediction capacity (Table 7).

4.2.2. Implemented Alerts Based on Moving Average. As presented in the *Related Work* section, weight-associated alerts have been improved, and hence, tendency rules for weight have been substituted for a more advanced method, based on moving average. Moving Average Convergence Divergence (MACD) algorithm calculates the difference between the average value taken from two windows and generates an alert when this difference exceeds a prespecified threshold. Following the same moving average (MA) concept, a similar method is implemented which consists on the following (Figure 1):

(i) a: immediate previous days (starting from the checking day and continuing backwards) over which the average is calculated

(ii) b: previous days (starting from at least the latest day from a and continuing backwards) over which the average should be calculated

(iii) d: distance between the last day of a and first day from b

(iv) Difference threshold (THR): size of difference between a and b average that should generate an alert

In Figure 2, different scores for each possible variable's value for the MA alert are presented. The tested and illustrated results are from all possible combinations of the following variable's values: $a = (2, 3, 7)$, $b = (3, 4, 7, 14)$, $d = (0, 1, 3, 7)$, and THR = (0.2, 0.5, 0.75, 1, 1.5, 1.8, 2, 3).

After representing all the results of the MA algorithm and applying the Youden index [26], the optimal value of these combinations is the one obtained with $a = 2$, $b = 3$, $d = 0$, and THR = 0.75 (green dot in Figure 2). This alert achieves Se value of 0.56 and FA/pt-y of 11.06 in the training set, similar to the results of the already alert-implemented weight alert. But based on the literature [6–8], this latest one is best.

4.3. Selection of Alerts for Instances Generation. To obtain the right dataset of instances, the best combination of alerts is sought. Once the alerts are selected, when at least one of these alerts is triggered, the patient data of that day are used to build the dataset for machine learning model building (see *Built Dataset for Machine Learning Classifiers*). In Table 8, the results of the combinations of different alerts are presented.

R1 refers to the sum of MA weight alert and the two best alerts from the questionnaire related to *ankle* (n5) and

Table 7: Questionnaire questions' performance.

n	Tag	Answer	Se	FA/pt-y
1	Well-being	Same	0.88	120.7
		Better	0.42	90.8
		Worse	0.37	2.7
2	Medication	Yes	1	210
		No	0.05	3.8
3	New medication	Yes	0.15	5.3
		No	1	209
4	Diet and exercise	Yes	1	203
		No	0.22	11.18
5	Ankle	Same	0.86	114
		Better	0.44	96
		Worse	0.35	2.9
6	Walks	Yes	0.99	196
		No	0.37	18
7	Shortness of breath	Yes	0.41	19.93
		No	0.96	194
8	Mucus	Yes	0.44	60.5
		No	0.84	153.5

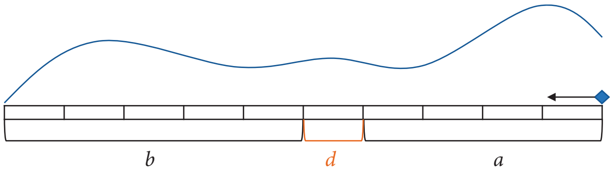


Figure 1: Representation of the applied MA algorithm.

well-being (n1) (Tables 4 and 7). If some of these alerts are triggered, R1 is also triggered. R2 refers to the R1 plus the yellow alerts of SBP, DBP, O2Sat, and HR (Table 5). Finally, R3 refers to R2 plus the questions n3, n4, n6, and n7 from Table 4. Since R2 (Table 8) detects almost all decompensations (95%), though FA/pt-y is quite high (FA/pt-y = 51.12), this is the one used to generate the instances for the machine-learning classifiers.

4.4. Built Dataset for Machine-Learning Classifiers.

Next, the attributes that are considered for each of the instances and that are applied in the classifiers are presented. Note that the applied attributes come from (i) the telemonitoring dataset, (ii) the baseline dataset, and (iii) the readmission dataset.

(i) *Telemonitoring dataset*:

(a) The value of SBP, DBP, HR, O2Sat attributes, and, in the case of the weight, the values of the MA algorithm

(b) The number of consecutive alerts for each type of alert:

(1) Yellow alerts: the number of yellow alerts that have been triggered in the previous consecutive days related to SBP, DBP, HR, and O2Sat (4 attributes)

(2) Red alerts: the number of red alerts that have been triggered in the previous consecutive days related to SBP, DBP, HR, and O2Sat (4 attributes)

(3) MA: the number of alerts that have been triggered in the previous consecutive days for the MA algorithm (1 attribute)

(c) Questionnaires: the answers of the 8 questions of the questionnaire, shown in Table 7 (8 attributes)

(ii) *Baseline dataset*: the baseline information of the patient shown in Table 2 (24 attributes)

(iii) *Readmissions dataset*: whether in the moment of the instance is about a readmission, i.e., if the last 30 days, the patient has discharged because HF (1 attribute)

4.5. Applied Machine-Learning Classifiers.

In this section, we briefly describe the main classification algorithms that were used during the experiments carried out in this work. Since classifier definitions are well known in the literature, we will provide just a summary overview about them.

4.5.1. Naïve Bayes.

Naive Bayes methods follow the "naive" assumption that the components of the feature vectors are statistically independent, so that the posterior probability of the class can be approximated as

$$p(y \mid x) = \frac{p(y)\prod_i^n (p(x_i \mid y))}{p(x)}, \quad (1)$$

where $p(x_i)$ is the likelihood of the i-th feature, and $p(y)$ the a priori probability of the class. The Gaussian Naive Bayes assumes that the likelihood follows a Gaussian distribution, where the mean and standard deviation of each feature are estimated from the sample. On the other hand, the Bernoulli Naive Bayes assumes Bernoulli's distribution in the parameters, and hence, it estimates the probability of $p(x_i \mid y)$ following this last distribution.

4.5.2. Decision Tree.

Decision Trees (DTs) [27, 28] are built by recursive partitioning of the data space using a quantitative criterion (e.g., mutual information, gain-ratio, gini index), maybe followed by a pruning process to reduce overfitting. Tree leaves correspond to the probabilistic assignment of data samples to classes. One of the most popular implementations of the algorithm is C4.5 [27], which is an extension of the previous ID3 [29] algorithm. At each node, the algorithm selects the feature that best splits the samples according to the normalized information gain.

4.5.3. Random Forest.

Random forest [30] is an ensemble classifier consisting of multiple decision trees trained using randomly selected feature subspaces. This method builds multiple decision trees at the training phase. Often, a pruning process is applied to reduce both tree complexity and training data overfitting. In order to predict the class of a new instance, it is put down to each of these trees. Each tree

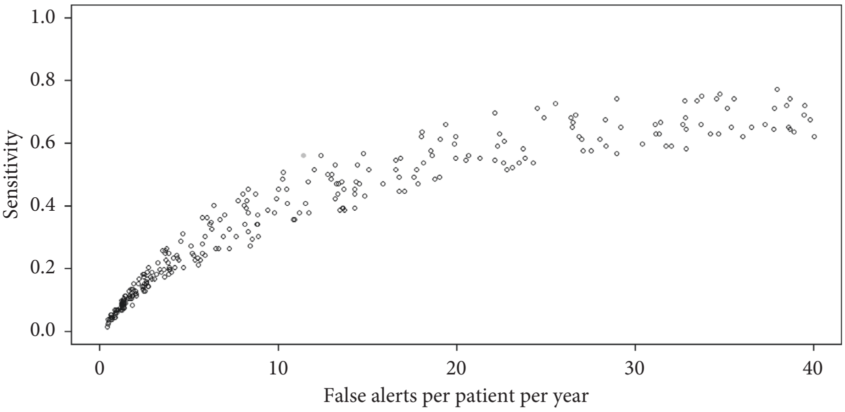


Figure 2: Representation of several MA to determine the Younden index.

Table 8: Inclusion Criterion performance.

Rules	Description of the rules	Se	FApy
R1	Weight + ankle + well-being	0.79	15.33
R2	R1 + yellow	0.95	51.12
R3	R2 + questionnaire alerts	1	84.5

gives a prediction (votes) and the class having most votes over all the trees of the forest will be selected (majority voting). The algorithm uses the bagging method [31], where each tree is trained using a random subset (with replacement) of the original dataset. In addition, each split uses a random subset of features.

4.5.4. Support Vector Machine. Support vector machines (SVMs) [32, 33] look for the set of support vectors that allow to build the optimal discriminating surface in the sense of providing the greatest margin between classes. In this way, the decision function can be expressed in terms of the support vectors only:

$$f(x) = \operatorname{sign}\left(\sum \alpha_i y_i K(s_i, x) + w_0\right), \qquad (2)$$

where $K(x_i, x_j) \equiv \emptyset(x_i) T \emptyset(x_j)$ is a kernel function, α_i is a weight constant derived from the SVM process, and the s_i is the support vector [33]. Nonlinear kernel functions filling some conditions allow to map a nonlinearly separable discrimination problem into a linearly separable equivalent problem in higher dimensional space.

4.5.5. Neural Network. Multilayer Perceptron (MLP) is a neural network that consists of at least three layers of nodes, namely: (i) an input layer, (ii) one or more hidden layers, and (iii) an output layer. The input layer consists of a set of neurons that represents input features. The hidden layer transforms the outputs of the input layer by means of nonlinear activation functions. The output layer collects the values of the hidden layer and builds the output value. The model is trained using backpropagation, and it can classify data that is not linearly separable.

4.5.6. Class Balancing. In this work, like in many other supervised classification problems, imbalanced class distribution leads to important performance evaluation issues and problems to achieve desired results. The underlying problem with imbalanced datasets is that classification algorithms are often biased towards the majority class and hence, there is a higher misclassification rate of the minority class instances. Although there are several methods that can be used to tackle the class imbalance problem, we have followed an oversampling approach. Random oversampling is the simplest oversampling method, which consists of randomly replicating minority class samples. Despite its simplicity, this method leads easily to overfitting, since it generates exact copies of existing instances [34]. In order to deal with such problems, we have used a more sophisticated technique, namely, synthetic minority oversampling technique (SMOTE). This method over samples the minority class by creating synthetic instances based on its nearest neighbours [35].

Depending on the percentage of synthetic samples that want to be generated (in respect to the original minority class instances), some, or all, minority samples are selected. Having specified beforehand the number of nearest neighbours *k*, for each sample, the *k* nearest neighbours are found using the Euclidean distance. Once the nearest samples are selected, a random value between 0 and 1 is generated and multiplied to the distance of each feature between the actual instance and the neighbour. In other words, the vector of coefficients of a random convex linear combination is generated and applied to the *k* nearest neighbours to create a new sample.

5. Results

This section presents the results obtained after the development of the machine learning classifiers presented in *Applied Machine-Learning Classifiers* and the final results of the selected classifier in the testing dataset.

5.1. Validation Method. Although there are many ways to assess the generalization ability of a ML model, such as cross-validation, time series can be problematic for such validation techniques, as there is an ordered timeline factor to be considered. Henceforth, we use cross-validation on a rolling basis [36], as it is explained in Figure 3.

The training set is separated in the five sets shown in Figure 3. The number written inside the blocks is the number of decompensations corresponding to that period, which is the reason why the dates (on top) are chosen. The splits are not exactly equitable, since all the predecessors of a decompensation must fit within the same block. In Step 1, the classifier is trained in the first block (55 decompensations) and tested in the next block (17 decompensations) getting the score for Step 1. Following, in Step 2, the classifier is trained in Step 1 and tested in the new one (19 decompensations), getting the score for Step 2. Repeating the same with Step 3 and Step 4, we get four scores. It is supposed that the first step is the more unstable, as there are less data to train the classifiers, but, while the training set increases, it is believed that the results will become stable, and the score will converge to its real testing value.

The score value used to test the classifiers is the area under the ROC curve (AUC) [37], a measure that evaluates the balance between sensitivity and specificity and that gets an accurate estimation even in moderately imbalanced datasets, which is our case. The AUC value is used to check how well the classifiers perform and consequently select the best one. To test the global predictive model, we use Se and FA/pt-y which are the ones used in the literature.

5.2. Classifiers Comparison. In this section, the results of the classifiers explained in *Applied Machine-Learning Classifiers* are presented applied for the training dataset. Additionally, the rolling cross-validation method, presented above, is applied to avoid the overfitting. This way, the classifier(s) with best outcomes and generalizable (and therefore, stable) can be selected for the predictive model.

In Figure 4, the AUC values of each classifier are illustrated for each of the steps defined in the rolling cross-validation method. The points are the mean of the AUCs achieved in each case, with its standard deviation drawn with whiskers. High standard deviation value indicates that the classifier is less generalizable, while low standard deviation hints a stable classifier.

It is expected that the AUCs values converge as the number of steps grow, although with the available dataset, there are a trend of significative improvement in the second step and a worsening trend in the third one. However, Figure 4 clearly shows that the best classifiers are Naïve Bayes (NB) with Bernoulli method and the random forest (RF). NB classifier has lower AUC value than RF, but the standard derivation is almost negligible, and the trend through the steps is more stable. Hence, it is expected that its performance will not vary significantly over time with new data. RF gets the best scores, but is unstable, and it has high standard derivations. Henceforth, NB with Bernoulli method and RF classifiers are selected to validate the models.

Decision tree and SGD classifiers give the lowest results. The other three classifiers (NB with Gaussian distribution, SVM, and MLP classifiers) perform better, but not as good as the selected two.

5.3. Final Results

5.3.1. Alerts Performance. Since the alerts are used to generate the instances for the machine-learning classifiers (see *Selection of Alerts for Instances Generation*), first, the performance of these in the testing dataset is presented (Table 9).

Comparing these results (Table 9) with the obtained in the training set (Table 5), the weight-associated alerts get worse result. In the case of the questionnaire alerts (Table 10), there is a general worsening comparing with the training set (Table 7). Hence, it is possible to get worse results than the expected when testing the predictive model in the testing set.

5.3.2. Validation Results. In the current medical practice, their alerts all together obtain the following results: Se = 0.76 and FA/pt-y = 28.64 with the red alerts plus the questionnaire alert, and Se = 1 and FA/pt-y = 88.41 with the yellow alerts plus the questionnaire alert.

After applying the R2 alerts in the testing dataset (see *Selection of Alerts for Instances Generation* section), the selected machine learning classifiers achieved the following AUC values: NB with Bernoulli, AUC = 0.67 and RF, AUC = 0.62. As it was expected, NB with Bernoulli maintains the AUC value in accordance with the results obtained in the training dataset, and in the case of RF, due to the classifier instability, the score deteriorates (Figure 4).

Once the classifier is selected and trained, the results are given depending on the probability of the patient to suffer a decompensation. For that, the probability given by the classifier (0 if none of the alerts of the inclusion criterion is triggered) is split in terciles. Each tercile is associated with a colour: if the probability is less than 0.33, "green" group; if the probability is between 0.33 and 0.66, "yellow" group; and if it is upper than 0.66, "red" group. This way, the clinicians can base their decisions on the risk group. Setting the probabilities of the classifiers to the risk groups, the results achieved are the next (Table 11).

As presented in Table 11, the RF classifier results in a poor predictive model. However, NB reaches acceptable scores comparing with literature studies that have similar attributes (see section *Alerts for HF Patients*). Comparing the obtained results in the "red" group to the current medical practice, though NB (in the red group) gets 38% less of Se (0.76 ⟶ 0.47) value, it achieves 72% less of FA/pt-y (28.64 ⟶ 7.8) value. Henceforth, this predictive model improves the results of the actual alerts method and it is more reliable.

6. Discussion

Current medical practice may use sensitive alerts, that although they detect most of the decompensations due to their high sensitivity, they also have too many false alerts. Therefore, the main goal of this study is the reduction of these false alerts. This study has shown an improvement from current alerts system implemented in the hospital. The system reduces the number of false alerts notably, from 28.64 FA/pt-y of the current medical practice to even 7.8 FA/pt-y

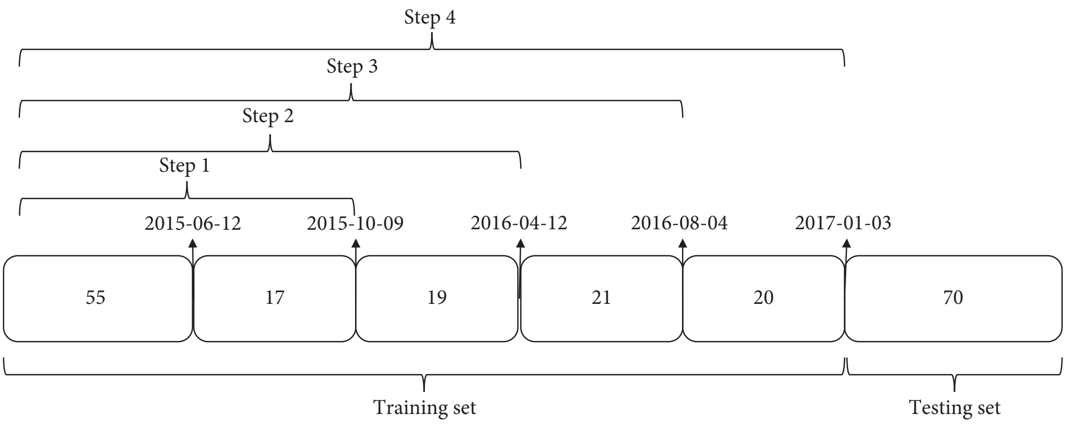


Figure 3: Cross-validation on a rolling basis applied in the study.

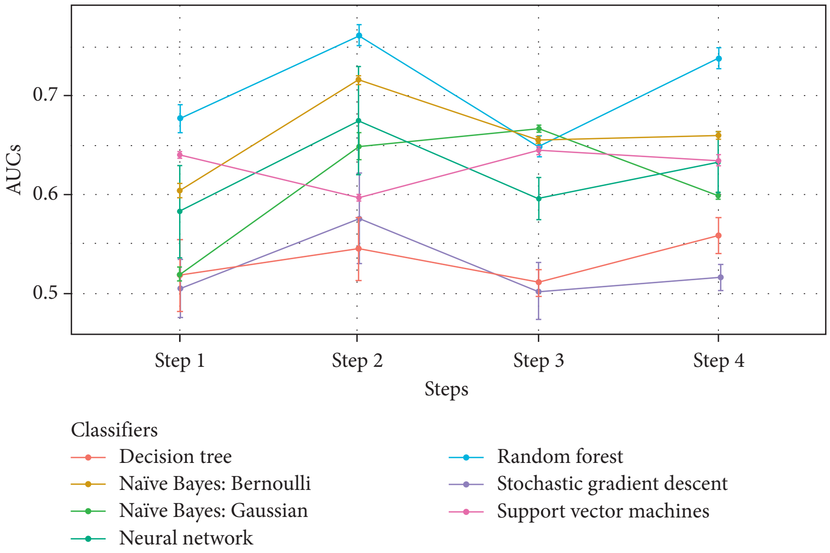


Figure 4: AUC values of the classifiers (colours) depending on the steps (axis x).

Table 9: Alerts' performance in the testing set.

Alert tag	Colour alert	Se	FA/pt-y
Weight	—	0.4	13.36
SBP	Yellow	0.49	18.9
	Red	0.1	2.44
DBP	Yellow	0.34	15.4
	Red	0.07	1.3
HR	Yellow	0.37	19.7
	Red	0.06	3.1
O2Sat	Yellow	0.5	27.27
	Red	0.2	4.2

Table 10: Questionnaire alerts' performance in the testing set.

Question tag	Answer	Se	FA/pt-y
Well-being	Worse	0.25	2.9
New Medication	Yes	0.13	5.2
Diet and exercise	No	0.16	8.75
Ankle	Worse	0.13	2.86
Walks	No	0.4	24.7
Shortness of breath	Yes	0.43	21.9

Table 11: Results of the predictive models.

Group	Random forest		Naïve Bayes	
	Se	FA/pt-y	Se	FA/pt-y
Green	1	79	0.75	59.4
Yellow	0.08	1.29	0.41	13.2
Red	0	0.11	0.47	7.8

for the "red" group, which is denoted as the most restrictive group. This last result is achieved with the predictive model built by applying NB with Bernoulli to the combination of telemonitoring alerts and questionnaire alerts (R2). However, as expected, the application of machine learning techniques entails a decrement on sensitivity values. The result obtained in this study for the "red" group is Se = 0.47, while the alerts used in the current medical practice applied to the same testing dataset achieve Se = 0.76. Despite this Se worsening, it is notorious that the FA/pt-y has much higher decrement, with which we conclude that this new predictive model improves the current medical practice. Moreover, when comparing the obtained results with the state of the art, the Se values are similar or better to these studies that do not consider transthoracic impedance (Table 1). Especially considering that in the SoA, most of the studies reduce the real FA/pt-y concatenating the neighbour alerts, since they assume that once an alert has been triggered, the clinician will take action, and hence, next consecutive alerts will not be triggered.

The current study also presents some limitations. Firstly, as presented in *Characteristics of Ambulatory Patients Dataset*, there are patients that did not monitor regularly. As a consequence, from 254 decompensations during this telemonitored period, only 202 of them had 3 or more measurements during the last week previous to the admission, and hence, could be used in our study. The rest did not have even 3 measurements, and hence, they were not predictable.

Secondly, as the clinical data used in the study are from Caucasian patients, the model may perform differently in different settings, such as in non-Caucasian population. Finally, we must stress that heart failure is a very complex disease with multiple factors, and its predictiveness is complex. Nevertheless, larger amount of data and the registration of all type of decompensations is key to improve the current model.

7. Conclusion and Future Work

This article presents the methodology to develop predictive models for HF decompensations prediction based on ambulatory patients' telemonitored data, extending the study for readmissions detection [9].

The results on these studies have been successfully implemented in a telemedicine system, called INCAR. This way INCAR provides the patient with the confidence of being monitored and guided with an advanced technology and clinical professionals' supervision.

Currently, new devices that monitor physical activity and sleeping quality are incorporated in the telemonitoring program in order to determine whether these features could have an impact in the results and improve the outcome. To finish, we will study the possibility of including in the telemonitoring plan a new device that monitors transthoracic impedance, and explore raising deep learning techniques, which have demonstrated their good performance and may improve the presented results.

Data Availability

The dataset used for this study contains personal information, and therefore, following the Research Data policy of Hindawi, it is not available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been funded by Basque Government by means of the RIS3 Program, under reference No 2017222015.

References

[1] O. Hamdi, M. A. Chalouf, D. Ouattara, and F. Krief, "eHealth: survey on research projects, comparative study of telemonitoring architectures and main issues," *Journal of Network and Computer Applications*, vol. 46, pp. 100–112, 2014.

[2] S. I. Chaudhry, J. A. Mattera, J. P. Curtis et al., "Telemonitoring in patients with heart failure," *New England Journal of Medicine*, vol. 363, no. 24, pp. 2301–2309, 2010.

[3] S. Mangini, P. V. Pires, F. G. M. Braga, and F. Bacal, "Decompensated heart failure," *Einstein*, vol. 11, no. 3, pp. 383–391, 2013.

[4] P. Ponikowski, A. A. Voors, S. D. Anker et al., "2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) developed with the special contribution of the Heart Failure Association (HFA) of the ESC," *European Heart Journal*, vol. 37, no. 27, pp. 2129–2200, 2016.

[5] W. T. Abraham, S. Compton, G. Haas et al., "Intrathoracic impedance vs daily weight monitoring for predicting worsening heart failure events: results of the Fluid Accumulation Status Trial (FAST)," *Congestive Heart Failure*, vol. 17, no. 2, pp. 51–55, 2011.

[6] I. C. Gyllensten, A. G. Bonomi, K. M. Goode et al., "Early indication of decompensated heart failure in patients on home-telemonitoring: a comparison of prediction algorithms based on daily weight and noninvasive transthoracic bio-impedance," *JMIR Medical Informatics*, vol. 4, no. 1, p. e3, 2016.

[7] J. Zhang, K. M. Goode, P. E. Cuddihy, J. G. F. Cleland, and TEN-HMS Investigators, "Predicting hospitalization due to worsening heart failure using daily weight measurement: analysis of the Trans-European Network-Home-Care Management System (TEN-HMS) study," *European Journal of Heart Failure*, vol. 11, no. 4, pp. 420–427, 2009.

[8] M. T. Ledwidge, R. O'Hanlon, L. Lalor et al., "Can individualized weight monitoring using the HeartPhone algorithm improve sensitivity for clinical deterioration of heart failure?," *European Journal of Heart Failure*, vol. 15, no. 4, pp. 447–455, 2013.

[9] J. Kerexeta, A. Artetxe, V. Escolar, A. Lozano, and N. Larburu, "Predicting 30-day readmission in heart failure using machine learning techniques," in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, Funchal, Portugal, January 2018.

[10] P. B. Adamson, M. R. Zile, Y. K. Cho et al., "Hemodynamic factors associated with acute decompensated heart failure: part 2—use in automated detection," *Journal of Cardiac Failure*, vol. 17, no. 5, pp. 366–373, 2011.

[11] F. R. Gilliam III, G. A. Ewald, and R. J. Sweeney, "Feasibility of automated heart failure decompensation detection using remote patient monitoring: results from the decompensation detection study," *Innovations in Cardiac Rhythm Management*, vol. 3, pp. 735–745, 2012.

[12] G. García-Sáez, M. Rigla, I. Martínez-Sarriegui et al., "Patient-oriented computerized clinical guidelines for mobile decision support in gestational diabetes," *Journal of Diabetes Science and Technology*, vol. 8, no. 2, pp. 238–246, 2014.

[13] M. Peleg, Y. Shahar, S. Quaglini et al., "Assessment of a personalized and distributed patient guidance system," *International Journal of Medical Informatics*, vol. 101, pp. 108–130, 2017.

[14] B. J. Mortazavi, N. S. Downing, E. M. Bucholz et al., "Analysis of machine learning techniques for heart failure

readmissions," *Circulation: Cardiovascular Quality and Outcomes*, vol. 9, no. 6, pp. 629–640, 2016.

[15] K. Zolfaghar, "Predicting risk-of-readmission for congestive heart failure patients: a multi-layer approach," 2013, http://arxiv.org/abs/1306.2094.

[16] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki, "Predictive modeling of hospital readmissions using metaheuristics and data mining," *Expert Systems with Applications*, vol. 42, no. 20, pp. 7110–7120, 2015.

[17] N. Meadem, N. Verbiest, K. Zolfaghar et al., "Exploring preprocessing techniques for prediction of risk of readmission for congestive heart failure patients," in *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD), Data Mining and Healthcare (DMH)*, vol. 150, Chicago, IL, USA, August 2013.

[18] H. M. Krumholz, Y.-T. Chen, Y. Wang, V. Vaccarino, M. J. Radford, and R. I. Horwitz, "Predictors of readmission among elderly survivors of admission with heart failure," *American Heart Journal*, vol. 139, no. 1, pp. 72–77, 2000.

[19] R. Amarasingham, B. J. Moore, Y. P. Tabak et al., "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data," *Medical Care*, vol. 48, no. 11, pp. 981–988, 2010.

[20] S. Sudhakar, W. Zhang, Y.-F. Kuo, M. Alghrouz, A. Barbajelata, and G. Sharma, "Validation of the readmission risk score in heart failure patients at a tertiary hospital," *Journal of Cardiac Failure*, vol. 21, no. 11, pp. 885–891, 2015.

[21] M. van der Heijden, B. Lijnse, P. J. F. Lucas, Y. F. Heijdra, and T. R. J. Schermer, "Managing COPD exacerbations with telemedicine," in *Artificial Intelligence in Medicine*, pp. 169–178, Westview Press, Boulder, CO, USA, 2011.

[22] M. S. Mohktar, *A Decision Support System for the Home Management of Patients with Chronic Obstructive Pulmonary Disease (COPD) Using Telehealth*, Graduate School of Biomedical Engineering, University of New South Wales, Kensington, NSW, Australia, 2012.

[23] M. Velikova, P. J. F. Lucas, and M. Spaanderman, "A predictive bayesian network model for home management of preeclampsia," in *Artificial Intelligence in Medicine*, pp. 179–183, Westview Press, Boulder, CO, USA, 2011.

[24] R. Lafta, J. Zhang, X. Tao et al., "An intelligent recommender system based on predictive analysis in telehealthcare environment," *Web Intelligence*, vol. 14, no. 4, pp. 325–336, 2017.

[25] M. T. Alarcón and J. I. González-Montalvo, "La escala sociofamiliar de Gijón, instrumento útil en el hospital general," *Revista Española de Geriatría y Gerontología*, vol. 33, no. 1, pp. 178-179, 1998.

[26] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[27] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, Burlington, MA, USA, 1993.

[28] L. Breiman, R. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees 1984*, Wadsworth and Brooks, Monterey, CA, USA, 1984.

[29] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[32] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

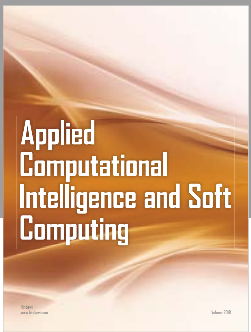
[33] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[34] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, 2013.

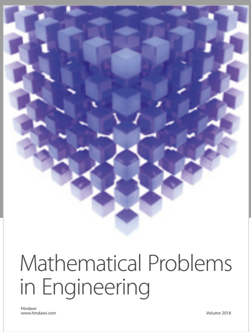
[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[36] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012.

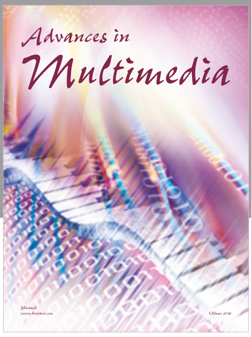
[37] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

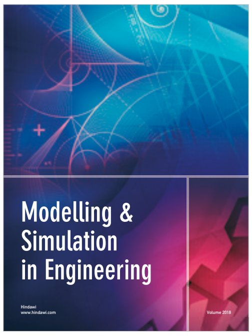

Applied Computational Intelligence and Soft Computing

The Scientific World Journal


Mathematical Problems in Engineering

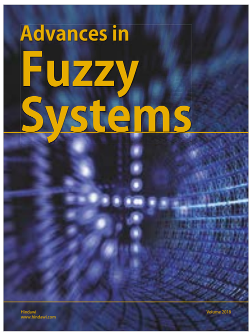
Journal of Engineering

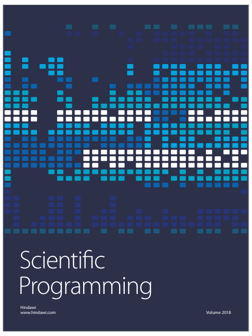

Advances in Multimedia

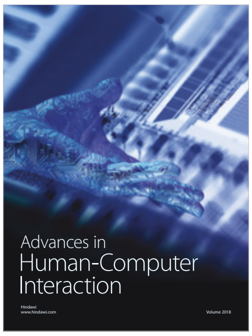

Modelling & Simulation in Engineering

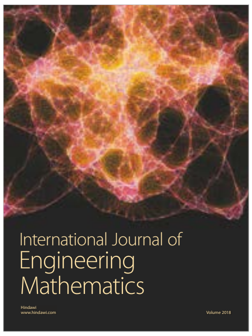
Advances in Artificial Intelligence

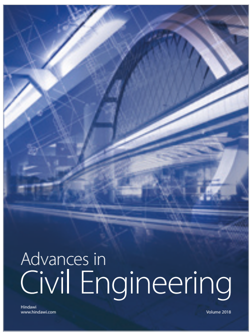
International Journal of Reconfigurable Computing

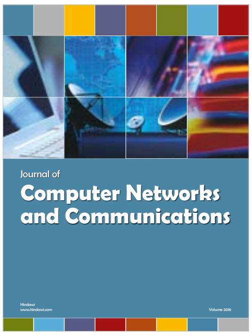

Advances in Fuzzy Systems


Scientific Programming


Advances in Human-Computer Interaction

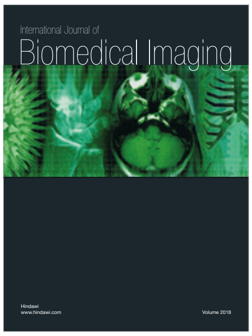

International Journal of Engineering Mathematics


Advances in Civil Engineering

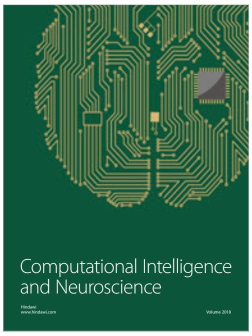

Journal of Computer Networks and Communications

Journal of Robotics

International Journal of Computer Games Technology


International Journal of Biomedical Imaging

Journal of Electrical and Computer Engineering


Computational Intelligence and Neuroscience