Performance Measures Evaluation for Highly Imbalanced Datasets: Application to Short Term Renal Failure Prediction

Ricardo Alexandre Silva Maia
ricardo.a.maia@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

June 2018

Abstract

Kidney failure is a serious health problem, often leading to fatalities. Early detection is essential, with urine production being one of the first warnings. Nevertheless, clinical data is characterized by imbalanced datasets, being the event to detect scarce. In this type of problems, where it is necessary to develop a robust classification model and choose the most relevant features, performance measures are essential. Typically, the AUC is used, although in situations of imbalance it favors the majority class. In this work, the influence of different performance measures on classifier development was studied, resorting to imbalanced benchmark datasets. The measures compared were AUC, AUK and F-Score, and it was concluded that F-Score presents better performance in feature selection, regardless of the imbalance. Regarding the definition of the classifier threshold, two conclusions were drawn: to maximize the sensitivity, the F-Score should be used; for a better trade-off between sensitivity and specificity, the AUK should be chosen. Relative to AUC, it was verified to be unsuitable for imbalanced data. The same methodology was applied to the real clinical case to predict urine outputs below 30 ml/h. To accomplish this, fuzzy models were developed using Fuzzy C-Means and Possibilistic Fuzzy C-Means (PFCM), along with logistic regression models. The study was extended to low and high sampling rate data. Overall, higher performance was observed using fuzzy models with PFCM, obtaining the best model using data of low sampling rate in the detection of the first critical event.

Keywords: Fuzzy Modeling, Urine Output, Feature Selection, Imbalanced Datasets, Performance Measures

1. Introduction

Acute Kidney Injury (AKI) is a concerning clinical situation on a global scale that can be described by an abrupt decrease in kidney function, often associated with the development of chronic kidney disease, as well as severe morbidity or mortality [14]. Although harmful, the early detection and treatment of this condition can reduce long-term consequences and prevent death. Moreover it mainly affects critically ill patients in intensive care, as well as the elderly and diabetic individuals. According to the International Society of Nephrology (ISN), nearly 13.3 million cases of AKI are identified per year, resulting in 1.7 million deaths, which highlights the need to develop and implement strategies to counteract its growth.

Currently, electronic records provide complete and up-to-date patient's information, promoting clinical interoperability and availability of data [12]. This, coupled with the high frequency with which values are recorded in the intensive care unit (ICU), provides a huge window of opportunity.

However, with the emergence of data, it is unfeasible to analyze it quickly and only with medical knowledge. To deal with this challenge, data mining has increasingly proved to be a suitable solution, allowing to analyze large datasets to find hidden relationships between features [17, 10].

Perhaps one of the most important applications is in prediction of short and long-term outcomes, working as a complementary tool to clinicians. For example, in [11], a classification system was developed using neural networks to predict hypertension by analyzing health conditions and medical records.

Over the past years, machine learning algorithms have been developed to explore problems related to AKI. In [2], Takagi-Sugeno models were developed using both Fuzzy C-Means (FCM) and Gustafson-Kessel (GK) clustering algorithms to predict short and long-term mortality in patients who have been diagnosed with AKI at ICU's admission. Furthermore, in [15], classifiers based on GK and support vector machines (SVM) were implemented using demographic data, comorbidities indexes and laboratory measures to predict urine output rates below 30 mL/h, acting as an indicator for AKI's diagnosis.

On the other hand, the development of a classifier is guided by a constant evaluation of its quality when facing new data [5]. Naturally, using countless variables

does not guarantee good outcomes. In fact, a feature selection should be applied, not only to reduce complexity, but also to maximize predictive ability. Within the wrapper methodology, performance measures are the engine to find the optimal subset of variables. The fundamental question is how to measure the classifier's performance.

Most of the score metrics are single scalar values that rely upon a threshold to distinguish classes. However, real-world problems are characterized by high imbalance, particularly when it comes to the clinical field. Dealing with such imbalance can be a sizable problem to classifiers. To overcome the bias towards the majority class, a more suitable choice of threshold can be made.

Focusing on just one class may lead to poor outcomes. In this sense, the area under the receiver operating characteristic curve (AUC) can be used, allowing to assess the overall quality of the model based on different thresholds [9]. However, David Hand stated that using AUC is equivalent to average the misclassification loss over an implicitly score-dependent distribution [4]. In addition, AUC is not suitable to handle skewness.

Hence, Kaymak et. al. [9] suggested a new performance measure based on Cohen's Kappa (AUK). This coefficient, proposed by Jacob Cohen, is a statistically robust measure that represents the proportion of agreement after disregarding chance agreement, favoring correct classifications of the minority class. In [20] it was used as objective function for feature selection, providing better outcomes than accuracy.

Although sensitivity and precision do not provide an overall view of the model, F-Score offers a different perspective that merges true and false positives with false negatives, thus favoring correct classifications of the positive class.

Therefore, this work aims to develop a classifier to predict short-term renal failures, by predicting urine outputs below 30 ml/h, using low and high frequency data and following the work done in [15]. Furthermore, to establish the groundwork for the evaluation of binary classifiers when facing imbalance, AUC, AUK and F-Score are studied either as objective functions in wrapper feature selection and to obtain the model's threshold.

2. Background

2.1. Logistic Regression

Logistic regression is a classical statistical modeling tool developed by David Cox [1]. This method aims to describe the relationship between the outcome and the categorical/continuous predictor variables, in which the output is a dichotomous variable. Hence, it is possible to predict the probability of occurrence of an event p by fitting data from input variables x through a logistic function. Therefore, the logistic regression equation is given by the natural logarithm of the odds (logit), which is a linear function of the predictors given by:

$$logit(p) = ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad (1)$$

where β_i is the regression coefficient for each of the input variables, while β_0 represents the value of the logit when the predictors are zero.

Having defined the values of the coefficients, it is possible to rewrite the regression equation in terms of the probability of occurrence of the event as:

$$p(x) = \frac{e^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}} \qquad (2)$$

2.2. Fuzzy Clustering

Clustering algorithms allow the discovery of hidden structures in a dataset constituted by, at first sight, unrelated elements where few prior knowledge is available [19]. The main goal of clustering is to separate data objects into distinct groups according to their degree of similarity.

2.2.1. Fuzzy C-Means

In Fuzzy C-Mean (FCM) each data point belongs to a group according to a certain degree specified by a membership grade. According to [6], this clustering algorithm attempts to partition the dataset $X = [x_1, \ldots, x_N]^T$ into n_c clusters, finding the centroid of each cluster such that a cost function of dissimilarity measure can be minimized.

The FCM algorithm focus in the minimization of the cost function J given by:

$$J = \sum_{i=1}^{n_c}\sum_{j=1}^{N} \mu_{ij}^m d_{ij}^2(x_j, v_i) \qquad (3)$$

where $1 \leq i \leq n_c$ represents the cluster index, n_c the number of clusters, N stands for the total number of samples, μ_{ij} represents the membership degree between sample x_j and i^{th} cluster, x_j is the j^{th} of N-dimensional measured data, v_i represents the center of the i^{th} cluster (prototype), d_{ij}^2 is the distance measure between data points and prototypes and $1 \leq m \leq \infty$ is the fuzziness parameter, which gives the degree of overlapping between clusters.

The partition matrix U gives all the membership degrees for every sample x_j:

$$U = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1N} \\ \vdots & \ddots & \vdots \\ \mu_{n_c 1} & \cdots & \mu_{n_c N} \end{bmatrix} \qquad (4)$$

However, in order to impose normalization, for each sample the sum of the membership degrees relative to every cluster must be equal to one.

$$\sum_{i=1}^{n_c} \mu_{ij} = 1, \quad \forall j \qquad (5)$$

Having defined the fuzzy parameter and the number of clusters, the partition matrix U is initialized, using the Euclidean distance to measure the similarity between data points, which is defined as:

$$d_{ij}^2(x_j, v_i) = (x_j - v_i)^T (x_j - v_i) \qquad (6)$$

Through successive iterations, the cluster centers are

computed as

$$v_i = \frac{\sum_{j=1}^{N} \mu_{ij}^m x_j}{\sum_{j=1}^{N} \mu_{ij}^m} \tag{7}$$

Having calculated the distances and prototypes, the values of the partition matrix are updated as

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{n_c} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \tag{8}$$

The algorithm stops if either one of two condition are met: if either the cost function improvement regarding the last iteration is less than a specified tolerance or the maximum number of iteration is reached.

2.2.2. Possibilistic Fuzzy C-Means

The Possibilistic Fuzzy C-Means (PFCM) is a clustering algorithm based on FCM, proposed by Nikhil Pal, Kuhu Pal, Keller and Bezdek [16], which aims to quantify the outliers influence in clusters. To do so, the concept of typicality degree is introduced, which states that each cluster should benefit the common features of data points, as well as weighing the absolute distance from each point to the cluster center. Therefore, to enhance the compactness and separability of clusters, points with lower value of typicality (outliers) must have a lower influence in the definition of centroid.

The PFCM focuses on the minimization of the cost function given as:

$$\begin{aligned} J_{m,\eta}(U,T,V,X) = \sum_{j=1}^{N}\sum_{i=1}^{n_c}(a\mu_{ij}^m + bt_{ij}^\eta)\times \\ \times \left\|x_j - v_i\right\|^2 + \sum_{i=1}^{n_c}\gamma_i \sum_{j=1}^{N}(1-t_{ij})^\eta \end{aligned} \tag{9}$$

where $a > 0$ represents the relative importance of fuzzy membership, $b > 0$ represents the weight of the typicality values in the objective function, $m > 1$ is the fuzziness parameter, $\eta > 1$ is the typicality matrix coefficient, c is the desired number of cluster, N stands for the total number of data samples, μ_{ij} represents the membership degree between sample x_j and i^{th} cluster, x_j is the j^{th} of N-dimensional measured data, v_i represents the center of the cluster (prototype), t_{ij} is the typicality matrix between sample x_j and i^{th} cluster and $\gamma_i > 0$ are user defined constants.

The fuzzy partition matrix has the exact same meaning of membership as described previously for FCM. In turn, γ_i values are weight coefficients that influence the typicality matrix and should be computed as:

$$\gamma_i = K\frac{\sum_{j=1}^{N} \mu_{ij}^m d_{ij}^2}{\sum_{j=1}^{N} \mu_{ij}^m}, \quad K > 0 \tag{10}$$

where the most common choice is $K = 1$. In addition, u_{ij} are the entries of the terminal FCM partition matrix for the specific dataset, while the distance matrix is computed using the same terminal FCM cluster centers.

Similarly, the distance matrix is given by the Euclidean distance between data object and cluster centers. On the other hand, the typicality matrix elements can be computing recurring to equation 11.

$$t_{ij} = \left[1 + \left(\frac{b}{\gamma_i} d_{ij}^2\right)^{1/(\eta-1)}\right]^{-1} \tag{11}$$

Finally, the cluster centers are updated according to the new fuzzy partition and typicality matrices as:

$$v_i = \frac{\sum_{j=1}^{N}(a\mu_{ij}^m + bt_{ij}^\eta)x_j}{\sum_{j=1}^{N}(a\mu_{ij}^m + bt_{ij}^\eta)} \tag{12}$$

2.3. Fuzzy Modeling

Most real-world processes are nonlinear. In those cases it is advantageous to use fuzzy logic to infer nonlinear relationships between inputs and outputs by means of if-then rules and logical connectives, allowing the development of systems that can handle uncertainty within the data [18, 6].

2.3.1. Takagi-Sugeno Fuzzy Models

This work uses Takagi-Sugeno (TS) fuzzy models, in which each rule describes a local input-output relation, and the consequent part of the rules is represented by a mathematical function instead of a fuzzy set, typically in an affine form [13]. Those are usually given by

$$\begin{aligned} R_i: \text{ If } x_1 \text{ is } A_{i1} \ldots \text{ and if } x_n \text{ is } A_{in} \text{ then} \\ y_i = a_{i1}x_1 + \ldots + a_{in}x_n + b, \quad i = 1, \ldots, Z \end{aligned} \tag{13}$$

where f_i is the consequent function of rule R_i, n is the total number of features, Z represents the total number of rules, x is the antecedent vector, A_{in} is the fuzzy set of the n^{th} part of the antecedent, y_i is the output value for the i^{th} rule, a_{in} is the parameter associated with feature n and, finally, b_i is a scalar offset.

In the case of a TS model, the firing strength for the i^{th} rule, θ_i, is given by the product of the membership values for the fuzzy set A_{in} in the antecedent part, $\mu_{A_{in}}$, as described in Equation 14.

$$\theta_i = \prod_{m=1}^{n} \mu_{A_{in}}(x) \tag{14}$$

Thus, resorting to Equation 15, the final model output can be computed through the weighted average of the individual outputs obtained by each rule

$$y = \frac{\sum_{i=1}^{K} \theta_i y_i}{\sum_{i=1}^{K} \theta_i} \tag{15}$$

where K denotes the number of rules.

2.4. Feature Selection

Real-world datasets are often characterized by high dimensionality, which may present redundant or even irrelevant features. Nonetheless, removing variables that appear to be useless by themselves can remove distinctive characteristics of the data. Therefore, feature selection aims to select a subset of relevant features for building robust models with generalization capability. Thus, it allows to counter the curse of dimensionality, as well as promotes a better cluster detection [20].

In this study a wrapper greedy search algorithm was

implemented, following the Sequential Forward Selection (SFS) approach. SFS is a greedy iterative process that progressively evaluates the performance of the model using different combinations of features. It starts by evaluating each of the features separately and, based on a specific criterion, the one that returns the best performance score is selected. Then, this feature is combined with each one of the remaining, proceeding to the development and evaluation of the model to find the subset that maximizes its performance, and so on. In this work, it was decided to stop the process only when all features were selected, avoiding sub-optimal solutions.

2.5. Performance Measures

In this section, a brief description of the score metrics used to assess the model's performance is presented.

2.5.1. Accuracy

Accuracy represents the proportion of true positive outcomes in the entire dataset, given by

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

Although it is easy to interpret, this metric does not account for the misclassification costs of the minority class.

2.5.2. Precision

Precision represents the proportion of outputs correctly classified as positive in the total set of samples classified as being positive.

$$Precision = \frac{TP}{TP+FP} \tag{17}$$

2.5.3. Sensitivity

Sensitivity, also know as recall, is the measure of correct classifications of the positive class among all the positive population, and is given by:

$$Sensitivity = \frac{TP}{TP+FN} \tag{18}$$

2.5.4. Specificity

The specificity, or true negative rate, measures the proportion of negative outcomes that truly belongs to the negative class:

$$Specificity = \frac{TN}{TN+FP} \tag{19}$$

where TP corresponds to positive outputs correctly classified, FN to positive outputs classified as negative, TN to negative outputs classified correctly and lastly, FP represents the negative outputs classified as positive.

2.6. F-Score

F-Score represents the harmonic average of precision and sensitivity, in which a higher value means a better performance on the positive class, being suitable for imbalanced data. This index is given by:

$$F_\gamma = (1+\gamma^2) \times \frac{precision \times recall}{(\gamma^2 \times precision) + recall} \tag{20}$$

where γ represents the relative importance given to precision/sensitivity. Usually, $\gamma = 1$ is the most common case, as used in this work. Henceforth, the F_1-Score will be referred to as F-Score.

However, in this work, a different approach is taken for the F-Score. Instead of using a predefined threshold, a free choice is implemented (similar to what happens for AUC and AUK) in order to set the threshold value that maximizes this performance measure.

2.6.1. Area Under the ROC Curve

The Receiver Operating Characteristic (ROC) curve is a commonly used technique for visualizing, organizing and selecting classifiers based on their performance by plotting the false positive rate against the true positive rate, while changing the threshold [3].

In order to compare different classifiers, a ROC curve must be reduced to a single scalar value. To achieve this, the area under the curve (AUC) is computed using a trapezoidal numerical integration. The higher the AUC value, the better the overall performance of the classifier.

Since each point in the curve represents a threshold value, it is then possible to determine a decision threshold that minimizes misclassications. Usually, this value is chosen by maximizing the TPR, while the FPR remains as lower as possible.

2.6.2. Area Under the Cohen's Kappa Curve

Cohen's kappa is a statistical measure of inter-rater agreement that can be used to assess classifiers' performance. It gives more emphasis to the minority class, improving sensitivity without disregarding specificity. Thus, Cohen's kappa can be defined as

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{21}$$

where $Pr(a)$ equals accuracy and $Pr(e)$ represents the hypothetical probability of chance agreement. In turn, to define $Pr(e)$ it is essential to clarify its components:

$$p_+ = \frac{TP+FN}{TP+TN+FP+FN} \tag{22}$$

$$p_- = \frac{FP+TN}{TP+TN+FP+FN} \tag{23}$$

$$\hat{p_+} = \frac{TP+FP}{TP+TN+FP+FN} \tag{24}$$

$$\hat{p_-} = \frac{TN+FN}{TP+TN+FP+FN} \tag{25}$$

where p_+ and p_- represent the percentage of positive and negative classes, while $\hat{p_+}$ and $\hat{p_-}$ stands for the percentage of data classified as positive and negative, respectively. Therefore, the hypothetical probability of chance agreement is computed as:

$$Pr(e) = p_+\hat{p_+} + p_-\hat{p_-} \tag{26}$$

Kappa can assume values between -1 and 1, with higher values indicating a good prediction. By plotting the values of κ against the false positive rates for different threshold, and computing the area under the

curve (equation 27), it is possible to reduce it to a single scalar that, similarly to AUC, measures the overall performance: the AUK.

$$AUK = \int_0^1 \kappa(KPR) dFPR \tag{27}$$

The optimal threshold to increase the performance of the classifier coincides with the maximum of the Kappa curve [9].

2.7. Data Reduction

One of the main problems of real-world databases is the imbalance of classes, leading to unsatisfactory model outcomes as a result of poor generalization. Therefore, to smooth the effects of highly skewed datasets, undersampling strategies can be employed, resulting in a more balanced class distribution.

2.7.1. Edited Nearest Neighbor (ENN)

ENN removes instances of the majority class when most of its K nearest neighbors belongs to the other class. This method allows to remove noisy samples and close border class cases, achieving a smoother decision.

2.7.2. Neighborhood Cleaning Rule (NCR)

NCR allows to improve the ENN method by combining data deletion with data cleaning. Therefore, this method starts by applying the ENN to identify noisy data, removing majority class samples if at least two of the three nearest neighbors belong to the minority class. Then, for each minority class sample, the three nearest neighbors are founded and those that misclassify it and belong to the majority class are also removed.

3. Benchmark Datasets

The choice of the performance measure that best fits a binary classifier is influenced by the level of balancing between classes. To study this issue, particular attention was given to AUC, AUK and F-Score, applied to several binary benchmark datasets. All the datasets refer to real cases and were taken from the UCI machine learning repository.

In the present study, 11 datasets from distinctive fields were used. Henceforth, BCWD stands for Breast Cancer Wisconsin Diagnostic, while BCWO represents the original version of the same dataset. On the other hand, Indian Liver Patient Dataset will hereafter be referred to as ILPD, the Liver Disorders as BUPA, while the South African Heart will be addressed as SAHeart. In Table 1 are presented the main characteristics of the datasets.

In order to analyze different percentages of imbalance, the initial distribution of each dataset was used, as well as proportions of 10% (high imbalance) and 50% (balanced data) for the positive class. To that end, samples of the minority and majority classes, respectively, were randomly removed.

Table 1: Benchmark datasets characteristics.

Dataset	Samples	Features	Positive Class [%]
Australian	690	14	44.48
BCWD	569	30	37.26
BCWO	683	9	34.99
SAHeart	462	9	34.63
ILPD	579	10	28.50
Ionosphere	351	23	35.90
Liver Disorders	345	6	42.03
Mammographic	830	5	48.55
Phoneme	5404	5	29.35
PIMA	768	8	34.90
Statlog Heart	270	13	44.44

4. Dataset Construction and Preprocessing

The data used in this work were collected from the Medical Information Mart for Intensive Care (MIMIC) III, which is a large and free health-related database composed by data collected from patients that were admitted in critical care units. Henceforth, the data extracted from this dataset will be referred as low sampling rate data.

In addition, there is a complementary category of this dataset, the MIMIC III Waveform Database Matched Subset, which contains records of physiologic signals and time series of vital signs collected with a high frequency of sampling. These records have been matched and time-aligned with some of the MIMIC III patients [7].

4.1. Low Sampling Rate Data

Following the previous work [8], the clinically approved and selected variables are as follows:

- Mean Arterial Blood Pressure (MAP) (f_1);
- Urine Output (UO) (f_2);
- Potassium (f_3);
- MAP 3 hours before (f_4);
- Systolic blood pressure difference 6h before (f_5);
- Oasis Score (f_6);
- Oasis Respiration Rate Score (ORRS) (f_7);
- Elixhauser Fluid Electrolyte Score (EFES) (f_8).

In this work, the 1064 patients that form the final cohort are also those that were selected in [8], since these are the ones that allow to minimize the number of missing values for each feature.

4.1.1. Critical Urine Output

The model to be developed aims to predict the critical urine output level, an event characterized by an urine output lower than 30ml/h. This can be seen as a binary classification problem whose objective is to predict, for each time sample, if the patient will have a critical drop in urine's production in the next collection. Therefore,

the classifier can be presented as follows:

$$y = \begin{cases} 0 & \text{if UO above 30 ml/h} \\ 1 & \text{otherwise} \end{cases}$$

4.1.2. Data Imputation Methodology

When dealing with different time-series it is frequent to find variables with uneven sampling times. In this dataset, urine output or blood pressure variables are measured in a hourly basis, whereas laboratory measurements (e.g. potassium, creatinine) are measured once a day.

In order to align all the features used, and since the objective is to predict critical urine outputs, UO will be used as time template to input the dynamic features (i.e. potassium and those related with blood pressure). To that end, records in which urine output values were null or duplicated were excluded.

Furthermore, to perform data imputation, the last observation carried forward (LOCF) method can be applied, imputing values using the last value known of a specific variable, similarly to what physicians do. However, bearing in mind that the features are time-series, two methods were compared with LOCF: linear and cubic interpolation.

The results obtained showed that linear interpolation for MAP and SAP achieved better performance, while for the potassium the LOCF method resulted in a lower average error. Therefore, the variables $f1$, $f4$ and $f5$ were imputed using a linear interpolation, whereas in the case of potassium the LOCF method was applied.

4.1.3. Analysis of Outliers

In order to exclude outliers, three approaches are applied, namely: clinical knowledge combined with visual inspection, interquartile range (IQR) method and Tukey's method. These methodologies were applied to all the dynamic variables.

Table 2 shows the limits applied, and the consequent percentage of data that is considered outliers for each of the three methods. As can be observed, both interquatile and Tukeys methods are quite conservative, classifying points that are within a plausible range as outliers. By removing data that could be useful to differentiate patients, the performance of the classifier can be seriously impaired. Therefore, the clinical knowledge/visual inspection was selected as the method to remove outliers.

Table 2: Removal of outliers using three methodologies

	Visual/Clinical Knowledge			Interquartile Method			Tukey's Method		
Feature	Min	Max	% Outliers	Min	Max	% Outliers	Min	Max	% Outliers
f_1	0.01	300	0.05	37	125	1.25	4	158	0.20
f_2	0	3000	0.94	-108.55	300.92	6.18	-262.11	454.47	2.13
f_3	0.5	14	0.02	2.55	5.35	1.07	1.5	6.4	0.07
f_4	0.01	300	0.05	37	125	1.25	4	158	0.20
f_5	-	-	-	-8	8	2.67	-14	14	0.19

4.1.4. Inclusion of Vasopressor Intakes

Vasopressors are usually used in the ICU to treat hypotension. By acting on cardiac contractility and heart rate, vasopressors promotes a higher blood pressure, preventing that MAP values decrease below the recommended value of 65 mmHg. Furthermore, the effects of this medication will damage the relationship between arterial blood pressure and urine output, which can affect the classifier. Therefore, it was decided to join this information to the model development stage.

For each instant of time associated with the urine output, the information regarding the vasopressors follows a predefined guideline, given as:

$$\text{Vasopressors} = \begin{cases} 0 & \text{if there is no administration} \\ 1 & \text{if time } i \text{ - time}_p < 6\text{h} \\ 2 & \text{otherwise} \end{cases}$$

where i is the ith time associated with each urine output and time$_p$ represents the time of the last vasopressor administration.

4.1.5. Prediction of the First Critical Event

The uncertain clinical evolution of a patient may lead to complications such as oliguria, which should be immediately addressed and closely monitored, allowing to prevent future consequences. From the medical point of view, the first drastic drop in the amount of urine to levels below 30 ml/h is the most important to detect, promoting an immediate action on the patient. In this way, a second approach to the problem is proposed, which only considers data until the end of the first critical event, allowing to map only the initial conditions that lead to the appearance of the first renal failure.

4.2. High Sampling Rate Data

High sampling rate data (HSRD) is characterized by a vast collection of physiologic signals and time-series of vital signs, which allows a deepen understanding of the clinical condition of a patient. The first step of data extraction consisted in mapping the MIMIC III Waveform Database Matched Subset records with the cohort of 1064 patients used in low frequency data, which resulted in 259 patients. It was possible to track 50 different variables although, given the purpose of the ongoing work, only 6 were selected due to their clinical relevance, namely:

- Heart Rate (X_1);
- Mean Arterial Blood Pressure (MAP) (X_2);
- Systolic Arterial Blood Pressure (SAP) (X_3);
- Diastolic Arterial Blood Pressure (DAP) (X_4);
- Respiration Rate (X_5);
- SpO_2 (X_6).

In order to obtain the final dataset, the patients were excluded if any of the following assumptions were verified: (1) there was no information for at least one of the variables mentioned above; (2) data collection starts after the last record in low sampling rate data; (3) the ICU

admission identifier does not match with its correspondent in the low sampling rate data. From these criteria, a final cohort of 126 patients was obtained.

4.2.1. Analysis of Outliers

Preliminary analysis of the data is essential to ascertain the existence of noisy and irrelevant samples. Similar to the approach for the low sampling rate data, three methods to deal with outliers were applied, namely: clinical knowledge/visual inspection, IQR method and Tukeys method. Table 3 shows the limits applied, as well as the percentage of data that is considered an outlier. Due to the excessive reduction of samples, the clinical knowledge/visual inspection was selected to deal with the outliers.

Table 3: Removal of outliers using three methodologies

	Visual/Expert Knowledge			Interquartile Method			Tukey's Method		
Feature	Min	Max	% Outliers	Min	Max	% Outliers	Min	Max	% Outliers
X_1	0.01	250	0	40.5	124.5	1.58	9	156	0.045
X_2	0.01	300	0.005	41.5	125.5	1.45	10	157	0.10
X_3	0.01	350	0.001	66	186	1.42	21	231	0.063
X_4	0.01	300	0.008	28.5	96.5	2.01	3	122	0.40
X_5	0.01	-	0	5.5	33.5	1.83	-5	44	0.27
X_6	0.01	100	0	90	106	2.71	84	112	1.51

4.2.2. Final Set of Features

For the cohort constituted by 126 patients, two types of sampling time were identified: 1 second and 1 minute. Using one second as sampling time results in a huge amount of data without a substantial increase on useful information, since the samples would become considerably similar. Such increase in dimensionality led to choosing one minute as the sampling time.

One the other hand, both hypertension and administration of vasopressors alter blood pressure values, which may hide the true physiological condition of the patient and, consequently, the development of AKI. In turn, age is also associated with a change in the clinical conditions of reference, as well as a greater susceptibility to diseases. Therefore, in addition to the features chosen from the high sampling rate dataset, it is interesting to analyze the influence that age, sepsis, history of hypertension or even the administration of vasopressors may have on the distinction between patients with and without acute renal failure development. Moreover, the complete list of features to be used is shown below:

- MAP
- SAP
- DAP
- Oasis Score
- Vasopressors
- HR
- RR
- SpO_2
- Hypertension
- EFES
- ORRS
- Sepsis
- Age

5. Results

5.1. Benchmark Datasets

5.1.1. Data Preprocessing and Model Assessment

The first step to deal with the benchmark datasets was the removal of all the missing values. Then, in order to analyze different percentages of the positive class, the initial distribution of each dataset was used, as well as proportions of 10% and 50% for the positive class. To that end, samples referring to the minority and majority classes, respectively, were randomly removed.

To divide the datasets, 5-fold cross validation was applied, aiming to maintain the proportion between both classes equal to the original.

To fix the parameters associated with the fuzzy modeling, a grid search was implemented for each dataset considering all the distributions of the positive class used. Finally, to evaluate the model performance, AUC, AUK and F-Score were used.

5.1.2. Feature Selection

To perform feature selection, both logistic regression and FM-FCM were applied as modeling strategies, using the SFS method together with 5 fold cross validation. For both models, the AUC, AUK and F-Score were used as objective function to all the benchmark datasets, being the entire process repeated 10 times. Lastly, to assess the best set of features for each of the objective functions, the Cohens Kappa and the Matthews Correlation Coefficient (MCC) were used as performance measures.

The results obtained allow to conclude that, for the majority of the cases, the best set of features was obtained using the F-Score as objective function, regardless the proportion of the positive class. On the other hand, when the data distribution is similar, AUC and AUK tend to evidence the same subset of features.

Nevertheless, the major conclusion is that the F-Score is a suitable and useful measure to deal with feature selection, highlighting the features with higher generalization capability.

5.1.3. Selection of the Optimal Threshold

Currently, one of the fundamental problems in classification models is the choice of the threshold that maximizes the performance of the model. To study this issue, the performance of the models after feature selection was evaluated using accuracy, sensitivity, specificity and precision in order to quantify the impact of each threshold. Additionally, 5×5 fold cross validation was applied.

According to the results of the study, some conclusions can be drawn in accordance with the percentage of samples of the minority class, namely:

- The optimal threshold obtained from AUC proves to be unsuitable when dealing with highly skewed datasets, evidenced by a low value of the sensitivity when compared to the specificity;
- The threshold obtained from AUK frequently promotes a more balanced model between sensitivity, specificity and precision;
- The F-Score allows to maximize the model sensitivity, revealing a high suitability when dealing

with highly imbalanced datasets.

5.2. Low Sampling Rate Data - All Critical Events

5.2.1. Division of Data

The division of the data was performed using 5-fold cross validation. Further, patients were organized according to vasopressor intakes, Oasis Respiration Rate Score and Elixhauser Fluid Electrolyte Score. Thus, in addition to guaranteeing a similar distribution of the critical event output, the data variance between each fold is minimized.

5.2.2. Parameter Selection

To define the fuzzy parameters, a grid search was performed varying the number of clusters between 2 and 8, and the fuzzy parameter between 1.1 and 3.5, with increments of 0.1. Undersampling strategies were implemented along with the original dataset.

The number of clusters was set at 4 and the fuzzy parameter at 2.9, whereas the ENN method with k = 2 was chosen to perform undersampling.

Further, to set the parameters for the PFCM algorithm, a grid search was again employed, using a range of values for the parameters a, b and η of [1,3], [1,5] and [1,4], respectively, with increments of 0.5.

5.2.3. Model Assessment

Three schemes were used to develop models: (1) fuzzy modeling using FCM (FM-FCM); (2) fuzzy modeling using PFCM (FM-PFCM); (3) logistic regression models. Their performance was evaluate in terms of AUC, AUK, and F-Score, and the results obtained are presented in Table 4. Although the difference between

Table 4: Classifier performance after 10×5 fold cross validation using FM-FCM, FM-PFCM and Logistic Regression.

Methods	AUC [%]	AUK [%]	F-Score [%]
FM-FCM	83.85 ± 0.83	19.82 ± 0.61	52.90 ± 1.94
FM-PFCM	84.82 ± 0.92	20.28 ± 0.63	53.60 ± 1.69
LR	85.00 ± 0.96	20.26 ± 0.74	52.90 ± 1.89

FM-PFCM and logistic regression is minimal, the former presents better results for AUK and F-Score. Such result show that the influence exerted by the persistent outliers can influence the model performance, which makes the FM-PFCM more suitable.

5.2.4. Comparison Between Different Thresholds

The main goal is to established the threshold value that maximizes the performance of the model, accounting for the misclassification costs of both classes. Therefore, the model was evaluated through accuracy, sensitivity, specificity and precision, using the optimal thresholds obtained by AUC, AUK and F-Score. The results are shown in the Table 5.

The threshold obtained from the AUK maximizes the correct classifications of the critical event, resulting in a higher number of true positives when compared to AUC. In fact, the AUC provides a more random model, not accounting for the misclassification errors.

Table 5: Performance indexes for FM-PFCM using the optimal threshold obtained from AUC, AUK and F-Score after 5×5 fold cross validation.

Threshold	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]
AUC	88.29 ± 0.45	33.24 ± 2.56	96.91 ± 0.46	62.97 ± 3.42
AUK	86.53 ± 0.55	57.01 ± 2.34	91.16 ± 2.34	50.32 ± 1.86
F-Score	85.64 ± 1.11	61.05 ± 3.08	89.50 ± 1.63	47.96 ± 3.22

Regarding the F-Score, there is an increase in sensitivity when compared to AUK, followed by a decrease in precision and specificity. However, the values reflect a slightly more balanced model, promoting the classification of the minority class.

5.3. Low Sampling Rate Data - First Critical Event

The detection of all the critical urine outputs showed a low prediction rate, as well as a low precision. The latter concept has high impact in medical context, since clinicians pursue a system that can provide a reliable outcome. Furthermore, the first critical event is of highest importance since an immediate action may prevent the development of the disease. Hence, a second approach was applied to predict only the first critical event.

The methodology used for data division, as well as for the parameter selection was the same as the one defined in the previous approach. In this case, logistic regression was not applied, focusing the study on FM-PFCM. Regarding the final dataset, the NCR undersampling method allowed to improve the performance of the model, increasing the value of the F-Score.

5.3.1. Model Assessment

To assess the advantages of the PFCM algorithm when compared to the traditional FCM approach, their performance was evaluate in terms of AUC, AUK, and F-Score. The results, presented in Table 4, show a clear increase of the F-Score, which evidences the influence exerted by outliers. Thus, the PFCM algorithm proves to be better suited for the current goal.

Table 6: Classifier performance after 5×5 fold cross validation using FM-FCM and FM-PFCM.

Methods	AUC [%]	AUK [%]	F-Score [%]
FM-FCM	85.84 ± 2.05	17.31 ± 1.00	63.54 ± 1.79
FM-PFCM	85.94 ± 1.92	17.29 ± 0.91	64.33 ± 1.90

5.3.2. Comparison Between Different Thresholds

Proceeding the study using FM-PFCM, the model capability to identify the critical urine drops below 30 ml/h was assessed, resorting to the optimal thresholds obtained from the AUC, AUK and F-Score. The results obtained can be observed in Table 7.

A similar performance between the F-Score and AUK is verified, whose thresholds allow for a slight improvement over sensitivity when compared to AUC, without significant decrease in specificity and precision.

Table 7: Performance indexes for FM-PFCM using the optimal threshold obtained from AUC, AUK and F-Score after 5×5 fold cross validation.

Threshold	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]
AUC	95.11 ± 0.12	50.30 ± 2.89	99.34 ± 0.21	88.01 ± 2.84
AUK	95.05 ± 0.11	52.30 ± 3.08	99.07 ± 0.23	84.28 ± 2.34
F-Score	95.02 ± 0.17	52.58 ± 2.83	99.00 ± 0.24	83.35 ± 2.74

However, regardless of the origin of the threshold, the models difficulty in detecting the first critical event is notorious. In relation to the approach to detect of all the critical events, there is a reduction in sensitivity, but at the same time an increase in specificity and especially in precision. Indeed, this improvement is remarkable and concedes a higher reliability, essential in the medical context.

5.4. High Sampling Rate Data - First Critical Event

Considering only the prediction of the first critical urine event, a new study was performed using data with one minute of sampling time. Regarding the data division, the dataset was partitioned into 5 folds according to the output classes, aiming to maintain the original class distribution within each fold. Similar to the low sampling rate approaches, the initial fuzzy parameters were defined using a grid search, and the performance of the model evaluated using AUC, AUK and F-Score.

5.4.1. Feature Selection Using Fuzzy Modeling

To extract the most relevant variables from the final dataset, feature selection was conducted using SFS. To do so, three objective functions (AUC, AUK and F-Score) were used and their performance compared, being the procedure repeated 10 times. The results for fuzzy modeling are presented in Table 8, whereas for logistic regression can be observed in Table 9.

Table 8: Features selected using FM-FCM with AUC, AUK and F-Score as objective functions.

OF	Features Selected	AUC [%]	AUK [%]	F-Score [%]
AUC/AUK	X_2, X_3, X_4, X_6, f_7, Age	68.07 ± 2.30	7.73 ± 0.73	27.67 ± 2.83
Without FS	All	57.87 ± 3.57	3.16 ± 1.41	23.32 ± 3.50

Table 9: Features selected using Logistic Regression with AUC, AUK and F-Score as objective functions.

OF	Features Selected	AUC [%]	AUK [%]	F-Score [%]
AUC	X_2, X_3, X_4, f_7, Age	68.75 ± 5.45	8.30 ± 2.66	31.18 ± 9.06
Without FS	All	57.87 ± 3.57	3.16 ± 1.41	23.32 ± 3.50

As can be observed for FM-FCM, both AUC and AUK highlight the same features, enhancing the quality of the model when compared to the results before feature selection. On the other hand, the F-Score did not not provide reliable outcomes, ending up selecting one static variable (Oasis Score) that has no influence on the evolution of the patients clinical condition.

Regarding the logistic regression models, the AUC highlights features related to blood pressure, as well as age and oasis respiration score, similar to the results of fuzzy modeling. In relation to AUK and F-Score, these only highlighted two categorical features, age and oasis respiration score, which are purely static variables incapable of accounting for the clinical evolution.

5.4.2. Comparison between Fuzzy Modeling and Logistic Regression

To compare the performance of both types of models, the AUC, AUK and F-Score were used. Further, it was verified that the application of the NCR method allows an increase in quality of both models. The results, presented in Table 10, evidence a higher performance of the logistic regression model, but also a higher standard deviation. Nevertheless, both models show low performances.

Table 10: Classifier performance after 5×5 fold cross validation using FM-FCM and Logistic Regression.

Methods	AUC [%]	AUK [%]	F-Score [%]
FM-FCM	68.11 ± 3.02	7.80 ± 1.01	30.10 ± 2.70
LR	68.79 ± 5.58	8.33 ± 2.73	31.67 ± 9.49

In fact, after evaluating the performance of the models in terms of accuracy, sensitivity, specificity and precision, it can be concluded that this approach is unsuitable for detecting the critical urine outputs which, from the clinical point of view, does not foster reliability.

Table 11: Performance indexes for FM-FCM using the optimal threshold obtained from AUC, AUK and F-Score after 5×5 fold cross validation.

Threshold	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]
AUC	90.49 ± 0.06	4.56 ± 8.23	99.62 ± 0.77	38.56 ± 35.86
AUK	77.28 ± 9.08	50.03 ± 18.23	80.19 ± 11.85	25.74 ± 11.66
F-Score	76.63 ± 9.44	51.67 ± 19.32	79.30 ± 12.35	25.27 ± 10.98

6. Conclusions

The analysis herein conducted allowed to conclude that F-Score allows the selection of the most relevant features, promoting the development of robust classifiers with generalization capability, regardless of the minority class size.

Regarding the selection of the optimal threshold, it is clear that the AUC tends to overlook the minority class when dealing with imbalanced datasets. On the other hand, in most cases, the AUK proves to be a more balanced and robust measure, properly weighing the error in both classes.

Lastly, it was concluded that the F-Score is the best measure to obtain the threshold when dealing with highly skewed datasets. In fact, the proposed change in the way this measure is calculated has clear benefits. Instead of using a specific threshold, an appropriate range of values with increments of 0.001 is used, allowing to select the value that maximizes models performance. This, coupled with the intrinsic properties of this measure, offers clear benefits in unbalanced data.

Concerning the prediction of the first critical urine output, the best results were obtained using the FM-PFCM approach. The final model shows a low performance in terms of the sensitivity, opposed to the values of specificity close to 1, which may arise from the reduced number of samples for the critical event. Even more significant is the increase in precision in comparison to the prediction of all the events. From the clinical point of view, it is still an imperfect system. However, it shows high reliability, revealing potential to be implemented as a clinical decision support system in real life situations.

Regarding the high sampling rate data, the models show poor performance, which may be explained by two factors: the high imbalance of the dataset and the similarity between samples. Therefore, it would be interesting to explore the high dimensionality of this type of data using autoregressive models to construct, for example, the time profile of the MAP.

In addition, the study of performance measures to define the optimal threshold should be extended to lower percentages of the minority class, as well as test the Matthews Correlation Coefficient, either to select the best set of features or to obtain the optimum threshold.

References

[1] D. R. Cox. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958.

[2] V. Cunha, C. Salgado, S. Vieira, and J. Sousa. Fuzzy modeling to predict short and long-term mortality among patients with Acute Kidney Injury. *2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016*, pages 148–153, 2016.

[3] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers ROC Graphs : Notes and Practical Considerations for Data Mining Researchers. *HP Invent*, page 27, 2003.

[4] D. J. Hand. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009.

[5] D. J. Hand. Assessing the Performance of Classification Methods. *International Statistical Review*, 80(3):400–414, 2012.

[6] J.-S. R. Jang, C.-T. Sun, and E. Mizutani. Neuro-Fuzzy And Soft Computing Jang: a computational approach to learning and machine intelligence, 1997.

[7] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:1–9, 2016.

[8] R. Jorge. Computational Intelligence for Short Term Kidney Function Prediction in the ICU, 2016.

[9] U. Kaymak, A. Ben-David, and R. Potharst. The AUK: A simple alternative to the AUC. *Engineering Applications of Artificial Intelligence*, 25(5):1082–1089, 2012.

[10] J. Labarère, R. Bertrand, and M. J. Fine. How to derive and validate clinical prediction models for use in intensive care medicine. *Intensive Care Medicine*, 40(4):513–527, 2014.

[11] D. LaFreniere, F. Zulkernine, D. Barber, and K. Martin. Using machine learning to predict hypertension from a clinical dataset. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2016.

[12] N. Menachemi and T. H. Collum. Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, 4:47–55, 2011.

[13] L. F. Mendonça, S. M. Vieira, and J. M. Sousa. Decision tree search methods in fuzzy modeling and classification. *International Journal of Approximate Reasoning*, 44(2):106–123, 2007.

[14] M. Ostermann and M. Joannidis. Acute kidney injury 2016: diagnosis and diagnostic workup. *Critical Care*, 20(1):1–13, 2016.

[15] R. Pacheco, C. M. Salgado, R. Deliberato, L. A. Celi, and S. M. Vieira. Short-term prediction of low kidney function in ICU patients. *IEEE International Conference on Fuzzy Systems*, (2), 2017.

[16] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 13(4):517–530, 2005.

[17] S. Patel and H. Patel. Survey of Data Mining Techniques Used in Healthcare Domain. *International Journal of Information Sciences and Techniques*, 6, 2016.

[18] J. M. C. Sousa and U. Kaymak. *Fuzzy decision making in modeling and control*, volume 27. 2002.

[19] R. D. Viegas. Feature Extraction for modeling patients ' outcomes : an application to readmissions in ICUs. (June), 2015.

[20] S. M. Vieira, U. Kaymak, and J. M. Sousa. Cohen's kappa coefficient as a performance measure for feature selection. *2010 IEEE World Congress on Computational Intelligence, WCCI 2010*, (May 2016), 2010.