Journal of Engineering Science *Vol. XXX, no. 1 (2023), pp. 106 - 117*
Fascicle *Electronics and Computer Science* *ISSN 2587-3474*
Topic *Biomedical Engineering* *eISSN 2587-3482*

https://doi.org/10.52326/jes.utm.2023.30(1).09
UDC 004:[612.13:616.94]

NEW APPROACHES TO MISSING BIOMEDICAL DATA RECOVERY FOR MACHINE LEARNING

Victor Iapăscurtă[1,2]*, ORCID: 0000-0002-4540-7045,
Ion Fiodorov[1], ORCID: 0000-0003-0938-3442

[1] *Technical University of Moldova, 168, Stefan cel Mare si Sfant Blvd., Chisinau, MD – 2004, Republic of Moldova*
[2] *N. Testemitanu State University of Medicine and Pharmacy, 165, Stefan cel Mare si Sfant Blvd., Chisinau, MD – 2004, Republic of Moldova*
*Corresponding author: Victor Iapăscurtă, *victor.iapascurta@doctorat.utm.md*

Received: 01. 18. 2023
Accepted: 03. 03. 2023

Abstract. Missing data is a common problem for medical data sets, especially large ones. This issue is of major importance since it can influence the analysis and further use of the data, e.g., for machine learning purposes. There are various methods for recovering missing data. One such method is to remove observations with missing values, but this is not very useful given the limited amount of data available. Another commonly used approach is the Last Observation Carried Forward (LOCF). But most such methods are not universal and may need adjustments to the data set at hand. This article describes the possibility of solving this problem in the case of multimodal time series of biomedical data coming from patients with sepsis. It describes and compares three approaches tailored to a sepsis dataset, which is analyzed and finally used to build a sepsis prediction system based on clinical data routinely recorded in an intensive care unit.

Keywords: *multimodal biomedical time series data; missing values; data recovery; sepsis; machine learning.*

Rezumat. Datele lipsă sunt o problemă comună pentru seturile de date medicale, în special pentru cele mari. Această problemă este de o importanță majoră, deoarece poate influența analiza și utilizarea ulterioară a datelor, de exemplu, în scopuri de învățare automată. Există abordări diferite pentru a trata datele lipsă. Una obișnuită este ștergerea observațiilor care conțin astfel de date, însă ea nu este aplicabilă atunci când volumul datelor este limitat. O altă abordare frecvent utilizată este "Last Observation Carried Forward (LOCF)". Dar majoritatea acestor metode nu sunt universale și pot necesita ajustări la setul de date la îndemână. Această lucrare descrie posibilitatea abordării acestei probleme în cazul seriilor temporale multimodale de date biomedicale provenite de la pacienții cu sepsis. Ea descrie și compară trei abordări adaptate setului de date care este analizat și utilizat în cele din urmă pentru construirea unui sistem de predicție a sepsisului bazat pe date clinice înregistrate în mod obișnuit într-o unitate de terapie intensivă.

Cuvinte cheie: *serii temporale de date biomedicale multimodale; valori lipsă; recuperare date; sepsis; învățare automată.*

1. Introduction

In research, missing data are frequently inevitable, but their ability to affect the findings is rarely explored. According to recent systematic reviews [1,2] many healthcare data sets are found to be incomplete, with missing values, and require cleaning and missing values imputation to enhance the effectiveness and accuracy of the data analysis. Along with an overview of the situation in the field, these articles present a review of methods for using machine learning to impute missing data.

The nature of "missingness" in this context - random vs nonrandom - is important to note. However, using observable data, it is sometimes impossible to tell apart between missing at random and missing not at random. As the data set employed in this study lacks a justification for the nature of the missingness, it is believed that the missing data are absent at random (during the time the patient is undergoing a medical procedure that requires the sensors used for data collection to be removed, human errors, equipment malfunction, etc.). The bias imposed by the recovery technique is thought to be less in the situation of data missing at random, even though data recovery may be achievable regardless of the kind of missingness [3].

When missing values occur in data, there are several ways to handle them. A lesser number of tools are useful for continuous data, such as time series, while the majority of these techniques are acceptable for static data. Complete case analysis (CCA) [4] is the simplest method, which involves deleting observations with missing values, but in many situations, especially when there is a limited amount of data, this technique may not be practical. In imbalanced sets, where each observation in the minority class is significant, this problem is of particular significance when using the data for machine learning techniques.

Data recovery techniques may often be classified into:

1. Based on single imputation - replace a missing data point with a single value using a single imputation approach, often using the Last Observation Carried Forward (LOCF), Baseline Observation Carried Forward (BOCF), and Next Observation Carried Backward (NOCB) procedures or data from other sources (e.g., mean value imputation, regression imputation, etc.).

2. Based on multiple imputations - using many plausible imputed data sets and properly integrating the findings from each, multiple recovery methods create plausible imputed data sets. There are several statistical software available for this (e.g., Amelia, FURIA, MICE in R programming language (R), etc.)

The Guideline on Missing Data in Confirmatory Clinical Trials [5] states that "if missing values are handled by simply excluding any patients with missing values from the analysis, this will result in a reduction in the number of cases available for analysis and consequently typically result in a reduction of the statistical power." Obviously, the likelihood of a power drop increases with the number of missing data. Thus, it is imperative to make every effort to reduce the quantity of missing data. Unluckily, there isn't a methodical technique to handle missing data that works in every circumstance. As a result, while preparing a trial, it is crucial to take into account how to reduce the quantity of missing data and how missing data will be treated in the analysis.

Three data recovery techniques are presented in the current work, each of which involves a number of steps and components: (a) conventional techniques (such as LOCF and NOCB); (b) a less popular Kalman filtering-based approach; and (c) regression imputation, which substitutes the predictions from a regression of the missing variables on the observed

variables. The idea of seeing the human body as a complex system in which the many characteristics that define its functioning (in health or sickness) are associated and missing data may be generated from existing data using approximated correlation is one of the underlying principles for the final technique.

This paper's research contribution specifically relates to the presentation of several approaches for missing value recovery in connection to the objectives of future usage of the recovered data.

As part of a bigger research project aimed at developing a machine learning-based software application for sepsis prediction, the data recovered via a variety of ways are eventually being used to build a machine learning system.

2. Research Data and Processing Methods

2.1. Data

The "Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019" [6] public database provided the information utilized in this study. The public portion of the data was sourced from two different US hospitals: Emory University Hospital (set A) and Beth Israel Deaconess Medical Center (set B). With the necessary Institutional Review Boards' consent, these data were gathered over the previous ten years, de-identified, and classified using Sepsis-3 clinical criteria. They include 8 vital sign variables, 26 laboratory variables, and 6 demographic factors. They are made up of hourly summaries of vital signs, lab results, and static patient descriptions for 40,336 patients. Each patient's characteristics were distilled into hourly bins (e.g., multiple heart rate measurements in an hourly time window were summarized as the median heart rate measurement).

The data in set B presents more than 80% of missing values. Thus, set A is selected for further study because it had fewer missing data (i.e., 79,4%) and a greater prevalence of sepsis (8,80% vs. 5,71% in set B). There are 1790 septic patients among the 20336 patients in this set, and 502 of those subsets have all the missing values for at least one parameter (out of 6 parameters of interest). Following the application of the initial selection criteria (such as the presence of at least 7 hourly observations prior to the diagnosis of sepsis, the lack of artifacts, etc.), there are 211 subsets that have missing data but may be able to have them recovered.

Table 1 depicts the look of an original sepsis file with one parameter (i.e., DBP) having all-missing values (NA) and other physiological parameters having some of their values missing. It describes observations on seven parameters of interest chosen for additional study, including age and labeling. These parameters include heart rate, peripheral blood oxygen saturation, temperature, systolic blood pressure, diastolic blood pressure, respiratory rate, patient's age, and the label of the observation.

Table 1

A fragment of an original sepsis file

HR	O_2Sat	Temp	SBP	DBP	Resp	Age	Sepsis label
NA	**NA**	**NA**	**NA**	**NA**	**NA**	43.55	0
83	99	**NA**	109	**NA**	13	43.55	0
89	99	**NA**	102	**NA**	22	43.55	0
82	98	36.56	108	**NA**	20	43.55	0
91	98	**NA**	108	**NA**	16	43.55	0

Continuation Table 1

93	99	37.1	106	**NA**	20	43.55	1
95	100	**NA**	130	**NA**	15	43.55	1
96	98	37.2	121	**NA**	18	43.55	1

Note. HR - heart rate, O2Sat - peripheral blood oxygen saturation, Temp - temperature, SBP - systolic blood pressure, DBP - diastolic blood pressure, Resp - respiratory rate. Sepsis label - the label of the observation (0 – for non-sepsis observations and 1 – for sepsis). NA – not available (missing value).

2.2. Missing Values Recovery Methods

Data recovery performed throughout this research is based on several simple common algorithms/methods, including their combination as well as less common techniques tailored to the available data. The first method is commonly used for similar data, while the next two methods are specifically designed for the data recovery task concerning the data at hand.

2.2.1 LOCF and NOCB

The last observation carried forward (LOCF) [7] is a missing value recovery method that uses the last measured value (per column) to fill in the next missing one(s). The next observation carried backward (NOCB) [8] is a "reversed version" of LOCF, by which the missing values are filled in backward. Examples of when these approaches can be applied are columns "HR", "O_2Sat", "Temp", "SBP" and "Resp" in Table 1. These techniques can evidently not be used for columns with all values missing (e.g., the "DBP" column in Table 1)

2.2.2 A custom technique

This approach was designed to take care of both cases: for columns with some values missing as well as for columns with all missing values.

The following procedures are included in this method when there is at least one value in each column: (a) each column is assessed for missing data (NA); (b) the first and last values (per column) are "recovered" in accordance with the value in the closest cell in the same column; (c) interpolation is completed, with the calculation of the values between two present values taking the trend into consideration (increase or decrease). For interpolation, the "zoo" package in R [9] (i.e., "na.approx()" function) is utilized (see, for instance, the "Temp" column in Table 1).

When all values in a column are missing, the recovery procedure is as follows: (a) from the complete cases set, the present values were extracted for each parameter/column separately by class (e.g., septic patients from set A). (b) the number of "non-missing" values (n) was determined and the mean (mean) and standard deviation (sd) were calculated for these values. Using the {rnorm()} function in R, and "n", "mean" and "sd" as arguments to it, Gaussian distributions were generated for each of the 6 parameters. The missing values in the original data are replaced with values extracted from generated distributions.

2.2.3 A technique with the use of Kalman filtering and machine learning algorithms

This approach uses Kalman filtering for columns with at least three non-missing values and Generalized Linear Models for data recovery in case of all missing values in a column.

The Kalman filter, which first appeared in [10], uses a discrete filtering model based on the dynamic and measurements models as linear Gaussian. The following is a description of the basics of this approach with elements of optimal control and dynamic linear models as in [11].

The Kalman filtering technique is used to estimate states based on linear dynamical systems in state space format and offers estimates of unknown variables given the observations over time. According to this model, the state is evolving as follows from time $t-1$ to time t:

$$x_t = Fx_{t-1} + Bu_{t-1} + w_{t-1}, \quad (1)$$

where F is the state transition matrix estimated using the previous state vector x_{t-1}, B is the control-input matrix applied to the control vector u_{t-1}, and w_{t-1} is the process noise vector that is assumed to be zero-mean Gaussian with the covariance Q, i.e., $w_{t-1} \sim N(0, Q)$.

The link between the state and the measurement at the current time step t is described by the process model in conjunction with the measurement model as:

$$z_t = Hx_t + v_t, \quad (2)$$

where z_t is the measurement vector, H is the measurement matrix, and v_t is the measurement noise vector that is assumed to be zero-mean Gaussian with the covariance C, i.e., $v_t \sim N(0, C)$.

Given the initial estimate of x_0, the sequence of measurements, $z_1, z_2, \ldots, z_t$, and the details of the system defined by F, B, H, Q and C, the Kalman filter's job is to offer an estimate of x_t at time t. Typically, Q and C are tuned parameters that the user may change to get the desired performance.

Kalman filter algorithm consists of two steps: (1) prediction and (2) update, and can be described as follows:

Prediction (step 1):

Estimating predicted state $\hat{x}_t^- = F\hat{x}_{t-1}^+ + Bu_{t-1}$ (3)

Calculating predicted error covariance $P_t^- = FP_{t-1}^+ + F^T + Q$ (4)

Update (step 2):

Estimating measurement residual $\tilde{y}_t = z_t - H\hat{x}_t^-$ (5)

Calculating Kalman gain $K_t = P_t^- H^T (R + HP_t^- H^T)^{-1}$ (6)

Updating state estimate $\hat{x}_t^+ = \hat{x}_t^- + K_t\tilde{y}$ (7)

Updating error covariance $P_t^+ = (I - K_t H)P_t^-$, (8)

where the „hat" operator denotes an estimate of a variable. Thus, $\hat{x}$ is an estimate of x. The superscripts "-" and "+" denote predicted (prior) and updated (posterior) estimates, respectively.

The software implementation used for the data recovery purpose in this research is "imputeTS" package in R [12] and represents an extended version of the algorithm described above (i.e., Eq. 1-8) with Kalman smoothing on structural time series models.

The algorithm used for data recovery purposes in columns with all values missing represents a Generalized Linear Model (GLM) [13] available on the H_2O platform [14].

The dependence between the response variable y and the covariates vector x is modelled as a linear function by the Gaussian technique (behind GLM):

$$y = x^T\beta + \beta_0 + \epsilon, \quad (9)$$

where, β is the parameter vector, β_0 denotes the intercept and ϵ is a gaussian random variable representing the noise in the model, $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

By maximizing the log-likelihood over the parameter vector for the observed data, the model's estimation is achieved. By resolving the following likelihood optimization with parameter regularization, the GLM [13] employed in this study fits the model:

$$\max_{\beta,\beta_0} (GLM\ Log \cdot likelihood - \ Regularization\ Penalty). \tag{10}$$

The weighted sum of the ℓ_1 (least absolute shrinkage parameter) and ℓ_2 (ridge regression) norms of the coefficients vector is the regularization penalty, and it is expressed as follows:

$$\lambda P_\alpha(\beta) = \lambda \left(\alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2 \right), \tag{11}$$

where α is the elastic net parameter, $\alpha \in [0, 1]$ and is λ a tuning parameter, and there is no penalty for the intercept.

The optimization task concerning observations can be described as follows:

$$\max_{\beta\beta_0} \sum_{i=1}^{N} \log f\,(y_i;\ \beta, \beta_0) - \lambda \left(\alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2 \right). \tag{12}$$

The Gradient Boosting Machine (GBM) [15], offered by the same H2O platform, turned out to be the top performing algorithm at the final machine learning (ML) stage.

R [16] is the programming language employed in the current study, and a number of packages from the same environment are used for a variety of tasks throughout the study. Plotting and interacting with the H_2O ML platform are done in the same language and environment.

3. Data Processing and Results

This study's data processing cycle includes several processes, such as missing value recovery, and tries to provide datasets appropriate for machine learning.

Machine learning algorithms are used in this research at different stages for two distinct purposes: (a) for missing values recovery (i.e., GLMs), and (b) for building the final sepsis prediction model (at this stage it was experimented with several algorithms and this is described in coming sections). The following is a description of the main processing steps through a Consolidated Standards of Reporting Trials (CONSORT)-like diagram.

3.1. Preprocessing Stage

At this stage files containing artifacts like human errors, equipment malfunction or failure, etc. were excluded. Due to the study design and the goal of building a sepsis prediction system with a prediction horizon of at least four hours (and three-hour observations needed to get the first prediction), sepsis files with less than seven observations were also excluded. A similar approach was used for non-sepsis subsets keeping only the files with seven and more consecutive observations without missing values concerning the parameters of interest. This is illustrated in Figure 1.

502 sepsis subsets are initially lacking all values for at least one important parameter, such as heart rate, peripheral blood oxygen saturation, temperature, systolic blood pressure, diastolic blood pressure, and respiratory rate.

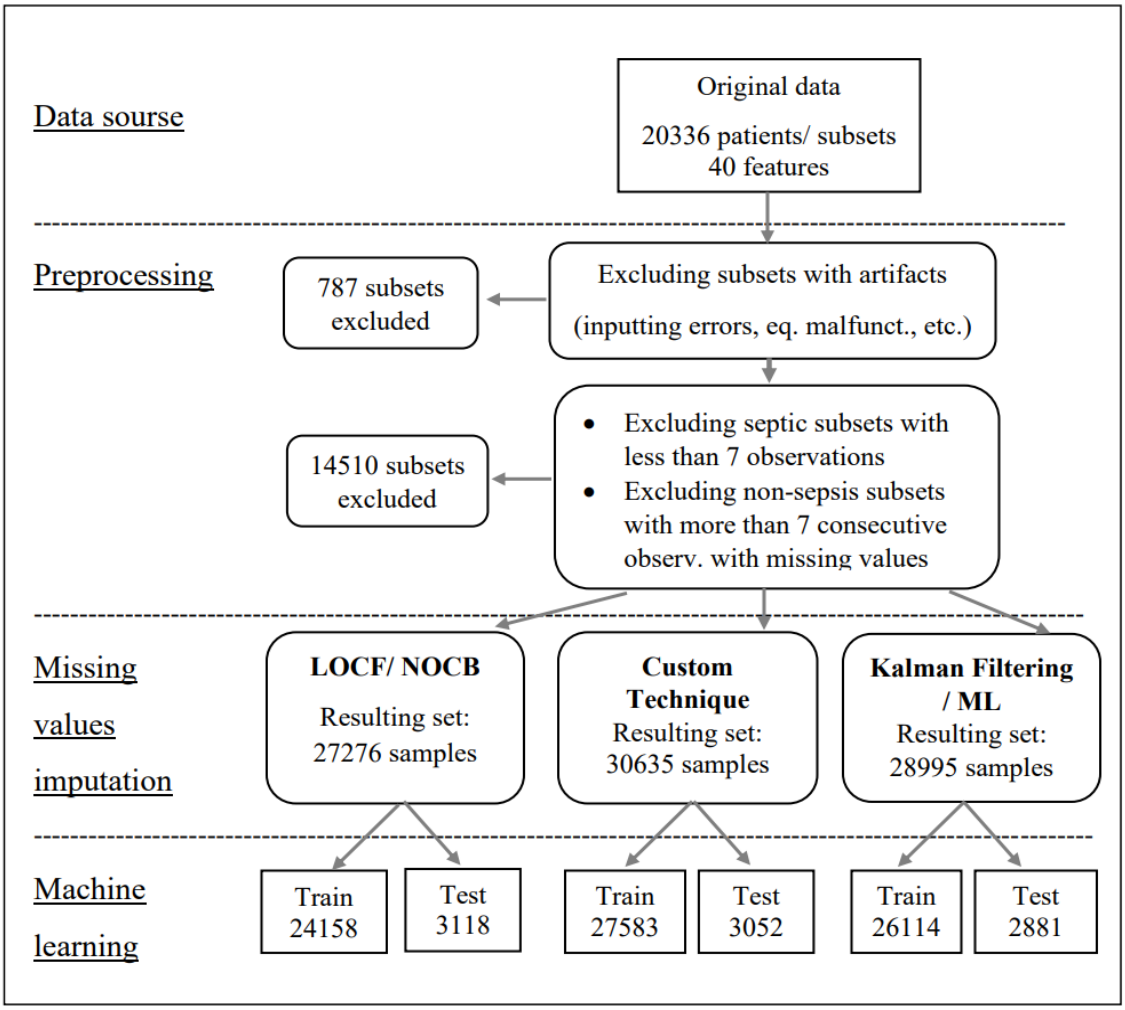


Figure 1. Data processing flow in the current research (CONSORT-like diagram).

Some of these files were reconstructed at the next stage with the custom approach described above or through ML algorithms used for data recovery purposes as part of the third method based on Kalman filtering.

3.2. Sepsis Data Reconstruction via Missing Values Recovery

3.2.1 Applying LOCF/NOCB

The first method (i.e., LOCF/NOCB) for data recovery in this research is LOCF (Last Observation Carried Forward, by "ImputeTS" package, R). The same package is used for NOCB as the second step. This is appropriate for columns in which some of the values are missing. It will not work for columns with all missing values, but it will recover lost data in columns when some of the values are missing.

3.2.2. Data recovery using the custom approach

Since this approach can deal with partially as well as totally (in a column) missing values it provides a final set of a larger size (i.e., 30635 samples) to be used for building the prediction system.

3.2.3 Kalman Filtering coupled with ML for data recovery

This approach also takes care of columns with partially missing values, on which Kalman filtering for data recovery can be used, should there be three and more non-missing values in the respective column.

The correlation between the six important factors listed above and the age was looked at in order to handle the all-missing values scenarios. The age has no missing data and exhibits a moderate correlation with some important metrics.

For each of the six parameters, the three most correlated parameters were chosen based on the correlation coefficients (e.g., for temperature the most correlated parameters are heart rate, systolic blood pressure, and age).

The correlation coefficients for seven parameters are shown in Table 2. The highest correlation coefficients are indicated by bold type.

Table 2

Correlation Coefficients

	HR	**O_2Sat**	**Temp**	**SBP**	**DBP**	**Resp**	**Age**
HR	1.00	-0.12	0.18	-0.02	**0.24**	**0.18**	**-0.18**
O_2Sat	**-0.12**	1.00	0.02	**0.06**	0.02	**-0.19**	-0.02
Temp	**0.18**	0.02	1.00	**0.09**	0.03	0.08	**-0.16**
SBP	-0.02	0.06	**0.09**	1.00	**0.53**	**0.07**	-0.01
DBP	**0.24**	0.02	0.03	**0.53**	1.00	0.06	**-0.33**
Resp	**0.18**	**-0.19**	**0.08**	0.07	0.06	1.00	0.07
Age	-0.18	-0.02	-0.16	-0.01	-0.33	0.07	1.00

Note. HR - heart rate, O2Sat - peripheral blood oxygen saturation, Temp - temperature, SBP - systolic blood pressure, DBP - diastolic blood pressure, Resp - respiratory rate. Sepsis label - the label of the observation (0 – for non-sepsis observations and 1 – for sepsis).

A series of GLMs were trained using this correlation data, and the models that performed the best were chosen for additional study. The primary traits of these models are displayed in Table 3.

Table 3

Logistic Regression Coefficients for Parameters of Interest (normalized)

Parameter for recovery	**Intercept**	**Coefficient (Parameter)**		
HR	89.4392	3.2658 (DBP)	3.2980 (Resp)	-2.3973 (Age)
O_2Sat	97.4894	-0.2305 (HR)	0.1840 (SBP)	-0.4954 (Resp)
Temp	37.2479	0.1233 (HR)	0.0720 (SBP)	-0.1001 (Age)
SBP	123.6418	1.6593 (Temp)	11.7255 (DBP)	0.6840 (Resp)
DBP	61.4999	2.5176 (HR)	6.8117 (SBP)	-3.7355 (Age)
Resp	20.3789	0.9001 (HR)	-1.0138 (O_2Sat)	0.3220 (Temp)

Note. HR - heart rate, O2Sat - peripheral blood oxygen saturation, Temp - temperature, SBP - systolic blood pressure, DBP - diastolic blood pressure, Resp - respiratory rate.

Together with Kalman filtering, these models are incorporated into the data recovery pipeline that reconstructs the missing value sepsis files. The look of a recovered file using this method is shown in Table 4. Bold values indicate recovered values. The resulting recovered subset is the one shown in Table 1 above.

Table 4

The recovered sepsis subset

HR	**O_2Sat**	**Temp**	**SBP**	**DBP**	**Resp**	**Age**	**Sepsis label**
83.37	**98.75**	**36.56**	**108.33**	**60.22**	**17.12**	43.55	0
83	99	**36.56**	109	**60.37**	13	43.55	0
89	99	**36.56**	102	**59.09**	22	43.55	0

Continuation Table 4

82	98	36.56	108	**59.93**	20	43.55	0
91	98	**36.83**	108	**61.2**	16	43.55	0
93	99	37.1	106	**60.87**	20	43.55	1
95	100	**37.15**	130	**68.45**	15	43.55	1
96	98	37.2	121	**65.86**	18	43.55	1

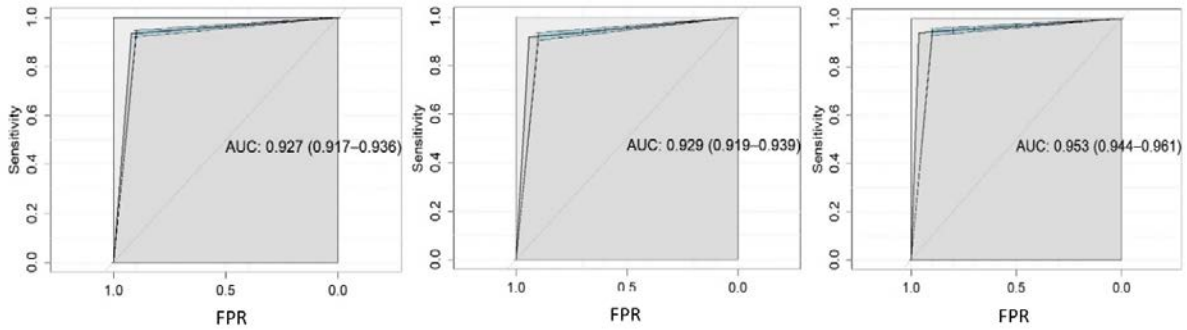
Note. HR - heart rate, O2Sat - peripheral blood oxygen saturation, Temp - temperature, SBP - systolic blood pressure, DBP - diastolic blood pressure, Resp - respiratory rate. Sepsis label - the label of the observation (0 – for non-sepsis observations and 1 – for sepsis).

3.3. Preparing Data Sets for Machine Learning

After being recovered, the data are divided into a training set (85–90% of the final samples) and a test set (10–15%). A sliding window method is used on each file (or subset) to aggregate observations into three-value-long chunks. Lastly, the algorithmic complexity (using the Block Decomposition Method) on each of the two 3x3 matrices is calculated, along with the difference between the parameter's value in three successive hourly samples [17]. The format of the data to be provided to the ML algorithm is represented by the 14L vector that is produced for each sample. The size of the final data sets is more than the number of initially selected files/subsets since each file comprises at least seven hourly observations of each of the six parameters of interest on which the sliding window approach is used. This procedure was applied to each of the final sets recovered through the approaches described earlier. The size of the final sets and train/test splits are presented in Figure 1.

3.4. Machine Learning Stage

A series of machine learning models are trained using the H2O platform with 10-fold cross-validation utilizing the training sets for ML produced with each of the three methods mentioned above. Gradient Boosting Machine (GBM), Generalized Linear Model (GLM), Distributed Random Forest (DRF), Stacked Ensemble (SE), and Deep Learning (DL) are the techniques utilized at this stage. Although GBM and SE had the best results, the GBM model was chosen for more study for explainability grounds.



a) b) c)

Figure 2. Classification performance by AUC of three GBM models built using the data recovered through three approaches: a) LOCF/NOCB (0.927); b) Custom method (0.929); c) Kalman filtering and ML (0.953).

Figure 2 displays the area under the ROC curve (AUC) for the classification (sepsis vs. non-sepsis) performance of these models on the test set of samples that did not take part in model training.

As one can see in the figure above the best performing model (i.e., AUC equal to 0.953) is model (c), or the model based on Kalman filtering for partially missing values coupled with ML approach through GLM for data recovery in columns with all-missing values. Table 5 summarizes further facts about the top-performing GBM.

Table 5

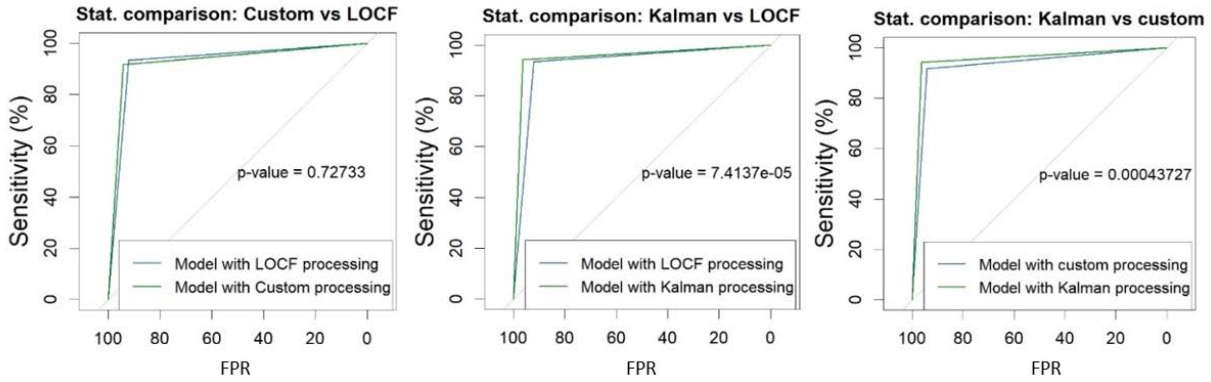
Confusion Matrices and Performance Statistics for Three Data Recovery Methods

		Reference (LOCF-NOCB)		Reference (Custom)		Reference (Kalman)	
		0	1	0	1	0	1
Prediction	0	1830	75	1874	89	1917	52
	1	159	1054	115	974	72	840

Accuracy	0.9250	0.9332	0.9570
95% CI	0.9151 - 0.9340	0.9237 - 0.9418	0.9489 – 0.9641
P-Value	< 2.2e-16	< 2.2e-16	< 2.2e-16
Cohen's Kappa	0.8401	0.8536	0.8999
Mcnemar's Test P-Value	5.767e-08	0.08006	0.08796
Sensitivity	0.9201	0.9422	0.9638
Specificity	0.9336	0.9163	0.9417

4. Discussion

It is possible to somewhat diminish the amount of missing data with thoughtful planning. This is significant because incomplete data might introduce bias into data analysis. This work uses a well-known viewpoint to handle the difficulty of recovering missing data during model construction. As far as we are aware, this is the first time the data recovery strategy based on the amalgamation of Kalman filtering and ML has been used for such or comparable datasets. One of key components of this method is the Kalman filter, used here in a less common way (in contrast to its more traditional application [11], e.g., for car cruise control, autopilot, dynamic positioning, target tracking, etc.). The second main component of this method is represented by six GLMs used to impute the values in columns with all-missing data. The latter is described in more details in [18].



a) b) c)

Figure 3. Statistical comparison of the performance of three ML models (by AUC): a) Custom method vs LOCF, b) Kalman vs LOCF, c) Kalman vs Custom method.

When used for missing values recovery, this approach provides the highest performance (AUC 0.953) of the sepsis prediction model and the difference is statistically

significant when compared with the other two approaches (p-values of 7.4137e-05 (b) and 4.3727e-04 (c) respectively). Figure 3 presents more details of the statistical comparison of the three methods described in the current work.

The suggested approach has some drawbacks. First off, it has not been evaluated on a variety of datasets, including datasets with categorical variables and those produced by other diseases (other than sepsis), among others. Additionally, it disregards data in which there is no association between the features and disregards the kind of variable distribution (normal, log-normal, logarithmic, etc.). So, identical datasets with continuous features and outcomes and a similar correlation between features could take into account our technique. Future research may have the opportunity to evaluate the method's resilience across various datasets.

Using the methods outlined in this study on sufficiently big and comparable data sets would be one of the future research paths. When analyzing various strategies for missing data recovery, particularly when evaluating the performance of classification ML models in differentiating between septic and non-septic cases, might serve as metrics of the method adequacy and dependability (e.g., comparing the results of the full set analysis to those of the complete case analysis).

5. Conclusions

The missing values in research data may be an issue, especially in case of large data sets with multiple observations and features. There are a number of methods for dealing with such issues, but a universally accepted approach is lacking, particularly when the data are used for machine learning purposes.

This paper tackles the missing value problem in a medical data set coming from patients in the intensive care unit. After recovery through three distinct approaches, these data are used to build a sepsis prediction system.

The authors' missing value imputation strategy, which is based on ML and Kalman filtering, offers the best classification performance (AUC 0.953) for the specific data utilized in this study.

As the missing values imputation method may have an impact on how well an ML model discriminates, it is worthwhile to test out several approaches before settling on the most appropriate one for the given set of data. It can result in a better diagnosis and course of therapy in some circumstances, potentially saving the patient's life (like septic patients in the current study).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ismail, A.R.; Abdin, N.Z.; Maen, M.K. Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare. *JRC* 2022, 3(2), pp. 143-152.
2. Oluwaseye, L.J.; Doorsamy, W.; Sena, B.P. A Review of Missing Data Handling Techniques for Machine Learning. *IJITIS* 2022, 5(3), pp. 971-1005.
3. Sterne, J.A.; White, I.R.; Carlin, J.B.; Spratt, M.; Royston, P.; Kenward, M.G.; Wood, A.M.; Carpenter, J.R.; Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009, 338:b2393.
4. Jain, R.; Xu, W. Dynamic model updating (DMU) approach for statistical learning model building with missing data. *BMC Bioinform* 2021, 22, 221.
5. Guideline on Missing Data in Confirmatory Clinical Trials, 2010 EMA/CPMP/EWP/1776/99 Rev. 1 Committee for Medicinal Products for Human Use (CHMP). Available online: https://www.ema.europa.eu/en (accessed on 14.02.2023).

6. Reyna, M.A.; Josef, C.S.; Jeter, R.; Shashikumar, S.; Westover, M.B.; Nemati, S.; Clifford, G.D.; Sharma, A. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. *Crit. Care Med.* 2020, 48(2), pp. 210-217.
7. Streiner, D.L. Last observation Carried Forward. In: *Encyclopedia of Research Design*, 1st ed.; Salkind, N.J. Ed.; SAGE Publications, Inc., 2010, 1776 p.
8. Single Imputation Methods for Missing Data: LOCF, BOCF, LRCF (Last Rank Carried Forward), and NOCB (Next Observation Carried Backward) 2021. Available online: http://onbiostatistics.blogspot.com/2021/01/single-imputation-methods-for-missing.html (accessed on 07.03.2023).
9. Zeileis, A.; Grothendieck, G. zoo: S3 Infrastructure for Regular and Irregular Time Series. *J. Stat. Soft. 2005*,14(6), pp. 1–27.
10. Kalman, R.E. A New Approach to Linear Filtering and Prediction Problems, Transactions of the ASME. *J. Basic Eng.* 1960, 82 (Series D), pp. 35-45.
11. Govaers, F. Introduction and Implementation of the Kalman Filter. *IntechOpen* 2019, 128 p., doi: 10.5772/intechopen.75731.
12. Moritz, S.; Bartz-Beielstein, T. ImputeTS: Time Series Missing Value Imputation in R. *The R Journal* 2017, 9(1), pp. 207-218.
13. Nykodym, T.; Kraljevic, T.; Wang, A.; Wong, W. *Generalized Linear Modeling with H2O,* 6ht ed.; H2O.ai, Inc.: Mountain View, CA, 2022; pp. 10-17.
14. H2O.ai. (3.32.1.1, 2021) h2o: R Interface for H2O. Available online: http://h2o-release.s3.amazonaws.com/h2o/rel-zipf/1/index.html (accessed on 07.03.2023).
15. Malohlava, M.; Candel, A. *Gradient Boosting Machine with H2O,* 7th ed.; H2O.ai, Inc.: Mountain View, CA, 2022; pp. 8-14.
16. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. R version 4.0.5, 2021. Available online: https://cran.r-project.org/bin/windows/base/old/4.0.5/ (accessed 07.03.2023).
17. Iapascurta, V. A less traditional approach to biomedical signal processing for sepsis prediction. In: *5th International Conference on Nanotechnologies and Biomedical Engineering*, Chisinau, Moldova, November 3-5, 2021; Tiginyanu, Sontea, Railean, Eds.; Springer IFMBE Proceedings Series, 2022; pp. 215-222.
18. Iapascurta, V. Dealing with Missing Continuous Biomedical Data: a Data Recovery Method for Machine Learning Purpose. In: *The 12th International Conference on Electronics, Communications and Computing*, Chisinau, Moldova October 20-21, 2022.

Citation: Iapăscurtă, V.; Fiodorov, I. New approaches to missing biomedical data recovery for machine learning. *Journal of Engineering Science* 2023, 30(1), pp. 106-117. https://doi.org/10.52326/jes.utm.2023.30(1).09.

Submission of manuscripts: jes@meridian.utm.md