New Approach based on Machine Learning for Short-Term Mortality Prediction in Neonatal Intensive Care Unit

Zaineb Kefi[1], Kamel Aloui[2], Mohamed Saber Naceur[3]
LTSIRS, National Engineering School of Tunisia, Tunis, Tunisia

***Abstract*—Mortality remains one of the most important outcomes to predict in Intensive Care Units (ICUs). In fact, the sooner mortality is predicted, the better critical decisions are made by doctors based on patient's illness severity. In this paper, a new approach based on Machine Learning (ML) techniques for short-term mortality prediction in Neonatal Intensive Care Unit (NICU) is proposed. This approach relies on many steps. At first, relevant features are selected from available data upon neonates' admission and from the time-series variables collected within the two first hours of stay in the NICU from the Medical Information Mart for Intensive Care III (MIMIC-III). After that, to predict mortality, many classifiers were tested which are Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB) and Random Forest (RF). The experimental results showed that LDA was the best performing classifier with an accuracy equal to 0.947 and AUROC equal to 0.97 with 31 features. The third step of this approach is mortality time prediction using the Galaxy-Random Forest method achieving an f-score equal to 0.871. The proposed approach compared favorably in terms of time, accuracy and AUROC with existing scoring systems and ML techniques. It is the first work predicting neonates mortality based on ML techniques and time-series data after only two hours of admission to the NICU.**

Keywords*—Mortality prediction; neonates; Intensive Care Units; machine learning

I. Introduction

Intensive Care Units (ICUs) are a hospital department [1] where doctors make the most important, complex and uncertain decisions. It's also the place where we find patients, with critical health conditions following accidents, prematurity, surgical complications, severe breathing problems, infections and serious injuries [2]. These patients need constant and close monitoring with sophisticated technology for body function maintain [3]. This sophisticated technology is expensive and not very abundant in ICUs which makes its good management a way for great gain in health care costs by reducing them remarkably. Then, predicting outcomes like mortality in similar contexts, helps doctors to do efficient resources allocation [4], to easily make critical decisions, to compare medication, to define care levels, to reduce health care costs and to discuss with patients' families about expected outcomes. The importance of mortality prediction resulted in many studies on relevant topics ranging from severity scoring systems and statistical methods to Machine Learning Techniques.

For years, scoring systems, also called severity scoring models or risk prediction models, were used by doctors to assess illness severity [5]. This estimate is based on fixed variables and coefficients that require the collection of a large amount of clinical information. However, for primary hospitals, this is not practical and even not very feasible in some cases. The second limitation of scoring systems appears when applied on a population different from that one on which they were developed. Therefore, because of these disadvantages and technological developments, statistical analysis and machine learning techniques have taken the place of classical methods. In fact, models developed locally are more flexible, quickly updated and improved which makes them more adapted for prediction than the standard severity scoring systems [6]. On another side, in terms of early mortality prediction, neonates scoring systems predict mortality from 12 and 24 hours since admission to the ICU except for Clinical Risk Index for Babies II (CRIBII) [7] which predicts mortality after one hour of admission. But, in addition to being purely statistically derived, the value of one of CRIB II variables used for the calculation, which is the maximum base deficit, covers the first 12 hours of admission. Likewise, many works based on machine learning techniques predicted neonates mortality after the first 12, 18, 24, 36, 48 and 72 hours of admission[8]. Even the model proposed by Veith et al. [9], predicting mortality at admission, is not specific to neonates and does not use clinical data.

So, because of mentioned limitations related to scoring systems and because the earliest work in mortality prediction for neonates using ML techniques and clinical data predicts mortality after 12 hours, the goal of the present paper is to provide a new approach for short-term mortality and death hour predictions based on time-series variables and data collected upon neonate's arrival to the NICU and 2 hours after admission. The proposed approach employs an ensemble of classifiers such as Support Vector Machine, Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Classification and Regression Trees, Naïve Bayes and Random Forest to predict in-hospital mortality. Galaxy-Random Forest is also used, for the first time in such context, to predict the time interval in which mortality occurs. The big health care MIMIC-III [10] database was used to design models for mortality estimation.

The rest of this paper is structured into five sections: the literature review is presented in Section 2. Section 3 presents the methodology in which work steps are described. In Section 4, results found during the performance evaluation of the proposed approach are exposed. Finally, in Sections 5 and 6, respectively, the results are discussed and the paper is concluded.

II. Related Works

In this section, works already done in neonates mortality prediction are reviewed, starting with scoring systems and going towards solutions based on machine learning techniques.

A. Scoring Systems

Neonates mortality may have several causes and the major one is prematurity [3]. Indeed, in case of prematurity, mortality risk increases especially when accompanied by a gestational age < 28 weeks and a birth weight < 1000 grams, which is generally the case [11]-[12]. Thus, birth weight was one of the most used indicators for mortality prediction which hampered developing scoring systems specific to neonates compared to those of older people. However, admitting the existence of other factors besides the birth weight which can have an effect on mortality and morbidity, a variety of neonate scoring systems, also known as risk adjustment scores, were developed in order to predict neonates mortality. Yet, they have not shown the same relevance. In addition to that, they work on data collected over different time intervals ranging from one hour after admission to the ICU to 24 hours and more.

The Clinical Risk Index for Babies (CRIB) [13] predicts mortality of infants with gestation less than 32 weeks at birth based on 6 variables collected in the first 12 hours after admission. Birth weight, gestation, congenital malformation, maximum base deficit in 12 hours, minimum and maximum appropriate FIO_2 in first 12 hours are these 6 most predictive factors according to logistic regression. Compared to other scoring systems, CRIB is less susceptible to treatment effects, the calculation takes five minutes per infant and the data required is easily collected. CRIB II [14] is an update of CRIB mainly affecting prediction time which went from 12 to the 1st hour of life. Birth defects and variables influenced by infant's care giving were excluded.

The Score for Neonatal Acute Physiology (SNAP) [15] predicts neonates' in-hospital mortality based on the least favorable first 24 hours physiologic measurements after admission counting 28 items. A score going from 0 to 5 is assigned to each item according to expert opinion. SNAP-Prenatal Extension (SNAP-PE) [15] uses SNAP scores in addition of birth weight, an APGAR score less than 7 at 5 minutes and small for gestational age to predict neonate's in-hospital mortality. SNAP II [16] and SNAP-PE II [5] are updates to previous scores which predict well mortality but raise the problem of data collection difficulty. This update saved 4 minutes in the scoring time. Such result was achieved following the data collection time interval change to 12 instead of 24 hours after admission. It was also reached after keeping only 6 variables which are strongly associated to death.

Other scores were developed but did not have the same success as those mentioned previously. The National Therapeutic Intervention Scoring System (NTISS) [7] is based on the treatment received by the patient, something that depends on the practices and unitary policy which makes this score unusual. Moreover, other scoring systems have suffered the same fate as The Berlin Score [7] because of its inclusion of subjective factors or the National Institute of Child Health and Human Development (NICHHD) and the Neonatal Mortality Prognosis Index (NMPI) that have not been widely used since development [7].

Although scoring systems have shown better mortality prediction compared to birth weight, they suffer from many issues. They remain not completely accurate due to the complexity of the neonatal clinical process and updates to these systems are a proof that they need adaptation to keep their effectiveness. They are not very flexible since they are based on specific data to be calculated. Also, individual outcomes prediction is not very feasible because of the variation in care approaches from one unit to another and even within the same unit. Moreover, in case of short-term mortality prediction in NICU, models meeting this need are not really suitable for this purpose. As a result, the need for other solutions to predict mortality has emerged.

B. Machine Learning Techniques

Machine learning techniques are more efficient than the standard severity scoring systems in the medical environment thanks to their ability to be improved and updated quickly. They are also more efficient than scoring systems when it's about a patient-specific prediction [6]. Indeed, the locally developed models have great flexibility and freedom to design customized models based on available data.

A Gaussian process classification was used by Rinta-Koski et al. [17] to predict preterm infant in-hospital mortality. They worked with 598 infants having weight under 1500 g which means very low birth weight. Data was collected from the Helsinki University Hospital database. SNAP-II and SNAPPE-II were calculated at arrival to the NICU. They also collected time series data for the first 72 hours of care to predict their outcomes. Evaluation also covered time periods ranging from 12 to 48 hours, passing by 18, 24 and 36 hours. With an AUC equal to 0.946, their proposed model outperformed clinical scores SNAP-II and SNAPPE-II.

To compare hospitals performance, Liu et al. [18] proposed an approach for mortality rate risk adjustment. It's based on Random Forest and Bart (Accuracies: Logistic Regression =0.93, Classification Random Forest =0.94, Regression Random Forest= 0.94, BART =0.95). This ensemble of tree methods performed better than logistic regression prediction accuracy for early born babies risk evaluation.

Artificial Neural Networks (ANNs) were used by Frize et al. [19] to build an automated tool predicting neonate's mortality 48 hours after admission (Sensitivity = 81%, Specificity = 98%). And by Saadah et al. [20] to predict mortality risk in case of nosocomial outbreaks of RSV(Sensitivity = 82%, Specificity = 100%). Cerqueira et al.

[21] also used ANN and Support Vector Machine (SVM) to design NICeSim which is an open-source simulator based on machine learning techniques to predict death probability of newborns(accuracy = 86.7%, AUC = 0.84).

C5.0 decision tree software was used by Gilchrist et al. [22] in order to develop a real-time mortality prediction model for 12, 24 and 48 hours. To validate the model, a 5-by-2 cross validation technique was used. And the F1-score was used to measure the performance of the model(Sensitivity = 63%, Specificity = 94%). According to authors, best predictors were serum pH, mean blood pressure, immature/total neutrophil ratio, respiratory rate, serum sodium, heart rate, serum glucose and pO2 blood oxygen level.

In their study, Townsend and Freize [23], worked on three outcomes which are Mortality, Length of Stay (LOS) and Ventilation Duration (DOV) 12 hours after admission. Models with risk estimation ranges were created using conjunction between the maximum likelihood (ML) approximation and a gradient descent artificial neural network (ANN). K-Nearest Neighbor (KNN) was used to solve the problem of missing values. The evaluation of the model's performance gave a sensitivity = 63% and a specificity = 99%.

A collaborative parent decision support system PADS, was proposed by Frize et al. [24] aiming to implicate parents in every step of decision-making with physicians in the NICU. Outcomes in this system, which are mortality, LOS (Length Of Stay) and Ventilation duration are estimated by Artificial Neural Network (ANN) (2n+1 hidden layers and n is the input number). After 24 hours of admission, estimations are delivered by the system. A kind of extension, named Physician-PArent Decision-Support PPADS, was proposed in 2013 [25]. It works with real-time data to predict mortality. The 5-by-2 cross validation was used to validate the model which was built based on "Multilayer Perceptron (MLP) feed-forward network with back-propagation" [25]. K-nearest neighbors was also used to deal with missing values.

In their work [9], not limited to neonates, Veith and Steele predicted patients mortality upon admission to the ICU based on machine learning techniques. They took patients data from the MIMIC-III database. Admission type, religion, ethnicity, marital status, the patient's insurance provider, language and previous location were attributes used in this work for the predictive model building. Their common point is the ease of their collection on admission. The best five performing algorithms were Logistic, Simple Logistic, LazyKStar, Bayes Net and Naïve Bayes. With 10-fold cross validation, they reached an AUC going from 0.721 to 0.689. With the training set they reached an AUC between 0.751 and 0.706.

The difference in the results, which is sometimes not too broad, reveals that there is no best machine learning technique over all situations and the performance of the model depends on the nature of the data and related problems as well as the context in which the prediction is conducted. However, some algorithms seem to perform better than others in certain aspects.

III. Materials and Methods

In this section, tools and methods used by the proposed approach for short-term mortality prediction in NICU are presented. Given the nature of the data monitored and collected, understanding data and dealing with probable problems related to it are the first challenges presented in this section. After that, features selection and the prediction model are presented.

A. DataBase

The data source in this study is the Medical Information Mart for Intensive Care III (MIMIC-III) [10]. It's a relational critical care database developed by the MIT Lab for Computational Physiology. MIMIC-III is more and more widely used thanks to its free accessibility. In this work, 1.4 is the used version (September 2016). MIMIC-III is composed by deidentified data coming from more than 40.000 critical care patients who have stayed in Beth Israel Deaconess Medical Center in Boston, Massachusetts. Many categories of data appear in the used database: demographics, laboratory measurements, vital signs, diagnosis and procedures codes, care givers observations and notes, fluid balance, medications and imaging reports. Data stored in MIMIC-III are related to 7.870 neonates, which means patients having an age under 28 days, and 38.597 distinct adults aged 16 years and over. The database has 26 tables and 324 attributes. Two million rows is the number of unstructured textual data existing in MIMIC-III and representing analyses and notes of various healthcare providers. In addition, we find 380 laboratory measurements and an average of 4.579 charted observations.

B. Data Preprocessing

After data collection, preprocessing is an important step in order to have data that strives for perfection and completeness which make it reliable for prediction tasks [26]. In order to reach such data, dealing with issues like missing values, duplication and normalization is a crucial task and it's the purpose of this section.

a) Data description and analysis: The path to the final dataset used in this study required a great preparation. It was long, packed with work and panoply of choices. Since neonates mortality prediction is the target in this work, only newborns with an age under 28 days were selected from MIMIC-III. 7.867 neonates was the size of the cohort comprising only 66 cases of death. Each patient from the selected cohort may have more than one admission and each admission may require several ICU stays, then we limited ourselves to patients' first stay in the ICU.

The final selection criterion was a kind of investigation about descriptors having impact on mortality based on previous studies and doctors' opinion. In fact, and as we mentioned above, prematurity is the major cause of neonatal mortality especially with a Gestational Age <28 weeks and a Birth Weight <1000 grams or less [11]-[12]. The APGAR score [27], which is a simple birth observation to evaluate newborn vitality, is also a good mortality predictor especially before 32 weeks of gestational age [28]. This score, can be estimated after 1, 5 and 10 minutes of birth. Only patients with APGAR scores under 10 minutes, which means scores for 1

and 5 minutes were kept, because the information was not available to all patients beyond this value at the time of this study. The final dataset consists of 800 patients counting 60 cases of in-hospital death and 740 survivals.

For mortality prediction, this work uses data collected at admission and during a time interval of 2 hours after admission to the ICU. Every patient record was composed by 7 general descriptors and a list of 164 time series variables. Static descriptors were collected automatically or manually from ADMISSIONS, ICUSTAYS, PATIENTS and NOTEEVENTS tables. Time series variables were collected from the CHARTEVENTS table, once, more than once or not at all. "ID" (a single integer identifying each admission to the ICU), "Gender", "Age" (days), "GA" (Gestational Age in weeks), "BW" (Birth Weight kg), "APGAR1" score for the first minute after birth and "APGAR5" score for the 5th minute since birth are the 7 descriptors. A timestamp indicating the time in hours and minutes is associated with each observation since admission to the ICU. From the list of time series variables, 58 appearing in 10% or more of the patient's records were kept. Table I shows the list of time series variables collected for this work, after deleting duplicated values, organized based on their coverage and appearing in 10% or more of the patient's records.

Finally, the 7 static descriptors ID, Age, Gender, BW, GA, APGAR1 and APGAR5 were kept as features. Similarly, attributes like Length, Present Weight, Head Circ, Day of Life were considered as features. On the other hand, we retained the average, the maximum and the minimum values for every time stamped variable in the time-series variables' final list.

a) Missing values: In a hospital environment and specifically in ICUs, a large amount of variables are measured. However, these measures are not always conducted and available at the same time for any patient and this is the reason for the frequent problem of missing data especially in early hours of admission. Unfair prediction or biased results can occur due to this issue. Thus, it's of interest to solve this problem from the preprocessing phase.

In the literature, three types of missing values exist: (1) Missing At Random (MAR), (2) Missing Completely At Random (MCAR) and (3) Not Missing At Random (NMAR) [29]-[30]. To address this problem, a multitude of approaches, which can be grouped into missing value imputation and case deletion, have emerged. Ignoring records containing missing values seems to be the most traditional approach and this is what is proposed by case deletion through its two deletion techniques working on MCAR only: Casewise and Listwise [29]. On the other hand, under the umbrella of missing value imputation, there are plenty of methods such as K-nearest Neighbor Imputation (KNNI), Concept Most Common Method, K-means Imputation, Regression Imputation, Expectation Maximization Imputation (EMI) and Multiple Imputation [29]. In the same context, Che et al. [31] used Recurrent Neural Networks (RNN) in their proposed solution to solve the problem of missing values in time series data. But, among all these techniques, using the mean or the average to replace missing values remains the Most Common Method in the case of Value Imputation.

TABLE. I. VARIABLES COLLECTED IN THE 2 FIRST HOURS OF ADMISSION TO ICU HAVING COVERAGE >= 10%

Variables	Coverage (%)
Heart Rate	98.5
SaO2	96.54
Resp Rate	94.81
HR Alarm [High]	88.81
HR Alarm [Low]	88.81
SaO2 Alarm [High]	87.43
SaO2 Alarm [Low]	87.43
BP Cuff [Diastolic]	83.85
BP Cuff [Mean]	83.85
BP Cuff [Systolic]	83.85
Temp (Skin temperature)	82.35
TIW(Temp/Iso/Warmer)	80.96
Glucometer (Glucose meter)	73.01
Temp Axillary	69.32
PH	57.09
FIO2 (Fractional Inspired Oxygen)	55.59
WBC (White Blood cell Count)	51.21
RBC(Red Blood cell Count)	50.4
BASOs	49.37
Eosinophils	49.37
LYMPHS	49.37
MONOs	49.37
NEUTS	49.37
Platelet	49.1
BANDS	48.9
Hematocrit	48.86
HGB	47.4
Flowrate	46.6
Humidity Temp [Meas]	44.75
PEEP	44.75
FIO2 [Meas]	44.41
Mean PAW	44.18
Polys	43.83
Head Circ	42.91
PEEP Alarm	42.45
High Pressure Relief	41.06
Mean PAW [Meas]	40.95
FIO2 Alarm [Low]	40.14
Total Fluids	37.37
Length	33.79
Temp Rectal	33.56
ETT Size (endotracheal tube size)	33.1
Present Weight	32.41
ETT Taped	31.26
PIP	25.84
Breath Rate	25.72
pCO2	25.61
Base Excess	25.49
pO2	25.49
Inspiratory Time	25.26
FIO2 Alarm [High]	25.14
PIP Alarm	23.53
TCO2	16.84
Day of Life	15.69
Sensitivity	13.49
Survanta	12.69
pH (Art)	12.34
Vt(Ventilator)	10.96

For the used dataset in this work, proceeding by two ways to solve the problem of missing data was the choice. Defining the rate of missing data for every variable and removing the ones having a rate over 50% was the first step. It allowed us to keep only 24 variables including the 7 initial descriptors. Then, for each remaining variable, the maximum, the minimum, the mean and the standard deviation were calculated and missing data were replaced by the mean value.

b) Class imbalance: Class imbalance is another quite common challenge in work with machine learning techniques. Indeed, in the medical field and when it comes to mortality prediction, we usually find ourselves in the case where only few instances belongs to the most significant class in learning task and this is what is named imbalanced classes. To solve this problem, several methods exist and they can be divided into three approaches [32]:

- Data level approach
- Algorithm level approach and
- Cost-sensitive approach.

Methods of the first approach, which is the followed one in this study, are under sampling, oversampling and hybrid. In this present work, given the size of dataset especially the number of cases of mortality and since some variables were eliminated in previous stages, oversampling was the solution. This technique increases the minority class records which are the class of dead patients.

C. Features Selection

After dealing with missing data and class imbalance, keeping only relevant attributes is a building stone in design of predictive models. Features selection aims to produce unaltered subset of attributes coming from the original variables for efficient prediction. The most important objectives of features selection are:

- Avoiding model overfitting
- Improving model performance (prediction performance and cluster detection).
- Reducing computation time
- Providing more cost-effective models
- Better generalizations models

According to [33]-[34], feature selection methods can be divided into three categories: filter methods, wrapper methods and embedded methods.

The first category, which is filter methods, includes CiS, Fish, Ttest, Info and Gini methods [34]. Variable selection is made regardless of the learning algorithms. Computation time effectiveness and robustness against overfitting are filter methods advantages [35]. On the other hand, selecting redundant variables is the disadvantages of these methods which make them more used in data preprocessing.

Under the umbrella of wrapper methods [34], which are the second category, we found search strategies like best-first search, genetic algorithms, hill-climbing search, sequential search and branch-and-bound search. They are proposed to reach optimal local learning performance. But, in practice, wrapper methods are rarely used despite allowing for the detection of possible interactions between variables thanks to variables subsets evaluation. Reasons of their rare use are computation time and overfitting [34]-[36]. In fact, with large variables number, computation time becomes important. The second disadvantage, increasing overfitting, takes place when we have insufficient observations.

The last category, which is embedded methods [35], is a kind of benefits combination of the previous techniques. In fact, there is no iterative feature evaluation although the inclusion of learning algorithm interactions. Regularization models are famous in embedded methods thanks to fitting errors minimization. They also force feature coefficients reduction to fit with learning models.

In this work, we tested two solutions for features selection in order to reach best features to use in classification to get a better performance. The first one is RFECV (Recursive Feature Elimination with Cross Validation). It's an iterative procedure and an instance of backward feature elimination aiming to train the classifier, compute the features' ranking criterion [37] and remove the ones having the smallest ranking criterion .The second one is tree based feature selection. This solution uses the feature importance method. Every feature has a feature importance attribute and the higher is the attribute the most important is the feature. Finally, to train the model, Random Forest (RF) was used as a classifier and its accuracy was used to compare the feature selection techniques.

D. Prediction Model

For the neonates' short-term mortality prediction after admission to the NICU, the proposed approach consists of three parts. Selecting the most relevant features is the first step in the mortality prediction model building. After that, classifying patients into mortals and survivals based on machine learning algorithms to select the best performing classifier is the second part. Tested algorithms were Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) [38]-[39].The third part is the build of the multi-label classification model to predict the hour of mortality of patients previously classified as mortals using Galaxy-Random Forest classification [40]. The global proposed approach is schematized in Fig. 1.

Galaxy-X Is a novel approach. Its purpose is multi-class classification. It's suitable for open-set recognition problems. Indeed, existing methods predicting outcomes in a closed-set of labels classify unknown instances based on known training classes. In such a case, misclassification of instances from unseen classes can occur. It's what is called open-set classification. In [40] this problem is treated by distinguishing instances that belongs to unknown classes from those similar to classes already seen. This distinction is possible thanks to the creation of a hyper-sphere having minimum bounding for each class of the training set. All instances of known classes will be included into this hyper-sphere [40]. Based on this

method, a given test instance will be classified in a space composed by known training classes and the unknown class. So, if the test instance belongs to unseen classes, it will take place in the unknown class.

To evaluate classifiers performance, Cross-Validation, Accuracy and Receiver Operating Characteristics (ROC) were the used metrics [41] in addition of Leave-P-Class-Out-CrossValidation [40] regarding the specificity of open-set classification.

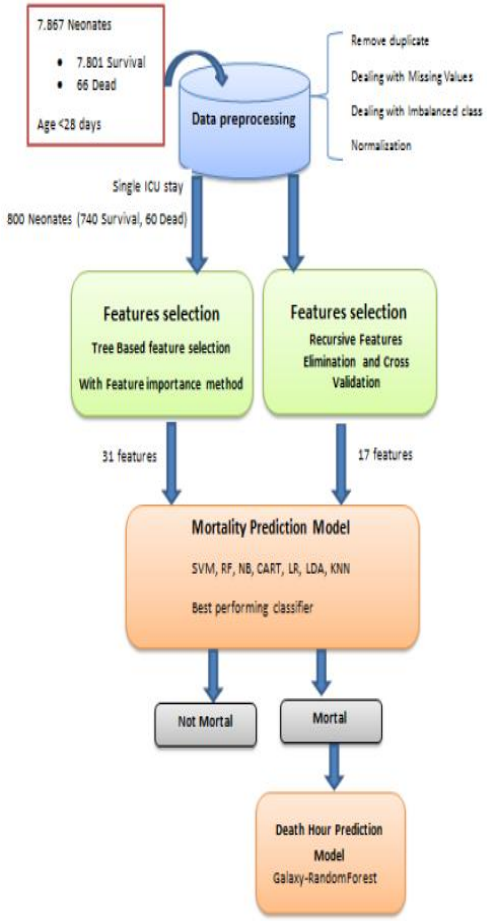

Fig. 1. Proposed Approach for Short-Term Mortality Prediction in NICU.

IV. Results

Fig. 2 shows the final list of 24 variables composed by 17 time series variables with missing values rates under 50% and 7 general descriptors. The number of attributes used in this work as a base for features selection goes from 24 to 58 because the 17 most frequent time series variables were presented each by minimum, maximum and average.

Based on the final list of 58 attribute and Random Forest classifier accuracy = 0.97 used in model training in order to compare features selection techniques, the first method of features selection, which is RFECV, allowed us to find the number of features we need for a better accuracy as well as the best ones among them. The accuracy scoring was proportional to the number of correct classifications. Then, the result was 17 as the optimal number of features was with 5 fold cross validation. Fig. 3 shows the number of features and the corresponding cross validation scores. The features selected are: Age, BW, GA, APGAR1, APGAR2, SAO2 (min, max and avg), Respiration Rate (min, max and avg), BP Cuff [Systolic] (max and avg), TIW (min and avg) and FIO2 (min and avg).

The second solution, which is the tree based feature selection, is more flexible in the final choice of features. In fact, a score specifying the importance of each feature is calculated and the choice is made based on this importance. Thanks to this principle, many datasets can be generated and used later with others classification algorithms. A further advantage of this solution is the ability to implicate physicians in decision making about relevant features apart from the importance score. Table II shows the list of the 31 selected features according to the importance for the purpose of classification and based on a threshold equal to 0.01 and physicians opinion.

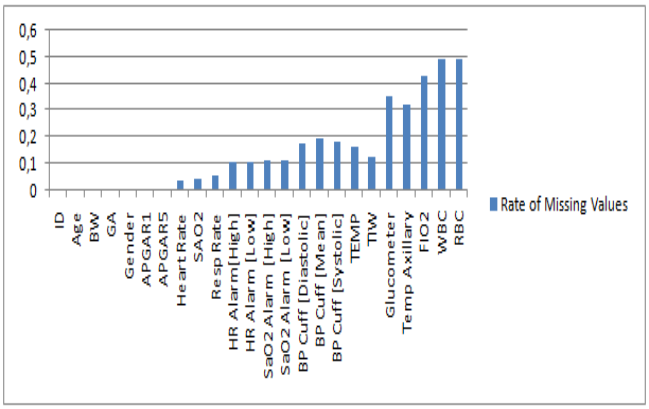

Fig. 2. Most Frequent Variables.

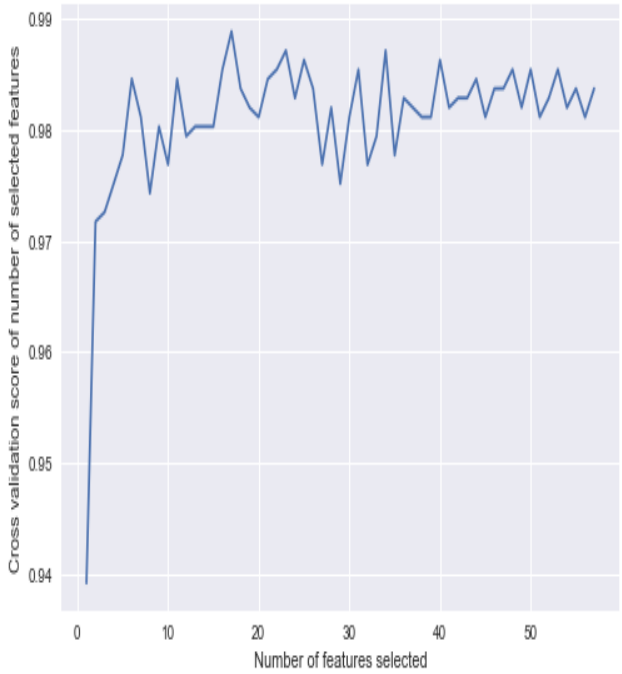

Fig. 3. Cross Validation Score of Number of Selected Features in RFECV.

TABLE. II. LIST OF 31 SELECTED FEATURES BASED ON FEATURES IMPORTANCE METHOD

Features	Description
AGE	AGE
GA	Gestational Age
BW	Birth Weight
APGAR1	APGAR for 1st 1 minute after birth
APGAR5	APGAR for 1st 5 minutes after birth
AVGHR	Average Heart Rate
MINSAO2	Minimum SAO2
MAXSAO2	Maximum SAO2
AVGSAO2	Average SAO2
MAXRR	Maximum Resp Rate
AVGRR	Average Resp Rate
MINSAO2AH	Minimum SAO2 Alarm [High]
MAXSAO2AH	Maximum SAO2 Alarm [High]
AVGSAO2AH	Average SAO2 Alarm [High]
MINBPCD	Minimum BP Cuff [Diastolic]
AVGBPCD	Average BP Cuff [Diastolic]
MINBPCM	Minimum BP Cuff [Mean]
MAXBPCM	Maximum BP Cuff [Mean]
AVGBPCM	Average BP Cuff [Mean]
MINBPCS	Minimum BP Cuff [Systolic]
MAXBPCS	Maximum BP Cuff [Systolic]
AVGTEMP	Average TEMP
MINTIW	Minimum TIW
AVGTIW	Average TIW
MINGLUCO	Minimum Glucometer
AVGGLUCO	Average Glucometer
MINTA	Minimum Temp Axilary
MAXTA	Maximum Temp Axillary
MINFIO2	Minimum FIO2
MAXWBC	Maximum WBC
MINRBC	Minimum RBC

To predict mortality based on data collected at admission and during the first 2 hours of stay into the NICU, several machine learning algorithms such as CART, LDA, KNN, LR, NB, RF and SVM were applied for patients' classification to identify mortal cases of those who are not. With 10-fold Cross Validation, for better representation of the whole dataset, LDA gave the best accuracy using the 31 most important features and LR gave the best accuracy using the 17 best features. The top three classifiers with 31 features are respectively LDA, LR, and KNN. The top three classifiers with 17 features are respectively LR, LDA and RF. But, according to accuracy, LR outperformed the others methods. Table III presents results of this work in terms of accuracy and standard deviation of each classifier based on sets of features generated by the two methods of features selection.

To summarize classifiers performance, another evaluation metric which is Receiver Operating Characteristics (ROC) is used. It's a kind of trade-offs between True Positive and False Positive error rates. And as observed in Fig. 4, LDA gave the best performance with 31 features (AUROC = 0.97) which was not the same case with the 17 best features, as shown in Fig. 5, KNN gave the best performance (AUROC = 0.97).

Finally, after classifying patients based on mortality, the Galaxy-X method was adapted to predict the patient death hours, of the classified cases as mortals, using 4 classes {1, 2, 3, 4} ∪ Unknown. The first class is that of patients whose mortality occurs between 8 and 24 hours after admission. Class 2 is composed by patients whose mortality takes place after one day of admission i.e. between 24 and 48 hours. Class 3 comprises patients whose mortality hour occurs after two days of admission meanings an interval equal to] 48, 72] hours. Class number 4 comprises patients whose mortality happens after 72 hours of admission to the ICU. The Unknown class is the last one. It encompasses patients whose mortality can happen at admission or during the first hours of admission (0-8).

It also contains misclassified patients as mortals. With Leave-P-Class-Out-CrossValidation (P=2), the f1-score of Galaxy Random Forest was 0.87.

TABLE. III. CLASSIFIERS ACCURACIES WITH 31 FEATURES AND WITH 17 FEATURES

	Classifiers with 31 important features		Classifiers with 17 best features	
Classifier	***Accuracy***	***StDev***	***Accuracy***	***StDev***
LR	0.940260	0.024331	0.956526	0.020104
LDA	0.947435	0.034886	0.954740	0.020186
KNN	0.931136	0.011005	0.922078	0.024524
CART	0.916688	0.028325	0.916623	0.033671
NB	0.793766	0.069147	0.931169	0.013546
SVM	0.931136	0.013683	0.923929	0.025387
SVM_RBF	0.931136	0.013683	0.923929	0.025387
SVM_POLY	0.896753	0.032194	0.918442	0.029643
RF	0.931071	0.022867	0.943864	0.028626

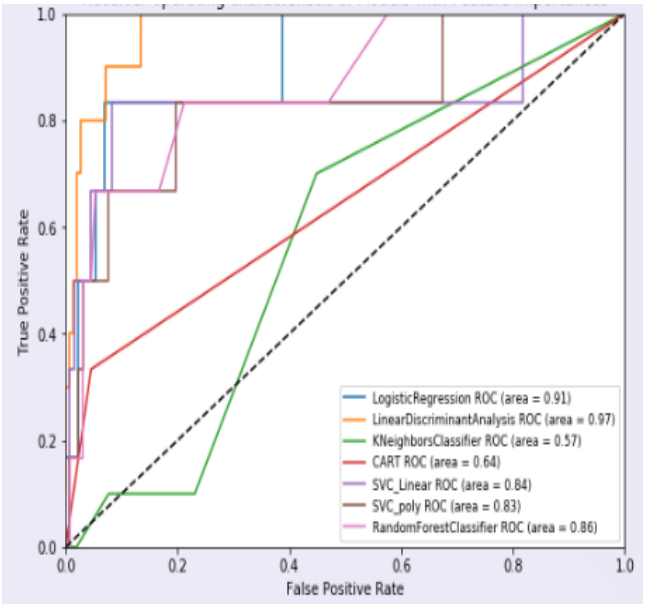

Fig. 4. ROC of Models with 31 Features.

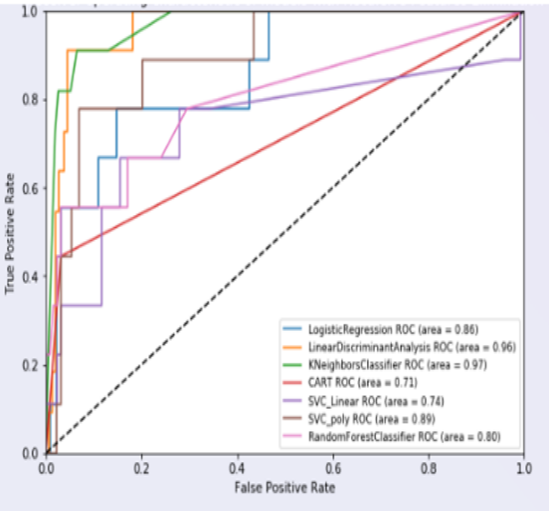

Fig. 5. ROC of Models with 17 Features.

V. Discussion

Scoring systems were helping doctors to calculate illness severity or mortality risk. In addition to these systems and for years, another alternative has emerged and has proven itself in the prediction of panoply of outcomes within the ICU. It is the machine learning techniques. They seek to help doctors make the right decisions at the right time.

Expecting CRIB II, which predicts neonate mortality after the first hour of admission to the ICU; these models predict mortality over a time span of 12 to 72 hours or more. It's in this context that the present work fits. Indeed, it aims to explore short-term mortality prediction after admission into the NICU. This prediction is important and has great impact on decision-making to reduce costs and to improve the resources management in these parts of hospitals where patients with the most critical health conditions reside.

The used database is the MIMIC-III, which houses different types of data of 7.867 neonates including 66 cases of mortality. From this initial dataset, data of 800 patients having age <28 days and only one admission to the ICU were selected. The data collected, include 7 general descriptors (ID, Gestational Age, Birth Weight, Gender, Age, Apgar 1 min and Apgar 5 min) collected at admission and time-series variables collected in the first two hours of NICU admission. Missing values and imbalanced classes were problems we encountered during the build of the model due to the very early time of data selection. After that, we used Features Importance and Recursive Features Elimination with Random Forest Classification to select the most relevant features. The second step of this work was the built of the mortality prediction model by testing an ensemble of machine learning techniques with the two sets of features generated in the previous step in order to find the best performing classifier. The best performing classifier with the 31 most important features was LDA. And KNN was the best performing classifier with the 17 best features. LDA was the one kept. The last step of the present approach for short-term mortality prediction was the built of the mortality hour prediction model based on a dataset containing only cases classified as mortal in the second step. Galaxy-RandomForest was used to build this model.

With an accuracy equal to 0.95, an AUROC equal to 0.96 with 17 variables and an AUROC equal to 0.97 with 31 variables LDA outperformed scoring systems such SNAP (AUROC=0.90), SNAP-PE (AUROC=0.93), CRIB (AUROC =0.90) and CRIB II (accuracy=0.867). Moreover, LDA outperformed the state of the art classifiers like SVM, RF, NB and CART and this was a very interesting outcome of this work from a machine learning perspective.

VI. Conclusion

Predicting outcomes in medical context based on machine learning techniques is the alternative that has gained ground compared with traditional solutions based on scoring systems. This is due to limitations of the latter and advantages and flexibility offered by recent solutions. Among these outcomes, mortality is one of the most predicted one especially in ICUs. Thus, this work fits into this context by proposing a new approach for short-term mortality prediction in NICU.

Through this work, three important results can be specified. First, to our knowledge, for short-term neonatal mortality prediction based on machine learning techniques and using time-series variables, it's the first work in this field after only 2 hours of admission into the NICU. Second, in terms of time and AUROC, it compares favorably with the state of the art classifiers (RF, SVM, CART and NB) and scoring systems (CRIB, SNAP, SNAP-PE, SNAP-II and SNAPPE-II). Finally, predicting death hours with Galaxy-Random Forest. This method can detect misclassifications and move them to the unknown class which can ameliorate the classification performance.

This work is an opportunity to consider the integration of prenatal data from mother and baby during pregnancy as these may be available at the child's birth and will not delay the prediction time. This is a perspective that can be implemented in future works in order to improve results. The establishment of a decision support system for short-term mortality prediction in NICU based on our proposed approach is another perspective for the present work.

References

[1] R. Davoodi and M. H. Moradi, "Mortality Prediction in Intensive Care Units (ICUs) Using a Deep Rule-basedFuzzy Classifier". J Biomed Inform. 2018 Mar;79:48-59. doi: 10.1016/j.jbi.2018.02.008. Epub 2018 Feb 19.

[2] S. M. Altawalbeh, R. Abu-Su'Ud, Q. Alefan, S. M. Momany, and S. L. Kane-Gill, "Evaluating Intensive Care Unit Medication Charges in a Teaching Hospital in Jordan ". Expert Review of Pharmacoeconomics & Outcomes Research 2019, DOI: 10.1080/14737167.2019.1571413

[3] T. J. Pollard and L. A. Celi, "Enabling Machine Learning in Critical Care". ICU management & practice. 2017;17(3):198-199.

[4] A. Belard, T. Buchman, J. Forsberg, B. K. Potter, C. J. Dente, A. Kirk, and E. Elster, "Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care". Journal of clinical monitoring and computing 2017. Volume 31, pages 261-271.

[5] S. S. Harsha and B. R. Archana, "SNAPPE-II (Score for Neonatal Acute Physiology with Perinatal Extension-II) in Predicting Mortality and Morbidity in NICU". J Clin Diagn Res. 2015;9(10):SC10–SC12.

[6] J. Xie, B. Su, C. Li, K. Lin, H. Li, Y. Hu, and G. Kong, "Review of modeling methods for predicting in-hospital mortality of patients in Intensive Care Unit". J. Emerg. Crit. Care Med. 2017 ; 1 (8), pp. 1-10.

[7] J. S. Dorling, D. J. Field, and B. N. Manktelow, "Neonatal disease severity scoring systems". Archives of disease in childhood. Fetal and neonatal edition 2005 ; 90 (1) :F11-F16.

[8] J. S. Malak, H. Zeraati, F. S. Nayeri, R. Safdari, and A. D. Shahraki, "Neonatal intensive care decision support systems using artificial intelligence techniques: a systematic review". Artificial Intelligence Review, 2018, pp. 1-20.

[9] N. Veith and R. Steele, "Machine Learning-based Prediction of ICU Patient Mortality at Time of Admission". Conference Paper · March 2018. DOI: 10.1145/3206098.3206116

[10] A. E. W. Johnson, T. J. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database". Scientific Data 2016, Volume 3.

[11] J. H. Park, Y. S. Chang, S. Sung, S. Y. Ahn, and W. S. Park, "Trends in Overall Mortality, and Timing and Cause of Death among Extremely Preterm Infants near the Limit of Viability". PloS one. 2017;12(1):e0170220. pmid:28114330

[12] R. M. Patel, S. Kandefer, M. C. Walsh, E. F. Bell, W. A. Carlo, A. R. Laptook, et al . , "Causes and timing of death in extremely premature infants from 2000 through 2011". The New England journal of medicine. 2015;372(4):331–40. pmid:25607427

[13] B. O. Valeri, C. M. Gasapardo, F. E. Martinez, and M. B. M. Linhares, "Effectiveness of sucrose used routinely for pain relief and neonatal clinical risk in preterm infants : a nonrandomized study". Clin.J.Pain. 2018; 34 (8), 713—722.

[14] Z. M. Ezz-Eldin, T. A. Hamid, M. R. Youssef, and Hel-D. Nabil, "Clinical Risk Index for Babies (CRIBII) scoring system in prediction of mortality in premature babies". J Clin Diagn Res, 9 (2015), pp. SCo8-SC11.

[15] G. Parry, J. Tucker, W. Tarnow-Mordi, and UK Neonatal Staffing Study Collaborative Group. "CRIB II: an update of the clinical risk index for babies score". Lancet. 2003;361(9371):1789–1791 [British edition]

[16] M. Beltempo, P. S. Shah, X. Y. Ye, J. Afifi, S. Lee, D. D. McMillan and on behalf of the Canadian Neonatal Network Investigators, "SNAP-II for prediction of mortality and mobidity in extremely preterm infants". The Journal of Maternal-Fetal & Neonatal Medicine 2019, 32:16, 2694-2701, DOI: 10.1080/14767058.2018.1446079

[17] O. P. Rinta-Koski, S. Särkkä, J. Hollmén, M. Leskinen, and S. Andersson , "Gaussian process classification for prediction of in-hospital mortality among preterm infants". Neurocomputing.Volume 298, 12 July 2018, Pages 134-141

[18] Y. Liu, M. Traskin, S. A. Lorch, E. I. George, and D. Small, "Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance". Health Care ManagSci.,2015; 18(1):58–66.

[19] M. Frize, J. Gilchrist, H. Martirosyan, and E. Bariciak, "Integration of outcome estimations with a clinical decision support system: application in the neonatal intensive care unit (NICU)". In: Paper presented at the medical measurements and applications (MeMeA), 2015 IEEE international symposium on. pp 175–179. https://doi.org/10.1109/memea.2015.7145194

[20] L. M. Saadah, F. D. Chedid, M. R. Sohail, Y. M. Nazzal, M. R. AlKaabi, and A. Y. Rahmani AY, "Palivizumab prophylaxis during nosocomial outbreaks of respiratory syncytial virus in a neonatal intensive care unit: predicting effectiveness with an artificial neural network model". Pharmacother J Hum Pharmacol Drug Ther 2015; 34(3):251–259.

[21] F. R. Cerqueira, T. G. Ferreira, A. de Paiva Oliveira, D. A. Augusto, E. Krempser, H. J. Corrêa Barbosa, and R. Siqueira-Batista, " NICeSim: an open-source simulator based on machine learning techniques to support medical research on prenatal and perinatal care decision making". ArtifIntell Med 2014;62(3):193–201. https://doi.org/10.1016/j.artmed. 2014.10.001

[22] J. Gilchrist, M. Frize, C. M. Ennett, and E. Bariciak, "Neonatal mortality prediction using real-time medical measurements". In: Paper presented at the medical measurements and applications proceedings (MeMeA), 2011 IEEE international workshop on. pp 65–70. https://doi.org/10.1109/memea.2011.5966653

[23] D. Townsend and M. Frize, "Complimentary artificial neural network approaches for prediction of events in the neonatal intensive care unit". In: Paper presented at the engineering in medicine and biology society, 2008. EMBS 2008.30th annual international conference of the IEEE.pp 4605–4608

[24] M. Frize, L. Yang, R. C. Walker, and A. M. Connor, "Conceptual framework of knowledge management for ethical decision-making support in neonatal intensive care". IEEE Trans InfTechnol Biomed 2005;9(2):205–215. https: //doi.org/10.1109/TITB.2005.847187

[25] M. Frize, E. Bariciak, and J. Gilchrist, "PPADS: physician-PArent decision-support for neonatal intensive care". In: Paper presented at the MedInfo. 2013; pp 23–27

[26] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects". Big Data Anal., 1 (2016), p. 9

[27] K. L. Watterberg, S. Aucott, W. E. Benitz, J. J. Cummings, E. C. Eichenwald, J. Goldsmith, et al., "The Apgar Score". Pediatrics 2015; 136:819–822

[28] T. A. Manuck, M. M. Rice, J. L. Bailit, W. A. Grobman, U. M. Reddy, R. J. Wapner, et al., "Preterm neonatal morbidity and mortality by gestational age: a contemporary cohort". Am J Obstet Gynecol. 2016;215:103e1–e14.

[29] S. Gupta and M. K. Gupta, "A Survey on Different Techniques for Handling Missing Values in Dataset". IJSRCSEIT 2018 ,Volume 4, Issue 1,ISSN : 2456-3307

[30] A. Awad, M. Bader-El-Den, J. McNicholas, and J. Briggs, "Early Hospital Mortality Prediction of Intensive Care Unit Patients Using an Ensemble Learning Approach". <![CDATA[International Journal of MedicalInformatics]]>(2017),https://doi.org/10.1016/j.ijmedinf.2017.10.002

[31] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu "Recurrent Neural Networks for Multivariate Time Series with Missing Values". Scientific Reports 2018,volume 8, Article number: 6085

[32] S. Maheshwari, R. C. Jain, and R. S. Jadon, "A Review on Class Imbalance Problem: Analysis and Potential Solutions". IJCSI International Journal of Computer Science Issues 2017, Volume 14, Issue 6.

[33] G. Chandrashekar and F. Sahin, "A survey on feature selection methods". Computer and Electrical Engineering 40(2014)16-28

[34] C. Zhang, J. Bi, and P. Soda, "Feature selection and resampling in class imbalance learning: Which comes first? An empirical study in the biological domain", 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 2017, pp. 933-938.

[35] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu H, "Feature Selection: A Data Perspective". ACM Computing Surveys (CSUR), 2017, 50(6): 94:1-94:45

[36] A. Jović, K. Brkić, and N. Bogunovi, "A review of feature selection methods with applications". In: Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on. IEEE, pp. 1200–1205

[37] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machine". Machine Learning 46(1-3)(2002) 389-422.

[38] M. I. Jordan and T. M. Mitchell, "Machine learning: trends, perspectives and prospects". Science 349 (2015), 255-260.

[39] V. K. Garg and A. Kalai, "Supervising Unsupervised Learning". arXiv preprint arXiv:1709.05262, 2017 - arxiv.org.

[40] W. Dhifli and A. B. Diallo, "Toward an Efficient Multi-class Classification in an Open Universe". International Conference on Machine Learning and Data Mining (MLDM 2016).

[41] K. J. Danjuma , "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients". IJCSI International Journal of Computer Science Issues 2015, Volume 12, Issue 2.