A Neonatal Sepsis Prediction Algorithm Using Electronic Medical Record (EMR) Data from Mbarara Regional Referral Hospital (MRRH)

Dennis Peace Ezeobi (helendennis4u@gmail.com)
Mbarara University of Science and Technology
Dr. William Wasswa
Mbarara University of Science and Technology
Dr. Angella Musimenta
Mbarara University of Science and Technology
Dr. Stella Kyoyagala
Mbarara National Referral Hospital

Research Article

Keywords: Neonatal sepsis prediction, Screening parameters, Predictive algorithm, Supervised Machine Learning, Electronic medical record (EMR), Cross-Industry Standard Process for Data Mining (CRISP-DM) model

Posted Date: July 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1353776/v2

Abstract

Background

Neonatal sepsis is a significant cause of neonatal death and has been a major challenge worldwide. The difficulty in early diagnosis of neonatal sepsis leads to delay in treatment. The early diagnosis of neonatal sepsis has been predicted to improve neonatal outcomes. The use of machine learning techniques with the relevant screening parameters provides new ways of understanding neonatal sepsis and having possible solutions to tackle the challenges it presents. This work proposes an algorithm for predicting neonatal sepsis using electronic medical record (EMR) data from Mbarara Regional Referral Hospital (MRRH) that can improve the early recognition and treatment of sepsis in neonates.

Methods

A retrospective analysis was performed on datasets composed of de-identified electronic medical records collected between 2015 to 2019. The dataset contains records of 482 neonates hospitalized in Mbarara Regional Referral Hospital, Uganda. The proposed algorithm implements Support Vector Machine (SVM), Logistic regression (LR), K-nearest neighbor (KNN), Naïve Bayes (NB), and Decision tree (DT) algorithms, which were trained, tested, and compared based on the acquired data. The performance of the proposed algorithm was evaluated by comparing it with the physician's diagnosis. The experiment used a Stratified K-fold cross-validation technique to evaluate the performance of the models. Statistical significance of the experimental results was carried out using the Wilcoxon Signed-Rank Test.

Results

The results of this study show that the proposed algorithm (with the lowest Sensitivity of 95%, lowest Specificity of 95%) outperformed the physician diagnosis (Sensitivity = 89%, Specificity = 11%). SVM model with radial basis function, polynomial kernels, and DT model (with the highest AUROC values of 98%) performed better than the other models in predicting neonatal sepsis as their results were statistically significant.

Conclusions

The study provides evidence that the combination of maternal risk factors, neonatal clinical signs, and laboratory tests effectively diagnose neonatal sepsis. Based on the study result, the proposed algorithm can help identify neonatal sepsis cases as it exceeded clinicians' sensitivity and specificity. A prospective study is warranted to test the algorithm's clinical utility, which could provide a decision support aid to clinicians.

Author Summary

About 2.5 million neonates die worldwide every year, and most of these deaths occur in low-resource settings. The estimated neonatal mortality rate (NMR) in Sub-Saharan Africa (SSA) is 28 per 1000 live births, with Uganda struggling with a high rate of 20 per 1000 live births. Sepsis is one of the major causes of this neonatal deaths worldwide, accounting for 17% of neonatal deaths in Uganda. Due to the difficulty in the early diagnosis of neonatal sepsis which leads to delay in treatment, there is a growing need to discover new and improved way of diagnosing neonatal sepsis. We have proposed a diagnostic algorithm for the prediction/diagnosis of sepsis in neonates. We have tested this algorithm using retrospective neonatal data and machine learning algorithms. The algorithm outperformed the physician’s diagnosis. Our algorithm has provided insight into the prediction and diagnosis of neonatal sepsis. It has shown that the combination of maternal risk factors, neonatal clinical signs, and laboratory tests effectively diagnose neonatal sepsis.

Background

About 2.5 million neonates die worldwide every year, and most of these deaths occur in low-resource settings (2,3). It is estimated that the neonatal mortality rate (NMR) in Sub-Saharan Africa (SSA) is 28 per 1000 live births, with Uganda struggling with a high rate of 20 per 1000 live births (2,4). The pediatric consensus definition of sepsis is systemic inflammatory response syndrome (SIRS) in the presence of or due to suspected or proven infection (5). The SIRS cause damage to the body and can quickly advance to severe sepsis, multi-organ system failure, and death (6,7). Therefore, early recognition and prompt treatment, which have been predicted to improve clinical management of sepsis, is required to reduce the morbidity and mortality of neonatal sepsis (8–12).

Neonatal sepsis is a significant cause of neonatal mortality and morbidity worldwide (13–15), and a majority of the morbidity and mortality from sepsis is preventable. Sepsis is one of the major causes of neonatal deaths in Uganda, like in other Sub-Saharan African countries, accounting for 17% of neonatal deaths in Uganda (1). Several authors classify neonatal sepsis as a community-and hospital-acquired instead of early-and late-onset in developing countries. Neonatal sepsis is usually classified as early-onset (<48–72h) and late-onset sepsis (>48–72h), depending on the age at onset (16,17). About 30-50% of survivors of neonatal sepsis end up with major long-term impairments and also faced with prolonged hospitalization, chronic lung disease, and neurodevelopmental disabilities (18–21). Recent data highlight the costs and burdens of sepsis, as it remains the most expensive cause of hospitalization (22–25). The development of clinical trials and global recommendations is hindered by the population's susceptibility, lack of consensus in definitions, and variability between regions (26). Multiple challenges in diagnostic and treatment decisions are faced by physicians caring for infected neonates. To date, there have been just modest improvements in terms of sepsis outcomes in neonates despite the increased understanding of its pathophysiology and efforts to improve clinical decision support in intensive care (27). The after-effect of sepsis-infected adults and children's impending intervention is receiving attention in recent studies (28,29).

Despite the explored significance of early treatment of sepsis, there are still unresolved challenges due to impeding recognition and intervention of sepsis (15,27,30–34). Neonates have a non-specific clinical presentation that overlaps with other neonatal disease processes. Also, laboratory tests have suboptimal diagnostic accuracy, which makes a rapid diagnosis of neonatal sepsis difficult. The blood culture, the standard gold test for neonatal sepsis diagnosis, faces the challenge of insufficient blood volume for blood culture and a low amount of invading microorganisms in the blood; this usually generates false-negative results (21,35). Despite negative culture results, neonates presumed to have sepsis are kept on a longtime antibiotic treatment. Studies have previously applied machine learning and statistical modeling techniques to tackle the problems related to sepsis recognition and intervention (36–39).

Several studies have used machine learning models to predict if a patient is at risk of developing sepsis or the onset time (36,38,40–42). Electronic health record (EHR) data have been used in recent studies to train models to enable early diagnosis of neonatal sepsis (21). HeRO score, which is a statistical prediction model, supported early recognition of neonatal sepsis as it was used to lessen deaths related to sepsis in very low birth weight neonates (<1500 grams) (43). However, in a retrospective study with a larger population, the HeRO score could not identify neonates with sepsis. This suggests that the predictive value is unreliable in clinical practice (44). In recent research, a machine learning model was developed using electronic health record data to recognize early neonatal sepsis in the neonatal intensive care unit. Though the model could predict neonatal sepsis, additional features are still required to improve its performance (21). This resulted from the uncertainty of the adequate screening parameters for the diagnosis of neonatal sepsis. Research that was carried out in India aimed at comparing the neonatal sepsis rapid diagnostic tests with blood culture for specificity and sensitivity. The study pointed out that the use of either two tests or three tests can rule out negative sepsis cases, which will help avoid antibiotics' overuse (35). In another research, the authors focused on assessing the performance of the Adult Sepsis Pathway and algorithm for the Modified St. John Rule, which were compared with qSOFA score algorithm using R scripts (45). Although this work identified algorithms that performed more than qSOFA, adults and children's immunology and physiology differ significantly that these algorithms may not be directly applicable to neonates. Hence, this study seeks to address the difficulty in early diagnosis of neonatal sepsis by developing an algorithm that combines maternal risk factors, neonatal clinical signs, and laboratory tests.

Methods

A Standard approach was implemented, a structured data mining project methodology defined by the Cross-Industry Standard Process for Data Mining (CRISP-DM) developed in 1996 (46,47). The experiment was executed in five phases, which include; business understanding, data understanding, data preparation, modeling, and evaluation. The experimental steps were performed in Python language using scikit-learn library. Scikit-learn (48) is an open-source machine learning library featuring various classification, regression, and clustering algorithms for the python programming language. This research aims to propose an algorithm that can improve the early recognition and treatment of sepsis in neonates using EMR data collected over a time span of four years.

The method is categorized into sub-sections adhering to the CRISP-DM framework, as shown in figure 1 above. The phases are performed based on the previous phase's accomplishments, and the subsections provide detailed information on the experiment.

Business Understanding

This study is a retrospective study, and the key focus of the study is to propose an algorithm that can improve the early recognition and treatment of sepsis in neonates. The design of the study is shown in figure 2 above.

Data Understanding

Secondary data was used in this study, based on the data reliability, suitability, and adequacy of the data. The dataset contains information about sepsis screening parameters from hospitalized neonates collected from the EMR of MRRH that covers a period of 4 years. The Mbarara Regional Referral Hospital is located in rural Uganda and is the main teaching hospital situated adjacent to the Mbarara University of Science and Technology campus. The hospital typically has about 21.3% cases of presumed neonatal sepsis per the pediatric ward's annual admissions, as it is the referral center for southwestern Uganda. A data abstraction tool was developed (see supplementary table 1) to retrieve information essential to the study. The dataset contains records of 482 neonates hospitalized from October 1st, 2015 to September 30th, 2019, that met the inclusion criteria with 38 different neonate screening features (Table 1). The category of neonates of concern is the group with early-onset sepsis (≤48-72h). The predictor variables are both continuous and categorical in nature.

Table 1: Screening parameters information

Variable	Attributes	Type	Description
Maternal risk factors	maternal_febrile	N	Maternal febrile episodes during pregnancy (Count)
	fever_during_labour	C	Fever during labor
	abnormal_vaginal_discharge	C	Abnormal vaginal discharge during pregnancy
	Antibiotic	C	Any antibiotic therapy received by mother in perinatal period
	gest_age	N	Gestation age at birth (weeks)
	place_of_delivery	C	Place of delivery
	mode_of_delivery	C	Mode of delivery
	duration_of_labour	N	Duration of labor
	duration_of_ROM	N	Duration from rupture of membranes to delivery of baby
	foul_smelling_liquor	C	Foul smelling liquor
	rupture_of_mem	C	Rupture of membrane
Neonatal signs	gender	C	Female, Male
	age_days	N	0-3 days old
	fever	C	Fever
	cold_body	C	Cold body
	poor_feeding	C	Poor feeding
	crying_excessively	C	Crying excessively
	weak_cry	C	Weak cry
	lethargy	C	Lethargy
	respiratory_difficulty	C	Respiratory difficulty
	weight	N	Weight
	temperature	N	Temperature
	respiratory_rate	N	Respiratory rate
	heart_rate	N	Heartrate
	tachypnoea	C	Tachypnoea
	apnoea	C	Apnea
Laboratory tests	wbc	N	White blood cells
	neu_count	N	Neutrophils
	lym_count	N	Lymphocytes
	mon_count	N	Monocytes
	eos_count	N	Eosinophils
	bas_count	N	Basophils
	Rbc	N	Red Blood Cells
	platelet_count	N	Platelets
	crp_count	C	C-reactive protein
	blood_culture	C	Blood culture

Table 1 above contains the summary of the dataset, where attributes are the screening parameters, and type represents the data type of each parameter, i.e., 'N' for numeric values and 'C' for categorical values.

Inclusion Criteria

The EMR data from MRRH was used base on the following conditions:

- Have gestational age (GA) of ≥37 weeks.
- Data of each neonate should have at least two observations from each of these variables;
    1. Maternal risk factors (fever during labor, maternal febrile during pregnancy, duration of rupture of membrane, duration of labor, foul odor of the amniotic fluid, and antibiotic treatment received by mother ≤4 hours prior to delivery).
    2. Neonatal clinical signs (heart rate, temperature, respiratory distress, apnea condition, lethargy condition, and feeding difficulty).
    3. Laboratory tests (C-reactive protein, white blood cell count, neutrophil count, and platelet count).
- Availability of a defined neonatal sepsis status report.
- The age at the time of onset should be less than or equal to 48-72h.

Exclusion Criteria

Excluded cases include:

- Bacteria cultures were positive from sources other than blood.
- Positive cultures for viral or fungal pathogens.
- Undefined results due to pending cultures at the time of data extraction.
- Cultures positive for known contaminants.

A detailed overview of the steps carried out for data preparation and feature engineering used in this study is provided in the following section.

Data Preparation

The data preparation phase covers every activity involved in transforming and cleaning the data to make it fit to be used in the modeling phase. The missing values, noise, and outliers present in the data identified during the data understanding phase were removed in data pre-processing. The identified missing values were assigned using the mean value of the feature.

SMOTE Algorithm - Balancing of Dataset

The main problem with the data is that it is highly imbalanced and small in size. There were approximately 22% records with neonatal sepsis as '0', and the rest 78% of the records have neonatal sepsis as '1'. Proceeding to modelling without balancing the data will cause the trained model to be biased and have a high cost of misclassifying minority class.

Sampling techniques such as under-sampling or over-sampling, should be implemented to balance the data set, as shown in figure 3 above.

Looking at the fact that the data set is relatively small in size, this makes all the instances to be highly important, and no information should be at the risk of loss. Therefore, under-sampling is eliminated, and the over-sampling technique will be preferable for the experiment.

The synthetic samples are created in the space, as shown in figure 4 by Synthetic Minority Over-sampling Technique (SMOTE) algorithm, which applies the KNN approach where it selects K-nearest Neighbors and joins them. The algorithm takes the feature vectors and its nearest neighbors, and it computes the distance between these vectors. A random number between (0, 1) is used to multiply the difference, and it is added back to the feature.

Normalization and Standardization: Z-score

Unscaled or unstandardized features are known to make the learning algorithms to predict recklessly. Normalization or standardization, an important step to be carried out before proceeding to model building, is required to ensure that all the feature values are on the same scale. The values of features are standardized from different dynamic ranges into specific ranges through standardization, a preprocessing step. All the parameters are scaled to have zero mean and unit variance, converted by standard score, also known as z-score.

One-Hot Encoding: Categorical Variables

The neonatal sepsis data set has 21 categorical variables out of 38 features. In order to build the SVM model, categorical variables need to be converted into numeric variables as vector machine works on numeric variables. This problem was overcome by constructing a dummy variable from the categorical variables using the Pandas library in python.

Random Forest Classifier: Feature Importance

Finally, each variable's importance in predicting neonatal sepsis was identified with the Random Forest algorithm, an ensemble modeling technique based on iteratively removing variables with low ranking and using cross-validation to assess the learning performance. Each variable was assigned a score to show the importance of the variables in the model. The higher the score, the higher the importance of that particular variable. While variables with a lower score are considered the least important. Data pre-processing techniques assist in extracting more useful information, which helps build a model with higher accuracy and performance.

Modelling

The study developed a diagnostic algorithm for predicting neonatal sepsis, which was used to train the machine learning (ML) algorithms used in this study. Therefore, five supervised ML algorithms, Support Vector Machine (SVM), Logistic regression (LR), K-nearest neighbor (KNN), Naïve Bayes (NB), and Decision tree (DT), were implemented to build the models. SVM is a relatively new classification method that resulted from the collaboration between statistical and machine learning developed by Vapnik et al. in the 1990s (50), whereas the most commonly used prognostic modeling method is LR. The KNN algorithm is used for classification and regression, and it is a non-parametric method. The NB is a probabilistic classifier that performs well in multi-class prediction. Furthermore, DT builds classification or regression models in the form of a tree structure.

SVM, which is also a supervised machine learning technique, is similar to LR, and they are both used for regression and classification problems. The dissimilarity is that SVM models input variables by finding a boundary for the classification of the target variable known as hyper-plane. When the hyper-plane has data points nearest to it, the data points are called support vectors. The removal of these points will lead to the alteration of the dividing hyperplane as they are the data set's critical elements. SVM functions for both regression and classification, respectively.

SVM algorithms find boundaries for classification when there is no possible separation within a high number of input variables, as shown in figure 5 above. The input variables are transformed by increasing the dimensionality of the variable space to generate the separation boundary.

The SVM Linear kernel model, SVM radial kernel model, and SVM polynomial kernel model were built as part of the experiment. Each model is tuned with different values of tuning parameter 'C' and 'γ'. SVM model separates classes that cannot be separated using line or plane but only using kernel function and requires a non-linear region to separate such classes. This transformation of the data into higher dimensional feature space to separate it linearly is known as the kernel trick.

Evaluation

In order to ascertain the performance of the proposed algorithm for this experiment, two steps were used. Firstly, a stratified K-fold cross-validation technique was used for the validation of the trained ML algorithms. In this validation technique, the folds are selected in a way that each class labels in each fold are equally distributed. The target variable is binary; therefore, each fold contains roughly the same proportions of the two types of class labels. The data set was split into k subsets where k =10, and each time one of these k subsets was used as the test set, and the k-1 subsets were used as a training set. This way, all data points are part of the test set exactly once and also gets to be in training set k-1 times. Single estimation was produced by taking the average results from the k folds. The algorithm takes time for training, which is the only disadvantage of using k-fold cross-validation.

Secondly, the performance of the proposed algorithm was compared with the physician's diagnosis. In order to achieve this, the sensitivity and specificity of the models were compared with that of the physician. By using the sepsis labels and blood culture information, the physician diagnosis matrix was created by assigning each of the 482 neonates to the appropriate cell in the 2×2 matrix. Table 9 below shows the physician diagnosis matrices for the study samples. To compare the proposed algorithm performance to the physician, first, the ML algorithms performance measures were generated such that their sensitivities and specificities are the same as that of the physician. This allowed us to deduce whether the proposed algorithm performs better or worse than the physician's diagnosis. Statistical significance of the experimental results was carried out using the Wilcoxon Signed-Rank Test.

The performance of the models was compared based on the accuracy obtained in the prediction of neonatal sepsis. Also, the evaluation parameters were obtained, such as average classification accuracy, receiver operation curve (ROC) (51), and area under the curve (AUC) (52). The mean accuracy of each model was visualized by generating the ROC-AUC plot of each model, the derived curve is called AUROC.

Another evaluation metric used to describe a classifier's performance is the confusion matrix, which involves calculating evaluation parameters. The confusion matrix is used to generate the values of true positive rate and false-positive rate.

Comparing the models will help determine the performance difference between the models in terms of classification accuracy.

Ethics approval and consent to participate

The study was approved by the Research Ethics Committee of Mbarara University of Science and Technology (Ref: MUREC 1/7), which waived the need for written informed consent given that the study was carried out retrospectively and made use of anonymized data. All methods were performed in accordance with the relevant guidelines and regulations.

Results

Data Understanding

Table 2: Statistical Description of data

S/No.	Parameters	Count	Mean	Standard Deviation	Minimum	25%	50%	75%	Maximum
1	Age in days	482.000000	1.746888	0.699154	0.000000	1.000000	2.000000	2.000000	3.000000
2	Gestation age at birth	482.000000	39.595643	1.747241	37.000000	38.000000	39.000000	41.000000	41.000000
3	Duration of labor	422.000000	19.594787	17.640458	0.000000	8.000000	14.000000	24.000000	72.000000
4	Duration of rupture of membrane	440.000000	15.328409	12.626562	0.000000	5.000000	13.000000	23.000000	72.000000
5	weight	482.000000	3.016979	0.539018	1.140000	2.690000	3.000000	3.340000 1.	
6	temperature	482.000000	38.611411	1.356349	33.700000	38.025000	38.700000	39.200000	50.000000
7	Respirator rate	477.000000	60.616352	17.778799	0.000000	50.000000	59.000000	69.000000	168.000000
8	Heart rate	474.000000	151.940928	23.925021	84.000000	138.000000	160.000000	166.000000	228.000000
9	WBC	396.000000	16.785253	12.369180	2.100000	4.767500	13.150000	30.725000	60.570000
10	Neutrophils count	178.000000	4.019719	4.966946	1.250000	1.580000	1.700000	2.500000	23.000000
11	Lymphocytes count	73.000000	5.341507	2.632397	1.300000	3.200000	4.400000	7.100000	13.200000
12	Monocytes count	24.000000	1.565417	0.672607	0.510000	1.135000	1.475000	1.787500	3.070000
13	Eosinophils count	24.000000	0.222500	0.319255	0.000000	0.050000	0.110000	0.217500	1.480000
14	Basophils count	24.000000	0.058333	0.099681	0.000000	0.010000	0.030000	0.052500	0.390000
15	RBC	87.000000	4.201839	1.021249	0.750000	3.705000	4.340000	4.855000	6.130000
16	Platelet count	360.000000	205.013889	131.820856	18.000000	113.750000	147.000000	283.250000	708.000000
17	Neonatal sepsis	482.000000	0.784232	0.411781	0.000000	1.000000	1.000000	1.000000	1.000000

The descriptive statistics of the data are shown above in table 2. The target variable (neonatal sepsis) is binary and has a value either '1,' i.e., neonatal sepsis is true or '0,' i.e., no neonatal sepsis. Information about the mean, standard deviation, maximum value, minimum value, and distribution (quartile range) of each numeric parameter are presented in the table above.

Table 3: Missing Value Analysis (numeric parameters)

Parameter	Missing Count	Missing Percent
Duration of labor	60	0.12
Duration of rupture of membrane	42	0.09
Respiratory rate	5	0.01
Heart rate	8	0.02
WBC	86	0.18
Neutrophils count	304	0.63
Lymphocytes count	409	0.85
Monocytes count	458	0.95
Eosinophils count	458	0.95
Basophils count	458	0.95
RBC	395	0.82
Platelet count	122	0.25

The Count column presents information about the total number of records of each feature. 12 parameters, duration of labor, duration of rupture of membrane, respiratory rate, heart rate, WBC, neutrophils count, lymphocytes count, monocytes count, eosinophils count, basophils count, RBC and platelet count have missing value out of 17 parameters which is given in table 3 above. Parameters with missing percent above 0.80 were dropped.

Figure 6 shows the distribution of the numeric features with respect to the target variable (neonatal sepsis). Age in days, gestation age at birth, weight, and respiratory rate, are normally distributed with the neonatal sepsis. Parameters such as duration of labor, duration of rupture of membrane, WBC, platelet count, and neutrophils count are positively skewed. While temperature and heart rate are negatively skewed.

The frequency plot of categorical variables; gender, maternal febrile episodes during pregnancy, fever during labor, abnormal vaginal discharge during pregnancy, antibiotic, place of delivery, mode of delivery, rupture of membrane, foul smelling liquor, fever, cold body, poor feeding, crying excessively, weak cry, lethargy, respiratory difficulty, respiratory distress, tachypnoea, apnea, CRP count, and blood culture are plotted as shown in figure 7 and figure 8 above. The ‘Target (neonatal sepsis)’ variable is highly biased as per the information provided by the bar graph. Only 22% of the values are ’0,’ and the rest of the records have ’1’ values. The balancing of this target feature will be addressed in the data preparation section. The categorical variables are binary, as shown in figures 7 and 8 above.

The Pearson correlation coefficient was used in the experiment to interpret the linear association between the numeric-continuous variables. The correlation coefficient range is from -1 to 1; the linear relationship is stronger as the absolute value increases. The correlation heatmap matrix shown in figure 9 shows the strength of the relationship between the features. The result deduced from the matrix is stated below:

- All the variable features have very little correlation with neonatal sepsis.
- Platelet count and temperature are highly negatively correlated.
- Heart rate has a weak positive correlation with respiratory rate and temperature.
- Duration of rupture of membrane is weakly positively correlated with duration of labor.
- Neutrophils count is weakly positively correlated with WBC.

Figure 10 shows the relationship strength and magnitude of relations between independent features (variables) and Target (neonatal sepsis). The features on the left side of the axis have a negative correlation with neonatal sepsis, i.e., the increase in the value of these features will decrease the risk of neonatal sepsis. Whereas the variables on the right side of the axis have a positive correlation, i.e., the increase in these features' value will increase neonatal sepsis's risk. In addition, the height of the bar graph from the center of the axis shows the magnitude of the correlation strength of each feature with neonatal sepsis.

Data Preparation

After the data has been analyzed; the first step is to remove the issues identified in the dataset to enable it to fit for the modeling. The missing values of duration of labor, duration of rupture of membrane, respiratory rate, heart rate, WBC, neutrophils count, and platelet count was imputed, respectively, with the mean value of each feature having missing values.

SMOTE Algorithm: Balancing of Dataset

Standard classifier algorithms like Logistic Regression have the likelihood to make results biased regarding classes with a higher number of instances. Base on this characteristic, classifiers most times ignore minority class features regarding them to be noise. Therefore, the probability of misclassification of the minority class as compared to the majority class is high.

Table 4: SMOTE oversampling

Target: Neonatal sepsis	Imbalanced Dataset	Balanced Dataset
1	378	378
0	104	378

The data set was balanced by creating synthetic records using the SMOTE algorithm, an over-sampling technique. Initially, the number of records belonging to neonatal sepsis as '1' is significantly higher than those belonging to class '0', as shown in table 4 above. The number of samples containing the '0' value is increased to 50% after running SMOTE oversampling algorithm.

Random Forest Classifier: Feature Importance

Finally, the Random Forest algorithm was used to identify each feature's importance in predicting neonatal sepsis, shown in figure 11. Each feature is assigned a score to show the importance of the feature in the model. The higher the score, the higher the importance of that particular feature. While features with a lower score are considered the least important. The height of the bar graph shows how important each feature is with neonatal sepsis.

Modelling

The Developed Algorithm for Neonatal Sepsis Prediction

The proposed algorithm consists of four phases: maternal condition, observational condition, laboratory condition, and neonatal sepsis.

Table 5: Pseudo code for the maternal condition

Step 1: Create a tuple M of the 6 parameters declared above, M = (a_0...a_n), $1 \leq n \leq 6$

Step 2: Initialize elements of tuple M; R = (b_0... b_i), $1 \leq i \leq 6$

Step 3: FOR each i in R DO

IF $i \leftarrow b_0$ = = "Yes" THEN

RETURN True

ELIF $i \leftarrow b_1$ = = "Yes" THEN

RETURN True

ELIF $i \leftarrow b_2$ = = "≥18 hours" THEN

RETURN True

ELIF $i \leftarrow b_3$ = = "≥18 hours" THEN

RETURN True

ELIF $i \leftarrow b_4$ = = "Yes" THEN

RETURN True

ELIF $i \leftarrow b_5$ = = "No" THEN

RETURN True

ELSE

RETURN False

END IF

END FOR

Step 4: IF True ≥ 1

RETURN "Maternal Condition"

ELSE

RETURN "No Maternal Condition"

END IF

Phase I: Maternal Condition

This phase checks if a neonate has a maternal condition. The algorithm looks through the maternal risk characteristics provided and determines based on the values if a neonate has a maternal condition or not. Table 5 shows the pseudo-code of the maternal condition phase.

The Parameters used (maternal risk characteristics):

a0 = Fever during labor.

a1 = Maternal febrile during pregnancy.

a2 = Duration of rupture of membrane

a3 = Duration of labor

a4 = Foul odor of the amniotic fluid.

a5 = Antibiotic treatment received by mother ≤4 hours prior to delivery.

Parameter's value:

b0 = (a0 = Yes)

b1 = (a1 = Yes)

b2 = (a2 = ≥18 hours)

b3 = (a3 = ≥18 hours)

b4 = (a4 = Yes)

b5 = (a5 = No)

Table 6: Pseudo code for the observational condition

Step 1: Create a tuple O of the 6 parameters declared above, O = (c_0... c_x), $1 \leq x \leq 6$

Step 2: Initialize elements of tuple O; S = (d_0... d_j), $1 \leq j \leq 6$

Step 3: FOR each j in S DO

IF j←d_0= = "≥160" OR "≤100" THEN

RETURN True

ELIF j← d_1 = = "≥38" OR "≤36.5" THEN

RETURN True

ELIF j← d_2 = = "Yes" THEN

RETURN True

ELIF j← d_3 = = "Yes" THEN

RETURN True

ELIF j← d_4 = = "Yes" THEN

RETURN True

ELIF j← d_5 = = "Yes" THEN

RETURN True

ELSE

RETURN False

END IF

END FOR

Step 4: IF True ≥ 2

RETURN "Observational Condition"

ELSE

RETURN "No Observational Condition"

END IF

Phase II: Observational Condition

This phase checks if a neonate has an observational condition. The algorithm looks through the neonatal clinical signs provided and determines based on the values if a neonate has an observational condition or not. Table 6 shows the pseudo-code of the observational condition phase.

The Parameters used (neonatal clinical signs):

c0 = Heart rate

c1 = Temperature

c2 = Respiratory distress

c3 = Apnea condition.

c4 = Lethargy condition.

c5 = Feeding difficulty

Parameter's value:

d0 = (c0 = ≥160 (tachycardia) or ≤100 (bradycardia) BPM)

d1 = (c1 = ≥38°C (fever) or ≤36.5°C (hypothermia))

d2 = (c2 = Yes)

d3 = (c3 = Yes)

d4 = (c4 = Yes)

d5 = (c5 = Yes)

Table 7: Pseudo code for the laboratory condition

Step 1: Create a tuple L of the 4 parameters declared above, L = (e_0...e_c), 1 ≤ c ≤ 4

Step 2: Initialize elements of tuple L; T = (f_0... f_k), 1 ≤ k ≤ 4

Step 3: FOR each k in T DO

IF k←f_0= = "≥10" THEN

RETURN True

ELIF k← f_1= = "≤5,000" OR "≥30,000" THEN

RETURN True

ELIF k← f_2= = "≤1,750" THEN

RETURN True

ELIF k← f_3= = "≤150,000" THEN

RETURN True

ELSE

RETURN False

END IF

END FOR

Step 4: IF True ≥ 2

RETURN "Laboratory Condition"

ELSE

RETURN "No Laboratory Condition"

END IF

Phase III: Laboratory Condition

This phase checks if a neonate has a laboratory condition. The algorithm looks through the laboratory tests provided and determines based on the values if a neonate has a laboratory condition or not. Table 7 shows the pseudo-code of the laboratory condition phase.

The Parameters used (laboratory tests):

e0 = C-reactive protein

e1 = White blood cell count

e2 = Neutrophil count

e3 = Platelet count

Parameter's value:

f0 = (e0 = ≥10mg/L)

f1 = (e1 = ≤5,000 or ≥30,000 per microL)

f2 = (e2 = ≤1,750 per microL)

f3 = (e3 = ≤150,000 per microL)

Table 8: Pseudo code for the neonatal sepsis

Step 1: Create a set N of the 3 parameters declared above, N = (g_0...g_e), 1 ≤ e ≤ 3

Step 2: Initialize elements of set N; P = (h_0... h_y), 1 ≤ y ≤ 3

Step 3: FOR each y in P DO

IF y= = "Yes" THEN

RETURN True

ELSE

RETURN False

END IF

END FOR

Step 4: IF True = = 3

RETURN "Neonatal Sepsis"

ELSE

RETURN "No Neonatal Sepsis"

END IF

Phase IV: Neonatal Sepsis

This phase checks if a neonate has neonatal sepsis. The algorithm looks through the maternal condition, observational condition, and laboratory condition and determines based on their outcomes if a neonate has sepsis or not. Table 8 shows the pseudo-code of the neonatal sepsis phase.

The Parameters used (neonatal sepsis variables):

g0 = Maternal condition

g1 = Observational condition

g2 = Laboratory condition

Parameter's value:

h0 = (g0 = Yes)

h1 = (g1 = Yes)

h2 = (g2 = Yes)

Support Vector Machine: Target Values

The linear SVM algorithm model obtained a minimum and maximum accuracy of 89% and 97%, with average classifier accuracy of 95%. From the ROC curve, it can be seen that almost half of the folds achieved accuracy above 85%, as shown in figure 12 above. It can be deduced from the results that the model's performance with linear SVM is slightly higher than KNN and NB models.

The ROC curves were plotted separately for SVM radial basis function and polynomial kernels. The average classifier accuracy and accuracy per iteration are shown in figure 13 and figure 14. Both models obtained a mean accuracy of 98%.

Logistic Regression: Target Values

Similarly, the Logistic Regression model is built using 10 k-fold stratified samplings to create training and test datasets. The model is a binary classification regression model. After the training of the model, then it is used to predict the target value. The model's score is created through this process, which then gives the prediction accuracy of the model. Finally, 10 scores are then created in which the mean of these scores gives the average accuracy of the LR classifier. The LR model obtained a minimum and maximum accuracy of 91% and 99%, respectively, with an average mean accuracy of 95%, as shown in figure 15 above.

K-Nearest Neighbor: Target Values

K-nearest neighbor model is built using 10 k-fold stratified samplings to create training and test datasets. KNN is preferably used when the features all have continuous value. Classification is achieved when the nearest neighbor is identified, which helps determine the class of an unknown sample. The model obtained a minimum and maximum accuracy of 86% and 97% with average classifier accuracy of 91%, as shown in figure 16 above.

Naïve Bayes: Target Values

Naïve Bayes model is built using 10 k-fold stratified samplings to create training and test datasets. The model uses all the attributes in the data and analyses these attributes individually as though they all have equal importance and independent of each other. The model obtained a minimum and maximum accuracy of 83% and 95% with average classifier accuracy of 90%, as shown in figure 17 above.

Decision Tree: Target Values

The decision tree model is built using 10 k-fold stratified samplings to create training and test datasets. For the model to classify a new item, it first needs to generate a decision tree based on the attribute values of the available training data. The model obtained a minimum and maximum accuracy of 95% and 100% with average classifier accuracy of 98%, as shown in figure 18 above. The classifiers' classification accuracy will be discussed further in the next section, 'Discussion.'

Evaluation

In order to evaluate the model performance, a ROC-AUC curve is required, which was created in the modeling section. The computing of True Positive Rate (TPR) and False Positive Rate (FPR) is the key requirement for plotting of ROC curve. roc_curve() and auc() are inbuilt functions in Sklearn, which returns TPR and FPR as output.

Table 9: Physician diagnosis matrix

Physician diagnosis versus gold standard	Blood culture +ve	Blood culture -ve
Septic	47	383
Not septic	6	46
Physician sensitivity	0.89	
Physician specificity	0.11	
Physician PPV	0.11	
Physician NPV	0.88	

The models were compared with an average accuracy achieved after each iteration from the stratified k-fold validation technique used to split train-test data. The proposed algorithm's performance was then compared with the physician's diagnosis shown in table 9 above.

Table 10: Comparing model prediction with Physician diagnosis

Algorithm	Sensitivity	Difference	Positive Predictive Value (PPV)	Difference	Negative Predictive Value (NPV)	Difference	Area under the ROC curve (AUC)
Fixed specificity (0.11)							
Physician	0.89		0.11		0.88		NA
SVM_L	0.97	0.08	0.8	0.69	0.97	0.09	0.95
SVM_RBF	1.0	0.11	0.95	0.84	1.0	0.12	0.98
SVM_POLY	1.0	0.11	0.93	0.82	1.0	0.12	0.98
LR	0.97	0.08	0.88	0.77	0.97	0.09	0.95
KNN	0.94	0.05	0.92	0.81	0.94	0.06	0.91
NB	0.95	0.06	0.95	0.84	0.95	0.07	0.90
DT	0.95	0.06	0.95	0.84	0.95	0.07	0.98

Table 10 above shows the performance measures generated by fixing specificity at 0.11.

Table 11: Comparing model prediction with Physician diagnosis

Algorithm	Sensitivity	Difference	Positive Predictive Value (PPV)	Difference	Negative Predictive Value (NPV)	Difference	Area under the ROC curve (AUC)
Fixed sensitivity (0.89)							
Physician	0.11		0.11		0.88		NA
SVM_L	0.97	0.86	0.97	0.86	0.90	0.02	0.95
SVM_RBF	0.95	0.84	0.95	0.84	0.95	0.07	0.98
SVM_POLY	0.97	0.86	0.97	0.86	0.95	0.07	0.98
LR	0.95	0.84	0.94	0.84	0.86	-0.02	0.95
KNN	1.0	0.89	1.0	0.89	0.76	-0.12	0.91
NB	0.97	0.86	0.97	0.86	0.95	0.07	0.90
DT	0.95	0.84	0.95	0.84	0.95	0.07	0.98

Table 11 above shows the performance measures generated by fixing sensitivity at 0.89.

Statistical Significance of the Experimental Results

The Wilcoxon Signed-Rank Test was performed on the accuracy scores recorded for each model, i.e., 10 accuracies per model, to test the statistical significance of the experimental results. The cut-off chosen to determine the significance of the results is '0.05'.

Table 12: Statistical significance of experimental results

Model	p-Value
SVML – SVM_RBF	**<0.01**
SVML - SVM_POLY	**0.02**
SVML - LR	0.92
SVML – KNN	0.06
SVML - NB	**<0.01**
SVM_RBF - SVM_POLY	0.92
SVM_RBF - LR	**<0.01**
SVM_RBF – NB	**<0.01**
SVM_RBF - DT	0.87
SVM_POLY - LR	**0.01**
SVM_POLY - NB	**<0.01**
SVM_POLY - DT	0.86
LR - NB	**0.02**
LR – DT	**<0.01**
NB - DT	**<0.01**

As shown in table 12 above, 10 out of 15 results are statistically significant. Support vector machine algorithms with radial basis function, polynomial kernels, and Decision tree algorithm performed better than the other algorithms in predicting neonatal sepsis as the results were statistically significant.

Discussion

This section gives a detailed analysis of the results of the experiment carried out in the previous section. The proposed algorithm and the ML algorithms' performance will be discussed, and there will be a conclusion of the experiment's strengths and limitations after a critical evaluation.

This research proposes an algorithm for neonatal sepsis prediction, which was used to train five supervised machine learning algorithms, and their performance was evaluated using the AUROC value. The classifiers are trained on a set of samples with balanced dependent variable values by applying the oversampling data technique. Before the training, data preprocessing steps such as imputation of missing values, feature standardization, and normalization, generation of dummy variables have been applied to the features.

Performance Evaluation of the Proposed Algorithm

The proposed algorithm is four-phased, consisting of maternal risk characteristics, neonatal clinical signs, and laboratory tests. In order to evaluate the diagnostic performance of the proposed algorithm, the performance of the trained ML algorithms was compared to the physician's diagnosis using the dataset from MRRH. The study used a representative set of ML algorithms. Their performance measures were generated so that their sensitivities and specificities are the same as that of the physician. The specificity of the ML algorithms was fixed at the physician's specificity while calculating the sensitivity. The ML algorithms' sensitivity was fixed at the physician's sensitivity while calculating the specificity, as shown in tables 13 and 14. This allowed deducing of whether the proposed algorithm performs better or worse than the physician diagnosis. This study's result shows that the proposed algorithm outperformed the physician diagnosis. The results also suggest that the proposed algorithm can be used for the early prediction of neonatal sepsis.

One of the studies that are closest to this study reported in the literature is a retrospective study for predicting neonatal late-onset sepsis (LOS) using the RALIS algorithm that consists of neonatal clinical signs (53). Mithal et al. (2018) reported an AUC of 0.90 for LOS prediction using linear regression based on a comparison between cases and controls (53). The second is also a retrospective study for predicting neonatal LOS using a diagnostic algorithm consisting of neonatal clinical signs and laboratory tests (54). Mani et al. (2014) explored a set of ML algorithms (SVM, NB, TAN, AODE, KNN, CART, RF, LR, and LBR) with the highest AUROC value been 0.65 based on a comparison with the physician's treatment (54). In contrast, this study focused on early-onset sepsis (EOS). It explored a set of ML algorithms with the highest AUROC value been 0.98 and the lowest being 0.90 based on a comparison with the physician's diagnosis. It included more variables in the proposed algorithm to distinguish neonates without sepsis to avoid subjecting neonates without sepsis to unnecessary antibiotics use.

The proposed algorithm with ML algorithms may also identify truly infected neonates before the availability of blood culture tests and, therefore, contribute to earlier detection and treatment. The improvement in the sensitivity of the proposed algorithm is not at the cost of its specificity. The proposed algorithm and the ML algorithms used in this study have significant real-time strengths. They could be used as an early warning system to alert physicians that neonatal sepsis may be present or developing. However, like the vital signs monitoring proposed by Gur et al. (2015) and clinically evaluated by Mithal et al.

(2018), these tools should be used as decision support tools and not as stand-alone decision-making expert systems (53,55). The proposed algorithm has to be tested in prospective settings and using data from other institutions (in future studies) to ascertain its clinical setting performance.

Strength and Limitations of the Study

The study proposed an algorithm that explores the combination of maternal risk factors, neonatal clinical signs, and laboratory tests as predictor variables in neonatal sepsis prediction. The study found the combination to be very efficient in the diagnosis of neonatal sepsis.

The research also studied the contribution of supervised machine learning techniques in clinical diagnosis. The experiment used five machine learning algorithms (SVM, LR, KNN, NB, and DT) belonging to different families and trained on the same dataset. The algorithms used are similar in a way that they can all be used for the classification of instances but also different as some of the algorithms are preferred where the data is linearly separable or have a single decision surface while some of the algorithms work best with non-linearly separable classification problems.

Lastly, Data pre-processing techniques, namely feature scaling using z-score, balancing of the dataset using SMOTE algorithm, and creating a dummy variable using one-hot encoding, are studied extensively throughout this research, and this was used on the data to improve the results. Multiple iterations are used in the modeling by applying stratified 10 k-fold validation. The mean accuracy of the accuracies derived from each fold is taken, which is the average accuracy of the classifiers.

Moving ahead to the limitations, this study made use of retrospective data and requires a follow-up prospective study. The proposed algorithm was developed based on the available screening parameters on the patient's records from MRRH. This limited the study from exploring some important screening parameters. The missing values in the dataset were higher with the laboratory tests, limiting the number of laboratory tests used. The proposed algorithm may function differently if modified with the identified screening parameters that are not currently in the algorithm. This can be explored further as part of the future study.

Another limitation of this study is that the ML algorithms' training and testing are based on a small-sized dataset. The dataset trends are biased; records containing neonatal sepsis as true are (3/4) of the records. If a relevant size of data is used for the experiment, the ML algorithms may function differently, and this can also be explored further as part of the future study. Lastly, the ML algorithms were not compared with an AUPRC value due to the time limit.

Conclusion

The proposed algorithm was developed based on three main variables, which include; maternal risk factors, neonatal clinical signs, and laboratory tests. The proposed algorithm was compared with the physician's diagnosis, and the proposed algorithm was found to outperform the physician's diagnosis. The study provides evidence that the combination of maternal risk factors, neonatal clinical signs, and laboratory tests can effectively diagnose neonatal sepsis. Based on the study result, the proposed algorithm can help identify neonatal sepsis cases as it exceeded clinicians' sensitivity and specificity. A prospective study is warranted to test the algorithm's clinical utility, which could provide a decision support aid to clinicians. This will undoubtedly improve the early recognition and treatment of neonatal sepsis. The study results suggest that ML algorithms can identify neonatal sepsis cases within a large and complex database.

Future Work & Recommendations

The proposed algorithm was developed on limited screening parameters. It was based on the available screening parameters on the patient’s records from MRRH, and the dataset used in the experiment is small in size. A sufficient number of screening parameters could be included in the algorithm to develop a more robust algorithm. Screening parameters such as chorioamnionitis, GBS status, heart rate variability, absolute neutrophil count, I/T ratio, M-ESR, and total leukocyte count can be used to modify the proposed algorithm. Hence, another area for future research would be to conduct the research prospectively by directly monitoring the patients, enabling the capturing of required patient’s information that will help develop a more generic algorithm, and validation of this algorithm is required to understand its functionality in a clinical setting.

This research focused on five algorithms: support vector, logistic regression, k-nearest neighbor, naïve bayes, and decision tree. However, ML algorithms such as random forests (RF) and neural networks can be further compared to find the best algorithm in relation to learning time, prediction accuracy, and size of data available. Due to time constraints, there was no much tuning of the SVM algorithm. Hence, future work can apply deep learning algorithms. Carry out a more enhanced tuning on the SVM algorithm to improve its prediction accuracy. Use a sufficient amount of data to train algorithms, and evaluate using the area under the precision-recall curve (AUPRC).

Abbreviations

AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristics
CRISP-DM	Cross Industry Standard Process for Data Mining
DT	Decision Tree
EHR	Electronic Health Records
EMR	Electronic Medical Record
EOS	Early-Onset Sepsis
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GBS	Group B Streptococcus
I/T ratio	Immature to Total Neutrophil Ratio
KNN	K-Nearest Neighbor
LOS	Late-Onset Sepsis
LR	Logistic Regression
M-ESR	Micro Erythrocyte Sedimentation Rate
ML	Machine Learning
MRRH	Mbarara Regional Referral Hospital
NB	Naive Bayes
NMR	Neonatal Mortality Rate
NPV	Negative Predictive Value
PPV	Positive Predictive Value
qSOFA	quick Sepsis-Related Organ Dysfunction Assessment Score
RF	Random Forests
SIRS	Systemic Inflammatory Response Syndrome
SMOTE	Synthetic Minority Oversampling Technique
SSA	Sub-Saharan Africa
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
WBC	White blood cell
WHO	World Health Organization

Declarations

The work presented in this manuscript is the result of our original research work. Where we have used the works of other persons, due acknowledgements are clearly stated. This work has not been submitted for publication in any journal before.

Consent for publication

Not applicable

Availability of data and materials

All code written in support of this publication is publicly available at https://github.com/Helenaden/Neonatal-Sepsis-Prediction

Competing interests

The authors declare that they have no competing interests.

Funding

No external funding was obtained for this study.

Authors' contributions

PDE made substantial contributions to the conception of the study, collected, analyzed, and interpreted the data, statistical analysis, and drafted the manuscript. WW contributed to the conception of the study, designed the study, analyzed and interpreted the data, and drafted the manuscript. AM contributed to the conception of the study, partly analyzed and interpreted the data, statistical analysis, and reviewed the manuscript. PDE, AM, and SK directed the acquisition of data at the hospital. SK supported data analysis and interpretation, added some literature, and reviewed the manuscript. All authors have reviewed and approved the manuscript.

Acknowledgements

The authors would like to thank the staff at Mbarara Regional Referral Hospital (MRRH) who made this study possible.

Authors' information

PDE

Department: Information Technology

Course: Masters in Health Information Technology

Authors' qualifications: BS.c, MS.c

Institution: Mbarara University of Science and Technology (MUST)

Position: Post-Graduate Student

Location: Mbarara, Uganda

WW

Authors' qualifications: BS.c, MS.c, PhD

Department: Biomedical Sciences and Engineering

Institution: Mbarara University of Science and Technology (MUST)

Position: Head of Department (HOD)

Location: Mbarara, Uganda

AM

Authors' qualifications: BS.c, MS.c, PhD

Department: Information Technology

Institution: Mbarara University of Science and Technology (MUST)

Position: Deputy Dean of the Faculty

Location: Mbarara, Uganda

SK

Authors' qualifications: MBChB, MMed

Department: Pediatrics and Child Health MRRH

Hospital: Mbarara Regional Referral Hospital (MRRH)

Position: Head of Department (HOD), Neonatology division

Location: Mbarara, Uganda

References

1. HNN. Leading causes of neonatal deaths in Uganda [Internet]. 2018 [cited 2021 Jan 12]. Available from: https://www.healthynewbornnetwork.org/country/uganda/

2. Hug L, Sharrow D, Sun Y, Marcusanu A, You D, Mathers C, et al. Levels and Trends in Child Mortality Report 2017 | UNICEF [Internet]. 2017 [cited 2021 Jan 11]. Available from: https://www.unicef.org/reports/levels-and-trends-child-mortality-report-2017

3. WHO. Newborns: improving survival and well-being [Internet]. 2020 [cited 2021 Jan 11]. Available from: https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality#

4. UN-IGME. CME Info - Child Mortality Estimates [Internet]. 2018 [cited 2021 Jan 11]. Available from: https://childmortality.org/data/Uganda

5. Goldstein B, Giroir B, Randolph A. International pediatric sepsis consensus conference: Definitions for sepsis and organ dysfunction in pediatrics. Pediatr Crit Care Med [Internet]. 2005 Jan [cited 2021 Jan 11];6(1). Available from: https://journals.lww.com/pccmjournal/Fulltext/2005/01000/International_pediatric_sepsis_consensus.2.aspx

6. Reinhart K, Bauer M, Riedemann NC, Hartog CS. New approaches to sepsis: Molecular diagnostics and biomarkers. Clin Microbiol Rev. 2012;25(4):609–34.

7. Balamuth F, Alpern ER, Abbadessa MK, Hayes K, Schast A, Lavelle J, et al. Improving Recognition of Pediatric Severe Sepsis in the Emergency Department: Contributions of a Vital Sign Based Electronic Alert and Bedside Clinician Identification. Ann Emerg Med [Internet]. 2017 Dec 1 [cited 2021 Jan 11];70(6):759. Available from: /pmc/articles/PMC5698118/

8. Smyth MA, Brace-Mcdonnell SJ, Perkins GD. Identification of adults with sepsis in the prehospital environment: a systematic review. BMJ Open [Internet]. 2016 Aug 1 [cited 2021 Jan 11];6(8). Available from: /pmc/articles/PMC4985978/

9. Olvera L, Dutra D. Early Recognition and Management of Maternal Sepsis. Nurs Womens Health [Internet]. 2016 Apr 1 [cited 2021 Jan 11];20(2):182–96. Available from: http://www.nwhjournal.org/article/S1751485116000738/fulltext

10. Jones SL, Ashton CM, Kiehne L, Gigliotti E, Bell-Gordon C, Disbot M, et al. Reductions in Sepsis Mortality and Costs After Design and Implementation of a Nurse-Based Early Recognition and Response Program. Jt Comm J Qual Patient Saf [Internet]. 2015 Nov 1 [cited 2021 Jan 11];41(11):483. Available from: /pmc/articles/PMC4880050/

11. Fell DB, Hawken S, Wong CA, Wilson LA, Malia SQ, Chakraborty P, et al. Using newborn screening analytes to identify cases of neonatal sepsis. 2017; (December):1–10.

12. Bonet M, Souza JP, Abalos E, Fawole B, Knight M, Kouanda S, et al. The global maternal sepsis study and awareness campaign (GLOSS): study protocol. Reprod Health [Internet]. 2018 Jan 30 [cited 2021 Jan 11];15(1). Available from: /pmc/articles/PMC5791346/

13. Sarkar AP, Dhar G, Das Sarkar M, Ghosh TK, Ghosh S. Early diagnosis of neonatal sepsis in primary health care unit. Bangladesh J Med Sci [Internet]. 2015 Apr 18 [cited 2021 Jan 12];14(2):169–72. Available from: https://www.banglajol.info/index.php/BJMS/article/view/21806

14. Liu L, Oza S, Hogan D, Perin J, Rudan I, Lawn JE, et al. Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. Lancet [Internet]. 2015 Jan 31 [cited 2021 Jan 12];385(9966):430–40. Available from: http://www.thelancet.com/article/S0140673614616986/fulltext

15. Voller SMB, Myers PJ. Neonatal Sepsis. Clin Pediatr Emerg Med. 2016 Jun 1;17(2):129–33.

16. Zea-Vera A, Ochoa TJ. Challenges in the diagnosis and management of neonatal sepsis. J Trop Pediatr. 2015;61(1):1–13.

17. Shane AL, Sánchez PJ, Stoll BJ. Neonatal sepsis. Lancet [Internet]. 2017 Oct 14 [cited 2021 Jan 12];390(10104):1770–80. Available from: http://www.thelancet.com/article/S0140673617310024/fulltext

18. Stoll BJ, Hansen NI, Bell EF, Shankaran S, Laptook AR, Walsh MC, et al. Neonatal outcomes of extremely preterm infants from the NICHD Neonatal Research Network. Pediatrics [Internet]. 2010;126(3):443–56. Available from: http://www.pediatrics.org/misc/reprints.shtml

19. Stephens BE, Vohr BR. Neurodevelopmental outcome of the premature infant. Pediatr Clin North Am [Internet]. 2009 Jun [cited 2021 Jan 12];56(3):631–46. Available from: https://pubmed.ncbi.nlm.nih.gov/19501696/

20. Stoll BJ, Hansen NI, Adams-Chapman I, Fanaroff AA, Hintz SR, Vohr B, et al. Neurodevelopmental and Growth Impairment Among Extremely Low-Birth-Weight Infants With Neonatal Infection. JAMA [Internet]. 2004 Nov 17 [cited 2021 Jan 12];292(19):2357–65. Available from: https://jamanetwork.com/journals/jama/fullarticle/199811

21. Masino AJ, Harris MC, Forsyth D, Ostapenko S, Srinivasan L, Bonafide CP, et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. PLoS One. 2019;14(2):1–23.

22. Hajj J, Blaine N, Salavaci J, Jacoby D. The “Centrality of Sepsis”: A Review on Incidence, Mortality, and Cost of Care. Healthcare [Internet]. 2018 Sep 1 [cited 2021 Jan 12];6(3). Available from: /pmc/articles/PMC6164723/

23. Gyang E, Shieh L, Forsey L, Maggio P. A Nurse-Driven Screening Tool for the Early Identification of Sepsis in an Intermediate Care Unit Setting. J Hosp Med [Internet]. 2015 Feb 1 [cited 2021 Jan 12];10(2):97. Available from: /pmc/articles/PMC4816455/

24. Cohen J, Vincent JL, Adhikari NKJ, Machado FR, Angus DC, Calandra T, et al. Sepsis: A roadmap for future research. Lancet Infect Dis [Internet]. 2015;15(5):581–614. Available from: http://dx.doi.org/10.1016/S1473-3099(15)70112-X

25. Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and Costs of Sepsis in the United States-An Analysis Based on Timing of Diagnosis and Severity Level. Crit Care Med [Internet]. 2018 [cited 2021 Jan 12];46(12):1889–97. Available from: https://pubmed.ncbi.nlm.nih.gov/30048332/

26. Oeser C, Lutsar I, Metsvaht T, Turner MA, Heath PT, Sharland M. Clinical trials in neonatal sepsis. J Antimicrob Chemother. 2013;68(12):2733–45.

27. James.L.Wynn. Defining neonatal sepsis. Curr Opin paediatr. 2016;28(2):135–40.

28. Weiss SL, Fitzgerald JC, Balamuth F, Alpern ER, Lavelle J, Chilutti M, et al. Delayed antimicrobial therapy increases mortality and organ dysfunction duration in pediatric sepsis. Crit Care Med. 2014;42(11):2409–17.

29. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA [Internet]. 2016 Feb 23 [cited 2021 Jan 12];315(8):762–74. Available from: https://jamanetwork.com/journals/jama/fullarticle/2492875

30. Shah BA, Padbury JF. Neonatal sepsis: An old problem with new insights. Virulence [Internet]. 2014 [cited 2021 Jan 13];5(1):170. Available from: /pmc/articles/PMC3916371/

31. Klingenberg C, Kornelisse RF, Buonocore G, Maier RF, Stocker M. Culture-Negative Early-Onset Neonatal Sepsis — At the Crossroad Between Efficient Sepsis Care and Antimicrobial Stewardship. Front Pediatr [Internet]. 2018 [cited 2021 Jan 13];6:285. Available from: /pmc/articles/PMC6189301/

32. Fuchs A, Bielicki J, Mathur S, Sharland M, Van Den Anker JN. Antibiotic Use for Sepsis in Neonates and Children: 2016 Evidence Update WHO-Reviews. 2016.

33. Deleon C, Shattuck K, Jain SK. Biomarkers of neonatal sepsis. Neoreviews [Internet]. 2015 [cited 2021 Jan 13];16(5):e297–308. Available from: https://researchexperts.utmb.edu/en/publications/biomarkers-of-neonatal-sepsis

34. Tank PJ, Omar A, Musoke R. Audit of Antibiotic Prescribing Practices for Neonatal Sepsis and Measurement of Outcome in New Born Unit at Kenyatta National Hospital. Int J Pediatr (United Kingdom). 2019;2019.

35. R. K, Manjunath S, Doddabasappa P, J. M. Evaluation of screening of neonatal sepsis. Int J Contemp Pediatr [Internet]. 2018 Feb 22 [cited 2021 Jan 13];5(2):580–3. Available from: https://www.ijpediatrics.com/index.php/ijcp/article/view/1378

36. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med. 2016;23(3):269–78.

37. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. JMIR Med Informatics. 2016;4(3):1–15.

38. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015;7(299).

39. Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. Comput Biol Med [Internet]. 2017;89:248–55. Available from: http://dx.doi.org/10.1016/j.compbiomed.2017.08.015

40. Mayhew MB, Petersen BK, Sales AP, Greene JD, Liu VX, Wasson TS. Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models. J Biomed Inform [Internet]. 2018;78:33–42. Available from: https://doi.org/10.1016/j.jbi.2017.11.015

41. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system. J Am Med Informatics Assoc. 2014;21(2):315–25.

42. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Crit Care Med. 2018;46(4):547–53.

43. Fairchild KD, Aschner JL. HeRO monitoring to reduce mortality in NICU patients. Res Reports Neonatol [Internet]. 2012 Aug 15 [cited 2021 Jan 14];2:65–76. Available from: https://www.dovepress.com/hero-monitoring-to-reduce-mortality-in-nicu-patients-peer-reviewed-fulltext-article-RRN

44. Coggins SA, Weitkamp JH, Grunwald L, Stark AR, Reese J, Walsh W, et al. Heart Rate Characteristic index monitoring for blood stream infection in an NICU: A 3-year experience. Arch Dis Child Fetal Neonatal Ed [Internet]. 2016 Jul 1 [cited 2021 Jan 15];101(4):F329. Available from: /pmc/articles/PMC4851911/

45. Li L, Walter SR, Rathnayake K, Westbrook JI. Evaluation and optimisation of risk identification tools for the early detection of sepsis in adult inpatients. G. Balint, Antala B, Carty C, Mabieme J-MA, Amar IB, Kaplanova A, editors. Uniw śląski [Internet]. 2018 [cited 2021 Jan 14];343–54. Available from: https://researchers.mq.edu.au/en/publications/evaluation-and-optimisation-of-risk-identification-tools-for-the-

46. Leventhal B. An introduction to data mining and other techniques for advanced analytics. J Direct, Data Digit Mark Pract. 2010;12(2):137–53.

47. Fernando KES, Mcgregor C, James AG. CRISP-TDM0 for standardized knowledge discovery from physiological data streams: Retinopathy of prematurity and blood oxygen saturation case study. 2017 IEEE Life Sci Conf LSC 2017. 2018 Jan 23;2018-Janua:226–9.

48. Scikit-learn. scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation [Internet]. [cited 2021 Jan 14]. Available from: https://scikit-learn.org/stable/

49. Rohit Walimbe. Handling imbalanced dataset in supervised learning using family of SMOTE algorithm. - DataScienceCentral.com [Internet]. 2017 [cited 2021 Jan 14]. Available from: https://www.datasciencecentral.com/handling-imbalanced-data-sets-in-supervised-learning-using-family/

50. Boser BE, Guyon IM, Vapnik VN. Training algorithm for optimal margin classifiers. Proc Fifth Annu ACM Work Comput Learn Theory. 1992;144–52.

51. Verplancke T, Van Looy S, Benoit D, Vansteelandt S, Depuydt P, De Turck F, et al. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. BMC Med Inform Decis Mak [Internet]. 2008 [cited 2021 Jan 14];8. Available from: https://pubmed.ncbi.nlm.nih.gov/19061509/

52. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. Med Care [Internet]. 2010 Jun [cited 2021 Jan 14];48(6 SUPPL.). Available from: https://journals.lww.com/lww-medicalcare/Fulltext/2010/06001/Prediction_Modeling_Using_EHR_Data__Challenges,.17.aspx

53. Mithal LB, Yogev R, Palac HL, Kaminsky D, Gur I, Mestan KK. Vital signs analysis algorithm detects inflammatory response in premature infants with late onset sepsis and necrotizing enterocolitis. Early Hum Dev [Internet]. 2018 Feb 1 [cited 2021 Jan 16];117:83. Available from: /pmc/articles/PMC5983899/

54. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Informatics Assoc [Internet]. 2014 [cited 2021 Jan 16];21(2):326–36. Available from: /pmc/articles/PMC3932458/

55. Gur I, Riskin A, Markel G, Bader D, Nave Y, Barzilay B, et al. Pilot study of a new mathematical algorithm for early detection of late-onset sepsis in very low-birth-weight infants. Am J Perinatol [Internet]. 2015 Jul 31 [cited 2021 Jan 16];32(4):321–30. Available from: http://www.thieme-connect.com/products/ejournals/html/10.1055/s-0034-1384645

Figures

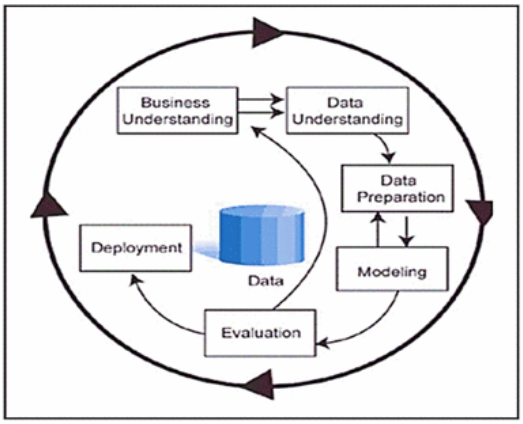

Figure 1

CRISP-data mining process model (47)

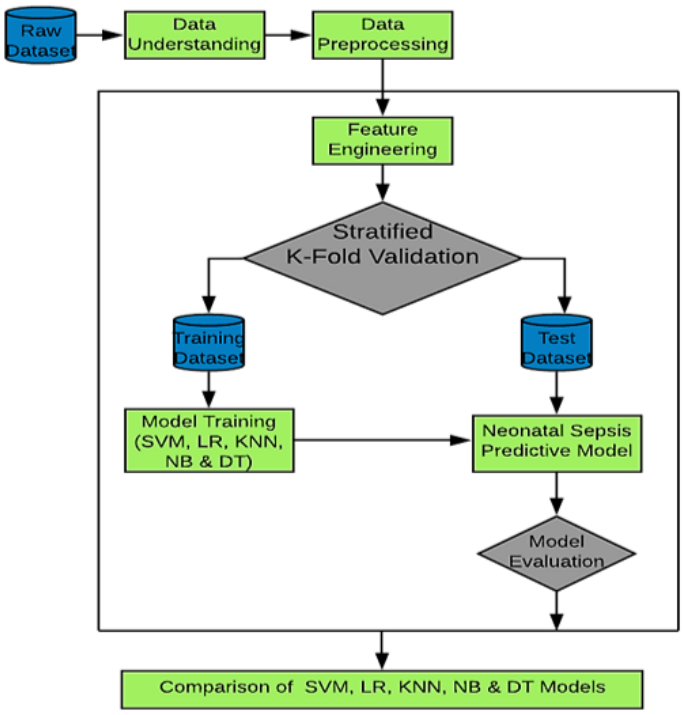

Figure 2

Design of the research experiment

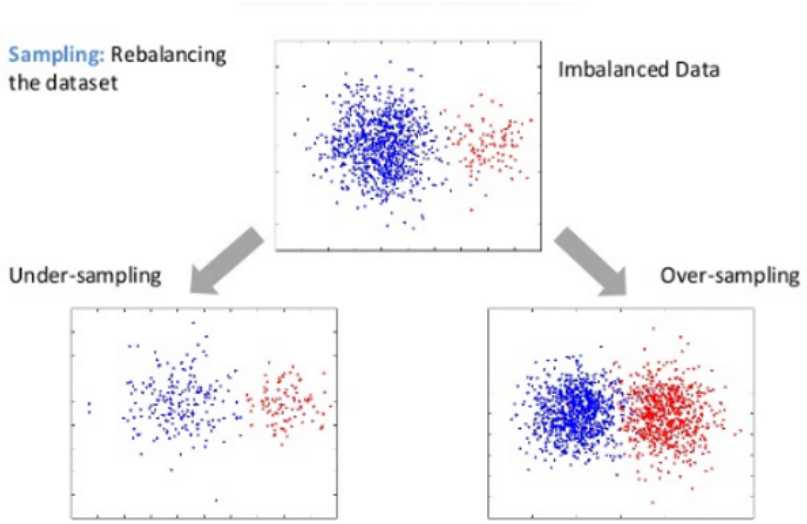

Figure 3

Sampling for Imbalanced data (49)

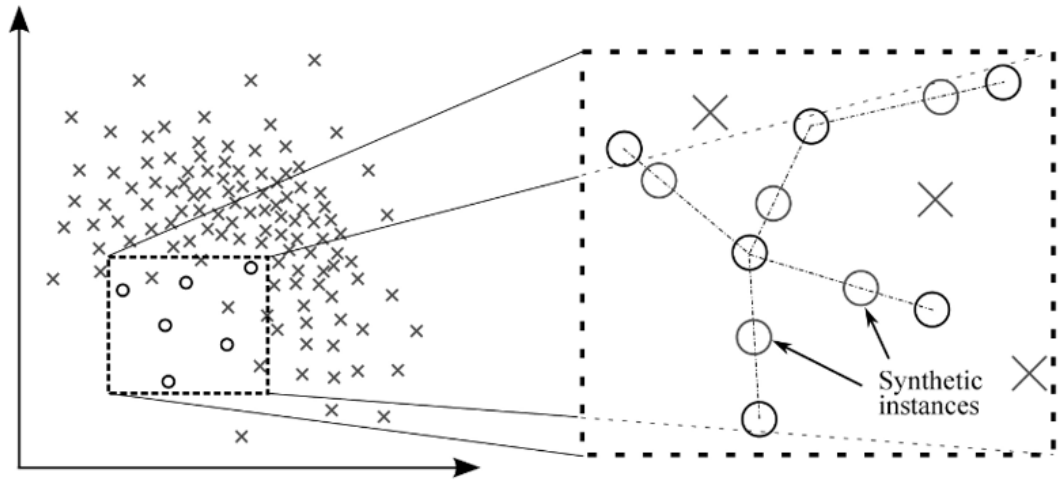

Figure 4

SMOTE algorithm KNN approach (49)

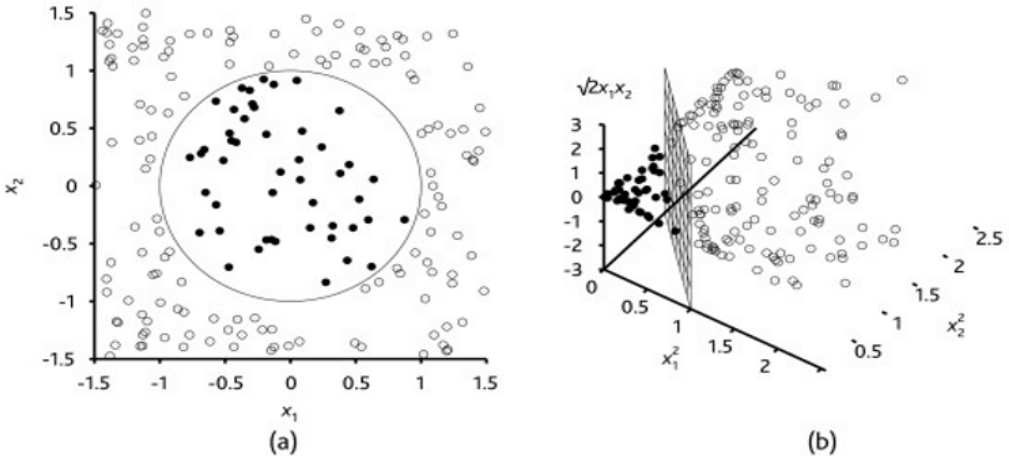

Figure 5

Classification by Support Vector Machine (51)

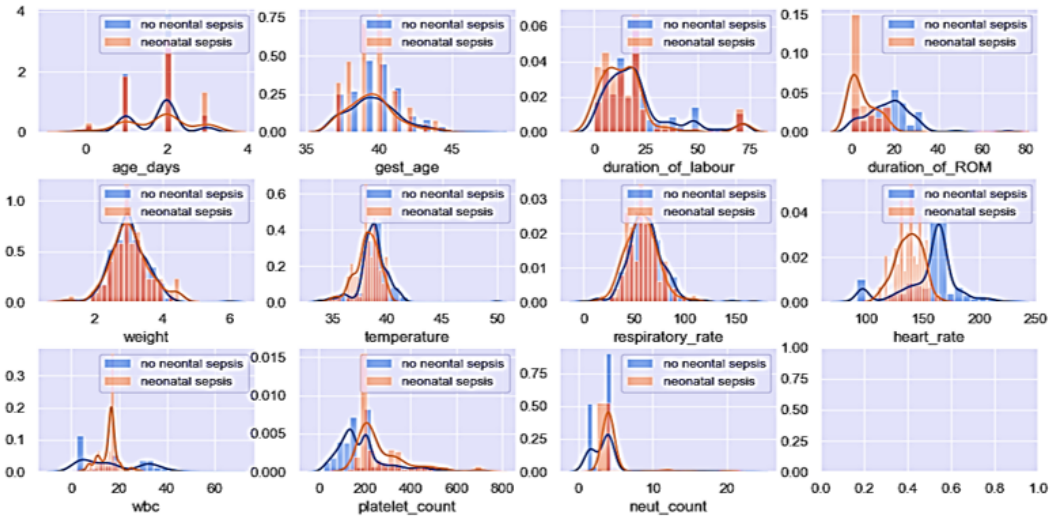

Figure 6

Distribution plot of numeric features with target (neonatal sepsis)

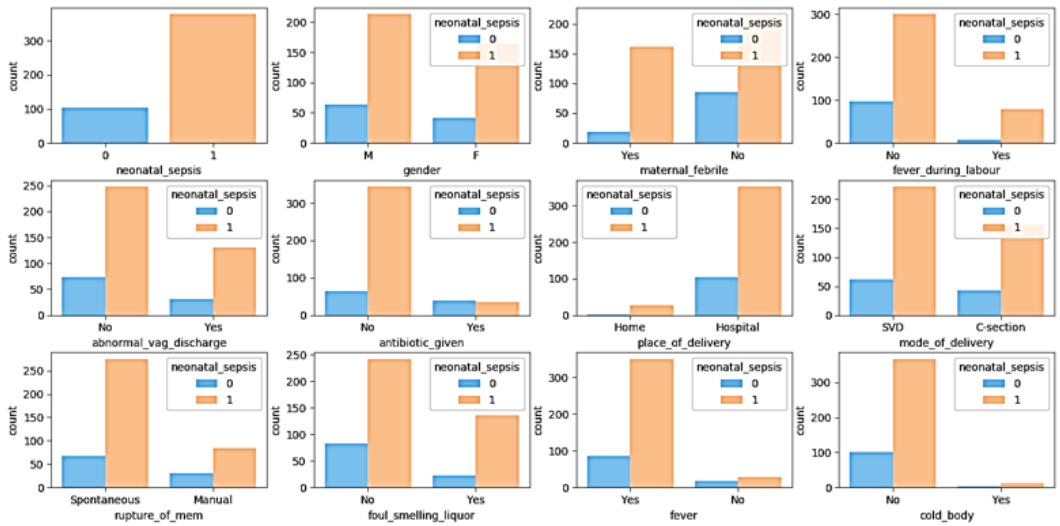

Figure 7

Distribution plot of categorical features with target (neonatal sepsis)

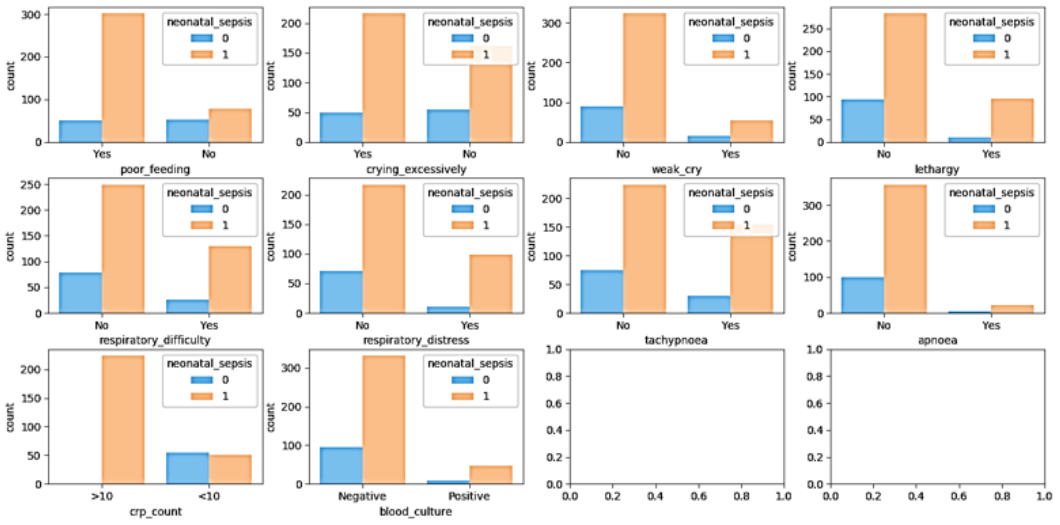

Figure 8

Distribution plot of categorical features with target (neonatal sepsis)

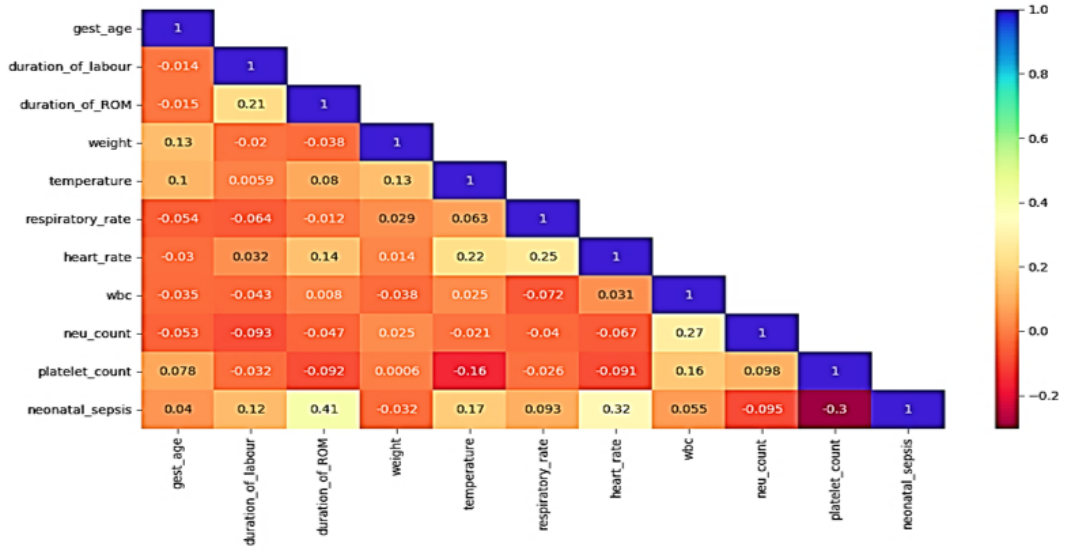

Figure 9

Heatmap matrix of features with Target (neonatal sepsis)

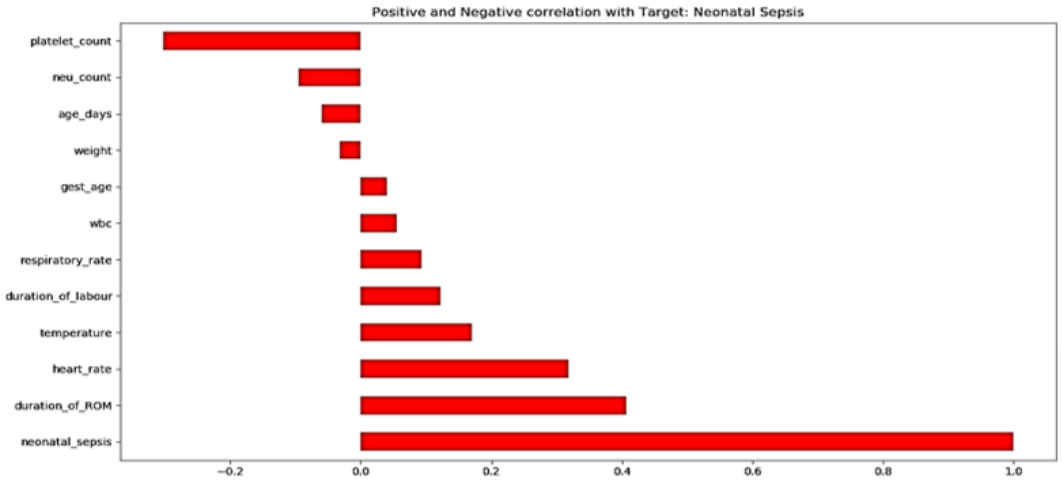

Figure 10

Positive-Negative correlation with Target

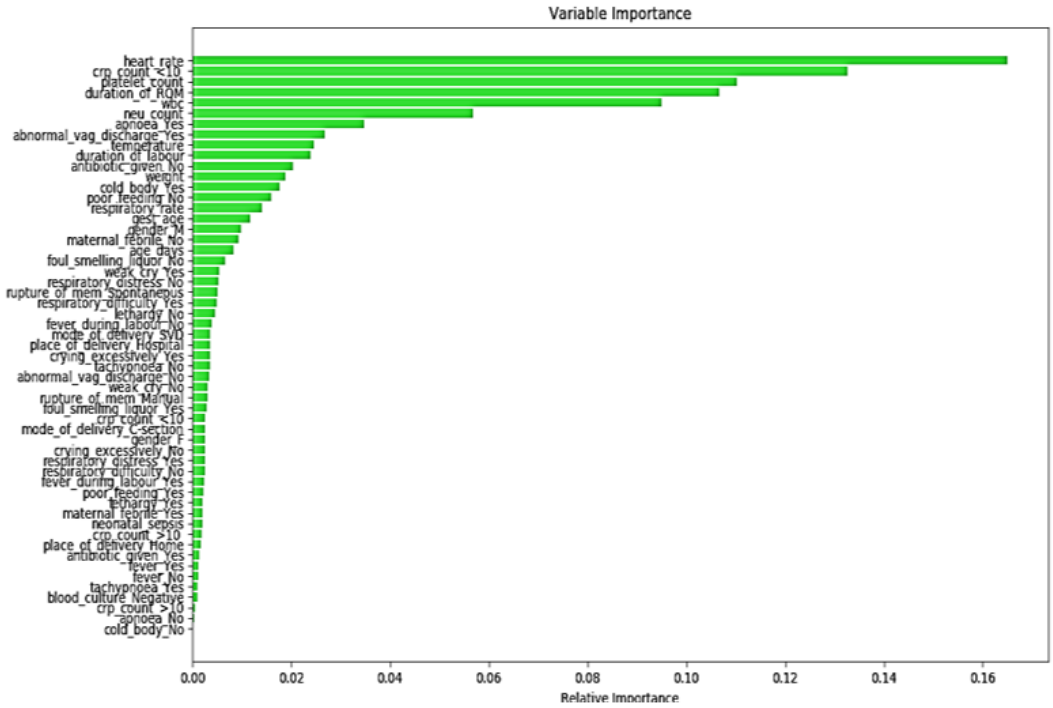

Figure 11

Random Forest Classifier: Feature importance

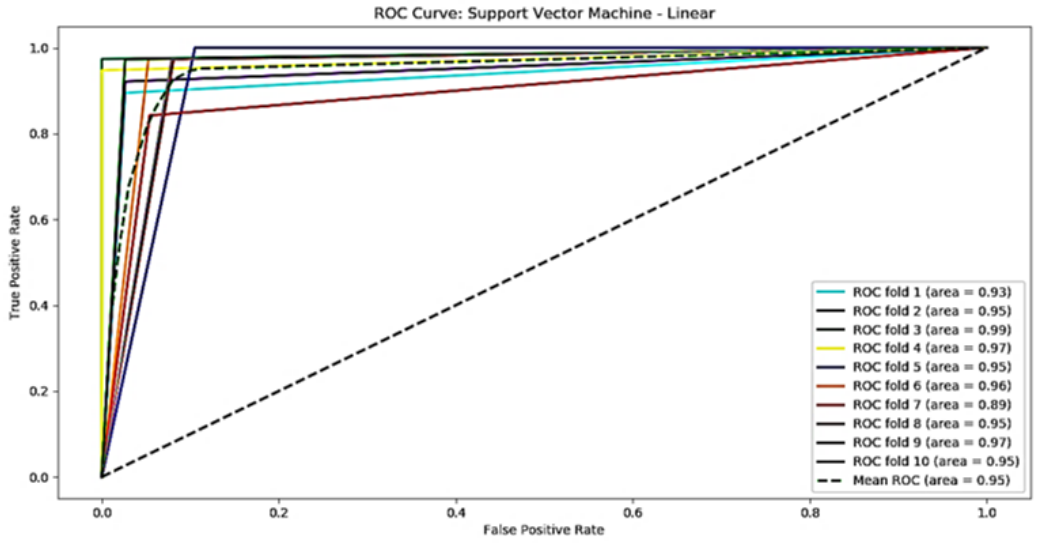

Figure 12

ROC curve: Support Vector Machine – Linear

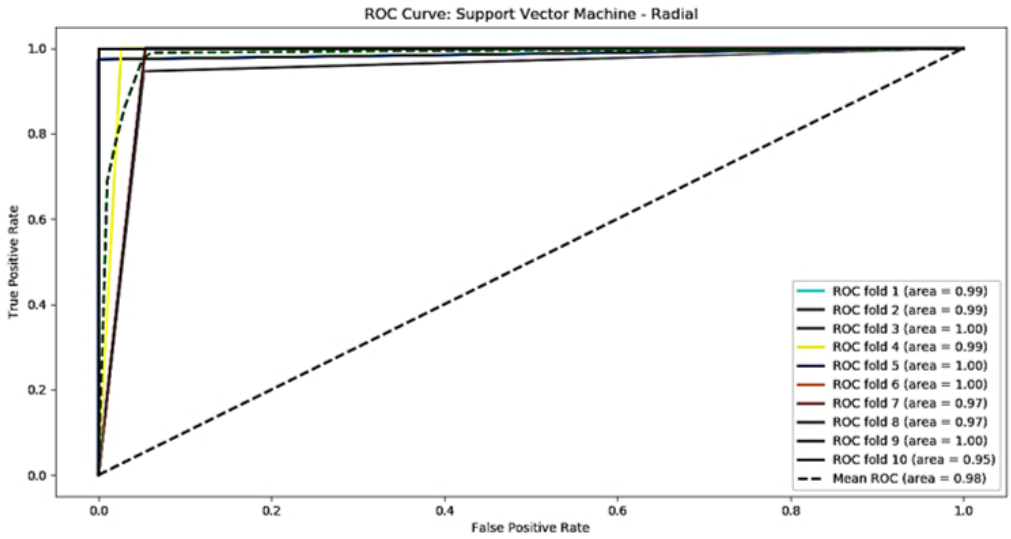

Figure 13

ROC curve: Support Vector Machine – Radial

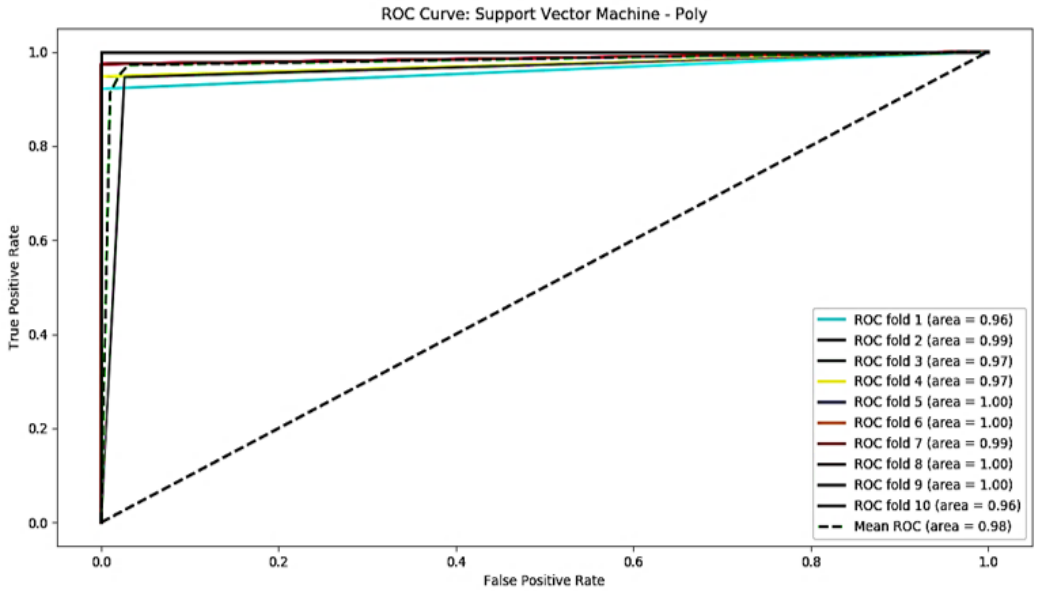

Figure 14

ROC curve: Support Vector Machine – Poly

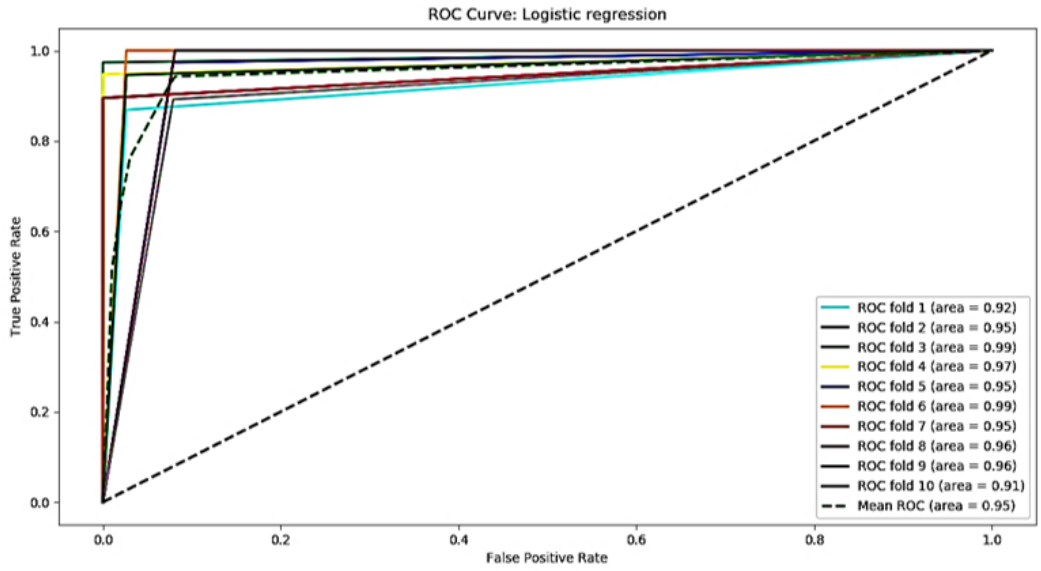

Figure 15

ROC curve: Logistic Regression

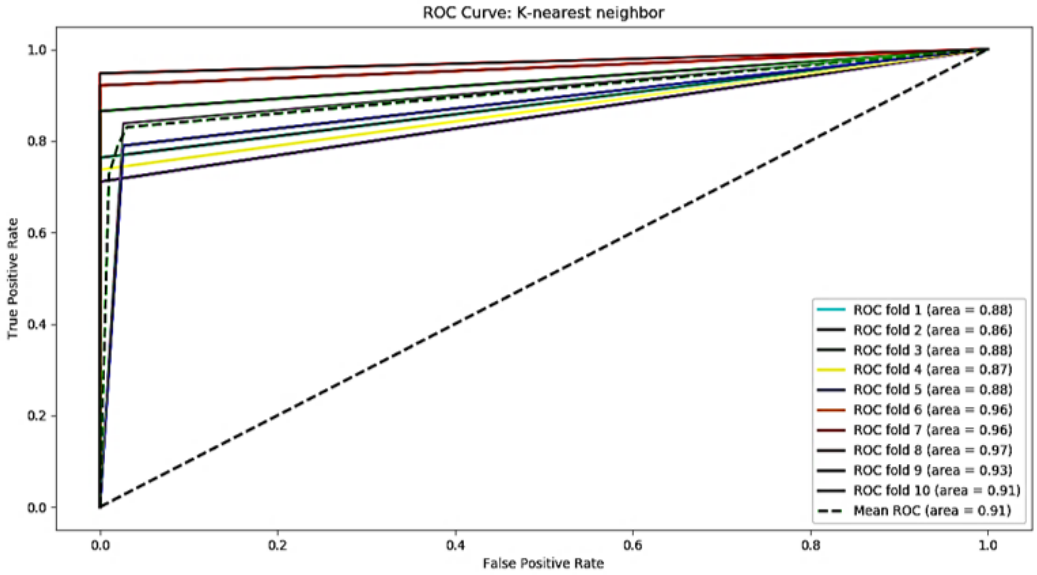

Figure 16

ROC curve: K-nearest neighbor

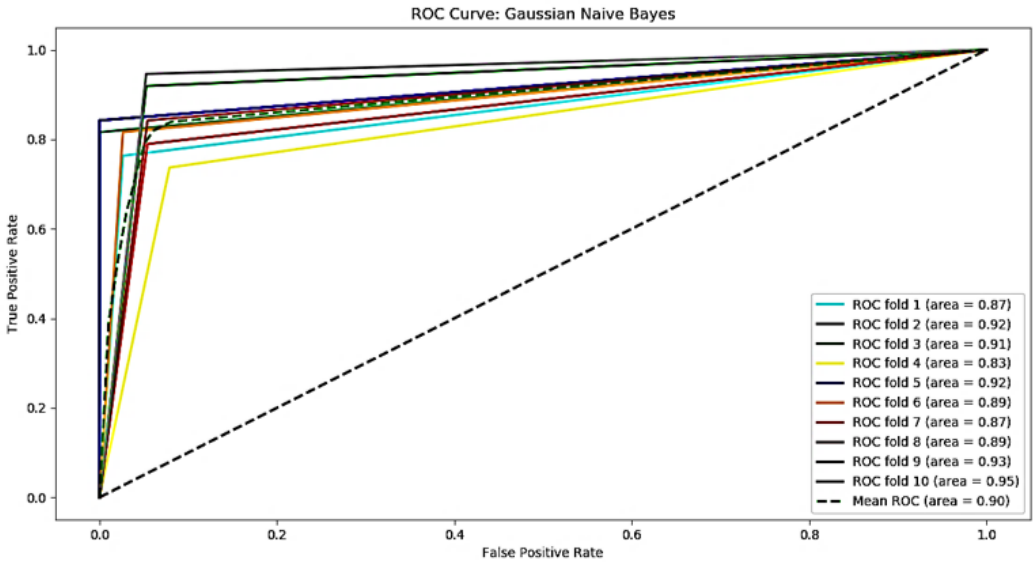

Figure 17

ROC curve: Naïve bayes

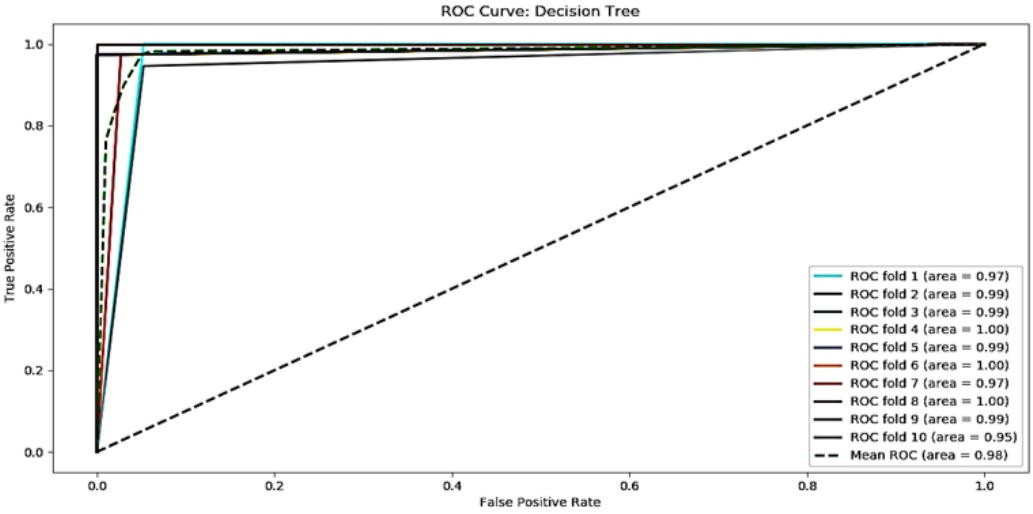

Figure 18

ROC curve: Decision tree

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInfofile.docx