

A Neonatal Sepsis Prediction Algorithm Using Electronic Medical Record Data

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: March 1st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1353776/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **Title**

2 A Neonatal Sepsis Prediction Algorithm Using Electronic Medical Record Data

3

4 **Authors**

5 Name: Dennis Peace Ezeobi (Corresponding Author)

6 Email address: helendennis4u@gmail.com, 2018mhit004@std.must.ac.ug

7 Institution: Mbarara University of Science and Technology (MUST)

8

9 Name: Dr. William Wasswa

10 Email address: wwasswa@must.ac.ug

11 Institution: Mbarara University of Science and Technology (MUST)

12

13 Name: Dr. Angella Musimenta

14 Email address: amusiiimenta@must.ac.ug

15 Institution: Mbarara University of Science and Technology (MUST)

16

17 Name: Dr. Stella Kyoyagala, MBChB, MMed

18 Email address: kstella20002000@yahoo.com

19 Institution: Mbarara Regional Referral Hospital (MRRH)

20

21 **Abstract**

22 **Background**

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

31 **Methods**

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

41 **Results**

42 The results of this study show that the proposed algorithm (with the lowest Sensitivity of 0.95,
43 lowest Specificity of 0.95) outperformed the physician diagnosis (Sensitivity = 0.89, Specificity =
44 0.11). SVM model with radial basis function, polynomial kernels, and DT model (with the highest

45 AUROC values of 0.98) performed better than the other models in predicting neonatal sepsis as
46 their results were statistically significant.

47 **Conclusions**

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

53

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

57

58 **Background**

59 About 2.5 million neonates die worldwide every year, and most of these deaths occur in low-
60 resource settings (1,2). It is estimated that the neonatal mortality rate (NMR) in Sub-Saharan
61 Africa (SSA) is 28 per 1000 live births, with Uganda struggling with a high rate of 20 per 1000
62 live births (1,3). The pediatric consensus definition of sepsis is systemic inflammatory response
63 syndrome (SIRS) in the presence of or due to suspected or proven infection (4). The SIRS cause
64 damage to the body and can quickly advance to severe sepsis, multi-organ system failure, and
65 death (5,6). Therefore, early recognition and prompt treatment, which have been predicted to
66 improve clinical management of sepsis, is required to reduce the morbidity and mortality of
67 neonatal sepsis (7–11).

68 Neonatal sepsis is a significant cause of neonatal mortality and morbidity worldwide (12–14), and
69 a majority of the morbidity and mortality from sepsis is preventable. Sepsis is one of the major
70 causes of neonatal deaths in Uganda, like in other Sub-Saharan African countries, accounting for
71 17% of neonatal deaths in Uganda (15). Several authors classify neonatal sepsis as a community-
72 and hospital-acquired instead of early-and late-onset in developing countries. Neonatal sepsis is
73 usually classified as early-onset (<48–72h) and late-onset sepsis (>48–72h), depending on the age
74 at onset (16,17). About 30-50% of survivors of neonatal sepsis end up with major long-term
75 impairments and also faced with prolonged hospitalization, chronic lung disease, and
76 neurodevelopmental disabilities (18–21). Recent data highlight the costs and burdens of sepsis, as
77 it remains the most expensive cause of hospitalization (22–25). The development of clinical trials
78 and global recommendations is hindered by the population's susceptibility, lack of consensus in
79 definitions, and variability between regions (26). Multiple challenges in diagnostic and treatment
80 decisions are faced by physicians caring for infected neonates. To date, there have been just modest
81 improvements in terms of sepsis outcomes in neonates despite the increased understanding of its
82 pathophysiology and efforts to improve clinical decision support in intensive care (27). The after-
83 effect of sepsis-infected adults and children's impending intervention is receiving attention in
84 recent studies (28,29).

85 Despite the explored significance of early treatment of sepsis, there are still unresolved challenges
86 due to impeding recognition and intervention of sepsis (14,27,30–34). Neonates have a non-
87 specific clinical presentation that overlaps with other neonatal disease processes. Also, laboratory
88 tests have suboptimal diagnostic accuracy, which makes a rapid diagnosis of neonatal sepsis
89 difficult. The blood culture, the standard gold test for neonatal sepsis diagnosis, faces the challenge
90 of insufficient blood volume for blood culture and a low amount of invading microorganisms in

91 the blood; this usually generates false-negative results (21,35). Despite negative culture results,
92 neonates presumed to have sepsis are kept on a longtime antibiotic treatment. Studies have
93 previously applied machine learning and statistical modeling techniques to tackle the problems
94 related to sepsis recognition and intervention (36–39).

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

114 diagnosis of neonatal sepsis by developing an algorithm that combines maternal risk factors,
115 neonatal clinical signs, and laboratory tests.

116

117 **Methods**

118 A Standard approach was implemented, a structured data mining project methodology defined by
119 the Cross-Industry Standard Process for Data Mining (CRISP-DM) developed in 1996 (46,47).

120 The experiment was executed in five phases, which include; business understanding, data

121 understanding, data preparation, modeling, and evaluation. The experimental steps were

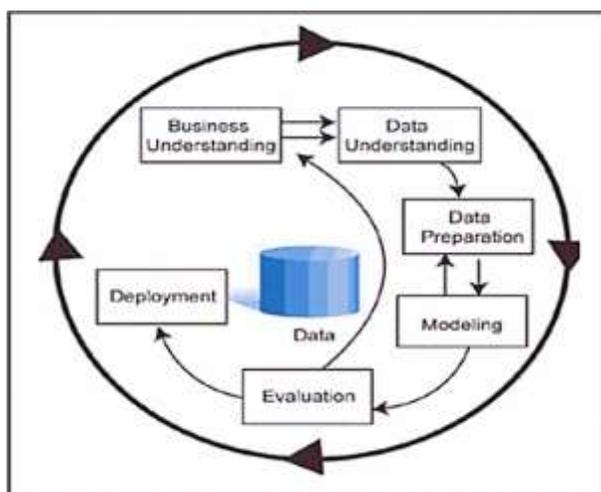
122 performed in Python language using scikit-learn library. Scikit-learn (48) is an open-source

123 machine learning library featuring various classification, regression, and clustering algorithms for

124 the python programming language. This research aims to propose an algorithm that can improve

125 the early recognition and treatment of sepsis in neonates using EMR data collected over a time

126 span of four years.



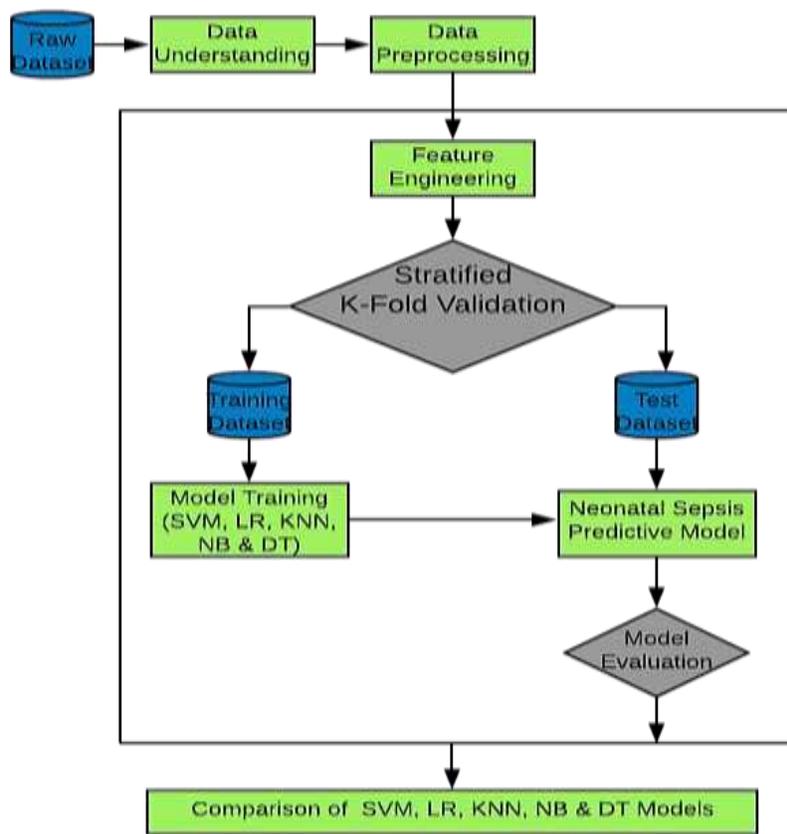
127

128

Figure 1: CRISP-data mining process model (47)

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

132 Business Understanding



133

134

Figure 2: Design of the research experiment

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

138 Data Understanding

139 Secondary data was used in this study, based on the data reliability, suitability, and adequacy of
140 the data. The dataset contains information about sepsis screening parameters from hospitalized
141 neonates collected from the EMR of MRRH that covers a period of 4 years. The Mbarara Regional

142 Referral Hospital is located in rural Uganda and is the main teaching hospital situated adjacent to
143 the Mbarara University of Science and Technology campus. The hospital typically has about
144 21.3% cases of presumed neonatal sepsis per the pediatric ward's annual admissions, as it is the
145 referral center for southwestern Uganda. A data abstraction tool was developed (see supplementary
146 table 1) to retrieve information essential to the study. The dataset contains records of 482 neonates
147 hospitalized from October 1st, 2015 to September 30th, 2019, that met the inclusion criteria with
148 38 different neonate screening features (Table 1). The category of neonates of concern is the group
149 with early-onset sepsis ($\leq 48-72h$). The predictor variables are both continuous and categorical in
150 nature.

151 The neonatal data experiment was performed using Python code, which includes the following
152 steps:

- 153 1. Statistical analysis: This step involves determining basic statistics for analysis such as
154 distribution, mean, median, mode, max, min, standard deviation, normalization, and
155 the skewness of features.
- 156 2. Missing value analysis: This step involves calculating the count and percent count of
157 the target variable's missing values and features.
- 158 3. Outlier analysis: This step was used to find out the values lying out of range, e.g.,
159 weight.
- 160 4. Exploratory data analysis: This step involves the plots of the frequency distribution of
161 numeric and categorical features with respect to the target variable.
- 162 5. Heatmap correlation matrix: This step involves using a heatmap matrix to analyze the
163 correlation between the target variable (neonatal sepsis) with the features. A separate

164 plot is generated to identify whether the features are positively or negatively correlated
 165 with the target variable.

166 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
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

167

168 Table 1 above contains the summary of the dataset, where attributes are the screening parameters,

169 and type represents the data type of each parameter, i.e., 'N' for numeric values and 'C' for

170 categorical values.

171 **Inclusion Criteria**

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

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

184 **Exclusion Criteria**

185 Excluded cases include:

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

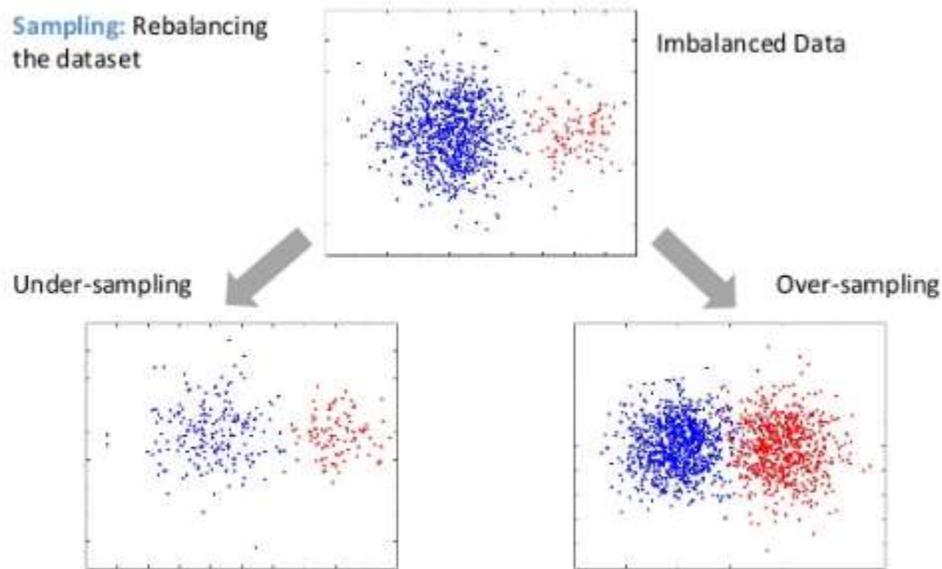
190 A detailed overview of the steps carried out for data preparation and feature engineering used in
191 this study is provided in the following section.

192 **Data Preparation**

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

197 **SMOTE Algorithm - Balancing of Dataset**

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

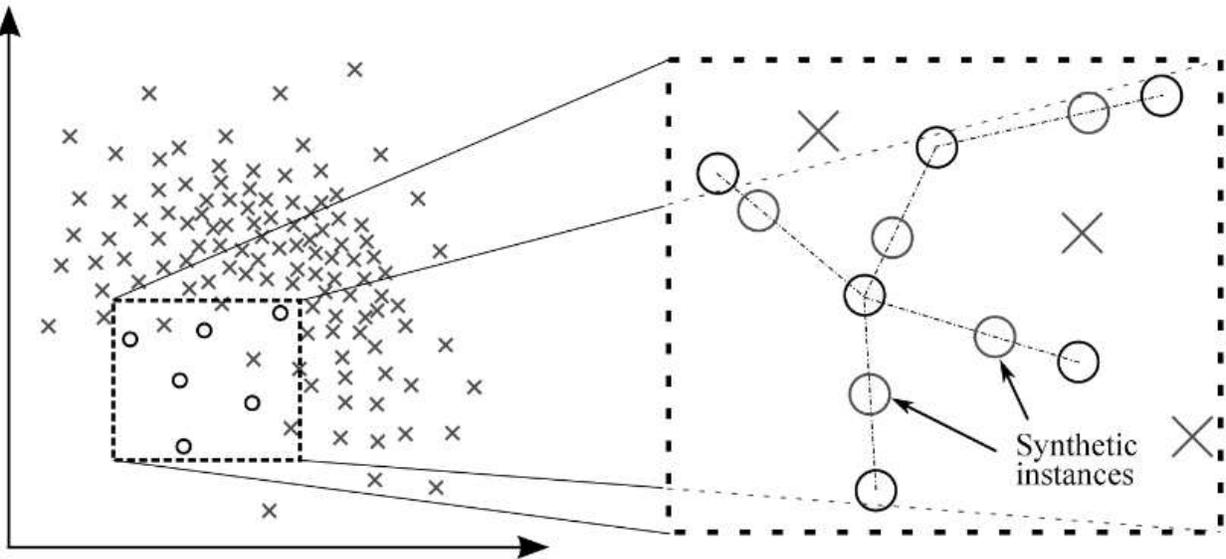


202

203 Figure 3: Sampling for Imbalanced data (49)

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

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



209

Figure 4: SMOTE algorithm KNN approach (49)

210

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

216 **Normalization and Standardization: Z-score**

217 Unscaled or unstandardized features are known to make the learning algorithms to predict
218 recklessly. Normalization or standardization, an important step to be carried out before proceeding
219 to model building, is required to ensure that all the feature values are on the same scale. The values
220 of features are standardized from different dynamic ranges into specific ranges through

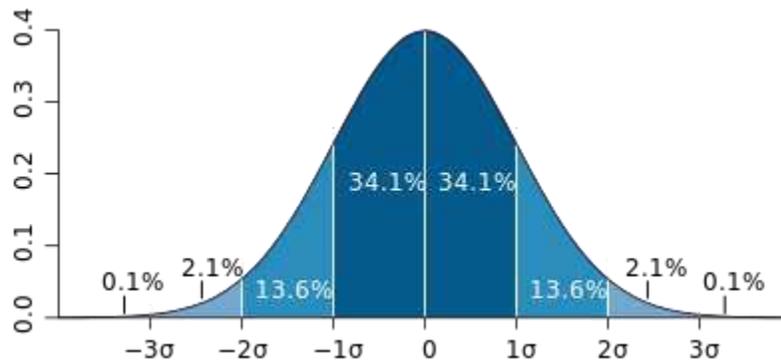
221 standardization, a preprocessing step. All the parameters are scaled to have zero mean and unit
222 variance, converted by standard score, also known as z-score.

223 *Equation 1: Z-score which is given as;*

224
$$z_i = \frac{x_i - \bar{x}}{s}$$

225 \bar{x} is the sample mean

226 s is the sample standard deviation



227

228 Figure 5: Normal distribution (Bell curve)

229 Z-scores, which range from -3 standard deviations to +3 standard deviations, can be placed on a
230 normal distribution curve, as shown in figure 5 above.

231 **One-Hot Encoding: Categorical Variables**

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

236 **Random Forest Classifier: Feature Importance**

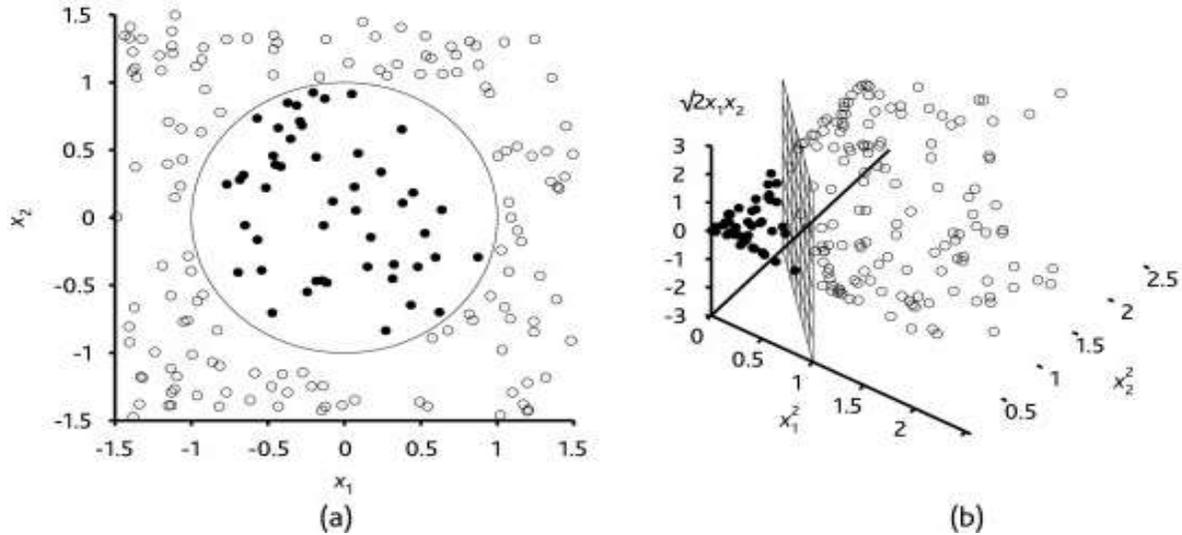
237 Finally, each variable's importance in predicting neonatal sepsis was identified with the Random
238 Forest algorithm, an ensemble modeling technique based on iteratively removing variables with

239 low ranking and using cross-validation to assess the learning performance. Each variable was
240 assigned a score to show the importance of the variables in the model. The higher the score, the
241 higher the importance of that particular variable. While variables with a lower score are considered
242 the least important. Data pre-processing techniques assist in extracting more useful information,
243 which helps build a model with higher accuracy and performance.

244 **Modelling**

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

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



261

262

Figure 6: Classification by Support Vector Machine (51)

263 SVM algorithms find boundaries for classification when there is no possible separation within a
 264 high number of input variables, as shown in figure 6 above. The input variables are transformed
 265 by increasing the dimensionality of the variable space to generate the separation boundary.

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

271 **Evaluation**

272 In order to ascertain the performance of the proposed algorithm for this experiment, two steps were
 273 used. Firstly, a stratified K-fold cross-validation technique was used for the validation of the
 274 trained ML algorithms. In this validation technique, the folds are selected in a way that each class
 275 labels in each fold are equally distributed. The target variable is binary; therefore, each fold
 276 contains roughly the same proportions of the two types of class labels. The data set was split into
 277 k subsets where k =10, and each time one of these k subsets was used as the test set, and the k-1

278 subsets were used as a training set. This way, all data points are part of the test set exactly once
 279 and also gets to be in training set k-1 times. Single estimation was produced by taking the average
 280 results from the k folds. The algorithm takes time for training, which is the only disadvantage of
 281 using k-fold cross-validation.



282

283 Figure 7: Stratified 5-fold cross validation technique (52)

284 In experiments, the ideal standard value used is k=10. The training and test split in 5-fold cross-
 285 validation is shown in figure 7 above:

286 Secondly, the performance of the proposed algorithm was compared with the physician's
 287 diagnosis. In order to achieve this, the sensitivity and specificity of the models were compared
 288 with that of the physician. By using the sepsis labels and blood culture information, the physician
 289 diagnosis matrix was created by assigning each of the 482 neonates to the appropriate cell in the
 290 2x2 matrix. Table 12 below shows the physician diagnosis matrices for the study samples. To
 291 compare the proposed algorithm performance to the physician, first, the ML algorithms
 292 performance measures were generated such that their sensitivities and specificities are the same as

293 that of the physician. This allowed us to deduce whether the proposed algorithm performs better
294 or worse than the physician's diagnosis. Statistical significance of the experimental results was
295 carried out using the Wilcoxon Signed-Rank Test.

296 The performance of the models was compared based on the accuracy obtained in the prediction of
297 neonatal sepsis. In each fold, the model's accuracy was computed, which gives 10 accuracies per
298 model.

299 *Equation 2: The accuracy of the model which is given as;*

$$300 \quad Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

301 where,

302 TP (True Positive); positive instances that are classified as positive,

303 FP (False Positive); negative instances that are classified as positive,

304 FN (False Negative); positive instances that are classified as negative,

305 TN (True Negative); negative instances that are classified as negative.

306 Also, the evaluation parameters were obtained, such as average classification accuracy, receiver
307 operation curve (ROC) (51), and area under the curve (AUC) (53). The mean accuracy of each
308 model was visualized by generating the ROC-AUC plot of each model. The ROC curve consists
309 of two metrics, True Positive Rate (TPR) and False Positive Rate (FPR).

310 True positive rate (TPR), also known as sensitivity, hit rate, or recall.

311 *Equation 3: Sensitivity, hit rate or recall, is defined as;*

$$312 \quad TPR = \frac{TP}{TP + FN}$$

313 This metric correlates with the proportion of positive data points that are correctly considered
314 positive with respect to all positive data points. In a simple term, it means the higher the TPR, the
315 fewer the positive data points that are missed.

316 False-positive rate (FPR) or fall-out is defined as

317 *Equation 4: False positive rate (FPR) or fall-out is defined as;*

$$318 \quad FPR = \frac{FP}{FP + TN}$$

319 *Equation 5: This metric can also be generated from specificity as;*

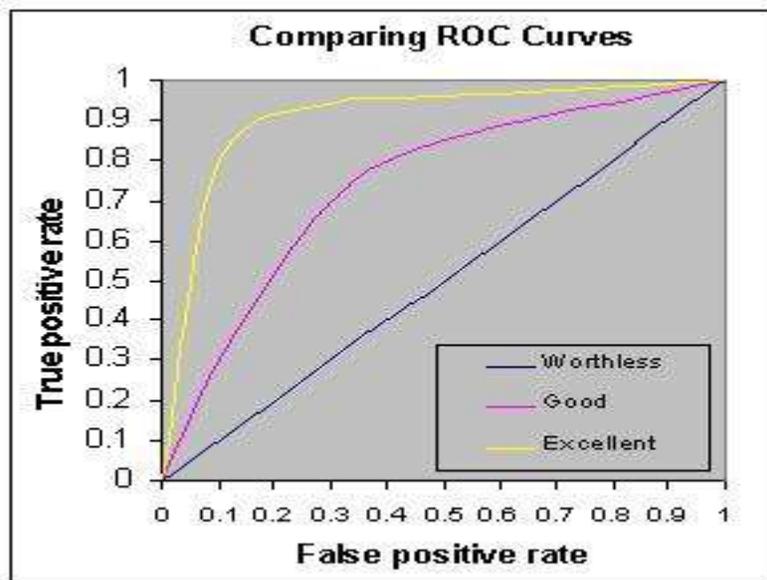
$$320 \quad FPR = 1 - \text{Specificity}$$

321 *Equation 6: Where specificity is defined as;*

$$322 \quad FPR = \frac{TN}{TN + FP}$$

323

324 This metric correlates with the proportion of negative data points that are mistakenly considered
325 positive with respect to all negative data points. In a simple term, it means the higher the FPR, the
326 more the negative data points that are misclassified.



327

328 Figure 8: ROC-AUC curve (Thomas, 2001)

329 In order to generate AUC, FPR and TPR will be combined into one metric, where a single graph
330 is plotted with the values of FPR on the x-axis and the values of TPR on the y-axis. The derived
331 curve is called AUROC, as shown in figure 8 above.

332 Table 2: Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

333
334 Another evaluation metric used to describe a classifier's performance is the confusion matrix,
335 which involves calculating evaluation parameters and is shown in table 2 above. The confusion
336 matrix is used to generate the values of true positive rate and false-positive rate.

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

339 **Summary of Design**

340 This chapter is committed to providing the breakdown of the experiment that was carried out for
341 the dissertation. This chapter begins with a short explanation of the dataset, including the variable
342 types and data source. One of the principal interests is the issues present in the raw data. Hence,
343 data pre-processing machine learning techniques for cleaning and normalizing the data made the
344 data fit for the modeling phase. These data pre-processing techniques include SMOTE algorithm
345 for balancing dataset, Z-score for standardization of data, and One-hot encoding for generating
346 dummy variables.

347 Furthermore, seven supervised machine learning algorithms, SVM with linear kernel, SVM with
 348 the radial kernel, and SVM with polynomial kernels, LR, KNN, NB, and DT, were trained. The
 349 chapter ends with the evaluation of the proposed algorithm.

350 The next chapter details the results of the study design and experiment.

351

352 **Results**

353 **Data Understanding**

354 Table 3: Statistical Description of data

S/ No	Parameters	Count	Mean	Standa rd Deviati on	Mini mum	25%	50%	75%	Maxim um
1	age_days	482.00 0000	1.7468 88	0.6991 54	0.000 000	1.0000 00	2.0000 00	2.0000 00	3.0000 00
2	gest_age	482.00 0000	39.595 643	1.7472 41	37.00 0000	38.000 000	39.000 000	41.000 000	41.000 000
3	duration_of _labour	422.00 0000	19.594 787	17.640 458	0.000 000	8.0000 00	14.000 000	24.000 000	72.000 000
4	duration_of _ROM	440.00 0000	15.328 409	12.626 562	0.000 000	5.0000 00	13.000 000	23.000 000	72.000 000

5	weight	482.00 0000	3.0169 79	0.5390 18	1.140 000	2.6900 00	3.0000 00	3.3400 00	6.0
6	temperature	482.00 0000	38.611 411	1.3563 49	33.70 0000	38.025 000	38.700 000	39.200 000	50.000 000
7	respiratory_ rate	477.00 0000	60.616 352	17.778 799	0.000 000	50.000 000	59.000 000	69.000 000	168.00 0000
8	heart_rate	474.00 0000	151.94 0928	23.925 021	84.00 0000	138.00 0000	160.00 0000	166.00 0000	228.00 0000
9	wbc	396.00 0000	16.785 253	12.369 180	2.100 000	4.7675 00	13.150 000	30.725 000	60.570 000
10	neu_count	178.00 0000	4.0197 19	4.9669 46	1.250 000	1.5800 00	1.7000 00	2.5000 00	23.000 000
11	lym_count	73.000 000	5.3415 07	2.6323 97	1.300 000	3.2000 00	4.4000 00	7.1000 00	13.200 000

12	mon_count	24.000 000	1.5654 17	0.6726 07	0.510 000	1.1350 00	1.4750 00	1.7875 00	3.0700 00
13	eos_count	24.000 000	0.2225 00	0.3192 55	0.000 000	0.0500 00	0.1100 00	0.2175 00	1.4800 00
14	bas_count	24.000 000	0.0583 33	0.0996 81	0.000 000	0.0100 00	0.0300 00	0.0525 00	0.3900 00
15	rbc	87.000 000	4.2018 39	1.0212 49	0.750 000	3.7050 00	4.3400 00	4.8550 00	6.1300 00
16	platelet_co unt	360.00 0000	205.01 3889	131.82 0856	18.00 0000	113.75 0000	147.00 0000	283.25 0000	708.00 0000
17	neonatal_se psis	482.00 0000	0.7842 32	0.4117 81	0.000 000	1.0000 00	1.0000 00	1.0000 00	1.0000 00

355

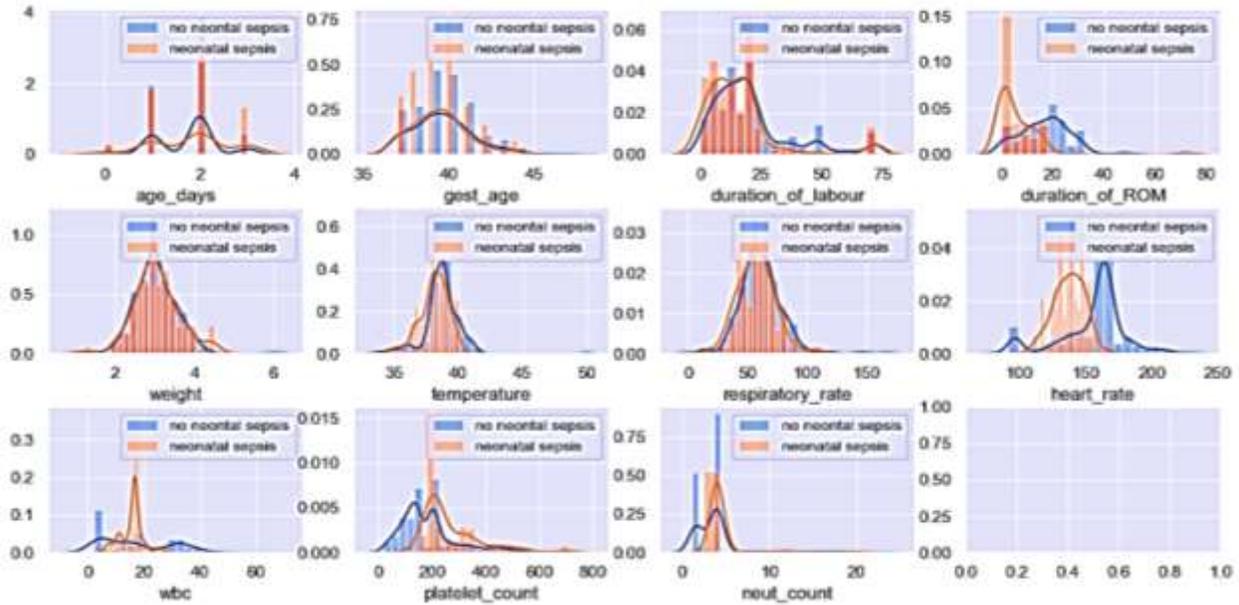
356 The descriptive statistics of the data are shown above in table 3. The target variable
357 (neonatal_sepsis) is binary and has a value either '1,' i.e., neonatal sepsis is true or '0,' i.e., no
358 neonatal sepsis. Information about the mean, standard deviation, maximum value, minimum value,
359 and distribution (quartile range) of each numeric parameter are presented in the table above.

360 Table 4: Missing Value Analysis (numeric parameters)

Parameter	Missing Count	Missing Percent
duration_of_labour	60	0.12
duration_of_ROM	42	0.09
respiratory_rate	5	0.01
heart_rate	8	0.02
wbc	86	0.18
neu_count	304	0.63
lym_count	409	0.85
mon_count	458	0.95
eos_count	458	0.95
bas_count	458	0.95
rbc	395	0.82
platelet_count	122	0.25

361

362 The Count column presents information about the total number of records of each feature. 12
363 parameters, duration_of_labour, duration_of_ROM, respiratory_rate, heart_rate, wbc, neu_count,
364 lym_count, mon_count, eos_count, bas_count, rbc and platelet_count have missing value out of
365 17 parameters which is given in table 4 above. Parameters with missing percent above 0.80 were
366 dropped.



367

368

Figure 9: Distribution plot of numeric features with target (neonatal_sepsis)

369

Figure 9 shows the distribution of the numeric features with respect to the target variable

370

(neonatal_sepsis). age_days, gest_age, weight, and respiratory_rate, are normally distributed with

371

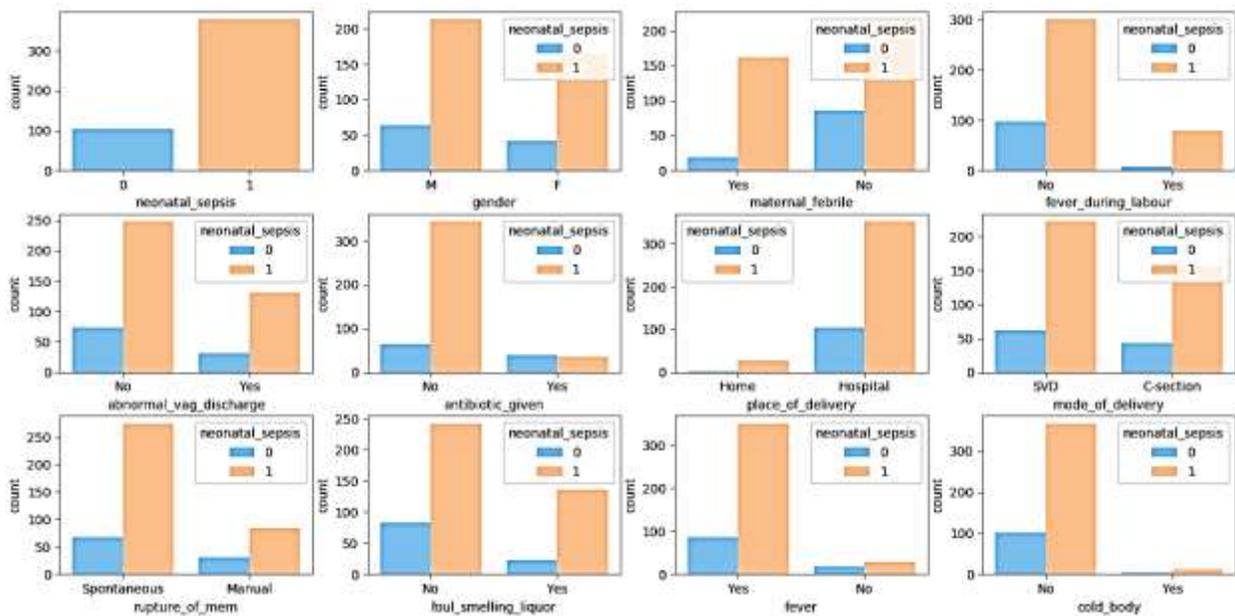
the neonatal sepsis. Parameters such as duration_of_labour, duration_of_ROM, wbc,

372

platelet_count, and neut_count are positively skewed. While temperature and heart_rate are

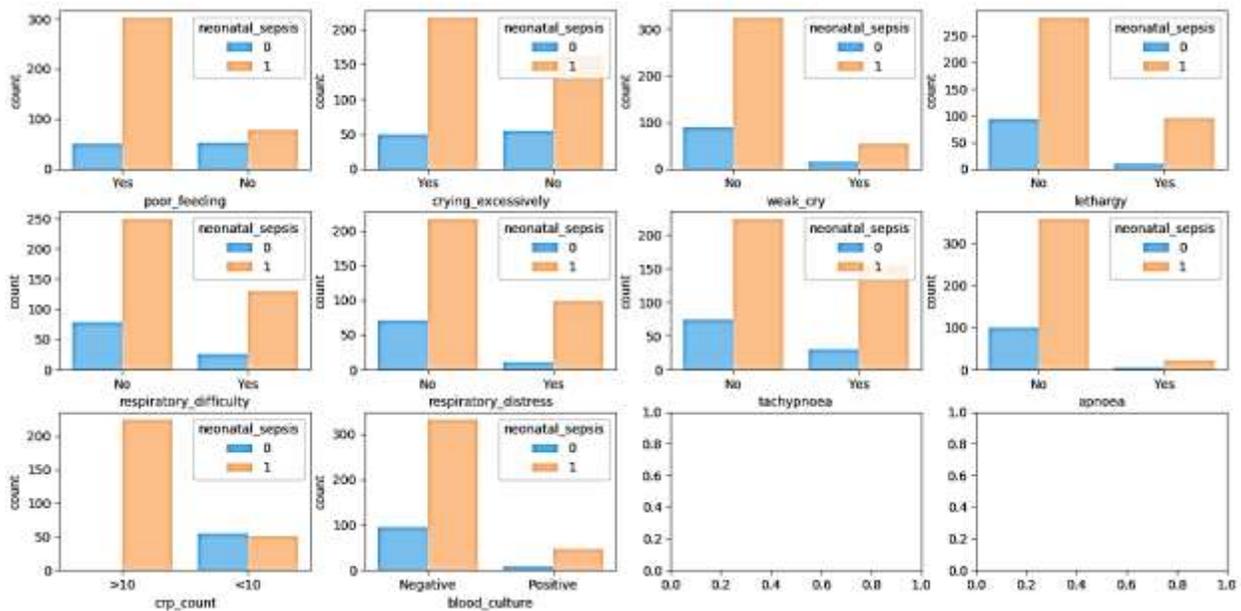
373

negatively skewed.



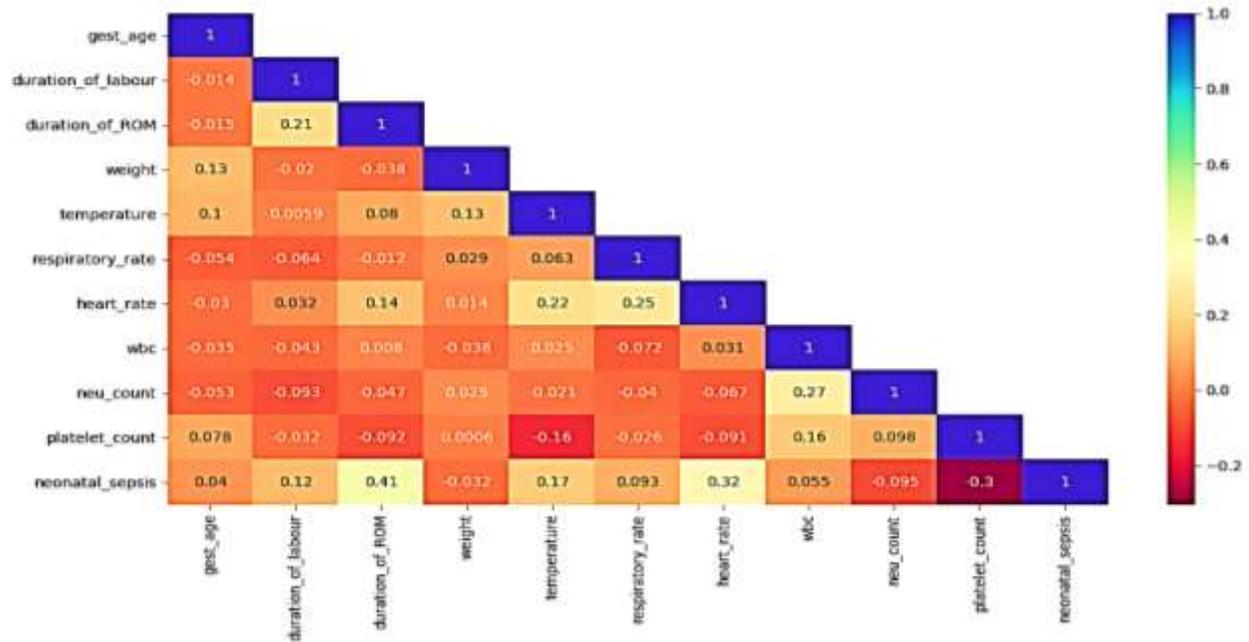
374

375 Figure 10: Distribution plot of categorical features with target (neonatal_sepsis)



376
377 Figure 11: Distribution plot of categorical features with target (neonatal_sepsis)

378 The frequency plot of categorical variables; gender, maternal_febrile, fever_during_labour,
379 abnormal_vaginal_discharge, antibiotic, place_of_delivery, mode_of_delivery, rupture_of_mem,
380 foul_smelling_liquor, fever, cold_body, poor_feeding, crying_excessively, weak_cry, lethargy,
381 respiratory_difficulty, respiratory_distress, tachypnoea, apnoea, crp_count, and blood_culture are
382 plotted as shown in figure 10 and figure 11 above. The 'Target (neonatal_sepsis)' variable is highly
383 biased as per the information provided by the bar graph. Only 22% of the values are '0,' and the
384 rest of the records have '1' values. The balancing of this target feature will be addressed in the
385 data preparation section. The categorical variables are binary, as shown in figures 10 and 11 above.



386

387

Figure 12: Heatmap matrix of features with Target (neonatal_sepsis)

388

The Pearson correlation coefficient was used in the experiment to interpret the linear association

389

between the numeric-continuous variables. The correlation coefficient range is from -1 to 1; the

390

linear relationship is stronger as the absolute value increases. The correlation heatmap matrix

391

shown in figure 12 shows the strength of the relationship between the features. The result deduced

392

from the matrix is stated below:

393

- All the variable features have very little correlation with neonatal sepsis.

394

- platelet_count and temperature are highly negatively correlated.

395

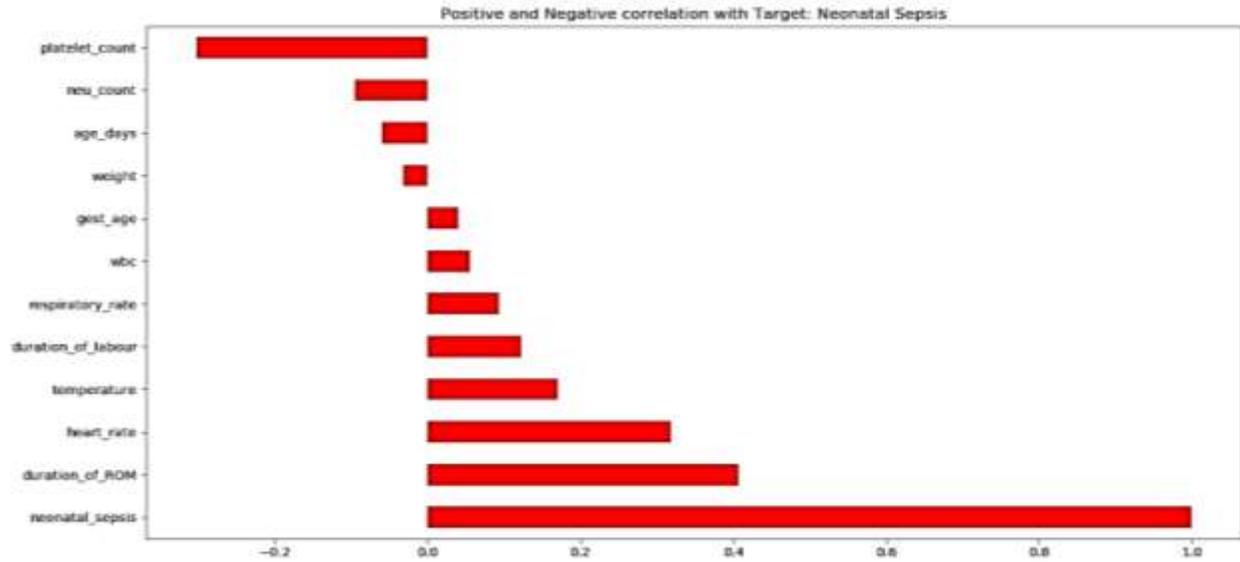
- heart_rate has a weak positive correlation with respiratory_rate and temperature.

396

- duration_of_ROM is weakly positively correlated with duration_of_labour.

397

- neu_count is weakly positively correlated with wbc.



398

Figure 13: Positive-Negative correlation with Target

399

400 Figure 13 shows the relationship strength and magnitude of relations between independent features
 401 (variables) and Target (neonatal_sepsis). The features on the left side of the axis have a negative
 402 correlation with neonatal sepsis, i.e., the increase in the value of these features will decrease the
 403 risk of neonatal sepsis. Whereas the variables on the right side of the axis have a positive
 404 correlation, i.e., the increase in these features' value will increase neonatal sepsis's risk. In addition,
 405 the height of the bar graph from the center of the axis shows the magnitude of the correlation
 406 strength of each feature with neonatal sepsis.

407 Data Preparation

408 After the data has been analyzed; the first step is to remove the issues identified in the dataset to
 409 enable it to fit for the modeling. The missing values of duration_of_labour, duration_of_ROM,
 410 respiratory_rate, heart_rate, wbc, neu_count, and platelet_count was imputed, respectively, with
 411 the mean value of each feature having missing values.

412 SMOTE Algorithm: Balancing of Dataset

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

417 Table 5: SMOTE oversampling

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

418
419 The data set was balanced by creating synthetic records using the SMOTE algorithm, an over-
420 sampling technique. Initially, the number of records belonging to neonatal sepsis as ‘1’ is
421 significantly higher than those belonging to class ‘0’, as shown in table 5 above. The number of
422 samples containing the ‘0’ value is increased to 50% after running SMOTE oversampling
423 algorithm.

424 **Normalization and Standardization: Z-score**

425 The distribution of the features shown in figure 9 provides information regarding the features'
426 skewness, which was eliminated by the predictor variables' feature scaling. Building the model
427 with normalized values was done by calculating the Z-score of each numeric variable.

428 **One-Hot Encoding: Categorical Features**

429 Since regression and support vector algorithms only work on numeric features and do not handle
430 features with string values, the next step after creating the balanced and normalized dataset is to
431 eliminate categorical features. The 21 categorical features are nominal, i.e., there is no particular

432 natural order in which their values follow. In order to handle these nominal features in
 433 classification, one-hot encoding was carried out.

434 Table 6: One-Hot Encoding

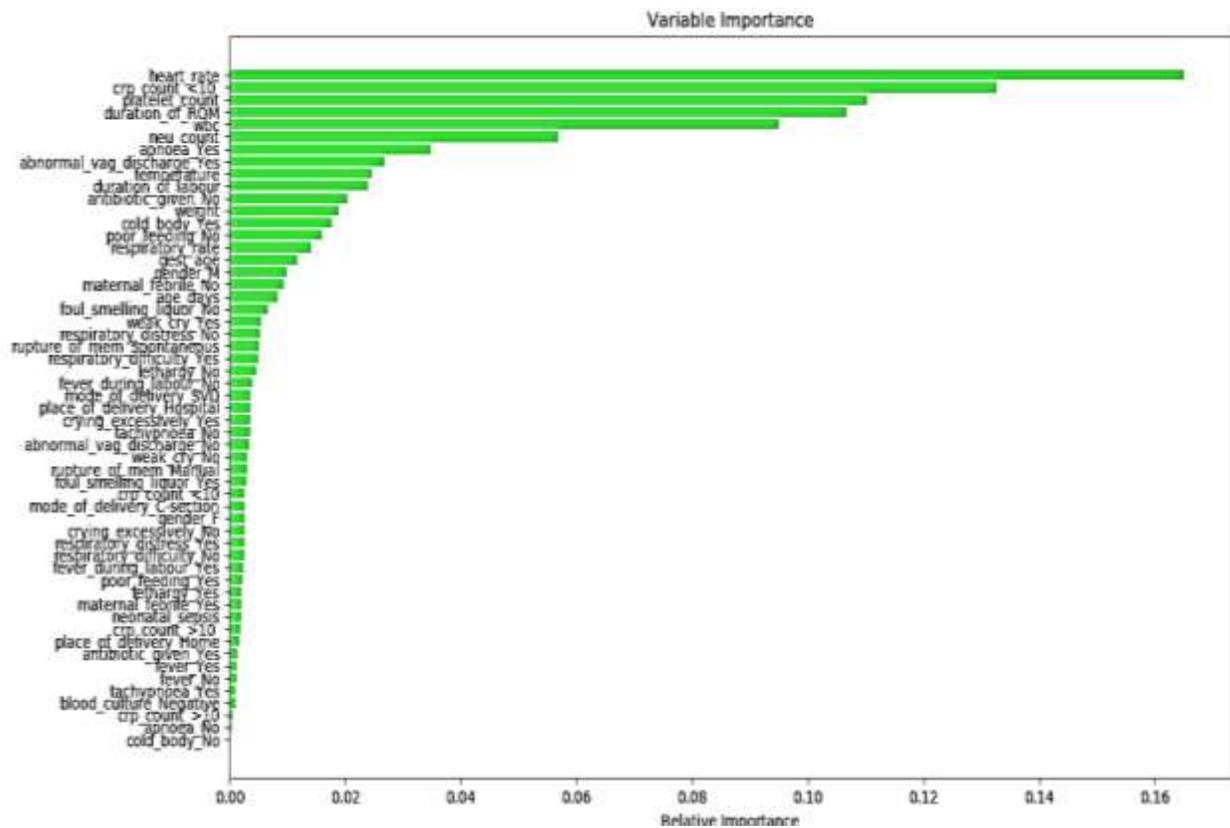
Categorical features	Dummy features
Gender	gender_F, gender_M
maternal_febriile	maternal_febriile_No, maternal_febriile_Yes
fever_during_labour	fever_during_labour_No, fever_during_labour_Yes,
abnormal_vaginal_discharge	abnormal_vaginal_discharge_No, abnormal_vaginal_discharge_Yes
Antibiotic	antibiotic_No, antibiotic_Yes
place_of_delivery	place_of_delivery_home, place_of_delivery_Hospital
mode_of_delivery	mode_of_delivery_C-section, mode_of_delivery_SVD
'rupture_of_mem	rupture_of_mem_Manual, rupture_of_mem_Spontaneous
foul_smelling_liquor	foul_smelling_liquor_No, foul_smelling_liquor_Yes
Fever	fever_No, fever_Yes
cold_body	cold_body_No, cold_body_Yes
poor_feeding	poor_feeding_No, poor_feeding_Yes
crying_excessively	crying_excessively_No, crying_excessively_Yes
weak_cry	weak_cry_No, weak_cry_Yes
Lethargy	lethargy_No, lethargy_Yes
respiratory_difficulty	respiratory_difficulty_No, respiratory_difficulty_Yes
respiratory_distress	respiratory_distress_No, respiratory_distress_Yes
Tachypnoea	tachypnoea_No, tachypnoea_Yes
Apnoea	apnoea_No, apnoea_Yes

crp_count	crp_count_<10, crp_count_>10
blood_culture	blood_culture_Negative, blood_culture_Positive

435

436 One-hot encoding involves creating dummy features where each possible value of the nominal
 437 feature has a binary value. The number of features increased to 53 after applying one-hot encoding
 438 to the 21 categorical features. Dummy features created from categorical features is shown above
 439 (Table 6).

440 **Random Forest Classifier: Feature Importance**



441

442 Figure 14: Random Forest Classifier: Feature importance

443 Finally, the Random Forest algorithm was used to identify each feature's importance in predicting
 444 neonatal sepsis, shown in figure 14. Each feature is assigned a score to show the importance of the
 445 feature in the model. The higher the score, the higher the importance of that particular feature.

446 While features with a lower score are considered the least important. The height of the bar graph
447 shows how important each feature is with neonatal sepsis.

448 **Modelling**

449 **The Developed Algorithm for Neonatal Sepsis Prediction**

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

452 Table 7: Pseudo code for the maternal condition

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

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

Step 3: FOR each i in R DO

 IF $i \leftarrow b_0 = \text{"Yes"}$ THEN

 RETURN True

 ELIF $i \leftarrow b_1 = \text{"Yes"}$ THEN

 RETURN True

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

 RETURN True

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

 RETURN True

 ELIF $i \leftarrow b_4 = \text{"Yes"}$ THEN

 RETURN True

 ELIF $i \leftarrow b_5 = \text{"No"}$ THEN

 RETURN True

ELSE

```
RETURN False
END IF
END FOR
Step 4: IF True  $\geq$  1
RETURN "Maternal Condition"
ELSE
RETURN "No Maternal Condition"
END IF
```

453

454 **Phase I: Maternal Condition**

455 This phase checks if a neonate has a maternal condition. The algorithm looks through the maternal
456 risk characteristics provided and determines based on the values if a neonate has a maternal
457 condition or not. Table 7 shows the pseudo-code of the maternal condition phase.

458 **The Parameters used (maternal risk characteristics):**

459 a0 = Fever during labor.

460 a1 = Maternal febrile during pregnancy.

461 a2 = Duration of rupture of membrane

462 a3 = Duration of labor

463 a4 = Foul odor of the amniotic fluid.

464 a5 = Antibiotic treatment received by mother \leq 4 hours prior to delivery.

465 **Parameter's value:**

466 b0 = (a0 = Yes)

467 b1 = (a1 = Yes)

468 b2 = (a2 = \geq 18 hours)

469 b3 = (a3 = ≥ 18 hours)

470 b4 = (a4 = Yes)

471 b5 = (a5 = No)

472 Table 8: Pseudo code for the observational condition

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

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

Step 3: FOR each j in S DO

 IF $j \leftarrow d_0 = \text{“}\geq 160\text{”}$ OR $\text{“}\leq 100\text{”}$ THEN

 RETURN True

 ELIF $j \leftarrow d_1 = \text{“}\geq 38\text{”}$ OR $\text{“}\leq 36.5\text{”}$ THEN

 RETURN True

 ELIF $j \leftarrow d_2 = \text{“Yes”}$ THEN

 RETURN True

 ELIF $j \leftarrow d_3 = \text{“Yes”}$ THEN

 RETURN True

 ELIF $j \leftarrow d_4 = \text{“Yes”}$ THEN

 RETURN True

 ELIF $j \leftarrow d_5 = \text{“Yes”}$ 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

473

474 **Phase II: Observational Condition**

475 This phase checks if a neonate has an observational condition. The algorithm looks through the
476 neonatal clinical signs provided and determines based on the values if a neonate has an
477 observational condition or not. Table 8 shows the pseudo-code of the observational condition
478 phase.

479 **The Parameters used (neonatal clinical signs):**

480 c0 = Heart rate

481 c1 = Temperature

482 c2 = Respiratory distress

483 c3 = Apnea condition.

484 c4 = Lethargy condition.

485 c5 = Feeding difficulty

486 **Parameter’s value:**

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

488 d1 = (c1 = $\geq 38^{\circ}\text{C}$ (fever) or $\leq 36.5^{\circ}\text{C}$ (hypothermia))

489 d2 = (c2 = Yes)

490 d3 = (c3 = Yes)

491 d4 = (c4 = Yes)

492 d5 = (c5 = Yes)

493 Table 9: Pseudo code for the laboratory condition

Step 1: Create a tuple L of the 4 parameters declared above, $L = (e_0 \dots e_c)$, $1 \leq c \leq 4$

Step 2: Initialize elements of tuple L; $T = (f_0 \dots f_k)$, $1 \leq k \leq 4$

Step 3: FOR each k in T DO

 IF $k \leftarrow f_0 = \text{“}\geq 10\text{”}$ THEN

 RETURN True

 ELIF $k \leftarrow f_1 = \text{“}\leq 5,000\text{”}$ OR $\text{“}\geq 30,000\text{”}$ THEN

 RETURN True

 ELIF $k \leftarrow f_2 = \text{“}\leq 1,750\text{”}$ THEN

 RETURN True

 ELIF $k \leftarrow f_3 = \text{“}\leq 150,000\text{”}$ 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

494

495 **Phase III: Laboratory Condition**

496 This phase checks if a neonate has a laboratory condition. The algorithm looks through the
497 laboratory tests provided and determines based on the values if a neonate has a laboratory condition
498 or not. Table 9 shows the pseudo-code of the laboratory condition phase.

499 **The Parameters used (laboratory tests):**

500 e0 = C-reactive protein

501 e1 = White blood cell count

502 e2 = Neutrophil count

503 e3 = Platelet count

504 **Parameter's value:**

505 f0 = (e0 = ≥ 10 mg/L)

506 f1 = (e1 = $\leq 5,000$ or $\geq 30,000$ per microL)

507 f2 = (e2 = $\leq 1,750$ per microL)

508 f3 = (e3 = $\leq 150,000$ per microL)

509 Table 10: Pseudo code for the neonatal sepsis

Step 1: Create a set N of the 3 parameters declared above, $N = (g_0 \dots g_e), 1 \leq e \leq 3$

Step 2: Initialize elements of set N; $P = (h_0 \dots h_y), 1 \leq y \leq 3$

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

510

511 **Phase IV: Neonatal Sepsis**

512 This phase checks if a neonate has neonatal sepsis. The algorithm looks through the maternal
513 condition, observational condition, and laboratory condition and determines based on their
514 outcomes if a neonate has sepsis or not. Table 10 shows the pseudo-code of the neonatal sepsis
515 phase.

516 **The Parameters used (neonatal sepsis variables):**

517 g0 = Maternal condition

518 g1 = Observational condition

519 g2 = Laboratory condition

520 **Parameter’s value:**

521 h0 = (g0 = Yes)

522 h1 = (g1 = Yes)

523 h2 = (g2 = Yes)

524 In this phase, classification models (SVM, LR, KNN, NB, and DT) were built based on the
525 proposed algorithm for predicting neonatal sepsis. The balanced dataset created after the addition
526 of dummy features is used as training data to build the models. The input data is first normalized

527 before training of the model. In order to achieve a better fit of the model, correlation and
 528 multicollinearity analysis is performed. The features were all very weakly correlated with each
 529 other, and therefore, none of the features were dropped while training the model. In total, seven
 530 supervised machine learning models are built, SVM_L, SVM_R, SVM_P, LR, KNN, NB, and DT.
 531 The models had 10 accuracies per model as each model is fitted by running 10 iterations with each
 532 iteration giving the model's accuracy.

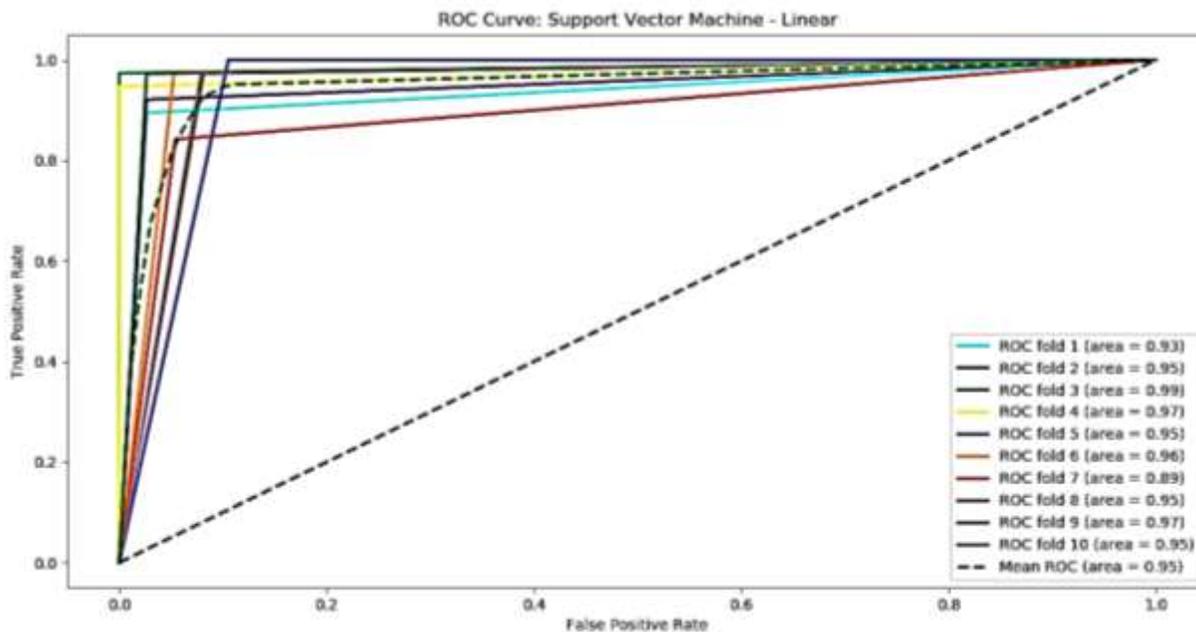
533 **Support Vector Machine: Target Values**

534 Table 11: SVM kernels and tuning parameters

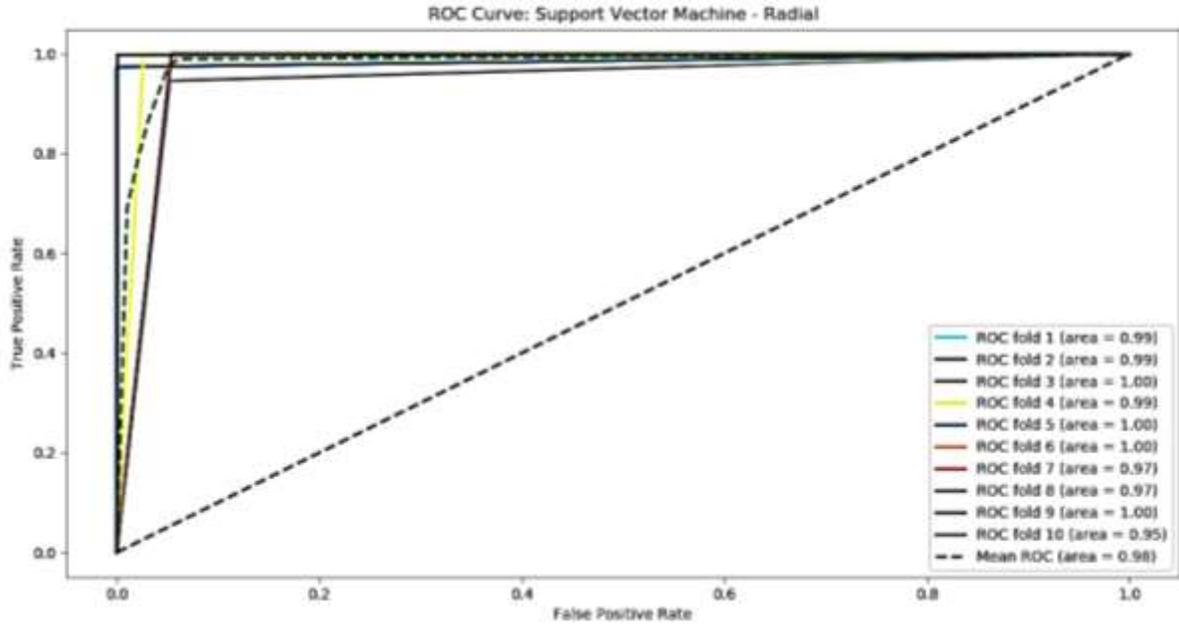
SVM kernels	C	γ
Linear	1,2,3,4,5,6,7,8,9,10	0.01,0.02,0.03,0.04,0.05,0.10,0.2,0.3,0.4,0.5
Radial basis function	1,2,3,4,5,6,7,8,9,10	0.01,0.02,0.03,0.04,0.05,0.10,0.2,0.3,0.4,0.5
Polynomial	1,2,3,4,5,6,7,8,9,10	0.01,0.02,0.03,0.04,0.05,0.10,0.2,0.3,0.4,0.5

535
 536 Firstly, the Support Vector Machine is built using 10 k-fold stratified validation technique to split
 537 data into training and test set. Three types of SVM models were used, SVM with linear kernel,
 538 radial kernel, and polynomial kernel tuned with different values of tuning parameters ‘C’ and ‘ γ ’
 539 as shown in table 11 above. Scikit-learn in python has inbuilt packages that hold functions for the
 540 stratified k-fold cross validator, SVM modeling, LR modeling, KNN modeling, NB modeling, and
 541 DT modeling. The ‘StratifiedKFold’ function is fed with all the data instances, having 10 as the
 542 number of splits used. The 10 folds are randomly created, of which 9 of these folds are used for
 543 model training, and one sample is set aside for model testing. The folds are created in a way that
 544 each fold contains an equal ratio of the target variable; let’s say if there are 70-30 ratio of neonatal
 545 sepsis and non-neonatal sepsis in the first fold, then other folds will also have 70-30 ratio. In total,

546 38 features were used in training the model, of which 21 are categorical features, and one-hot
547 encoding was used to convert them into binary vectors as expressed in data preparation.



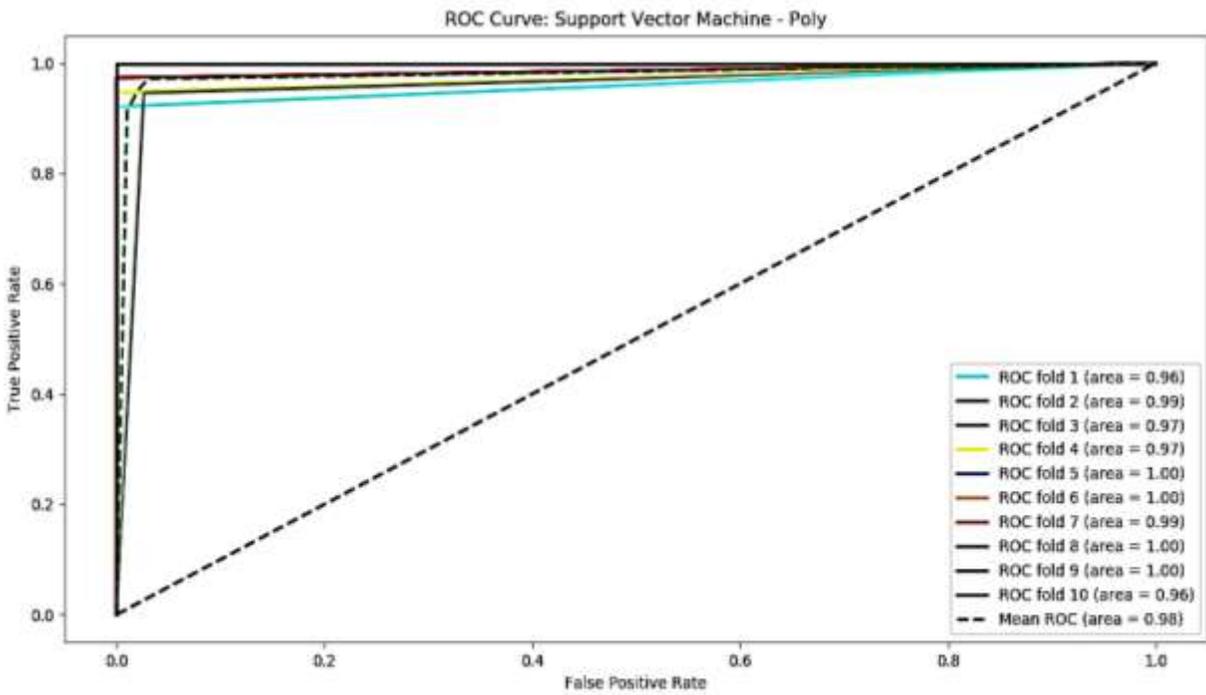
548
549 Figure 15: ROC curve: Support Vector Machine – Linear
550 The linear SVM algorithm model obtained a minimum and maximum accuracy of 0.89% and
551 0.97%, with average classifier accuracy of 0.95%. From the ROC curve, it can be seen that almost
552 half of the folds achieved accuracy above 0.85%, as shown in figure 15 above. It can be deduced
553 from the results that the model's performance with linear SVM is slightly higher than KNN and
554 NB models.



555

556

Figure 16: ROC curve: Support Vector Machine – Radial



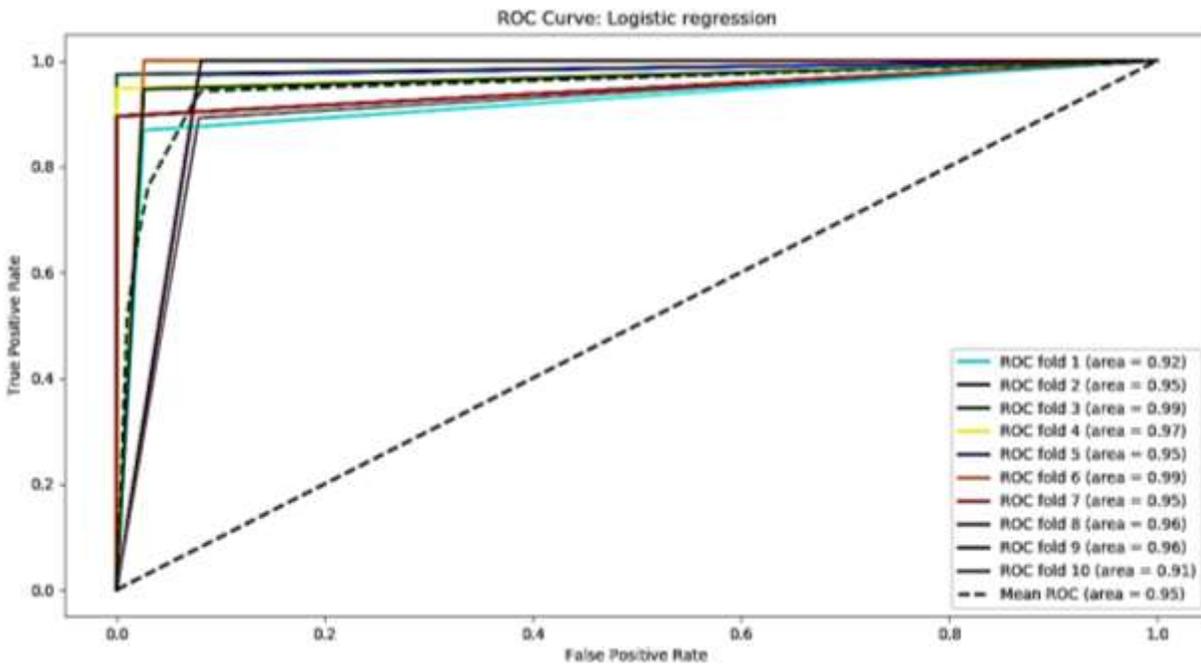
557

558

Figure 17: ROC curve: Support Vector Machine – Poly

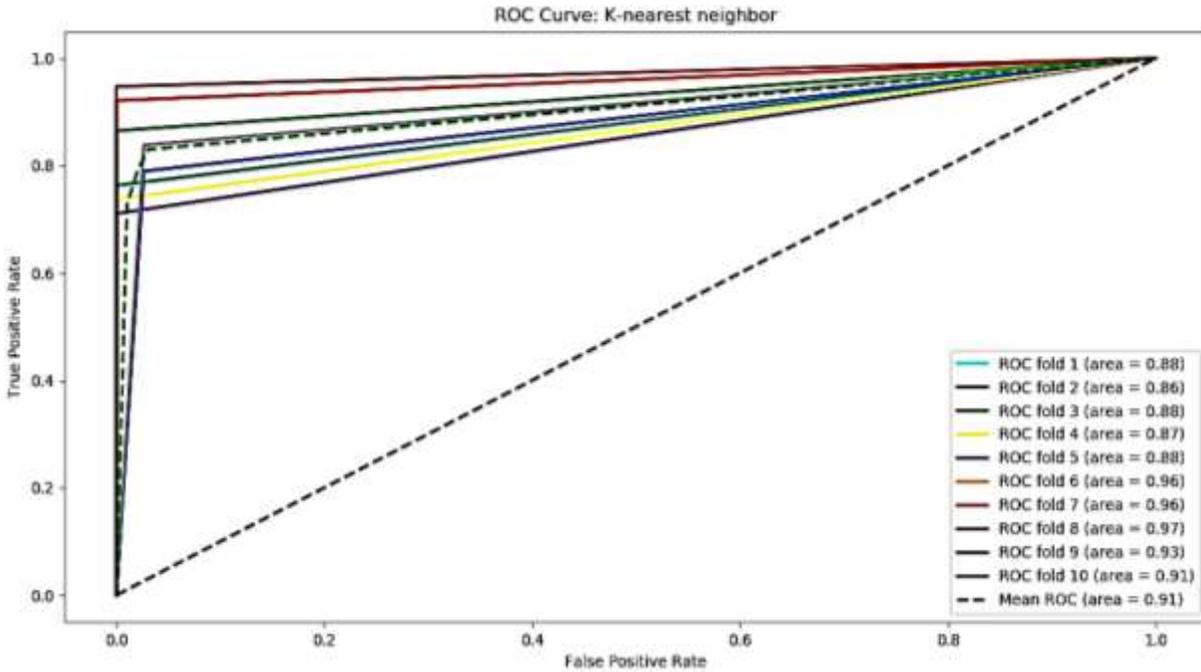
559 The ROC curves were plotted separately for SVM radial basis function and polynomial kernels.
560 The average classifier accuracy and accuracy per iteration are shown in figure 16 and figure 17.
561 Both models obtained a mean accuracy of 0.98%.

562 **Logistic Regression: Target Values**



563
564 Figure 18: ROC curve: Logistic Regression
565 Similarly, the Logistic Regression model is built using 10 k-fold stratified samplings to create
566 training and test datasets. The model is a binary classification regression model. After the training
567 of the model, then it is used to predict the target value. The model's score is created through this
568 process, which then gives the prediction accuracy of the model. Finally, 10 scores are then created
569 in which the mean of these scores gives the average accuracy of the LR classifier. The LR model
570 obtained a minimum and maximum accuracy of 0.91% and 0.99%, respectively, with an average
571 mean accuracy of 0.95%, as shown in figure 18 above.

572 **K-Nearest Neighbor: Target Values**



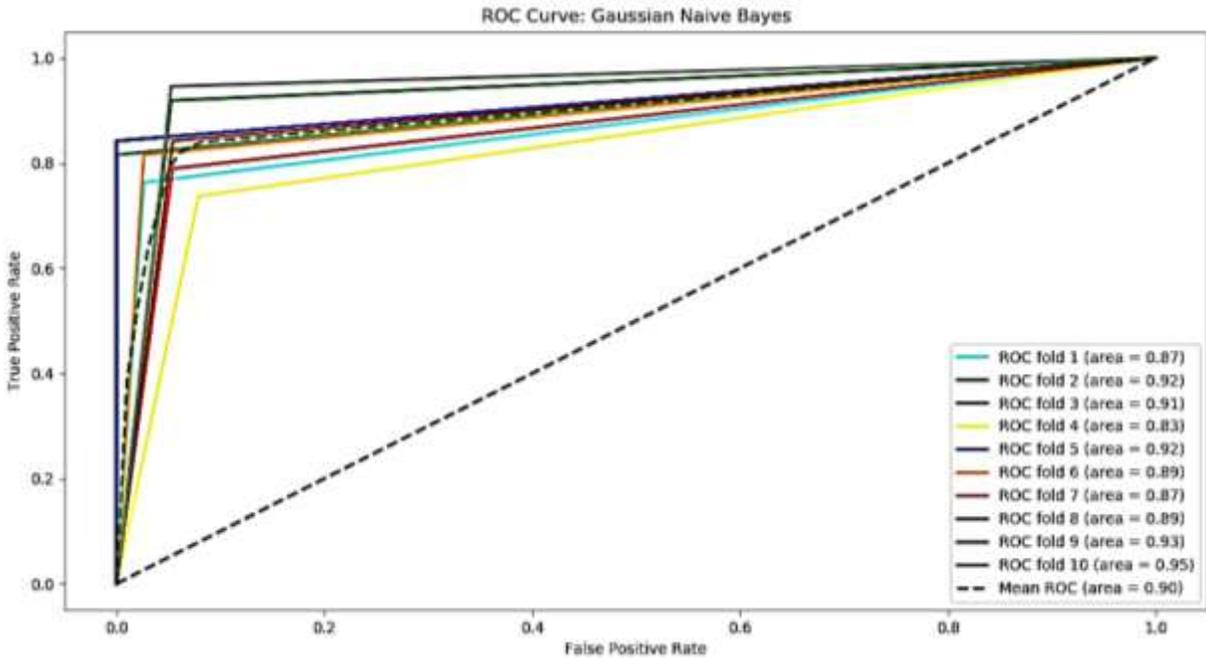
573

574

Figure 19: ROC curve: K-nearest neighbor

575 K-nearest neighbor model is built using 10 k-fold stratified samplings to create training and test
 576 datasets. KNN is preferably used when the features all have continuous value. Classification is
 577 achieved when the nearest neighbor is identified, which helps determine the class of an unknown
 578 sample. The model obtained a minimum and maximum accuracy of 0.86% and 0.97% with average
 579 classifier accuracy of 0.91%, as shown in figure 19 above.

580 **Naïve Bayes: Target Values**



581

582

Figure 20: ROC curve: Naïve bayes

583 Naïve Bayes model is built using 10 k-fold stratified samplings to create training and test datasets.

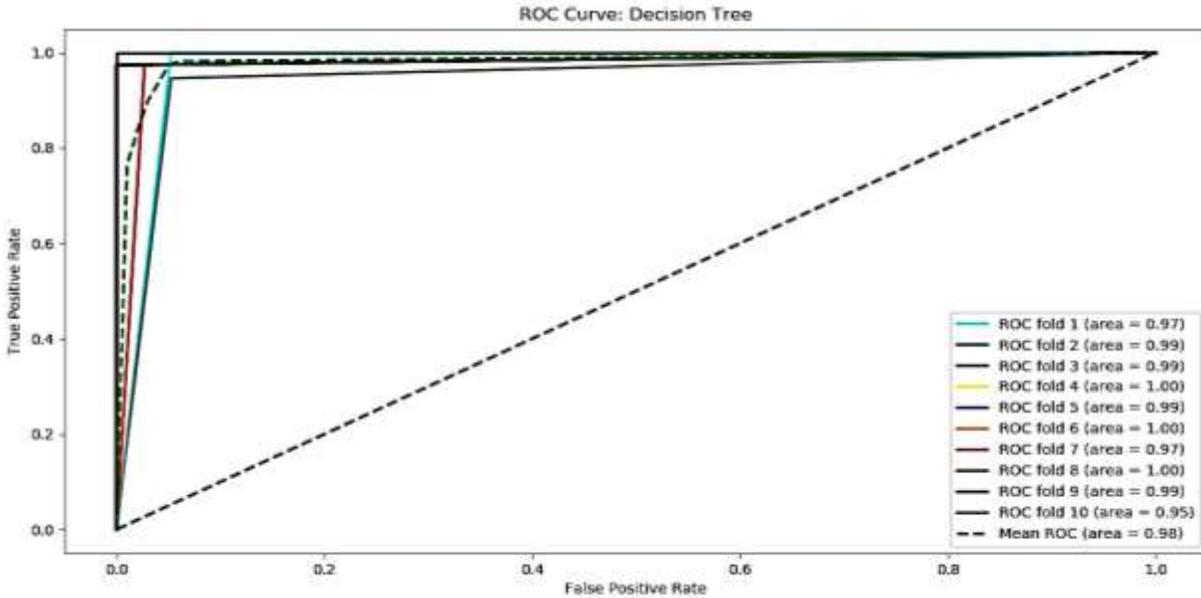
584 The model uses all the attributes in the data and analyses these attributes individually as though

585 they all have equal importance and independent of each other. The model obtained a minimum

586 and maximum accuracy of 0.83% and 0.95% with average classifier accuracy of 0.90%, as shown

587 in figure 20 above.

588 **Decision Tree: Target Values**



589

590

Figure 21: ROC curve: Decision tree

591 The decision tree model is built using 10 k-fold stratified samplings to create training and test

592 datasets. For the model to classify a new item, it first needs to generate a decision tree based on

593 the attribute values of the available training data. The model obtained a minimum and maximum

594 accuracy of 0.95% and 1.00% with average classifier accuracy of 0.98%, as shown in figure 21

595 below. The classifiers' classification accuracy will be discussed further in the next chapter,

596 'Discussion.'

597 **Evaluation**

598 In order to evaluate the model performance, a ROC-AUC curve is required, which was created in

599 the modeling section. The computing of True Positive Rate (TPR) and False Positive Rate (FPR)

600 is the key requirement for plotting of ROC curve. roc_curve() and auc() are inbuilt functions in

601 Sklearn, which returns TPR and FPR as output.

602 Table 12: 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	

603

604 The models were compared with an average accuracy achieved after each iteration from the
605 stratified k-fold validation technique used to split train-test data. The proposed algorithm's
606 performance was then compared with the physician's diagnosis shown in table 12 above.

607 Table 13: 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

608

609 Table 13 above shows the performance measures generated by fixing specificity at 0.11.

610

611 Table 14: 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_POL	0.97	0.86	0.97	0.86	0.95	0.07	0.98
Y							
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

612

613 Table 14 above shows the performance measures generated by fixing sensitivity at 0.89.

614

615 **Summary of Implementation**

616 This chapter outlines the practical application of the experiments to answer the research questions.
617 The exploratory data analysis is carried out on the dataset to understand the distribution of features
618 in the dataset with respect to the target variable. In order to explore the strength of the relationship
619 between each feature, a heatmap matrix was created. Except for features having missing value
620 above 0.80%, none of the features were dropped because all the features were weakly correlated
621 with each other.

622 To make the data fit for modeling, data-preprocessing techniques that involve imputation of
623 missing value, balancing of the values of the target variable using SMOTE algorithm,
624 normalization of variables using Z-score, and creation of dummy variables were all used. Further,
625 seven neonatal sepsis prediction models were built using SVM, LR, KNN, NB, and DT based on
626 the proposed algorithm. The training-test dataset was derived from stratified K-fold cross-
627 validation techniques giving a result of 10 accuracies per model. ROC curve for the built models
628 was created, and the AUC value was derived. Finally, to evaluate the performance of the proposed
629 algorithm, the sensitivity and specificity were compared with that of the physician diagnosis.

630

631 **Discussion**

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

635 This research proposes an algorithm for neonatal sepsis prediction, which was used to train five
636 supervised machine learning algorithms, and their performance was evaluated using the AUROC

637 value. The classifiers are trained on a set of samples with balanced dependent variable values by
638 applying the oversampling data technique. Before the training, data preprocessing steps such as
639 imputation of missing values, feature standardization, and normalization, generation of dummy
640 variables have been applied to the features.

641 **Performance Evaluation of the Proposed Algorithm**

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

654 One of the studies that are closest to this study reported in the literature is a retrospective study for
655 predicting neonatal late-onset sepsis (LOS) using the RALIS algorithm that consists of neonatal
656 clinical signs (54). Mithal et al. (2018) reported an AUC of 0.90 for LOS prediction using linear
657 regression based on a comparison between cases and controls (54). The second is also a
658 retrospective study for predicting neonatal LOS using a diagnostic algorithm consisting of neonatal
659 clinical signs and laboratory tests (55). Mani et al. (2014) explored a set of ML algorithms (SVM,

660 NB, TAN, AODE, KNN, CART, RF, LR, and LBR) with the highest AUROC value been 0.65
 661 based on a comparison with the physician's treatment (55). In contrast, this study focused on early-
 662 onset sepsis (EOS). It explored a set of ML algorithms with the highest AUROC value been 0.98
 663 and the lowest being 0.90 based on a comparison with the physician's diagnosis. It included more
 664 variables in the proposed algorithm to distinguish neonates without sepsis to avoid subjecting
 665 neonates without sepsis to unnecessary antibiotics use.

666 The proposed algorithm with ML algorithms may also identify truly infected neonates before the
 667 availability of blood culture tests and, therefore, contribute to earlier detection and treatment. The
 668 improvement in the sensitivity of the proposed algorithm is not at the cost of its specificity. The
 669 proposed algorithm and the ML algorithms used in this study have significant real-time strengths.
 670 They could be used as an early warning system to alert physicians that neonatal sepsis may be
 671 present or developing. However, like the vital signs monitoring proposed by Gur et al. (2015) and
 672 clinically evaluated by Mithal et al. (2018), these tools should be used as decision support tools
 673 and not as stand-alone decision-making expert systems (54,56). The proposed algorithm has to be
 674 tested in prospective settings and using data from other institutions (in future studies) to ascertain
 675 its clinical setting performance.

676 **Statistical Significance of the Experimental Results**

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

680 Table 15: 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

681

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

686 **Strength and Limitations of Results**

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

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

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

702 Moving ahead to the limitations, the proposed algorithm was developed based on the available
703 screening parameters on the patient's records from MRRH. This limited the study from exploring
704 some important screening parameters. The missing values in the dataset were higher with the
705 laboratory tests, limiting the number of laboratory tests used. The proposed algorithm may function
706 differently if modified with the identified screening parameters that are not currently in the
707 algorithm. This can be explored further as part of the future study.

708

709 Another limitation of this study is that the ML algorithms' training and testing are based on a small-
710 sized dataset. The dataset trends are biased; records containing neonatal sepsis as true are (3/4) of
711 the records. If a relevant size of data is used for the experiment, the ML algorithms may function

712 differently, and this can also be explored further as part of the future study. Lastly, the ML
713 algorithms were not compared with an AUPRC value due to the time limit.

714 **Summary of Analysis**

715 In this chapter, the breakdown and evaluation of the whole experiment are discussed. The five
716 algorithms are built on a dataset with balanced values. The proposed algorithm outperformed the
717 physician's diagnosis.

718 The SVM algorithms and DT outperformed the other ML algorithms in the prediction of neonatal
719 sepsis. The 'Wilcoxon Signed-Rank Test' was used to calculate each result's statistical significance
720 with a p-value <0.05 . The results show that the proposed algorithm with SVM and DT algorithms
721 is appropriate and efficient for predicting neonatal sepsis.

722 The results' strength and the limitation are explained, focusing on the proposed algorithm
723 performance and the data preprocessing techniques used to enhance the ML algorithms'
724 performance. The small-sized imbalanced data and the limited number of screening parameters
725 used in developing the proposed algorithm were the experiments' major limitations. A detailed
726 summary of the research, its contribution, impact, and future research areas are given in the
727 following chapter.

728

729 **Conclusion**

730 The proposed algorithm was developed based on three main variables, which include; maternal
731 risk factors, neonatal clinical signs, and laboratory tests. The proposed algorithm was compared
732 with the physician's diagnosis, and the proposed algorithm was found to outperform the physician's
733 diagnosis. The study provides evidence that the combination of maternal risk factors, neonatal
734 clinical signs, and laboratory tests can effectively diagnose neonatal sepsis. Based on the study

735 result, the proposed algorithm can help identify neonatal sepsis cases as it exceeded clinicians'
736 sensitivity and specificity. A prospective study is warranted to test the algorithm's clinical utility,
737 which could provide a decision support aid to clinicians. This will undoubtedly improve the early
738 recognition and treatment of neonatal sepsis. The study results suggest that ML algorithms can
739 identify neonatal sepsis cases within a large and complex database.

740 **Future Work & Recommendations**

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

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

757

758 **List of 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

759

760 **Declarations**

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

764 **Ethics approval and consent to participate**

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

769 **Consent for publication**

770 Not applicable

771 **Availability of data and materials**

772 Data is available in: <https://github.com/Helenaden/Neonatal-Sepsis-Prediction>

773 **Competing interests**

774 The authors declare that they have no competing interests.

775 **Funding**

776 No external funding was obtained for this study.

777 **Authors' contributions**

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

785 **Acknowledgements**

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

788 **Authors' information**

789 DPE

790 Department: Information Technology

791 Course: Masters in Health Information Technology

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

793 Institution: Mbarara University of Science and Technology (MUST)

794 Position: Post-Graduate Student

795

796 WW

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

798 Department: Biomedical Sciences and Engineering

799 Institution: Mbarara University of Science and Technology (MUST)

800 Position: Head of Department (HOD)

801

802 AM

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

804 Department: Information Technology

805 Institution: Mbarara University of Science and Technology (MUST)

806 Position: Deputy Dean of the Faculty

807 Location: Mbarara, Uganda

808

809 SK

810 Authors' qualifications: MBChB, MMed

811 Department: Pediatrics and Child Health MRRH

812 Hospital: Mbarara Regional Referral Hospital (MRRH)

813 Position: Head of Department (HOD), Neonatology division

814

815 **References**

816 1. Hug L, Sharrow D, Sun Y, Marcusanu A, You D, Mathers C, et al. Levels and Trends in

- 817 Child Mortality Report 2017 | UNICEF [Internet]. 2017 [cited 2021 Jan 11]. Available from:
818 <https://www.unicef.org/reports/levels-and-trends-child-mortality-report-2017>
- 819 2. WHO. Newborns: improving survival and well-being [Internet]. 2020 [cited 2021 Jan 11].
820 Available from: [https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-](https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality#)
821 [mortality#](https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality#)
- 822 3. UN-IGME. CME Info - Child Mortality Estimates [Internet]. 2018 [cited 2021 Jan 11].
823 Available from: <https://childmortality.org/data/Uganda>
- 824 4. Goldstein B, Giroir B, Randolph A. International pediatric sepsis consensus conference:
825 Definitions for sepsis and organ dysfunction in pediatrics. *Pediatr Crit Care Med* [Internet].
826 2005 Jan [cited 2021 Jan 11];6(1). Available from:
827 [https://journals.lww.com/pccmjournal/Fulltext/2005/01000/International_pediatric_sepsis](https://journals.lww.com/pccmjournal/Fulltext/2005/01000/International_pediatric_sepsis_consensus.2.aspx)
828 [_consensus.2.aspx](https://journals.lww.com/pccmjournal/Fulltext/2005/01000/International_pediatric_sepsis_consensus.2.aspx)
- 829 5. Reinhart K, Bauer M, Riedemann NC, Hartog CS. New approaches to sepsis: Molecular
830 diagnostics and biomarkers. *Clin Microbiol Rev*. 2012;25(4):609–34.
- 831 6. Balamuth F, Alpern ER, Abbadessa MK, Hayes K, Schast A, Lavelle J, et al. Improving
832 Recognition of Pediatric Severe Sepsis in the Emergency Department: Contributions of a
833 Vital Sign Based Electronic Alert and Bedside Clinician Identification. *Ann Emerg Med*
834 [Internet]. 2017 Dec 1 [cited 2021 Jan 11];70(6):759. Available from:
835 </pmc/articles/PMC5698118/>
- 836 7. Smyth MA, Brace-McDonnell SJ, Perkins GD. Identification of adults with sepsis in the
837 prehospital environment: a systematic review. *BMJ Open* [Internet]. 2016 Aug 1 [cited 2021
838 Jan 11];6(8). Available from: </pmc/articles/PMC4985978/>
- 839 8. Olvera L, Dutra D. Early Recognition and Management of Maternal Sepsis. *Nurs Womens*

- 840 Health [Internet]. 2016 Apr 1 [cited 2021 Jan 11];20(2):182–96. Available from:
841 <http://www.nwhjournal.org/article/S1751485116000738/fulltext>
- 842 9. Jones SL, Ashton CM, Kiehne L, Gigliotti E, Bell-Gordon C, Disbot M, et al. Reductions
843 in Sepsis Mortality and Costs After Design and Implementation of a Nurse-Based Early
844 Recognition and Response Program. *Jt Comm J Qual Patient Saf* [Internet]. 2015 Nov 1
845 [cited 2021 Jan 11];41(11):483. Available from: </pmc/articles/PMC4880050/>
- 846 10. Fell DB, Hawken S, Wong CA, Wilson LA, Malia SQ, Chakraborty P, et al. Using newborn
847 screening analytes to identify cases of neonatal sepsis. 2017;(December):1–10.
- 848 11. Bonet M, Souza JP, Abalos E, Fawole B, Knight M, Kouanda S, et al. The global maternal
849 sepsis study and awareness campaign (GLOSS): study protocol. *Reprod Health* [Internet].
850 2018 Jan 30 [cited 2021 Jan 11];15(1). Available from: </pmc/articles/PMC5791346/>
- 851 12. Sarkar AP, Dhar G, Das Sarkar M, Ghosh TK, Ghosh S. Early diagnosis of neonatal sepsis
852 in primary health care unit. *Bangladesh J Med Sci* [Internet]. 2015 Apr 18 [cited 2021 Jan
853 12];14(2):169–72. Available from:
854 <https://www.banglajol.info/index.php/BJMS/article/view/21806>
- 855 13. Liu L, Oza S, Hogan D, Perin J, Rudan I, Lawn JE, et al. Global, regional, and national
856 causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an
857 updated systematic analysis. *Lancet* [Internet]. 2015 Jan 31 [cited 2021 Jan
858 12];385(9966):430–40. Available from:
859 <http://www.thelancet.com/article/S0140673614616986/fulltext>
- 860 14. Voller SMB, Myers PJ. Neonatal Sepsis. *Clin Pediatr Emerg Med*. 2016 Jun 1;17(2):129–
861 33.
- 862 15. HNN. Leading causes of neonatal deaths in Uganda [Internet]. 2018 [cited 2021 Jan 12].

- 863 Available from: <https://www.healthynewbornnetwork.org/country/uganda/>
- 864 16. Zea-Vera A, Ochoa TJ. Challenges in the diagnosis and management of neonatal sepsis. *J*
865 *Trop Pediatr*. 2015;61(1):1–13.
- 866 17. Shane AL, Sánchez PJ, Stoll BJ. Neonatal sepsis. *Lancet* [Internet]. 2017 Oct 14 [cited 2021
867 Jan 12];390(10104):1770–80. Available from:
868 <http://www.thelancet.com/article/S0140673617310024/fulltext>
- 869 18. Stoll BJ, Hansen NI, Bell EF, Shankaran S, Laptook AR, Walsh MC, et al. Neonatal
870 outcomes of extremely preterm infants from the NICHD Neonatal Research Network.
871 *Pediatrics* [Internet]. 2010;126(3):443–56. Available from:
872 <http://www.pediatrics.org/misc/reprints.shtml>
- 873 19. Stephens BE, Vohr BR. Neurodevelopmental outcome of the premature infant. *Pediatr Clin*
874 *North Am* [Internet]. 2009 Jun [cited 2021 Jan 12];56(3):631–46. Available from:
875 <https://pubmed.ncbi.nlm.nih.gov/19501696/>
- 876 20. Stoll BJ, Hansen NI, Adams-Chapman I, Fanaroff AA, Hintz SR, Vohr B, et al.
877 Neurodevelopmental and Growth Impairment Among Extremely Low-Birth-Weight Infants
878 With Neonatal Infection. *JAMA* [Internet]. 2004 Nov 17 [cited 2021 Jan 12];292(19):2357–
879 65. Available from: <https://jamanetwork.com/journals/jama/fullarticle/199811>
- 880 21. Masino AJ, Harris MC, Forsyth D, Ostapenko S, Srinivasan L, Bonafide CP, et al. Machine
881 learning models for early sepsis recognition in the neonatal intensive care unit using readily
882 available electronic health record data. *PLoS One*. 2019;14(2):1–23.
- 883 22. Hajj J, Blaine N, Salavaci J, Jacoby D. The “Centrality of Sepsis”: A Review on Incidence,
884 Mortality, and Cost of Care. *Healthcare* [Internet]. 2018 Sep 1 [cited 2021 Jan 12];6(3).
885 Available from: </pmc/articles/PMC6164723/>

- 886 23. Gyang E, Shieh L, Forsey L, Maggio P. A Nurse-Driven Screening Tool for the Early
887 Identification of Sepsis in an Intermediate Care Unit Setting. *J Hosp Med* [Internet]. 2015
888 Feb 1 [cited 2021 Jan 12];10(2):97. Available from: [/pmc/articles/PMC4816455/](#)
- 889 24. Cohen J, Vincent JL, Adhikari NKJ, Machado FR, Angus DC, Calandra T, et al. Sepsis: A
890 roadmap for future research. *Lancet Infect Dis* [Internet]. 2015;15(5):581–614. Available
891 from: [http://dx.doi.org/10.1016/S1473-3099\(15\)70112-X](http://dx.doi.org/10.1016/S1473-3099(15)70112-X)
- 892 25. Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and Costs of Sepsis
893 in the United States-An Analysis Based on Timing of Diagnosis and Severity Level. *Crit*
894 *Care Med* [Internet]. 2018 [cited 2021 Jan 12];46(12):1889–97. Available from:
895 <https://pubmed.ncbi.nlm.nih.gov/30048332/>
- 896 26. Oeser C, Lutsar I, Metsvaht T, Turner MA, Heath PT, Sharland M. Clinical trials in neonatal
897 sepsis. *J Antimicrob Chemother*. 2013;68(12):2733–45.
- 898 27. James.L.Wynn. Defining neonatal sepsis. *Curr Opin paediatr*. 2016;28(2):135–40.
- 899 28. Weiss SL, Fitzgerald JC, Balamuth F, Alpern ER, Lavelle J, Chilutti M, et al. Delayed
900 antimicrobial therapy increases mortality and organ dysfunction duration in pediatric sepsis.
901 *Crit Care Med*. 2014;42(11):2409–17.
- 902 29. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment
903 of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis
904 and Septic Shock (Sepsis-3). *JAMA* [Internet]. 2016 Feb 23 [cited 2021 Jan
905 12];315(8):762–74. Available from:
906 <https://jamanetwork.com/journals/jama/fullarticle/2492875>
- 907 30. Shah BA, Padbury JF. Neonatal sepsis: An old problem with new insights. *Virulence*
908 [Internet]. 2014 [cited 2021 Jan 13];5(1):170. Available from: [/pmc/articles/PMC3916371/](#)

- 909 31. Klingenberg C, Kornelisse RF, Buonocore G, Maier RF, Stocker M. Culture-Negative
910 Early-Onset Neonatal Sepsis — At the Crossroad Between Efficient Sepsis Care and
911 Antimicrobial Stewardship. *Front Pediatr* [Internet]. 2018 [cited 2021 Jan 13];6:285.
912 Available from: [/pmc/articles/PMC6189301/](https://pubmed.ncbi.nlm.nih.gov/3189301/)
- 913 32. Fuchs A, Bielicki J, Mathur S, Sharland M, Van Den Anker JN. Antibiotic Use for Sepsis
914 in Neonates and Children: 2016 Evidence Update WHO-Reviews. 2016.
- 915 33. Deleon C, Shattuck K, Jain SK. Biomarkers of neonatal sepsis. *Neoreviews* [Internet]. 2015
916 [cited 2021 Jan 13];16(5):e297–308. Available from:
917 <https://researchexperts.utmb.edu/en/publications/biomarkers-of-neonatal-sepsis>
- 918 34. Tank PJ, Omar A, Musoke R. Audit of Antibiotic Prescribing Practices for Neonatal Sepsis
919 and Measurement of Outcome in New Born Unit at Kenyatta National Hospital. *Int J Pediatr*
920 (United Kingdom). 2019;2019.
- 921 35. R. K, Manjunath S, Doddabasappa P, J. M. Evaluation of screening of neonatal sepsis. *Int*
922 *J Contemp Pediatr* [Internet]. 2018 Feb 22 [cited 2021 Jan 13];5(2):580–3. Available from:
923 <https://www.ijpediatrics.com/index.php/ijcp/article/view/1378>
- 924 36. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al.
925 Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local
926 Big Data-Driven, Machine Learning Approach. *Acad Emerg Med*. 2016;23(3):269–78.
- 927 37. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in
928 the intensive care unit with minimal electronic health record data: A machine learning
929 approach. *JMIR Med Informatics*. 2016;4(3):1–15.
- 930 38. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score
931 (TREWScore) for septic shock. *Sci Transl Med*. 2015;7(299).

- 932 39. Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep
933 neural networks. *Comput Biol Med* [Internet]. 2017;89:248–55. Available from:
934 <http://dx.doi.org/10.1016/j.combiomed.2017.08.015>
- 935 40. Mayhew MB, Petersen BK, Sales AP, Greene JD, Liu VX, Wasson TS. Flexible, cluster-
936 based analysis of the electronic medical record of sepsis with composite mixture models. *J*
937 *Biomed Inform* [Internet]. 2018;78:33–42. Available from:
938 <https://doi.org/10.1016/j.jbi.2017.11.015>
- 939 41. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to
940 clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision
941 support system. *J Am Med Informatics Assoc*. 2014;21(2):315–25.
- 942 42. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable
943 Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Crit Care Med*.
944 2018;46(4):547–53.
- 945 43. Fairchild KD, Aschner JL. HeRO monitoring to reduce mortality in NICU patients. *Res*
946 *Reports Neonatol* [Internet]. 2012 Aug 15 [cited 2021 Jan 14];2:65–76. Available from:
947 [https://www.dovepress.com/hero-monitoring-to-reduce-mortality-in-nicu-patients-peer-](https://www.dovepress.com/hero-monitoring-to-reduce-mortality-in-nicu-patients-peer-reviewed-fulltext-article-RRN)
948 [reviewed-fulltext-article-RRN](https://www.dovepress.com/hero-monitoring-to-reduce-mortality-in-nicu-patients-peer-reviewed-fulltext-article-RRN)
- 949 44. Coggins SA, Weitkamp JH, Grunwald L, Stark AR, Reese J, Walsh W, et al. Heart Rate
950 Characteristic index monitoring for blood stream infection in an NICU: A 3-year
951 experience. *Arch Dis Child Fetal Neonatal Ed* [Internet]. 2016 Jul 1 [cited 2021 Jan
952 15];101(4):F329. Available from: [/pmc/articles/PMC4851911/](https://pubmed.ncbi.nlm.nih.gov/26711111/)
- 953 45. Li L, Walter SR, Rathnayake K, Westbrook JI. Evaluation and optimisation of risk
954 identification tools for the early detection of sepsis in adult inpatients. G. Balint, Antala B,

- 955 Carty C, Mabieme J-MA, Amar IB, Kaplanova A, editors. Uniw śląski [Internet]. 2018
956 [cited 2021 Jan 14];343–54. Available from:
957 [https://researchers.mq.edu.au/en/publications/evaluation-and-optimisation-of-risk-](https://researchers.mq.edu.au/en/publications/evaluation-and-optimisation-of-risk-identification-tools-for-the-)
958 [identification-tools-for-the-](https://researchers.mq.edu.au/en/publications/evaluation-and-optimisation-of-risk-identification-tools-for-the-)
- 959 46. Leventhal B. An introduction to data mining and other techniques for advanced analytics. J
960 Direct, Data Digit Mark Pract. 2010;12(2):137–53.
- 961 47. Fernando KES, Mcgregor C, James AG. CRISP-TDM0 for standardized knowledge
962 discovery from physiological data streams: Retinopathy of prematurity and blood oxygen
963 saturation case study. 2017 IEEE Life Sci Conf LSC 2017. 2018 Jan 23;2018-Janua:226–9.
- 964 48. Scikit-learn. scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation
965 [Internet]. [cited 2021 Jan 14]. Available from: <https://scikit-learn.org/stable/>
- 966 49. Rohit Walimbe. Handling imbalanced dataset in supervised learning using family of
967 SMOTE algorithm. - DataScienceCentral.com [Internet]. 2017 [cited 2021 Jan 14].
968 Available from: [https://www.datasciencecentral.com/handling-imbalanced-data-sets-in-](https://www.datasciencecentral.com/handling-imbalanced-data-sets-in-supervised-learning-using-family/)
969 [supervised-learning-using-family/](https://www.datasciencecentral.com/handling-imbalanced-data-sets-in-supervised-learning-using-family/)
- 970 50. Boser BE, Guyon IM, Vapnik VN. Training algorithm for optimal margin classifiers. Proc
971 Fifth Annu ACM Work Comput Learn Theory. 1992;144–52.
- 972 51. Verplancke T, Van Looy S, Benoit D, Vansteelandt S, Depuydt P, De Turck F, et al. Support
973 vector machine versus logistic regression modeling for prediction of hospital mortality in
974 critically ill patients with haematological malignancies. BMC Med Inform Decis Mak
975 [Internet]. 2008 [cited 2021 Jan 14];8. Available from:
976 <https://pubmed.ncbi.nlm.nih.gov/19061509/>
- 977 52. Mark Peng. General Tips for participating Kaggle Competitions [Internet]. 2015 [cited 2021

978 Jan 14]. Available from: [https://www.slideshare.net/markpeng/general-tips-for-](https://www.slideshare.net/markpeng/general-tips-for-participating-kaggle-competitions)
979 [participating-kaggle-competitions](https://www.slideshare.net/markpeng/general-tips-for-participating-kaggle-competitions)

980 53. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: Challenges, strategies, and
981 a comparison of machine learning approaches. *Med Care* [Internet]. 2010 Jun [cited 2021
982 Jan 14];48(6 SUPPL.). Available from: [https://journals.lww.com/lww-](https://journals.lww.com/lww-medicalcare/Fulltext/2010/06001/Prediction_Modeling_Using_EHR_Data__Challenges,.17.aspx)
983 [medicalcare/Fulltext/2010/06001/Prediction_Modeling_Using_EHR_Data__Challenges,.1](https://journals.lww.com/lww-medicalcare/Fulltext/2010/06001/Prediction_Modeling_Using_EHR_Data__Challenges,.17.aspx)
984 [7.aspx](https://journals.lww.com/lww-medicalcare/Fulltext/2010/06001/Prediction_Modeling_Using_EHR_Data__Challenges,.17.aspx)

985 54. Mithal LB, Yogev R, Palac HL, Kaminsky D, Gur I, Mestan KK. Vital signs analysis
986 algorithm detects inflammatory response in premature infants with late onset sepsis and
987 necrotizing enterocolitis. *Early Hum Dev* [Internet]. 2018 Feb 1 [cited 2021 Jan 16];117:83.
988 Available from: [/pmc/articles/PMC5983899/](https://pubmed.ncbi.nlm.nih.gov/35983899/)

989 55. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision
990 support using machine learning for early detection of late-onset neonatal sepsis. *J Am Med*
991 *Informatics Assoc* [Internet]. 2014 [cited 2021 Jan 16];21(2):326–36. Available from:
992 [/pmc/articles/PMC3932458/](https://pubmed.ncbi.nlm.nih.gov/263932458/)

993 56. Gur I, Riskin A, Markel G, Bader D, Nave Y, Barzilay B, et al. Pilot study of a new
994 mathematical algorithm for early detection of late-onset sepsis in very low-birth-weight
995 infants. *Am J Perinatol* [Internet]. 2015 Jul 31 [cited 2021 Jan 16];32(4):321–30. Available
996 from: <http://www.thieme-connect.com/products/ejournals/html/10.1055/s-0034-1384645>
997

Supplementary Files

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

- [SupplementaryInfofile.docx](#)