Engineered Science
DOI: https://dx.doi.org/10.30919/es976

Early Detection of Late Onset Neonatal Sepsis Using Machine Learning Algorithms

Roshan David Jathanna,[1,*] Dinesh Acharya,[1] Leslie Edward Lewis[2] and Krishnamoorthi Makkithaya[1]

Abstract

Reducing newborn mortality by 2030 is a Sustainable Development Goals target 3.2. Neonatal sepsis is the third major cause of neonatal death after prematurity and birth asphyxia. Late-onset neonatal sepsis (LOS) refers to sepsis in neonates between the ages of 3 and 28 days and is likely to be acquired from the environment rather than maternal transmission. Since clinical features are not evident in the initial stages of infection, early diagnosis of LOS is challenging. Studies have shown that physiological parameters can predict LOS before prominent clinical features. Clinicians can use these parameters as early warning signs to monitor neonates closely and intervene earlier to prevent complications and provide effective treatment. This paper compares various machine learning algorithms to predict the onset of neonatal sepsis using vital signs, laboratory measurements, and observations captured within 24 hours of admission from the MIMIC III dataset. Experimental results show that adaptive boosting, light gradient boosting and random forest with Synthetic Minority Oversampling Technique give the highest area under the receiver operating characteristic (AUROC) of 0.9248, 0.9245, and 0.9238, respectively, among all the algorithms evaluated using 10-fold stratified cross-validation. The soft voting classifier trained on an ensemble of the top three models predicted the onset of neonatal sepsis with an AUROC of 0.9266, accuracy of 0.8553, F1 score of 0.7829, and Matthew's correlation coefficient of 0.6995.

Keywords: Neonatal Sepsis; Machine Learning; MIMIC III; SDG 3.2.
Received: 21 August 2023; Revised: 29 September 2023; Accepted: 03 October 2023.
Article type: Research article.

1. Introduction

A neonate is an infant less than four weeks old. According to WHO Report 2021, out of 5 million deaths of under-five children worldwide, 2.3 million are neonates. In India, over half of the under-five deaths occurred during the first four weeks.[1] Reducing newborn mortality to 12 per 1,000 live births by 2030 is a Sustainable Development Goals (SDG) target 3.2.[2] Babies born before 37 weeks are considered premature or preterm. Neonatal Intensive Care Unit (NICU) specializes in treating sick or premature neonates. Medical practitioners working at NICUs have been saving the lives of many thousands of children born ill or premature.

Neonatal sepsis is the third major cause of neonatal death after prematurity and birth asphyxia-related conditions.[1] Sepsis is a life-threatening medical condition where the immune system overreacts to infection by releasing chemicals into the blood to fight infection, which results in inflammation and clotting of blood, causing less blood flow to internal organs and limbs. In severe cases, infection leads to a life-threatening drop in blood pressure called septic shock, which can fail several organs like the lungs, kidneys, and liver and even result in death.[3] Early onset neonatal sepsis (EOS) and late onset neonatal sepsis (LOS) are two types of sepsis in neonates based on the timing of infection. In EOS, the onset of symptoms of infection occurs within three days after birth. Vertical transmission from bacteria in the maternal genital tract is the primary reason for EOS. Horizontal transmission of pathogens acquired after birth is the primary reason for LOS, and it occurs gradually and subtly after three days of life but with very harmful effects.[4] Since clinical features are not evident in the initial stages of infection, early diagnosis of LOS sepsis is challenging. Studies have shown that

[1] *Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India.*
[2] *Department of Paediatrics, Kasturba Medical College, Manipal Academy of Higher Education, Manipal 576104, India.*
**Email:* roshan.jathanna@manipal.edu (R. D. Jathanna)

physiological parameters like heart rate variability and respiratory traits can predict LOS before obvious clinical symptoms.[5-9] Fig. 1 shows the traditional method of identification of LOS.

Medical Information Mart for Intensive Care III (MIMIC-III)[10-12] is a publicly available dataset that researchers widely used to build models predicting sepsis,[13,14] mortality,[15,16] and length of stay.[17] Researchers have built clinical decision systems[18] which suggest appropriate treatment based on patient data. Researchers have used the dataset to evaluate the safety of drugs used in the Intensive Care Unit (ICU).[19]

SNAPPE-II (Score for Neonatal Acute Physiology with Perinatal Extension-II)[20] is a scoring system used to assess the severity of illness and predict mortality risk in neonates admitted to the NICU. The score incorporates both physiological and perinatal factors to provide an estimation of the neonate's health status. Neonates with higher SNAPPE-II scores are susceptible to sepsis due to compromised health. Clinical Risk Index for Babies (CRIB) II[21] is a scoring system used to predict mortality risk in very low birth weight infants, combining birth weight, gestational age, and other clinical factors obtained during the first 12 hours of admission to the NICU.

Machine Learning is the field of computer science that can act without being explicitly programmed. Healthcare, financial banking, education, manufacturing engineering, fraud detection, intrusion detection, customer segmentation, and bioinformatics widely use machine learning to uncover hidden patterns, correlations, and other insights.[22]

Shirwaikar *et al*., have used supervised machine learning techniques to diagnose of neonatal diseases and found that the ensemble technique has better predictive power than Support Vector Machine (SVM), decision trees, and neural networks.[23] Parvin *et al*. systematically reviewed Scopus, Web of Science, and PubMed databases from 2015 to 2022 to analyze the performance of machine learning and deep learning algorithms for predicting sepsis and neonatal sepsis. Eleven papers selected were from different medical care units, with area under the receiver operating characteristic (AUROC) ranging from 0.68 to 0.95. The studies evaluated a variety of algorithms, including logistic regression, random forest, support vector machines, and neural networks.

Robi and Sitote have used the classification stacking model to predict four primary neonatal diseases: sepsis, birth asphyxia, necrotizing enterocolitis (NEC), and respiratory distress syndrome, accounting for 75% of neonatal deaths. The dataset was collected from Asella Comprehensive Hospital between 2018 and 2021. Comparisons were made with three other machine learning models (Xtreme Gradient Boosting, Random Forest, and SVM), and the developed stacking model demonstrated superior performance.[24] The findings suggest that the machine learning approach can significantly contribute to early detection and accurate diagnosis of neonatal diseases, particularly in resource-limited healthcare facilities.

In this study, various machine learning algorithms are compared to predict the onset of neonatal sepsis using vital signs, laboratory measurements, and observations captured within 24 hours of admission. The voting classifier, trained on the ensemble of adaptive boosting, light gradient boosting without Synthetic Minority Oversampling Technique (SMOTE) and random forest with SMOTE was able to perform with the highest AUROC, Accuracy, Recall, MCC and F1 score. Early detection of a medical complication results in effortless and effective treatment.

2. Methodology

Figure 2 illustrates the systematic procedure adopted for early detection of LOS using the MIMIC-III dataset. The data acquired is subjected to a data preprocessing pipeline that encompasses strategies for addressing missing values through imputation, standardizing data via normalization techniques, and conducting feature selection to enhance the quality and relevance of the dataset. Subsequently, the dataset is split into training and testing sets, ensuring robust model performance. In the realm of model development, a range of supervised machine learning algorithms undergo intensive training on the training set. The subsequent model evaluation employs metrics cross validated based on the training set, enabling a comprehensive performance assessment and facilitating the selection of the most suitable model. The final step involves evaluating the final trained models with a testing set to verify generalizability. Once the optimal model is finalised, data from the initial 24 hours post-admission, gathered from bedside monitors and the hospital information system, are preprocessed, deidentified, and input into the model for predicting the occurrence of LOS. In cases where sepsis in the neonate is predicted, stakeholders are notified to make informed decisions.

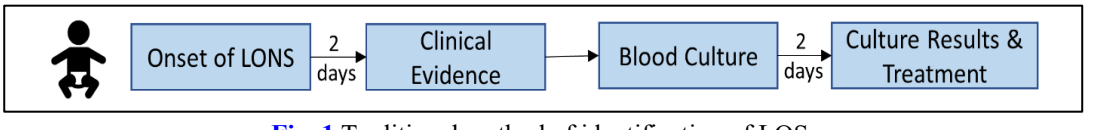

Fig. 1 Traditional method of identification of LOS.

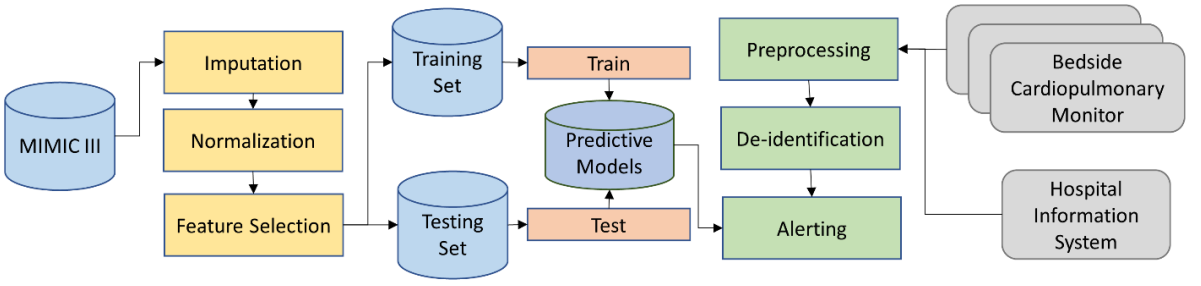

Fig. 2 Overall methodology.

2.1 Dataset

MIMIC-III[10-12] comprises 61,532 ICU admissions to Beth Israel Deaconess Medical Center from 2001 to 2012. It contains data on 7870 neonates. Access to MIMIC III was obtained after successfully completing an accredited course on protecting human research subjects, which includes adhering to the requirements of the Health Insurance Portability and Accountability Act (HIPAA) and signing data use agreement that specifies permissible data usage procedures, upholds strict security requirements, and forbids any attempts to determine the identities of specific patients.

2.2 Data preprocessing

The MIMIC-III dataset is distributed in the form of comma-separated value (CSV) files, accompanied by Structured Query Language (SQL) scripts to facilitate importing data into database systems. For convenient access, secure server with PostgresSQL 13.1 was created, where a copy of the entire database (Around 49 G.B.) was downloaded. PostgreSQL was selected because it is open-source relational database management system (RDBMS) that ensures data reliability, integrity, and provides advanced query optimization capabilities. The enitre MIMIC III database is installed in PostresSQL and using the Jupter notebook, sepsis-3-get-data.ipynb, various MIMIC III tables such as admissions, chartevents, icustays, inputevents, outputevents, labevents, microbiologyevents, noteevents, patients, prescriptions, procedureevents and services were queried to create a dataset.[25] There were 53550 records related to patients who were admitted after five days of birth, which had to be excluded because they may be an adult or admitted because of infection. Since this study is limited to neonates, adult records are excluded. Secondary admissions are excluded because this study aims to predict neonatal sepsis using the first 24-hour admission data. Multiple admissions for the same individual may introduce dependencies in the data, leading to biased results and over-estimating the model's performance hence 138 records of second admission were omitted. To ensure data completeness and reliability in the analysis, 7 records that did not have corresponding charted data in chartevents table were excluded. 5 records of newborns with EOS are excluded because they have different risk factors and clinical presentations compared to those with LOS, and the presence of EOS might confound the analysis of risk factors for LOS. The response variable for this study was whether the neonate would be infected with sepsis or not. The most widely used methods to mark the neonate as septic are: 1) Entry of one of the International Classification of Diseases (ICD-9) codes (995.92 for severe sepsis and 785.52 for septic shock). 2) Suspected infection with associated organ dysfunction (SOFA $\geq$ 2) score calculated at the temporal context of suspected infection (Sepsis-3)[26] 3) Martin *et al.*[27] proposed ICD-9 codes. Table 1 shows the number of records identified in MIMIC III for each criterion. For this study, the sepsis-3 criteria were used, and out of 7832 neonates, 2184 (27.88%) were labeled septic. Studies[28-34] have shown that sepsis-3 criteria is accurate and reliable in identifying neonates with sepsis.

Table 1. Class-wise distribution of records.

Criteria	Not Septic	Septic
Explicit	7831	1
Sepsis-3[26]	5648	2184
Martin *et al.*[27]	7790	42

The database was queried using Jupyter notebook[25] to obtain the values from tables for 118 features on the neonate's first 24 hours of admission. Since 35 features did not have value for any neonate, they were dropped. The data from MIMIC III had a lot of missing values for neonates. Handling missing values is crucial as it directly affects the accuracy and reliability of diagnostic or predictive models. Inappropriately imputing missing data can introduce biases, noise, and compromise model performance, potentially leading to misdiagnoses or inaccurate predictions, which will severely affect patient safety and treatment decisions. A robust imputation approach is essential to ensure trustworthy model predictions and provide clinicians with better insights. In our

study, the missing values in nominal features were imputed by different methods such as mean, median, mode, k nearest neighbours and iterative imputation with 10 iterations of Light Gradient Boosting Machine (LightGBM).[35,36]

Normalization of data ensures that a particular feature will not intrinsically influence the result because of its large value. This study explored three common normalization techniques, namely z-score, Min-Max, and Maximum Absolute Value Scalar, to preprocess our input data. The goal was to investigate their impact on model performance when predicting the outcome variable.

Our results showed that the Maximum Absolute Value Scalar normalization technique outperformed the other methods, exhibiting the highest AUROC value among the three. Hence for this study, the Maximum Absolute Value Scalar as the normalization technique was selected due to its ability to achieve the highest AUROC, indicating better discrimination and predictive accuracy.

Table 2. Features Selected for training.

Sl.	Features	F1 score	p-value < 0.05
1	heartrate_max	2038.55	0
2	wbc_max	1065.46	0
3	bilirubin_max	446.60	0
4	hematocrit_max	382.46	0
5	platelet_max	318.32	0
6	heartrate_mean	208.81	0
7	heartrate_min	171.51	0
8	po2_max	95.20	0
9	po2a_max	79.21	0
10	v_glucose_max	38.20	0
11	pco2a_max	34.85	0
12	totalco2_max	31.77	0
13	pco2_max	23.22	0
14	fio2_max	17.86	0
15	urineoutput	15.01	0.0001
16	potassium_max	11.93	0.0006
17	aniongap_max	11.08	0.0009
18	fio2a_max	10.23	0.0014
19	sodium_max	9.30	0.0023
20	fio2_chartevents_max	2.84	0.0321
21	bicarbonate_max	0.67	0.0411
22	ph_max	0.30	0.0486
23	chloride_max	0.27	0.0496

In this study, the response feature is categorical (septic/not septic), and the predictor is continuous, hence ANOVA f-test is used to select features.[37,38] ANOVA computes the F-statistic, a ratio of between-group variance to within-group variance. A high F-statistic indicates the feature is relevant for classification because of the large difference in means between the groups. The corresponding p-value indicates the probability of observing such a significant difference by chance. The 23 features with low p-values (<0.05) are selected for inclusion in the binary classification model because they are statistically significant, are shown in Table 2. For training the model, 70% of the data was used, and 30% was reserved for testing.

2.3 Evaluation metrics

A true positive (TP) indicates the model correctly classified that the neonate will be septic. False positive (FP) shows the model incorrectly predicts the onset of sepsis. A false negative (FN) indicates that the model predicts that the neonate will be healthy but becomes septic. True negative (TN) shows the model correctly classified the neonate as healthy. Accuracy (Acc) measures the closeness of the classified value to the actual value and is calculated by the total number of correct predictions out of the total number of predictions. ROC curve graphically represents the true positive rate and false positive rate trade-off at various categorization thresholds. AUROC measures the model's capacity to differentiate between positive and negative classes. Recall gauges the ability of the model to correctly identify neonates who will be septic among all the neonates who will become septic. Precision (Prec) focuses on minimizing misclassifying healthy neonates as having sepsis and is crucial because falsely identifying a neonate as septic will lead to the unnecessary use of antibiotics. F1-score combines precision and recall into a single metric. Cohen's kappa evaluates the agreement between two raters who divide observations into healthy and septic categories. Matthews Correlation Coefficient (MCC) is a balanced measure considering all four possible outcomes (true positive, false positive, true negative, false negative). According to the literature, AUROC is the most widely used metric for evaluating binary classification algorithms in medical research and is more discriminating than MCC for imbalanced datasets.[39] The limitation of AUROC is that it treats all misclassifications equally, regardless of the class distribution. Hence, an overly optimistic evaluation of the model's performance may be obtained in case of an imbalanced dataset. Therefore, other metrics like F1 score, Kappa, Recall, Precision and MCC are used metrics to provide a more comprehensive assessment of the model's performance.

2.4 Algorithms

This section discusses the various supervised machine learning techniques used in this study to perform the binary classification task.[40-50] Algorithms such as logistic regression,[40] naive bayes,[41] random forest, decision tree, k-nearest neighbors,[42,43] SVM,[44] linear discriminant analysis,

quadratic discriminant analysis,[45] adaptive boosting,[46] light gradient boosting machine,[47] gradient boosting machine,[48] and extra trees classfiers[49] were used for analysis.

To avoid overfitting and deliver good generalization performance, stratified 10-fold cross validation was used for training and evaluation of all estimators.[51] Stratified 10-fold cross validation prevents overfitting by maintaining class distribution consistency across data subsets and robust model generalization is ensured by evaluating performance across multiple splits, efficiently utilizing available data for both training and evaluation, and reducing variance in performance estimates through result averaging. Algorithm-specific parameters were fine-tuned using the Optuna hyperparameter optimization framework[52] to optimize the model and achieve the highest AUROC score. Optuna is a Python library that streamlines the process of finding the best hyperparameters using Bayesian hyperparameter tuning. The model with the highest cross-validated AUROC score per fold was selected. In this study, the optimization process was constrained with number of iterations configured to 100.

The machine learning models were trained and evaluated on Google Cloud using a Compute Engine virtual machine with 16 vCPUs and 60 GB of RAM and 1 NVIDIA Tesla T4 GPU. Google cloud's high performance system enabled parallel processing, efficient handling of large datasets, and rapid model training and evaluation.The operating system used was Ubuntu 20, while the programming language that drove the development was Python 3.9.16. The key python libraries and framework used are pandas, numpy, scikit-learn,[43] imblanced-learn, pycaret,[50] xgboost, lightgbm, optuna,[52] and gardio.[53]

2.4.1 Adaptive Boosting (AdaBoost)

Adaptive Boosting[46] is an ensemble learning method that works on the stepwise addition principle, using multiple weak classifiers to get a strong classifier. The value of the alpha parameter is inversely proportional to the weak classifier error. AdaBoost classifier assigns weights to training samples and trains a sequence of weak classifiers, where each subsequent weak classifier focuses more on the misclassified samples from the previous classifiers. The final classification is determined by combining the predictions of all weak classifiers, weighted by their respective strengths as shown in Equation (1):

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t \, h_t(x)\right) \quad (1)$$

where,

T is the number of weak classifiers.

$h_t(x)$ is the output of weak classifier t for input x.

α_t is the weight assigned to classifier t and is based on the error rate.

In this study, we aimed to optimize the AUROC for the AdaBoost classifier by tuning hyperparameters such as the base estimator, considering it as a decision tree, and then performed a search to find the optimal values for the learning rate and the number of estimators.

2.4.2 Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine[47] belongs to the family of gradient boosting frameworks and uses tree-based learning algorithms. The algorithm is designed to be efficient and scalable and builds the model in a boosting manner, where each newly trained tree corrects the errors made by the previous trees. The splitting process involves finding the best feature and threshold combination, resulting in the highest reduction in binary log loss.[54] The binary log loss is given in Equation (2):

$$Log\ Loss = \frac{1}{N}\sum_{i=1}^{N} -(y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)) \quad (2)$$

where,

N is the number of training samples.

y_i is the target label for i^{th} sample.

p_i is the predicted probability of ith sample being septic.

The hyperparameters considered for tuning include the maximum number of leaves in each decision tree, learning rate, number of weak learners that are sequentially added to the ensemble during the boosting process, minimum gain required for a split to occur, and minimum number of samples required to create a new split within a leaf node.

2.4.3 Gradient Boosting Classifier (GBC)

Gradient Boosting Classifier[48] produces a stronger ensemble model by iteratively fitting the weak classifiers to the negative gradients of the loss function for the current model's predictions. This technique combines weak predictive models, often decision trees, using the gradient boosting approach. It operates by repeatedly training new models to fix mistakes committed by the earlier ones by multiplying the prediction by a small learning rate as shown in Equation (3):

$$F_t(x) = F_{t-1}(x) + \eta . h_t(x) \quad (3)$$

where,

$F_{t-1}(x)$ represents the ensembles prediction up to iteration t-1.

η is the learning rate.

$h_t(x)$ is the output of classifier t for input x.

The weighted sum from all the models is used for the final prediction. Hyperparameters such as number of boosting trees in the ensemble, learning rate, minimum number of samples required to split an internal node during the tree-building

process, minimum number of samples required to be at a leaf node, maximum depth of individual decision trees, minimum impurity decrease required for a split to occur were tuned to achieve the best AUROC.

2.4.4 Random Forest (RF)

Random Forest[42] is an ensemble learning technique which builds a large number of decision trees using a random subset of the training data, and the predictions of all the individual trees are averaged to provide the final prediction. Decision trees have a hierarchical tree structure, where each inner node represents a test on a feature, each branch represents the test result, and each leaf node represents a class label or predicted value. To classify a new instance, the decision tree traverses down the tree from the root node using the values of the instance's features. At each node, the features and threshold of the node decide the next path. The process continues until a leaf node is reached, and the class associated with that leaf node is the predicted class for the instance. The ensemble prediction is obtained by computing the sign of the average of individual tree predictions as shown in Equation (4):

$$H(x) = sign\left(\frac{1}{N}\sum_{i=1}^{N} T_i(x)\right) \quad (4)$$

where,

$T_i(x)$ represents the prediction of the ith decision tree for input x.

N represents total number of decision trees.

In this study, we aimed to optimize the AUROC by tuning several crucial hyperparameters such as number of estimators, which controls the number of trees in the forest; the maximum depth of each tree, which limits the depth of the individual trees; the minimum impurity decrease threshold, used to determine whether a split is performed during the tree building process; the maximum number of features considered for splitting; the use of bootstrap samples for tree construction and the split criterion used to assess the quality of a split.

2.4.5 Extra Trees (ET)

Extra Trees[49] is similar to random forest but uses random thresholds rather than looking for the best split points to randomize the feature selection process. The randomness results in shorter training time but may slightly increase the variance. Key hyperparameters, including the number of decision trees in the ensemble, criteria to measure the quality of the split, maximum depth of individual decision trees, minimum impurity decrease required to perform a split during the tree-building process, the minimum number of samples required to split an internal node and minimum number of samples required to be at a leaf node were systematically explored to enhance the model's performance.

2.5 Synthetic Minority Oversampling Technique (SMOTE)

The records labeled septic and non-septic were 2184 and 5648, respectively. This imbalance can have implications for model performance, leading to a biased model favoring the majority class and poor predictive performance for the minority class. The SMOTE method addresses class imbalance by generating synthetic samples for the minority class.[55] The number of records labeled septic and non-septic after applying SMOTE to septic class using the five nearest neighbors is 4608 and 5648, respectively.

3. Results and discussion

Models based on all the algorithms discussed in the Methodology section were constructed and assessed using cross-validated results based on the training set. Tables S1 and S2 show the result of all the models based on the testing set without and with SMOTE, respectively sorted on AUROC. Table 3 shows the top ten machine learning algorithms with the highest AUROC sorted based on their AUROC. Experimental results show that adaptive boosting without SMOTE could predict sepsis for features extracted within the first 24 hours of admission with a best AUROC of 0.9248, Accuracy of 0.8494 and F1 score of 0.7277 on the testing set. Light gradient boosting without SMOTE and random forest with SMOTE performed with an AUROC of 0.9245 and 0.9238, respectively. Random forest with SMOTE had the best Recall of 0.9252 and F1 score of 0.7774.

ROC curve and feature importance plot of the top three models selected based on the AUROC score are shown in Figs. 3(a-f). Features with the minimum absolute Pearson correlation higher than the threshold 0.95 are removed. In all three models, urine production is a critical component. Other features include bilirubin, platelet count, partial pressure of oxygen (po2), sodium, haemoglobin, white blood cells (wbc) count, and heart rate. Heatmaps for the classifiers are provided in Figs. 4, S1 and S2.

The soft voting classifier[56] was trained on an ensemble of the top three models selected based on AUROC. The average of the probabilities for each class by all three models is calculated. The class with the highest average is selected as output. The soft voting classifier uses the equation shown in Equation (5):

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{3} p_{ij} \quad (5)$$

The advantage of using a soft voting classifier is that it can improve predictions by combining the strengths of the top three models and reduce the variance of predictions, which will lead to more robust models.

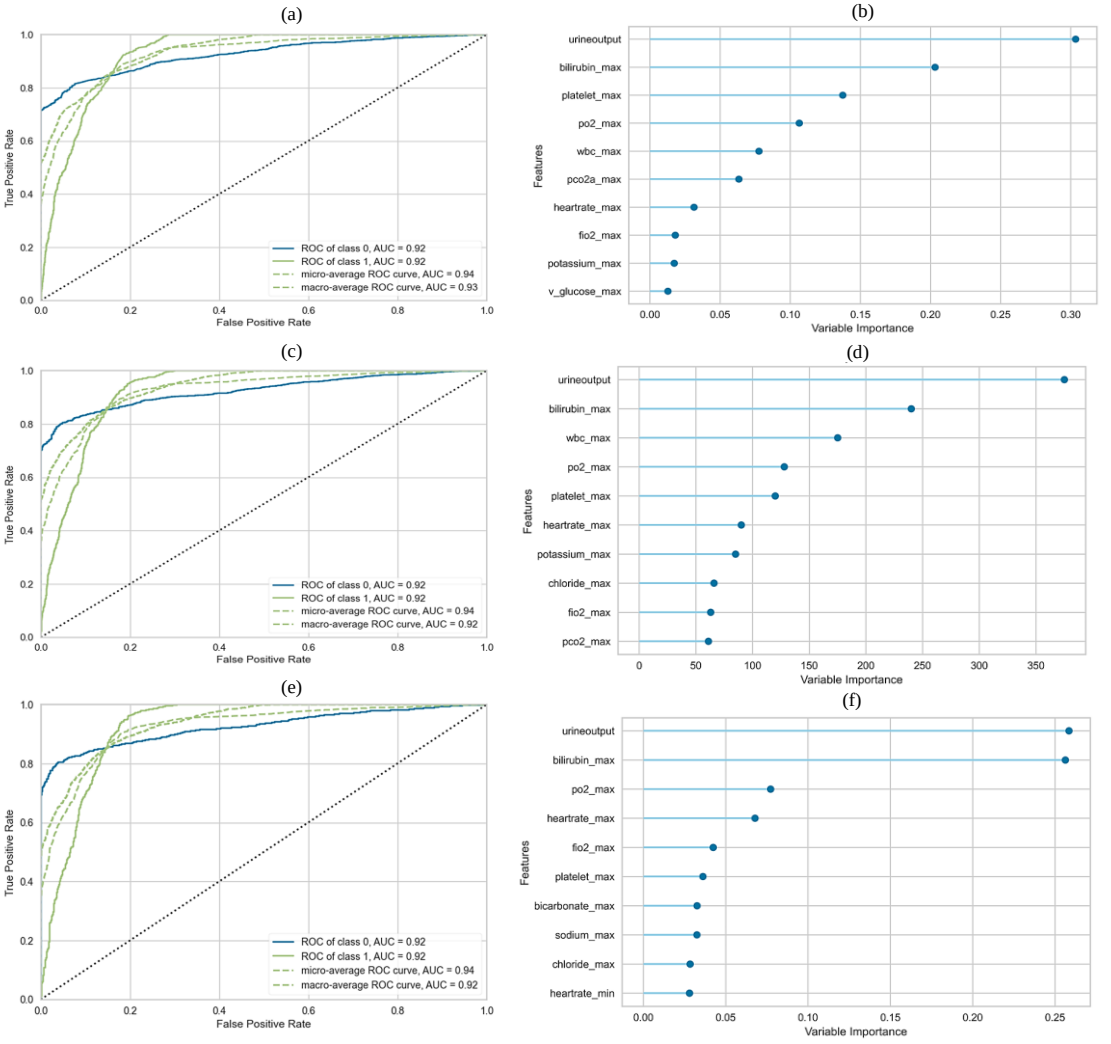

Fig. 3 (a) ROC Curves and (b) Feature Importance Plot for Adaptive Boosting Classifier; (c) ROC Curves and (d) Feature Importance Plot for Light Gradient Boosting Machine Classifier; (e) ROC Curves and (f) Feature Importance Plot for Random Forest Classifier (SMOTE).

A stacking classifier[57] was created by stacking the output of the top three models selected based on AUROC and using a logistic regression classifier to compute the outcome. ROC curve for voting and stacking classifiers are shown in Fig. 5. The result of data analysis for the voting and stacking classifier is shown in Table 4. It is evident that the voting classifier has the best AUROC of 0.9266, Accuracy of 0.8553, F1 score of 0.7829 and Recall of 0.9359 on holdout data, indicating a highly promising classifier. The classifier can differentiate between septic and non-septic cases, assisting healthcare professionals in prioritizing patients for prompt intervention because of its strong discriminatory power, balanced precision and recall, and high overall accuracy. The MCC also emphasises how well the classifier can identify underlying data trends, which is essential when working with imbalanced datasets.

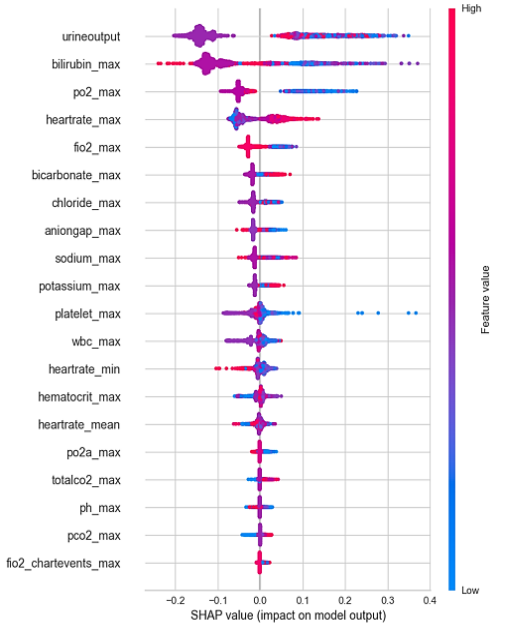

Fig. 4 Heatmap for AdaBoost Classifier without SMOTE.

Table 3. Data Analysis Results: Adaptive Boosting (AdaBoost), Light Gradient Boosting Machine (LightGBM), Random Forest (RF), Gradient Boosting (GBC), Extra Trees (ET).

Model	TP	FP	FN	TN	Acc.	AUROC	Recall	Prec.	F1	Kappa	MCC
AdaBoost	473	172	182	1523	0.8494	0.9248	0.7221	0.7333	0.7277	0.6236	0.6236
LightGBM	508	202	147	1493	0.8515	0.9245	0.7756	0.7155	0.7443	0.6399	0.6409
RF (SMOTE)	606	298	49	1397	0.8523	0.9238	0.9252	0.6704	0.7774	0.6711	0.6906
AdaBoost (SMOTE)	545	265	110	1430	0.8404	0.9236	0.8321	0.6728	0.744	0.63	0.6375
RF	521	206	134	1489	0.8553	0.9229	0.7954	0.7166	0.754	0.6519	0.6537
GBC	506	208	149	1487	0.8481	0.9226	0.7725	0.7087	0.7392	0.6323	0.6335
LightGBM (SMOTE)	572	261	83	1434	0.8536	0.9221	0.8733	0.6867	0.7688	0.6639	0.6742
GBC (SMOTE)	595	301	60	1394	0.8464	0.9218	0.9084	0.6641	0.7672	0.6567	0.6746
ET	562	289	93	1406	0.8374	0.9167	0.858	0.6604	0.7463	0.6297	0.6414
ET (SMOTE)	540	258	115	1437	0.8413	0.9155	0.8244	0.6767	0.7433	0.63	0.6365

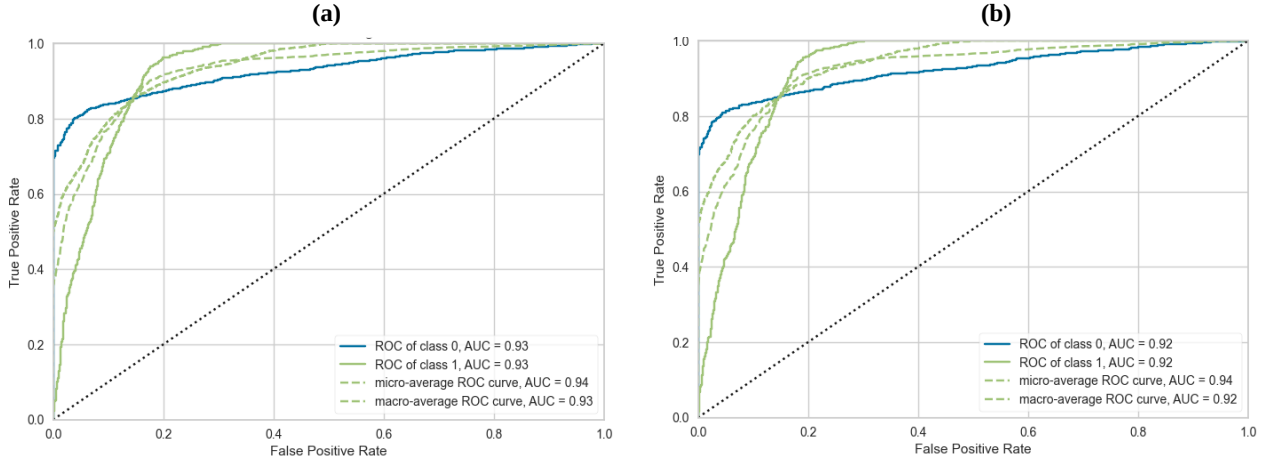

Fig. 5 (a) ROC for voting classifier; (b) ROC curve for stacking classifier.

Table 4. Data analysis results.

Model	TP	FP	FN	TN	Acc.	AUROC	Recall	Prec.	F1	Kappa	MCC
Voting Classifier	613	298	42	1397	0.8553	0.9266	0.9359	0.6729	0.7829	0.6787	0.6995
Stacking Classifier	566	263	89	1432	0.8502	0.9228	0.8641	0.6828	0.7628	0.6555	0.6653

The generalizability of findings from the MIMIC III dataset should be interpreted with caution due to several potential biases and limitations. First, the dataset primarily consists of patients from a single hospital, which may not reflect the diversity of patient populations and healthcare practices in other settings. Additionally, the dataset includes patients admitted between 2001 and 2012, potentially limiting the relevance of findings to current medical practices. Furthermore, privacy concerns and data anonymization procedures may introduce challenges in accurately linking patient records and limiting the generalizability of findings. Therefore, while the MIMIC III dataset provides valuable insights, researchers and practitioners should be cautious in extrapolating its findings to broader populations and contexts. Prospective studies and randomized controlled trials are warranted to strengthen the evidence base and support informed medical decision-making.

An interactive web interface shown in Fig. S3(a) was created using an open-source python package, Gradio[52] for the voting classifier with highest AUROC. Medical practitioners will use the web interface to enter values for the 23 features shown in Table 2 captured during the first 24 hours of admission of the neonate. The output is a prediction score for the septic and non-septic class label, as shown in Fig. S3(b). The web interface developed will be used in a NICU setup in India to validate the model's applicability and generalizability.

Introducing machine learning algorithms into clinical decision-making raises important ethical considerations and potential challenges. First, ensuring patient privacy and data security is paramount, as using of sensitive medical information risks unauthorized access and misuse. Additionally, the transparency and interpretability of machine learning models are a challenge, as their complex algorithms can compromise the ability of physicians to understand and

trust their decision-making process. Bias in the data used to train algorithms can also lead to inequalities in healthcare if not carefully addressed. Careful consideration of these ethical considerations and addressing potential challenges are essential to responsibly and beneficially integrating machine learning algorithms into clinical decision making.

4. Conclusions & future work

Sepsis is the third major cause of death in neonates, and appropriate measures should be taken at an early stage of infection to prevent morbidity or mortality. In this study, the performance of various machine learning algorithms was evaluated to predict the onset of neonatal sepsis using data from the first 24 hours of admission from MIMIC III dataset. The voting classifier, which was trained on the ensemble of adaptive boosting, light gradient boosting without SMOTE and random forest with SMOTE was able to perform with the highest AUROC, Accuracy, Recall, MCC and F1 score. The web interface developed as the result of this study can be used to assess the risk of developing sepsis during neonatal admission. The information obtained from the tool can aid practitioners at NICU in making informed decisions while treating neonates. Early detection of neonatal sepsis has significant implications for achieving SDG 3.2 and reducing neonatal mortality rates. Early identification and prompt treatment increase survival rates, prevent long-term complications and optimize resource allocation. In the future, we plan to study the potential impact of the model's results on clinical decision-making or neonatal care in NICU setups in India.

Acknowledgment

We want to acknowledge the creators and contributors of the MIMIC-III database for their hard work and dedication in making this valuable resource freely available to the research community.

Conflict of Interest

There is no conflict of interest.

Supporting Information

Applicable.

References

[1] "Levels & Trends in Child Mortality," United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME), 2023, https://childmortality.org/wp-content/uploads/2023/01/UN-IGME-Child-Mortality-Report-2022.pdf.

[2] "End preventable deaths of newborns and children under 5 years of age," World Health Organization, 2023, https://www.who.int/data/gho/data/themes/topics/sdg-target-3_2-newborn-and-child-mortality.

[3] B. A. Shah, J. F. Padbury, Neonatal sepsis, *Virulence*, 2014, **5**, 170-178, doi: 10.4161/viru.26906.

[4] A. L. Shane, P. J. Sánchez, B. J. Stoll, Neonatal sepsis, *The Lancet*, 2017, **390**, 1770-1780, doi: 10.1016/s0140-6736(17)31002-4.

[5] J. R. Moorman, W. A. Carlo, J. Kattwinkel, R. L. Schelonka, P. J. Porcelli, C. T. Navarrete, E. Bancalari, J. L. Aschner, M. W. Walker, J. A. Perez, C. Palmer, G. J. Stukenborg, D. E. Lake, T. M. O'Shea, Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: A randomized trial, *The Journal of Pediatrics*, 2011, **159**, 900-906, doi: 10.1016/j.jpeds.2011.06.044.

[6] C. McGregor, C. Catley, A. James, Variability analysis with analytics applied to physiological data streams from the neonatal intensive care unit. 2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS). June 20-22, 2012, Rome, Italy. IEEE, 2012, 1-5, doi: 10.1109/CBMS.2012.6266385.

[7] H. Khazaei, N. Mench-Bressan, C. McGregor, J. E. Pugh, Health informatics for neonatal intensive care units: an analytical modeling perspective, *IEEE Journal of Translational Engineering in Health and Medicine*, 2015, **3**, 1-9, doi: 10.1109/JTEHM.2015.2485268.

[8] K. D. Fairchild, R. L. Schelonka, D. A. Kaufman, W. A. Carlo, J. Kattwinkel, P. J. Porcelli, C. T. Navarrete, E. Bancalari, J. L. Aschner, M. W. Walker, J. A. Perez, C. Palmer, D. E. Lake, T. M. O'Shea, J. R. Moorman, Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial, *Pediatric Research*, 2013, **74**, 570-575, doi: 10.1038/pr.2013.136.

[9] L. B. Mithal, R. Yogev, H. Palac, I. Gur, K. K. Mestan, Computerized vital signs analysis and late onset infections in extremely low gestational age infants, *Journal of Perinatal Medicine*, 2016, **44**, 491-497, doi: 10.1515/jpm-2015-0362.

[10] A. Johnson, T. Pollard, R. Mark, MIMIC-III Clinical Database (version 1.4), PhysioNet, 2016, doi: 10.13026/C2XW26.

[11] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, *Scientific Data*, 2016, **3**, 160035, doi: 10.1038/sdata.2016.35.

[12] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet, *Circulation*, 2000, **101**, 215-220, doi: 10.1161/01.cir.101.23.e215.

[13] K. E. Henry, D. N. Hager, P. J. Pronovost, S. Saria, A targeted real-time early warning score (TREWScore) for septic shock, *Science Translational Medicine*, 2015, **7**, eaab3719, doi: 10.1126/scitranslmed.aab3719.

[14] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, R. Das, Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach, *JMIR Medical Informatics*, 2016, **4**, e28, doi: 10.2196/medinform.5909.

[15] M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, D. P. Edelson, Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards, *Critical Care Medicine*, 2016, **44**, 368-374, doi: 10.1097/ccm.0000000000001571.
[16] F. Li, H. Xin, J. Zhang, M. Fu, J. Zhou, Z. Lian, Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database, *BMJ Open*, 2021, **11**, e044779, doi: 10.1136/bmjopen-2020-044779.
[17] T. Gentimis, A. J. Alnaser, A. Durante, K. Cook, R. Steele, Predicting hospital length of stay using neural networks on MIMIC III data. 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech). November 6-10, 2017, Orlando, FL, USA. IEEE, 2018, 1194-1201, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191.
[18] C. J. McWilliams, D. J. Lawson, R. Santos-Rodriguez, I. D. Gilchrist, A. Champneys, T. H. Gould, M. J. Thomas, C. P. Bourdeaux, Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK, *BMJ Open*, 2019, **9**, e025925, doi: 10.1136/bmjopen-2018-025925.
[19] H. Weng, Y. Li, X. Nie, C. He, P. Feng, F. Zhao, Q. Chen, W. Sun, J. Jiang, Y. Zhang, Y. Huo, J. Li, Comparative effectiveness and safety of bolus *vs*. continuous infusion of loop diuretics: results from the MIMIC-III Database, *The American Journal of the Medical Sciences*, 2023, **365**, 353-360, doi: 10.1016/j.amjms.2022.12.013.
[20] S. S. Harsha, SNAPPE-II (score for neonatal acute physiology with perinatal extension-II) in predicting mortality and morbidity in NICU, *Journal of Clinical and Diagnostic Research*, 2015, **9**, SC10, doi: 10.7860/jcdr/2015/14848.6677.
[21] G. Parry, J. Tucker, W. Tarnow-Mordi, CRIB II: an update of the clinical risk index for babies score, *The Lancet*, 2003, **361**, 1789-1791, doi: 10.1016/s0140-6736(03)13397-1.
[22] V. Aithal, R. David Jathanna, Credit risk assessment using machine learning techniques, *International Journal of Innovative Technology and Exploring Engineering*, 2019, **9**, 3482-3486, doi: 10.35940/ijitee.a4936.119119.
[23] R. D. Shirwaikar, N. Mago, A. U. Dinesh, K. Makkithaya, H. K. Govardhan, Supervised Learning techniques for analysis of neonatal data. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). July 21-23, 2016, Bangalore, India. IEEE, 2017, 25-31, doi: 10.1109/ICATCCT.2016.7911960.
[24] Y. G. Robi, T. M. Sitote, Neonatal disease prediction using machine learning techniques, *Journal of Healthcare Engineering*, 2023, **2023**, 1-16, doi: 10.1155/2023/3567194.
[25] A. Johnson, T. Pollard, sepsis3-mimic, *URL*, 2018, doi: 10.5281/zenodo.1256723.
[26] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J.-L. Vincent, D. C. Angus, The third international consensus definitions for sepsis and septic shock (sepsis-3), *JAMA*, 2016, **315**, 801, doi: 10.1001/jama.2016.0287.
[27] G. S. Martin, D. M. Mannino, S. Eaton, M. Moss, The epidemiology of sepsis in the United States from 1979 through 2000, *New England Journal of Medicine*, 2003, **348**, 1546-1554, doi: 10.1056/nejmoa022139.
[28] A. E. W. Johnson, J. Aboab, J. D. Raffa, T. J. Pollard, R. O. Deliberato, L. A. Celi, D. J. Stone, A comparative analysis of sepsis identification methods in an electronic database, *Critical Care Medicine*, 2018, **46**, 494-499, doi: 10.1097/ccm.0000000000002965.
[29] T. J. Matics, L. N. Sanchez-Pinto, Adaptation and validation of a pediatric sequential organ failure assessment score and evaluation of the sepsis-3 definitions in critically ill children, *JAMA Pediatrics*, 2017, **171**, e172352, doi: 10.1001/jamapediatrics.2017.2352.
[30] C. W. Seymour, V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, A. Scherag, G. Rubenfeld, J. M. Kahn, M. Shankar-Hari, M. Singer, C. S. Deutschman, G. J. Escobar, D. C. Angus, Assessment of clinical criteria for sepsis, *JAMA*, 2016, **315**, 762, doi: 10.1001/jama.2016.0288.
[31] L. J. Schlapbach, L. Straney, R. Bellomo, G. MacLaren, D. Pilcher, Prognostic accuracy of age-adapted SOFA, SIRS, PELOD-2, and qSOFA for in-hospital mortality among children with suspected infection admitted to the intensive care unit, *Intensive Care Medicine*, 2018, **44**, 179-188, doi: 10.1007/s00134-017-5021-8.
[32] S. Coggins, M. C. Harris, R. Grundmeier, E. Kalb, U. Nawab, L. Srinivasan, Performance of pediatric systemic inflammatory response syndrome and organ dysfunction criteria in late-onset sepsis in a quaternary neonatal intensive care unit: a case-control study, *The Journal of Pediatrics*, 2020, **219**, 133-139.e1, doi: 10.1016/j.jpeds.2019.12.064.
[33] S. C. van Nassau, R. H. van Beek, G. J. Driessen, J. A. Hazelzet, H. M. van Wering, N. P. Boeddha, Translating sepsis-3 criteria in children: prognostic accuracy of age-adjusted quick SOFA score in children visiting the emergency department with suspected bacterial infection, *Frontiers in Pediatrics*, 2018, **6**, 266, doi: 10.3389/fped.2018.00266.
[34] J. K. Valik, L. Mellhammar, J. Sundén-Cullberg, L. Ward, C. Unge, H. Dalianis, A. Henriksson, K. Strålin, A. Linder, P. Nauclér, Peripheral oxygen saturation facilitates assessment of respiratory dysfunction in the sequential organ failure assessment score with implications for the sepsis-3 criteria, *Critical Care Medicine*, 2021, **50**, e272-e283, doi: 10.1097/ccm.0000000000005318.
[35] Y. Luo, Evaluating the state of the art in missing data imputation for clinical data, *Briefings in Bioinformatics*, 2022, **23**, bbab489, doi: 10.1093/bib/bbab489.
[36] Q. Yu, X. Guan, Y. Zhai, Z. Meng, The missing data filling method of the industrial Internet platform based on rules and

lightGBM, *IFAC-PapersOnLine*, 2020, **53**, 152-157, doi: 10.1016/j.ifacol.2021.04.094.
[37] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, J. J. Nair, F-test feature selection in Stacking ensemble model for breast cancer prediction, *Procedia Computer Science*, 2020, **171**, 1561-1570, doi: 10.1016/j.procs.2020.04.167.
[38] P. Dhal, C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, *Applied Intelligence*, 2022, **52**, 4543-4581, doi: 10.1007/s10489-021-02550-9.
[39] C. Halimu, A. Kasem, S. H. Shah Newaz, Empirical comparison of area under ROC curve (AUC) and mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification, Proceedings of the 3rd International Conference on Machine Learning and Soft Computing. January 25 - 28, 2019, Da Lat, Viet Nam. New York: ACM, 2019, 1-6, doi: 10.1145/3310986.3311023.
[40] X. Zou, Y. Hu, Z. Tian, K. Shen, Logistic regression model optimization and case analysis. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). October 19-20, 2019, Dalian, China. IEEE, 2020, 135-139, doi: 10.1109/ICCSNT47585.2019.8962457.
[41] H. Zhang, The optimality of Naive Bayes, In Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, 2004, 17, 562-567.
[42] T. Hastie, R. Tibshirani, J. Friedman, Overview of supervised learning. The Elements of Statistical Learning. New York, NY: Springer New York, 2008: 9-41, doi: 10.1007/978-0-387-84858-7_2.
[43] O. Kramer, Scikit-Learn, *Machine Learning for Evolution Strategies*, 2016, 37-53, doi: 10.1007/978-3-319-33383-0_5.
[44] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, *IEEE Intelligent Systems and Their Applications*, 1998, **13**, 18-28, doi: 10.1109/5254.708428.
[45] A. Araveeporn, Comparing the linear and quadratic discriminant analysis of diabetes disease classification based on data multicollinearity, *International Journal of Mathematics and Mathematical Sciences*, 2022, **2022**, 1-11, doi: 10.1155/2022/7829795.
[46] Yoav, Freund, Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 1997, **55**, 119-139, doi: 10.1006/jcss.1997.1504.
[47] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, *Advances in neural information processing systems*, 2017, **30**, 3146-3154.
[48] J. H. Friedman, Greedy function approximation: a gradient boosting machine, *The Annals of Statistics*, 2001, **29**, 1189-1232, doi: 10.1214/aos/1013203451.
[49] A. Sharaff, H. Gupta, Extra-tree classifier with metaheuristics approach for email classification. Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2019: 189-197, doi: 10.1007/978-981-13-6861-5_17.
[50] M. Ali, PyCaret: An open source, low-code machine learning library in Python, PyCaret version 3.0.0, 2023, https://www.pycaret.org.
[51] G. Marques, D. Agarwal, I. de la Torre Díez, Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network, *Applied Soft Computing*, 2020, **96**, 106691, doi: 10.1016/j.asoc.2020.106691.
[52] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. August 4 - 8, 2019, Anchorage, AK, USA. New York: ACM, 2019, 2623-2631, doi: 10.1145/3292500.3330701.
[53] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, J. Zou, Gradio: Hassle-free sharing and testing of ml models in the wild, *arXiv preprint*, 2019, doi: 10.48550/arXiv.1906.02569.
[54] Y. Ho, S. Wookey, The real-world-weight cross-entropy loss function: modeling the costs of mislabeling, *IEEE Access*, 2019, **8**, 4806-4813, doi: 10.1109/ACCESS.2019.2962617.
[55] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, 2002, **16**, 321-357, doi: 10.1613/jair.953.
[56] Saloni, Kumari, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, *International Journal of Cognitive Computing in Engineering*, 2021, **2**, 40-46, doi: 10.1016/j.ijcce.2021.01.001.
[57] E. S. Lee, Exploring the performance of stacking classifier to predict depression among the elderly. 2017 IEEE International Conference on Healthcare Informatics (ICHI). August 23-26, 2017, Park City, UT, USA. IEEE, 2017, 13-20, doi: 10.1109/ICHI.2017.95.