Risk Management and Healthcare Policy

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Combining Resampling Strategies and Ensemble Machine Learning Methods to Enhance Prediction of Neonates with a Low Apgar Score After Induction of Labor in Northern Tanzania

Clifford Silver Tarimo[1,2]
Soumitra S Bhuyan[3]
Quanman Li[1]
Weicun Ren[4]
Michael Johnson Mahande[5]
Jian Wu[1]

[1]Department of Epidemiology and Health Statistics, Zhengzhou University, Zhengzhou, People's Republic of China; [2]Department of Science and Laboratory Technology, Dar es Salaam Institute of Technology, Dar es Salaam, Tanzania; [3]Edward J. Bloustein School of Planning and Public Policy, Rutgers University, New Brunswick, NJ, USA; [4]College of Sanquan, Xinxiang Medical University, Xinxiang, People's Republic of China; [5]Department of Epidemiology and Applied Biostatistics, Kilimanjaro Christian Medical University College, Moshi, Tanzania

Objective: The goal of this study was to establish the most efficient boosting method in predicting neonatal low Apgar scores following labor induction intervention and to assess whether resampling strategies would improve the predictive performance of the selected boosting algorithms.

Methods: A total of 7716 singleton births delivered from 2000 to 2015 were analyzed. Cesarean deliveries following labor induction, deliveries with abnormal presentation, and deliveries with missing Apgar score or delivery mode information were excluded. We examined the effect of resampling approaches or data preprocessing on predicting low Apgar scores, specifically the synthetic minority oversampling technique (SMOTE), borderline-SMOTE, and the random undersampling (RUS) technique. Sensitivity, specificity, precision, area under receiver operating curve (AUROC), F-score, positive predicted values (PPV), negative predicted values (NPV) and accuracy of the three (3) boosting-based ensemble methods were used to evaluate their discriminative ability. The ensemble learning models tested include adoptive boosting (AdaBoost), gradient boosting (GB) and extreme gradient boosting method (XGBoost).

Results: The prevalence of low (<7) Apgar scores was 9.5% (n = 733). The prediction models performed nearly similar in their baseline mode. Following the application of resampling techniques, borderline-SMOTE significantly improved the predictive performance of all the boosting-based ensemble methods under observation in terms of sensitivity, F1-score, AUROC and PPV.

Conclusion: Policymakers, healthcare informaticians and neonatologists should consider implementing data preprocessing strategies when predicting a neonatal outcome with imbalanced data to enhance efficiency. The process may be more effective when borderline-SMOTE technique is deployed on the selected ensemble classifiers. However, future research may focus on testing additional resampling techniques, performing feature engineering, variable selection and optimizing further the ensemble learning hyperparameters.

Keywords: low Apgar score, labor induction, machine learning, ensemble learning, resampling methods, imbalanced data

Background

Labor induction (IOL) is the artificial stimulation of uterine contractions during pregnancy prior to the onset of labor in order to promote a vaginal birth.[1] Recent advances in obstetric and fetal monitoring techniques have resulted in the majority

Correspondence: Jian Wu
Email jianwu17@163.com

Received: 29 July 2021
Accepted: 26 August 2021
Published: 7 September 2021

of induced pregnancies having favorable outcomes; however, adverse health outcomes resulting in low Apgar scores in neonates continue to exist.[2] The Apgar score tool, developed by Virginia Apgar, is a test administered to newborns shortly after birth. This examination analyzes the heart rate, muscle tone, and other vital indicators of a baby to determine if extra medical care or emergency care is required.[3] The test is usually administered twice: once at 1 minute after birth and again at 5 minutes.[4] Apgar scores obtained 5 minutes after birth have become widely used in the prediction of neonatal outcomes such as asphyxia, hypoxic-ischemic encephalopathy, and cerebral palsy.[5] Additionally, recent research has established that Apgar values <7 five minutes after birth are related with impaired cognitive function, neurologic disability, and even subtle cognitive impairment as determined by scholastic achievement at the age of 16.[6] Perinatal morbidity and death can be decreased by identifying and managing high-risk newborns effectively.[7] Accurate detection of low Apgar scores at 5 minutes following labor induction is hence one among the ways to ensure optimal health and survival of the newborn.[8] Several studies based on statistical learning have shown relationship and the interplay of maternal and neonatal variables for low Apgar scores.[9,10] However, no studies have been conducted to date that focus exclusively on modeling neonatal Apgar scores following IOL intervention. As machine learning is applied to increasingly sensitive tasks and on increasingly noisy data, it is critical that these algorithms are validated against neonatal healthcare data.[11] In addition, myriad studies have reported the potential of ensemble learning algorithms in predictive tasks.[12,13] In the current study, we assessed the performance metrics of the three powerful ensemble learning algorithms. Due to skewed or imbalanced distribution of the outcome of interest, we further assessed whether the synthetic minority oversampling technique (SMOTE), Borderline-SMOTE and random undersampling (RUS) techniques would impact the learning process of the models.

Methods

Study Setting and Data Source

We analyzed data from the Kilimanjaro Christian Medical Centre (KCMC) birth registry for women who gave birth to singleton infants between 2000 and 2015. This facility serves a population of around 11 million people from the region and neighboring areas. The register collects data on the mother's health prior to and during pregnancy, as well as complications and the infant's status. All induced women who delivered singleton infants vaginally during the study period and had complete birth records were eligible for this study. Women with multiple gestations, stillbirths were excluded. These exclusions were necessary to offset the effect of possible overestimation of the prevalence of low Apgar scores (Figure 1). More information about the KCMC birth registry database can be found elsewhere.[14] The final sample comprised 7716 induced deliveries.

Description of the Response and the Predictor Variables

The response variable was "Apgar score" at 5 minutes (coded 0 for "normal", and 1 for "low") which was computed using five criteria. The first criterion included the strength and regularity of newborn's heart rate where babies with 100 beats per minute or more scored 2 points while those with less than 100 scored 1 point and those with 0 heart rate scored 0 points. The second criterion assessed lung maturity or breathing effort, awarding 2 points to newborns with regular breathing, 1 point to those with irregular breathing with 30 breaths per minute, and 0 points to those with no breath at all. Muscle tone and mobility make the third component, for which active neonates received 2 points, moderately active ones received 1 point, and those who limped received no point. The fourth factor is skin color and oxygenation, where infants with pink color receiving 2 points, those with bluish extremities receiving 1 point, and those with completely bluish color receiving 0 points. The final component assesses reflex responses to irritating stimuli, with crying receiving 2 points, whimpering receiving 1 point, and silence receiving 0 points. The investigator then added the scores for each finding and defined a number less than seven (<7) as low and >7 as normal Apgar score. The current study examined the predictors of low Apgar scores previously reported in literature such as parity, maternal age, gestational age, number of prenatal visits, induction method used, body mass index (BMI). The gestational age at birth was calculated using the last menstrual period date and expressed in whole weeks, with deliveries of less than 37 weeks classified as preterm, those between 37 and 41 weeks as term, and those of 41 weeks or more as postterm. Additional behavioral and neonatal risk factors included child sex, smoking and alcohol consumption during

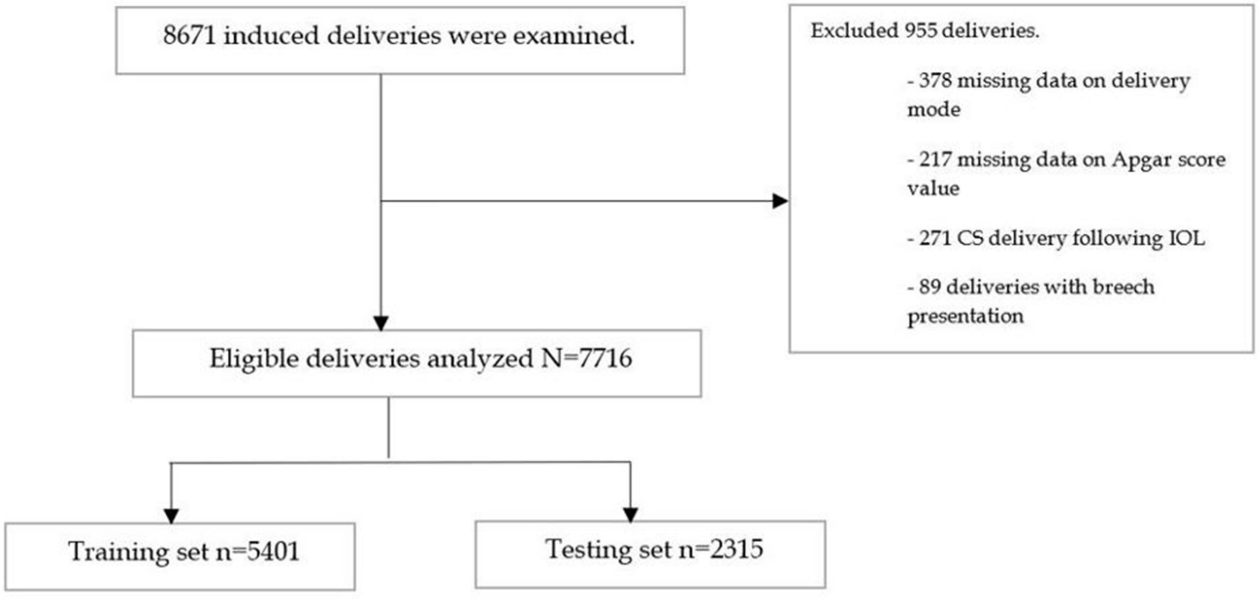

Figure 1 Schematic diagram for sample size estimation.
Abbreviations: CS, cesarean section; IOL, induction of labor.

pregnancy, as well as the history of using any form of family planning method were also examined. These factors were categorized as yes or no, with yes indicating the occurrence of these outcomes. The categories of the covariates for some factors variables were selected following a preliminary examination of the data.

The Boosting-Based Algorithms

Boosting algorithms have received significant attention in recent years in data science and machine learning. Boosting algorithms combine several weak models to produce a strong or more accurate model.[15,16] Boosting techniques such as AdaBoost, Gradient boosting, and extreme gradient boosting (XGBoost) are all examples of ensemble learning algorithms that are often employed, particularly in data science contests.[17] AdaBoost is designed to boost the performance of "weak learners." The algorithm constructs an ensemble of weak learners iteratively by modifying the weights of misclassified data in each iteration. It gives equal weight to each training set sample when training the initial weak learner.[18] Weights are revised for each succeeding weak learner in such a way that samples misclassified by the current weak learner receive a larger weight. Additionally, the family of boosting algorithms are said to be advantageous for resolving class imbalance problems since they provide a greater weight to the minority class with each iteration, as data from this class is frequently misclassified in other ML algorithms.[19] Gradient boosting (GB) constructs an additive model incrementally and it enables optimization of arbitrary differentiable loss functions. It makes use of the gradient descent algorithm to reduce the number of errors in sequential models.[20] In contrast to conventional gradient boosting, XGBoost employs its own way of tree construction, with the similarity score and gain determining the optimal node splits. So, it is a decision-tree-based ensemble method that utilizes a gradient boosting framework.[21] Figure 2 shows the basic mechanism of boosting-based algorithm in modelling process.

Resampling Techniques and Imbalanced Data

Our dataset was imbalanced in terms of class frequency, as the positive class (low Apgar score newborns) had only 733 individuals (9.5%). If one of the target classes contains a small number of occurrences in comparison to the other classes, the dataset is said to be imbalanced.[22,23] Numerous ways to deal with unbalanced datasets have been presented recently.[24–26] This paper presents two approaches for balancing the dataset including synthetic minority oversampling technique (SMOTE) and random undersampling (RUS) technique. In contrast to traditional boosting, which assigns equal weight to all misclassified cases, resampling methods (SMOTE or RUS) and boosting

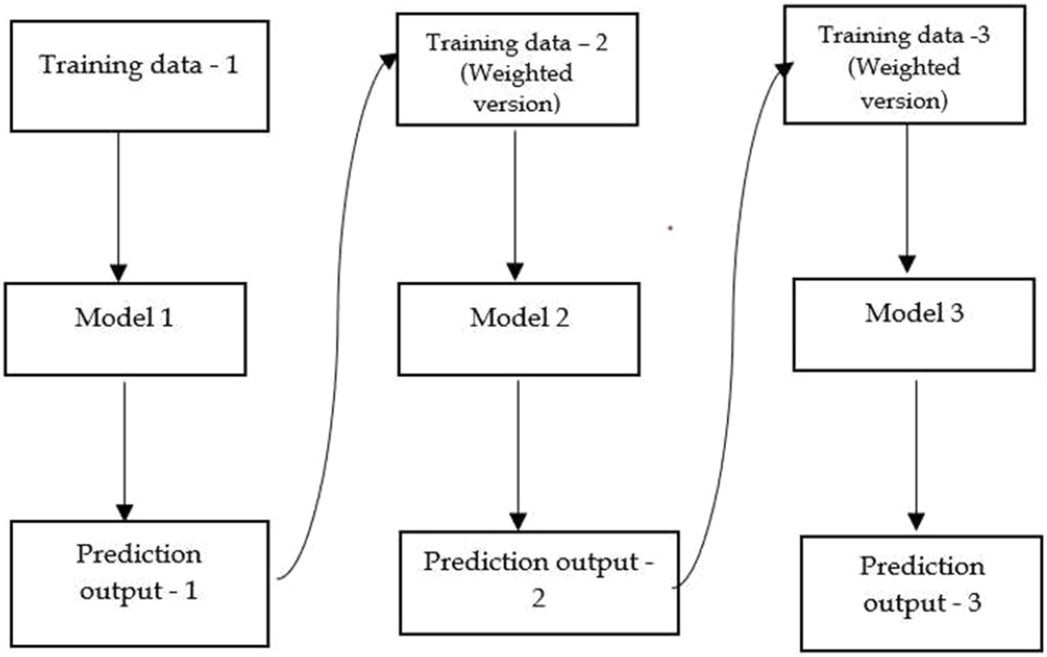

Figure 2 Basic mechanism for boosting-based algorithms.

algorithms (AdaBoost, Gradient boosting, XGBoost) applied to several highly and somewhat imbalanced datasets have been shown to improve prediction on the minority class.[27,28] Re-sampling is a preprocessing approach that balances the distribution of an unbalanced dataset before it is sent to any classifiers.[29] Resampling methods are designed to change the composition of a training dataset for an imbalanced classification task. SMOTE begins by randomly selecting an instance of a minority class and determining its k nearest minority class neighbors. The synthetic instance is then formed by selecting one of the k closest neighbors at random in the feature space to form a line segment.[30] Borderline-SMOTE begins by classifying observations belonging to the minority class. It considers any minority observation to be noise if all of its neighbors are members of the majority class and the minority observation is discarded while constructing synthetic data. Additionally, it resamples entirely from a few places designated as border points with both majority and minority class. Additionally, it resamples entirely from a few places designated as border points with both majority and minority class instances. Undersampling (RUS) approaches eliminate samples from the training dataset that belong to the majority class in order to more evenly distribute the classes. The strategy reduces the dataset by removing examples from the majority class with the goal of balancing the number of examples in each class.[31] Figure 3 indicates the basic mechanism for both RUS and SMOTE techniques.

Implementation and Data Analysis

Descriptive statistics were obtained using STATA version 14. Data preprocessing and the main analyses were performed using Python programming (version 3.8.0). The predictive models for low Apgar scores were generated with test and training sets using Python scikit-learn (version 0.24.0) packages for machine learning. The parameters to assess the predictive performance of the selected ensemble machine learning algorithms have been evaluated in equations (1) through (8). The dataset was firstly converted to comma-separated values (CSV) file and imported to Python tool. We used open-source libraries in Python including Scikit-learn, Numpy and Pandas. The python codes used to generate the results along with the outputs are attached herein (Supplementary File 1).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

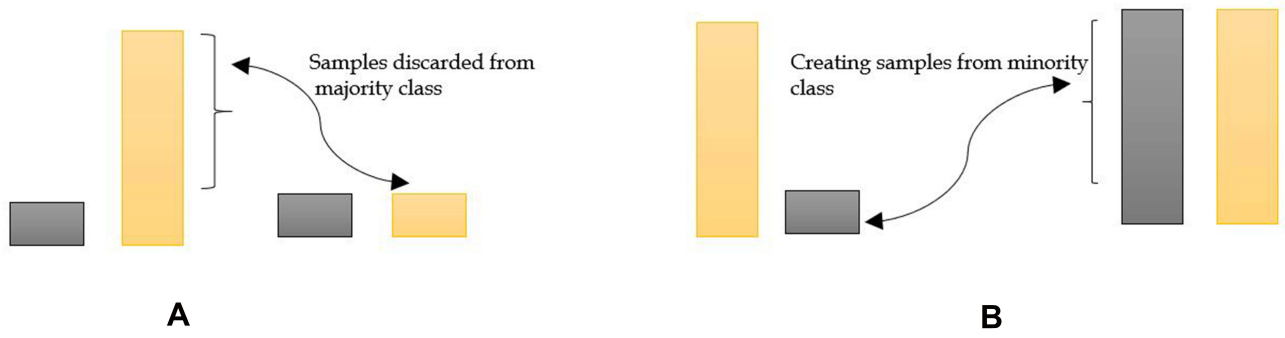

Figure 3 Mechanisms of resampling techniques used: (**A**) RUS – random undersampling (**B**) SMOTE – synthetic minority oversampling techniques.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (4)$$

$$\text{F1 score} = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \quad (5)$$

$$\text{AUC} = \frac{1 + Sensitivity - FPrate}{2} \quad (6)$$

$$\text{NPV} = \frac{TN}{TN + FN} \quad (7)$$

$$\text{PPV} = \frac{TP}{FP + TP} \quad (8)$$

where TP, FP, TN, FN, FPrate, PPV and NPV represent true positive, false positive, true negative, false negative, false-positive rate, positive predictive value and negative predictive value respectively.

Results

The sociodemographic and obstetric characteristics of the participants are summarized in Table 1. A total of 7716 Singleton births were analyzed. Of these, 55% of the deliveries were from nulliparous women while majority (88%) of study participants were aged <35 years and about 80% of the total deliveries were at term. The proportion of neonates with low Apgar scores (<7) was found to be 9.5%.

Prior to the use of resampling techniques, all models performed nearly identically. Of all the resampling techniques considered in the current study, borderline-SMOTE was shown to significantly improve the performance of all the models in terms of all the metrics under observation (Table 2). RUS and SMOTE exhibited little or no improvement on baseline performance in all instances of their respective ensemble models. Performance in terms of AUC metrics for AdaBoost, GB, and XGBoost has been shown in Figure 4.

Discussion

In this paper, we trained and evaluated the performance of three ensemble-based ML algorithms on a rare event (9.5% for <7 Apgar score versus 90.5% for >7 Apgar score). We then demonstrated how the resampling techniques can affect the learning process of the selected models on the imbalanced data. Kubat et al proposed a heuristic under-sampling method for balancing the data set by removing noise and redundant instances of the majority class.[32] Chawla et al oversampled the minority class using the SMOTE (Synthetic Minority Oversampling Technique) technique, which generated new synthetic examples along the line between the minority examples and their chosen nearest neighbors.[33] In the current study, both sampling techniques (SMOTE and RUS) were seen to slightly improve the "sensitivity" of the minority class, with the largest improvement seen from using borderline-SMOTE technique. Improvement of sensitivity means the ratio of correct positive predictions, that is, neonates with <7 Apgar score, to the total positive examples is relatively high. In other words, with the improvement shown by XGBoost following the Borderline-SMOTE resampling techniques, the model was able to correctly identify 93% (an improvement from 20% baseline performance) of the neonates with a low Apgar score, while missing 7% only. On the other hand, all the models performed well (Specificity = 99%) in correctly identifying neonates with normal (>7) Apgar score without the application of resampling methods. This could be because the number of neonates with a normal Apgar score was significantly

Table 1 Demographic Information of the Study Participant (N=7716)

Attributes	Low (<7) Apgar Score	Normal (≥7) Apgar Score	χ^2 p-value
Parity			
Nulliparous	409 (55.8)	3817 (54.66)	0.556
Multiparous	324 (44.2)	3166 (45.34)	
Maternal age (years)			
<25	273 (37.24)	2575 (36.88)	0.214
25–35	361 (49.25)	3606 (51.64)	
>35	99 (13.51)	802 (11.49)	
Gestational age			
Term	463 (63.17)	5683 (81.38)	<0.001
Preterm	209 (28.51)	593 (8.49)	
Post term	61 (8.32)	707 (10.12)	
PROM			
No	709 (96.73)	6829 (97.79)	0.067
Yes	24 (3.27)	154 (2.21)	
Gestational diabetes			
No	730 (99.59)	6974 (99.87)	0.067
Yes	3 (0.41)	9 (0.13)	
Prenatal visits			
<3	296 (40.38)	1796 (25.72)	<0.001
3–6	365 (49.80)	3997 (57.24)	
>6	72 (9.82)	1190 (17.04)	
Induction method			
Oxytocin	591 (80.63)	6361 (91.09)	<0.001
Prostaglandins	142 (19.37)	622 (8.91)	
Referred for delivery			
No	453 (61.80)	5573(79.81)	<0.001
Yes	280 (38.20)	1410 (20.19)	
Ever use of family planning			
No	344 (46.93)	2896 (41.47)	0.004
Yes	389 (53.07)	4087 (58.53)	
Smoking during pregnancy			
No	729 (99.45)	6966(99.76)	0.135
Yes	4 (0.55)	17 (0.24)	
Alcohol during pregnancy			
No	550 (75.03)	4977 (71.27)	0.032
Yes	183 (24.97)	2006 (28.73)	
Child sex			
Female	412 (56.21)	3563 (51.02)	0.008
Male	321 (43.79)	3420 (48.98)	
Body mass index			
Underweight	2 (0.27)	27 (0.39)	0.169
Normal	109 (14.87)	1262 (18.07)	
Overweight	455 (62.07)	4133 (59.19)	
Obese	167 (22.78)	1561 (22.35)	
Epilepsy			
No	732 (99.86)	6961 (99.68)	0.399
Yes	1 (0.14)	22 (0.32)	
Preeclampsia			
No	717 (97.82)	6873 (98.42)	0.217
Yes	16 (2.18)	110 (1.58)	

Abbreviation: PROM, premature rupture of membranes.

greater than those with a low Apgar score in this database (n=6983 vs n=733), making the negative class more likely to be predicted. Notable is the Positive Predicted Value (PPV) obtained with XGBoost using the Borderline-SMOTE resampling method, which indicates that 94% of neonates predicted to have a low Apgar score actually had one. Numerous studies have demonstrated the critical importance of maximizing model's sensitivity as well as PPV particularly when dealing with class imbalanced datasets.[34] Precision and sensitivity make it possible and desirable to evaluate a classifier's performance on the minority class, resulting in another metric called the *F*-score.[35] The *F*-score is high when both sensitivity and precision are high.[36] Again, the best F-score was obtained in all models when borderline-SMOTE was used. However, the best *F*-score was reached by borderline-SMOTE applied specifically on XGBoost classifier. In terms of AUROC, borderline-SMOTE demonstrated a considerable improvement in the ensemble learners' learning process. Neither SMOTE nor RUS techniques could improve the learning process in this occasion. Numerous studies have identified reasons for ineffectiveness in these resampling techniques, the most frequently cited being class overlap in feature space, which makes it more difficult for the classifier to learn the decision

Table 2 Predictive Performance for of Low Apgar Score Following Labor Induction Using Ensemble Learning

Algorithm	Resampling Technique	Sensitivity	Specificity	Precision	*F*-Score	Accuracy	AUROC	PPV	NPV
Adaptive boosting (AdaBoost)	Baseline	0.18	0.99	0.75	0.29	0.91	0.73	0.75	0.92
	SMOTE	0.46	0.80	0.19	0.27	0.75	0.67	0.19	0.93
	Borderline SMOTE	0.75	0.80	0.78	0.76	0.77	0.86	0.78	0.76
	RUS	0.52	0.76	0.19	0.28	0.74	0.69	0.19	0.94
Gradient boosting methods (GB)	Baseline	0.19	0.99	0.80	0.31	0.92	0.72	0.80	0.92
	SMOTE	0.42	0.85	0.22	0.29	0.80	0.68	0.22	0.93
	Borderline SMOTE	0.80	0.84	0.83	0.81	0.81	0.89	0.83	0.80
	RUS	0.49	0.79	0.20	0.28	0.76	0.70	0.20	0.94
Extreme gradient boosting (XGBoost)	Baseline	0.20	0.99	0.69	0.30	0.91	0.70	0.69	0.92
	SMOTE	0.25	0.95	0.39	0.30	0.89	0.68	0.39	0.92
	Borderline SMOTE	0.93	0.94	0.94	0.93	0.93	0.97	0.94	0.93
	RUS	0.59	0.69	0.17	0.26	0.64	0.68	0.16	0.94

Abbreviations: AUROC, area under receiver operating curve; NPV, negative predictive value; PPV, positive predictive value; RUS, random undersampling; SMOTE, synthetic minority oversampling technique.

boundary. Studies have established that, if there is an overlapping between the classes given the variables in the dataset, SMOTE would be generating synthetic points affecting the separability.[37,38] In addition, studies have pointed out that "Tomek links", which are pairs of opposing instances that are very close together prior to model building, could be generated as well as other points, therefore harming the classification.[39,40]

Study's Utility and Importance

Researchers working on artificial intelligence particularly on computer-assisted decision-making in healthcare as well as developers who are interested in developing predictive models for decision support system for neonatal healthcare can obtain clues on the efficiency of the ensemble learners particularly when the data is imbalanced and the respective resampling techniques that are likely to improve such prediction and hence make an informed decision. In totality, based on historical registry data, these model predictions enable healthcare informaticians to make highly accurate guesses about the likely outcomes of the intervention.

Study Limitations

As we examined data from a single tertiary institution, our findings may have good internal validity but limited generalizability or external validity. It is possible that the study will show different results for datasets collected from other tertiary hospitals in north Tanzania; thus, caution should be exercised when concluding the specific finding. Furthermore, because we only looked at AUROC, F-scores, precision, NPV, PPV, sensitivity and specificity as performance indicators for boosting-based algorithms, our findings may be rather limited. Future research may shed light on other performance metrics, particularly those for unbalanced data, such as informedness, markedness, and Matthew's correlation coefficient (MCC). Additionally, the current study did not conduct variable selection or feature engineering, nor did it address confounding variables, which could have limited or reduced classifier performance by increasing the likelihood of model overfitting. It would have been interesting to investigate whether or not the impact of feature engineering and confounding effects would result in improved results for both the SMOTE and RUS methods.

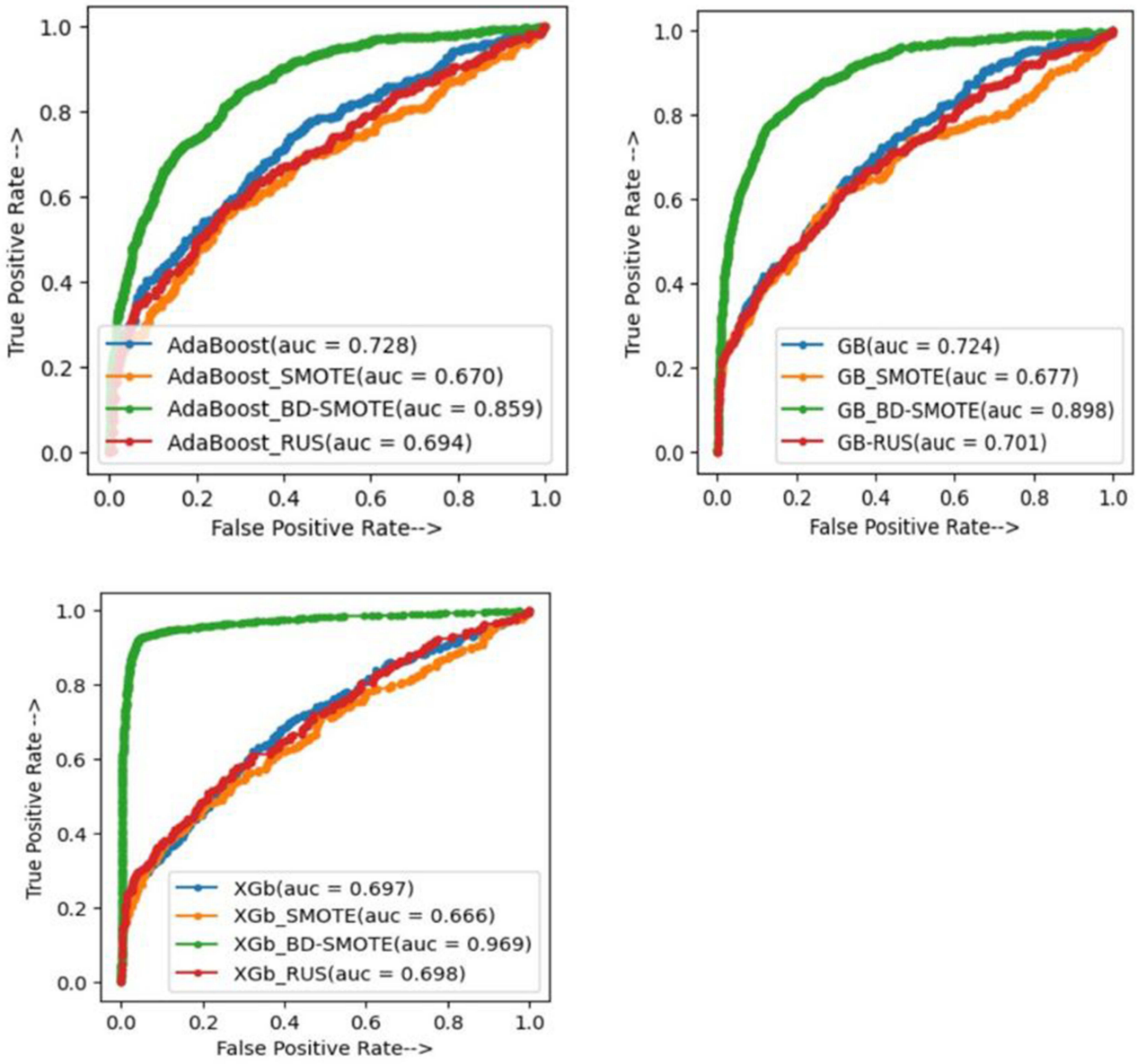

Figure 4 Receiver operating characteristic (ROC) curve diagrams for boosting-based ensemble classifiers comparing the performance by resampling methods.

Exploration and Future Works

We encourage further research into other strategies for improving the learning process in this neonatal outcome, such as the ADASYN (ADAptive SYNthetic) sampling approach and the use of other SMOTE variants such as Safe-Level-SMOTE, SVM-SMOTE and KMeans-SMOTE. The combination of hybrid methods, that is, executing SMOTE and RUS methods concurrently on these ensemble methods, is also worth trying.

Conclusion

Predicting neonatal low Apgar scores after labor induction using this database may be more effective and promising when borderline-SMOTE is executed along with the ensemble methods. Future research may focus on testing additional resampling techniques mentioned earlier, performing feature engineering or variable selection, and optimizing further the ensemble learning hyperparameters.

Ethics Approval and Consent to Participate

This study was approved by the Kilimanjaro Christian Medical University College (KCMU-College) research ethics committee (reference number 985). Because the interview was conducted shortly after the mother had given birth, consent was only obtained verbally before the interview and enrollment. Trained nurses provided the information to the participants about the birth registry project and the information that they would need from them. However, following the consent, the woman could still choose whether or not to respond to specific questions. The KCMC hospital provided administrative clearance to access the data, and the Kilimanjaro Christian Medical College Research Ethics and Review Committee (KCMU-CRERC) approved all consent procedures. The database used in the current study contained no personally identifiable information in order to protect the study participants' confidentiality and privacy.

Acknowledgment

The Birth Registry, the Obstetrics & Gynecology Department, and Epidemiology & Applied Biostatistics Department of the Kilimanjaro Christian Medical University College provided invaluable assistance during this investigation. Thanks to the KCMC birth registry study participants and the Norwegian birth registry for supplying the limited dataset utilized in this investigation.

Funding

This work was supported by the Research on CDC-Hospital-Community Trinity Coordinated Prevention and Control System for Major Infectious Diseases, Zhengzhou University 2020 Key Project of Discipline Construction [XKZDQY202007], 2021 Postgraduate Education Reform and Quality Improvement Project of Henan Province [YJS2021KC07], and National Key R&D Program of China [2018YFC0114501].

Disclosure

The authors declare that they have no competing interest.

References

1. Rayburn WF, Zhang J. Rising rates of labor induction: present concerns and future strategies. *Obstet Gynecol*. 2002;100(1):164–167.
2. Grobman WA, Gilbert S, Landon MB, et al. Outcomes of induction of labor after one prior cesarean. *Obstet Gynecol*. 2007;109(2):262–269. doi:10.1097/01.AOG.0000254169.49346.e9
3. Casey BM, McIntire DD, Leveno KJ. The continuing value of the Apgar score for the assessment of newborn infants. *New Eng J Med*. 2001;344(7):467–471. doi:10.1056/NEJM200102153440701
4. Finster M, Wood M, Raja SN. The Apgar score has survived the test of time. *J Am Soc Anesthesiol*. 2005;102(4):855–857.
5. Leinonen E, Gissler M, Haataja L, et al. Low Apgar scores at both one and five minutes are associated with long-term neurological morbidity. *Acta Paediatrica*. 2018;107(6):942–951. doi:10.1111/apa.14234
6. Ehrenstein V, Pedersen L, Grijota M, Nielsen GL, Rothman KJ, Sørensen HT. Association of Apgar score at five minutes with long-term neurologic disability and cognitive function in a prevalence study of Danish conscripts. *BMC Pregnancy Childbirth*. 2009;9(1):1–7. doi:10.1186/1471-2393-9-14
7. Manning FA, Harman CR, Morrison I, Menticoglou SM, Lange IR, Johnson JM. Fetal assessment based on fetal biophysical profile scoring: IV. An analysis of perinatal morbidity and mortality. *Am J Obstet Gynecol*. 1990;162(3):703–709. doi:10.1016/0002-9378(90)90990-O
8. Yeshaneh A, Kassa A, Kassa ZY, et al. The determinants of 5th minute low Apgar score among newborns who delivered at public hospitals in Hawassa City, South Ethiopia. *BMC Pediatr*. 2021;21:266. doi:10.1186/s12887-021-02745-6
9. Lai S, Flatley C, Kumar S. Perinatal risk factors for low and moderate five-minute Apgar scores at term. *Eur J Obstet Gynecol Reprod Biol*. 2017;210:251–256. doi:10.1016/j.ejogrb.2017.01.008
10. Rogers JF, Graves WL. Risk factors associated with low Apgar scores in a low-income population. *Paediatr Perinat Epidemiol*. 1993;7(2):205–216. doi:10.1111/j.1365-3016.1993.tb00394.x
11. Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 15 August 2018. 559–560.
12. Mung PS, Phyu S. Effective analytics on healthcare big data using ensemble learning. In: 2020 IEEE Conference on Computer Applications (ICCA); February 27, 2020; IEEE. 1–4.
13. Liu N, Li X, Qi E, Xu M, Li L, Gao B. A novel ensemble learning paradigm for medical diagnosis with imbalanced data. *IEEE Access*. 2020;8:171263–171280. doi:10.1109/ACCESS.2020.3014362
14. Bergsjø P, Mlay J, Lie RT, Lie-Nielsen E, Shao JF. A medical birth registry at Kilimanjaro Christian Medical Centre. *East Afr J Public Health*. 2007;4(1):1–4.
15. Robinson JW. Regression tree boosting to adjust health care cost predictions for diagnostic mix. *Health Serv Res*. 2008;43(2):755–772. doi:10.1111/j.1475-6773.2007.00761.x
16. Park Y, Ho J. Tackling overfitting in boosting for noisy healthcare data. In: IEEE Transactions on Knowledge and Data Engineering; December 16, 2019.
17. Joshi MV, Kumar V, Agarwal RC. Evaluating boosting algorithms to classify rare classes: comparison and improvements. In *Proceedings 2001 IEEE International Conference on Data Mining*, 29 November 2001. IEEE; 257–264.
18. Ying C, Qi-Guang M, Jia-Chen L, Lin G. Advance and prospects of AdaBoost algorithm. *Acta Autom Sin*. 2013;39(6):745–758. doi:10.1016/S1874-1029(13)60052-X
19. Lee W, Jun CH, Lee JS. Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification. *Inf Sci (Ny)*. 2017;381:92–103. doi:10.1016/j.ins.2016.11.014
20. Lusa L. Gradient boosting for high-dimensional prediction of rare events. *Comput Stat Data Anal*. 2017;113:19–37. doi:10.1016/j.csda.2016.07.016
21. Wang H, Liu C, Deng L. Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci Rep*. 2018;8(1):1–3.

22. Zhao Y, Wong ZS, Tsui KL. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *J Healthc Eng*. 2018;2018. doi:10.1155/2018/6275435
23. Li J, Liu LS, Fong S, et al. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PLoS One*. 2017;12(7):e0180830. doi:10.1371/journal.pone.0180830
24. Zhu B, Baesens B, Vanden Broucke SK. An empirical comparison of techniques for the class imbalance problem in churn prediction. *Inf Sci*. 2017;408:84–99. doi:10.1016/j.ins.2017.04.015
25. Gosain A, Sardana S. Handling class imbalance problem using oversampling techniques: a review. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI); September 13, 2017; IEEE. 79–85.
26. Amin A, Anwar S, Adnan A, et al. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. *IEEE Access*. 2016;26(4):7940–7957. doi:10.1109/ACCESS.2016.2619719
27. Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Inf Sci*. 2019;1(505):32–64. doi:10.1016/j.ins.2019.07.070
28. Prusa J, Khoshgoftaar TM, Dittman DJ, Napolitano A. Using random undersampling to alleviate class imbalance on tweet sentiment data. In: 2015 IEEE International Conference on Information Reuse and Integration; August 13, 2015; IEEE. 197–202.
29. Chernick MR. Resampling methods. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2012;2(3):255–262.
30. Cheng K, Zhang C, Yu H, Yang X, Zou H, Gao S. Grouped SMOTE with noise filtering mechanism for classifying imbalanced data. *IEEE Access*. 2019;7:170668–170681. doi:10.1109/ACCESS.2019.2955086
31. Triguero I, Galar M, Merino D, Maillo J, Bustince H, Herrera F. Evolutionary undersampling for extremely imbalanced big data classification under apache spark. In: 2016 IEEE Congress on Evolutionary Computation (CEC); July 24, 2016; IEEE. 640–647.
32. Kubat M, Matwin S. Addressing the course of imbalanced training sets: one-sided selection. In: *ICML*. Vol. 97. Citeseer; 1997:179–186.
33. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357. doi:10.1613/jair.953
34. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Australasian Joint Conference on Artificial Intelligence; December 4, 2006; Springer, Berlin, Heidelberg. 1015–1021.
35. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: European Conference on Information Retrieval; March 21, 2005; Springer, Berlin, Heidelberg. 345–359.
36. Guns R, Lioma C, Larsen B. The tipping point: f-score as a function of the number of retrieved items. *Inf Process Manag*. 2012;48(6):1171–1180. doi:10.1016/j.ipm.2012.02.009
37. Alahmari F. A comparison of resampling techniques for medical data using machine learning. *J Inf Knowl Manag*. 2020;19:1–13.
38. Vuttipittayamongkol P, Elyan E, Petrovski A. On the class overlap problem in imbalanced data classification, knowledge-based systems 212; 2021. Available from: http://www.sciencedirect.com/science/article/pii/S0950705120307607. Accessed August 31, 2021.
39. Zeng M, Zou B, Wei F, Liu X, Wang L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS); May 28, 2016; IEEE. 225–228.
40. Ning Q, Zhao X, Ma Z. A novel method for Identification of Glutarylation sites combining Borderline-SMOTE with Tomek links technique in imbalanced data. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics; July 8, 2021.