Comparative Analytics of Classifiers on Resampled Datasets for Pregnancy Outcome Prediction

Udoinyang G. Inyang [1], Imo J. Eyoh[3], Chukwudi O. Nwokoro[5]
Department of Computer Science
Faculty of Science, University of Uyo, Nigeria

Francis B. Osang[2], Adenrele A. Afolorunso[4]
Department of Computer Science, Faculty of Science,
National Open University of Nigeria, Abuja

Abstract—**The main challenges of predictive analytics revolve around the handling of datasets, especially the disproportionate distribution of instances among classes in addition to classifier-suitability issues. This unequal spread causes imbalance learning and severely obstructs prediction accuracy. In this paper, the performances of six classifiers and the effect of data balancing (DB) and formation approaches for predicting pregnancy outcome (PO) were investigated. Synthetic minority oversampling technique (SMOTE), resampling with and without replacement, were adopted for data imbalance treatment. Six classifiers including random forest (RF) were evaluated on each resampled dataset with four test modes using Waikato Environment for Knowledge Analysis and R programming libraries. The results of analysis of variance performed separately using F-measure and root mean squared error showed that mean performance of classifiers across the datasets varied significantly (F=117.9; p=0.00) at 95% confidence interval, while turkey multi-comparison test revealed RF(mean=0.78) and SMOTE (mean=0.73) as having significantly different means. The RF model on SMOTE produced each PO class accuracy ≥0.89, area under the curve ≥ 0.96 and coverage of 97.8% and was adjudged the best classifier-DB method pair. However, there was no significant difference (F=0.07, 0.01; p=1.000) in the mean performances of classifiers across test data modes respectively. It reveals that train/test data modes insignificantly affect classification accuracy, although there are noticeable variations in computational cost. The methodology significantly enhance the predictive accuracy of minority classes and confirms the importance of data-imbalance treatment, and the suitability of RF for PO classification.**

Keywords—Imbalance learning; pregnancy outcome; random forest; SMOTE; imbalance data

I. Introduction

Complications among pregnant women occur frequently and are the obvious sources of maternal mortality (MM) in addition to poor or undesirable pregnancy outcomes (POs). The frequency of MM in developing economies is 50 to 100% higher than those witnessed in developed countries [1]. Pregnancy complications serve as predictors of MMs and other POs (i.e. stillbirth, miscarriage, preterm birth, full term birth etc). Miscarriage, which is an unexpected vaginal flow of blood before twenty-eight (28) weeks of pregnancy, is one of the anomalies noticed among pregnant women especially in Nigeria and other developing countries. Globally, around eighty percent (80%) of maternal deaths and about ninety eight percent (98%) of stillbirths have been linked to direct obstetric complications, like haemorrhage, sepsis, side effects of abortion, preeclampsia and eclampsia, and prolonged obstructed labour [1]. Childbirth complications, maternal infections in pregnancy, maternal syndromes (as pre-eclampsia and diabetes), foetal growth limit and inherited complications are the main reason for the occurrence of stillbirths. Preterm births are associated with multiple pregnancy complications and occurs in 5 to 18% of pregnancies and is also the adjudged cause of infant morbidity and mortality [2].

Improvements in maternal health care systems largely depend on the availability of pieces of knowledge required for the understanding of the effect of pregnancy risks factors, and greatly impact on the future of obstetric health care while attempting to curb maternal morbidity. Although, a significant progress has been recorded in the development of statistical predictive models for PO classification, with better results than clinical tests, there is still room for enhancements in terms of accuracy, interpretability of results and sensitivity to adverse outcomes [3]. Feature ranking and selection, and machine learning (ML) approaches are progressively being utilized for obstetrics outcome classification. However, the suitability of an algorithm to a particular problem domain may affect its performance — accuracy and computational costs. In addition, data from real-world domains are hardly perfect. Some are characterized by uneven distribution of target classes (i.e. some examples of classes may appear more frequently than others) and poses a challenge to data mining (DM) algorithms, as the effectiveness of any DM algorithm is reasonably dependent on the sensitivities to the less frequent (minority) target class [4,5]. Generally, DM algorithms are by default tailored for datasets with equal target class distribution (i.e, they were designed with the assumption of an evenly distributed target class samples), therefore producing poor or below optimal predictive results for the minority target class(es) when imbalanced datasets are encountered. This is because the built model was skewed towards the majority class because of their dominance in the training dataset [4]. The consequences of the class imbalance manifest when the built model is deployed to classify new sets of examples. External influences like missing data, inconsistencies or other forms of noise impact greatly on the imbalanced data distribution, than those that are balanced or near balanced, and produces a noisy classification model [5, 6].

The main focus of predictive modelling, especially in medical researches, is the prediction of the minority target class because of the vital and very useful pieces of knowledge it conveys, despite its paucity in the dataset. Hence the need to adopt methodologies capable of overcoming the class bias issues. Authors in [5], [7] and [8] describe three methods for correcting data imbalance anomaly: (1) data level through resampling, (2) algorithm modification-based approaches, and (3) the cost-sensitive approach. The widely adopted resampling approaches (data level approaches) are based on oversampling and under sampling techniques. This paper aims at determining the best classifier-resampling pair for the prediction of PO using maternal risk factors as predictors. The objectives of this work were twofold; firstly, to compare different resampling techniques based on their ability to address class imbalance and guarantee high accuracy of individual PO classification. Secondly, to assess and perform comparative analysis on six ML algorithms based on their ability to correctly classify PO instances, especially those of the minority class labels. This is achieved by evaluating and comparing classifiers' performances on resampled dataset for the purpose of predicting PO. The remainder of this paper is organized into four sections. Section II gives related works associated with classification methods, dataset imbalance and resampling methods. In Section III, the experimental workflow is described with emphasis on dataset description, pre-processing and resampling, and predictive modelling. The results of the best performing models are described in Section IV while conclusions and future directions are given in Section V.

II. Literature Review

A. Classification and Prediction Models

Classification is a data mining (DM) technique that assigns objects to targeted clusters. Although there are many types of algorithms available in DM for solving medical problems, random forests (RF), k-nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), naïve Bayes (NB), and multi-layer perceptron (MLP) are considered in this paper for pregnancy outcome prediction (POP). SVM has been known to outperform many ML algorithms in many applications, in terms of prediction accuracy and computational cost [9]. Reference [10] employed SVM-based decision support system for preterm birth risks prediction. The model predicted when the birth is likely to occur and the possible outcome for the babies. The authors pointed out that SVM provided an excellent intelligent and comprehensive inference mechanism capable of enhancing the healthcare provided to pregnant women who are at risk with a true positive rate (TPR) of 83.9%, a false positive rate (FPR) of 0.27, and receiver operating characteristic (ROC) area of 0.79. Reference [11] utilized SVM-based decision support system for monitoring the process of child delivery. The data collected include data on heart rate, blood pressure, pulse, uterine contraction, cervical opening and urine volume from pregnant mothers in Indonesia. A total of 40 records were collected based on the earlier listed indicators and tested on the proposed SVM model. For all the selected indicators, an average accuracy of 97.5% was obtained. Author in [12], SVM was used to predict fetal distress using fetal heart rate parameters. A total of 909 data examples with nine parameters were collected and partitioned into 332 normal fetuses, and 577 diagnosis of various pathological conditions. Analysis of results showed that SVM was able to detect fetal distress with an accuracy of 83.0%.

Author in [13] compared two ML algorithms namely; SVM (with linear and non-linear kernels and logistic regression model for the prediction of preterm births. Data for the analysis were collected from a local hospital in India and included age, number of times pregnant, obesity, diabetes mellitus and hypertension with a 10-fold cross validation (10-FCV) for each run. The authors concluded that SVM provided a more accurate prediction with accuracy of 86% compared to the logistic regression model.

Author in [15] proposed a neonatal mortality prediction system using real time medical measurement data based on C4.5 model. The adopted indicators included mean blood pressure, serum, pH, immature/total neutrophil ratio, serum sodium, serum glucose, respiratory rate, heart rate, and pO_2 blood oxygen level. The C4.5 was applied to two sets of data; the summary observations obtained during the initial 12 hours of admission into the neonatal intensive health-care unit by the Canadian Neonatal Network from multiple NICU, and second was the data collected from Children's Hospital of Eastern Ontario (CHEO), Canada and consists of real time medical measurement from a single, out born-only NICU. Analysis of findings revealed that summary data for the first 48 hours of NICU admission provided the best results in the overall with mean sensitivity of 63% and mean specificity of 94%. The authors noted that the results obtained were very significant as the values exceeded the minimum requirement of their clinical partners. Author in [16] DT (C4.5) was applied for the prediction of levels of risk in pregnant women. According to the authors, the C4.5 was adopted because it is powerful, popular, and efficient and can handle the delicate nature of pregnancy problems. For the analysis, 600 pregnant women who went for monthly check-up in Bangalore district hospital were interviewed and two sets of data were obtained namely; unstandardized and standardized pregnancy datasets. The authors concluded that C4.5 classifier provided better results on standardized pregnancy dataset than unstandardized dataset with accuracy of 71.3% and 66.1% respectively. Author in [17] proposed a preterm birth prediction in symptomatic women using DT modelling for biomarkers. The purpose of their study was to use recursive partitioning to identify gestational age-specific and threshold values for infectious and endocrine biomarkers of every pending delivery. The preterm birth predictors considered were white blood cell count, cortisol, maternal age and corticotrophin-releasing hormone. Analysis of results from the DT showed that white blood cell greater than 12,000/mL prior to gestation of 28 weeks and corticotrophin-releasing hormone beyond 28 weeks provided more accurate biomarkers for the prediction of preterm birth within the first 48 hours.

Author in [18] utilized DT, NB, kNN, ANN and SVM for the prediction of high-risk pregnancy cases. The purpose of this study was the timely detection and provision of immediate intervention for these at-risk pregnancy women. In their analysis, DT outperformed other classifiers with accuracy of 97.01%, followed by ANN with accuracy of 93.40%, other classifiers performed in the average with SVM giving the worst accuracy of 76.39% in this context. Author in [19] adopted some DM tools to predict neonatal jaundice caused by hyperbilirubinemia. The aim of the work was to accurately identify neonates at risk of developing severe hyperbilirubinemia in order to offer early medical attention and treatment. Two hundred and twenty seven (227) healthy new-born infants with gestational age ≥ 35 weeks were enrolled for the experiment while bilirubin meter was used for capturing bilirubin levels from the time of birth to hospital discharge. An input space of 72 variables were collected and pruned to 62 via pre-processing. An interval of 8 hours was allowed between measurements for two months (February to March 2011) in Obstetrics Department of the Centro Hospitalar Tâmega e Sousa, E.P.E., North Portugal. The classifiers selected for the analysis included J48, simple classification and regression trees, NB, MLP, sequential minimal optimization (SMO) algorithm and simple logistic available in Waikato Environment for Knowledge Analysis (WEKA). The authors pointed out that only three classifiers namely; NB, MLP and simple logistic correctly predicted neonatal hyperbilirubinemia.

Author in [20] used nine ML tools for the prediction of fetal health status based on maternal clinical history. Ninety six (96) pregnant women between 18 and 41 years in Istanbul were involved for the experiment between January 17, 2015 and February 21, 2017 with 97 fetuses (95 single and 1 twins) and 23 input features. A 10-FCV was employed for training and testing using nine ML algorithms available in Azure ML system. These included averaged perceptron, boosted DT, Bayes point machine, decision forest, decision jungle, logistic regression, ANN and SVM. Result showed that features such as fetal age, age of mother, blood stereotype, test results, number of abortus, number of delivery and any illnesses of mother regarding pregnancy were significant factors that influenced fetal health status. Out of the selected algorithms, the authors pointed out that boosted DT, decision forest and decision jungle produced the best results with accuracy of 89.5%. In conclusion, the authors noted prediction systems are vital tools that could be employed by both clinicians and pregnant women to remotely predict fetal health status in an early stage.

Author in [21] proposed a hybrid system consisting of bijective soft set and back propagation ANN for the prediction of neonatal jaundice. The neonatal jaundice dataset comprising 808 instances with 16 attributes collected from January to December, 2007 in neonatal intensive care unit in Cairo, Egypt, was used for the experiment. The proposed system was compared with bijective soft set, back propagation ANN, MLP, decision table and NB and found to provide the best accuracy of 99.1%. Author in [22] utilized MLP to predict risk of diabetes mellitus that causes several complications during pregnancy. The experimental setting consisted of 394 pregnant women aged 21 years and above, eight attributes and 10-FCV test mode. Results revealed that MLP attained a precision of 0.74, Recall of 74.1%, F-measure (F_m) of 74.1%, and ROC area of 77.9%. The authors concluded that MLP is an excellent tool for predicting gestational diabetes mellitus.

The work reported in [23] employed SVM-based decision support system for preterm birth risks prediction. The SVM model predicted the likely to time birth occur and the possible outcome of babies' status. The results of the empirical experiment showcased SVM as the most performing in terms of intelligent and comprehensive inference mechanism regarding decision support for at risk pregnant women. The result produced true positive rate of 83.9%, a false positive rate of 0.27, and receiver operating characteristic (ROC) area of 0.79. Refs. [24-27] deployed decision support tools that would provide needed assistance to practitioners in ensuring safety of vaginal births after cesarean delivery for women of child bearing age and in the general management of PO. Refs. [28-29] have demonstrated the effectiveness of decision support systems in handling associations between two or more obstetric and neonatal emergencies. A comparative study of machine learning tools and statistical models was reported in [30] for the prediction of postpartum hemorrhage (PH) risks during labour with the aim of minimizing maternal morbidity and mortality. The experiments on data from 12 sites showed that all the models adopted in the study produced satisfactory results, although the extreme gradient boosting model (XGboost) had the best ability to discriminate among PH followed by random forests (RF) and lasso regression model. The effectiveness of ML methods in mining of electronic health data in the domain of atrial fibrillation (AF) induced risks prediction was reported in [31]. Out of a total of 2,252,219 women used for the study, 1,225,533 developed AF during a selected 6-month interval. Two hundred (200) widely used electronic health record features, (age and sex inclusive), and random oversampling approach implemented with a single-layer, fully connected ANN yielded the optimal prediction of six-month incident AF, with an area under the receiver operating characteristic curve (AUC) of 80.0% and an F1 score of 11.0%. The ANN model performed only slightly better than the basic logistic regression consisting of known clinical risk factors for AF, which had 79.4% and 79.0% as AUC and F1 value respectively. The results confirmed the effectiveness of machine learning algorithms in the prediction of AF in patients. The performance of Fuzzy approach, SVM, RF and Naïve Bayes (NB) for the prediction of cardiotocograph‑based labour stage classification from patients with uterine contraction pressure during ante‑partum and intra ‑partum period, the proposed algorithm tend to be efficient and effective in terms of visual estimation to incorporate automated decision support system, which will help to reduce high risk of hospitalized patients. Author in [32] experimental results of the impact computational intelligence on the precision of cardiovascular medicine was presented. The method was applied to neonatal coarctation classification and prediction by analyzing genome-wide DNA methylation of newborn blood DNA using in 24 isolated, non-syndromic cases. Six machine learning algorithms including deep learning was used for detection. Deep learning achieved the optimal performance with an AUC and sensitivity of 95% and 98% specificity at 95% confidence interval. The related works considered were based on a single dataset test mode. The significance of this work is the assessment of each classifier on varying dataset test modes.

B. Data Resampling Approaches

Real-world modelling problems are characterized by uneven target class spread. These domains include but not limited to fraud, medicine, spam, web, telecommunications, education and churn customers. In the medical domains like obstetrics, the frequency of desirable outcomes is usually higher than the adverse ones, thereby resulting in a data imbalance problem (DIP). During model building, the infrequent target class(es) have limited representation in the built model because of paucity of training samples, and therefore lacks the classification and prediction competences regarding such class(es). The degree of imbalance is measured by the imbalance ratio (IR) — the ratio of the frequency of observations in the majority class to the tally of instances in the minority class [33], [34]. The notational description of DIP is as follows [35]. Given a dataset Q with m examples and *n* attributes, where $Q = \{x_i, y_i\}$, $i = 1,2, \dots, m$, and where $x_i \in X$ is a data-point in the attribute set $X = \{b_1, \dots, b_n\}$, and $y_i \in Y$ is an element in the set of target classes $Y = \{1, \dots, c\}$. A subset of the desirable (majority class) instances $G \subset$X, and subsets of minority class (adverse instances) $U \subset X$, where $|G| < |U|$. The preprocessing via resampling applied the maternal dataset has the goal of balancing the training and testing sets Q such that$|G| \equiv |U|$.

Since DM algorithms were designed to learn from balanced class training representatives, they produce models that are less equipped for classification of instances in minority class(es) whereas a good coverage is recorded for majority class elements, when confronted with DIP[5], [8]. Although three approaches — resampling, algorithm modification and cost-sensitive approaches, are recommended for imbalance anomaly correction [5], this paper investigated the effect of resampling methodologies on predictive performance of some selected DM classifiers. The rationale for choosing resample approach is due to its simplicity, cost efficiency and classifier independence. Resampling methods operate either by adding elements to the minority class (oversampling) or reducing representatives of the majority class (undersampling). It can also combine both oversampling and undersampling approaches [34],[36]. Synthetic minority oversampling technique (SMOTE), is an oversampling approach that increases the elements of the minor class(es) by generating simulated data items in the nearness of the existing minority class instances, with the goal of flattening IR. Author in [36] describes two key stages for SMOTE implementation: (1) Clustering data-points based on class labels and finding kNN using euclidean distance between every minority data-point with respect to all other minority data items. The *k* least distance examples are chosen as the nearest neighbours. Euclidean distance (D) between one object with the minority class label (x) and another sample with the minority class label (y) for all features is defined by Eq. 1 [36]. (2) New data-points are constructed by inserting points between any two elements belonging to the minority class. One of its *kNN* will be randomized to be candidates in new data construction process. Thereafter, original minor data element (*x*) and one chosen candidate (*y*) will be used to generate new values among *x* and *y*. The process is defined by Eq. 2

$$D_{x,y} = \sqrt{\sum_{c=1}^{n} (x_c - y_c)^2} \tag{1}$$

$$N_c(x, y) = x_c + t.(x_c - y_c)\ for\ 0 \le t \le 1 \tag{2}$$

where $\mathrm{N_c(x, y)}$ is the new data-point, *n* is the number of attributes and *r* is a random number between 0 and 1.

Undersampling approaches generate a subgroup of the original dataset by deleting instances with the majority class label. Random undersampling takes place when observations that are deleted are arbitrarily picked from majority class until the data set becomes balanced whereas informative undersampling adopts available rules to mark items for deletion [37]. However, undersampling techniques seemingly impact the multi-class imbalanced data classification performance negatively if useful instances in each majority class are eliminated [38],[39].

C. *Classifier Evaluation Metrics*

In predictive analytics, it is an essential task to assess the quality of the predictions in order to guide in classifier modelling for the specified problem domain. A contingency or confusion matrix (CM) is usually applied for such purposes, providing not only classification errors and accuracy, but also parameters to compute other measures [8],[35]. CM is actually not a performance measure as such, but the basis for deriving other measures. The basic CM for a binary classifier (Table I) uses four indicators (true positive (TP), false positive (FP), true negative (TN) and false-negative (FN)) to measure the classification performance of both classes independently.

TABLE I. CM FOR A BINOMIAL CLASSIFICATION PROBLEM

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

TP is the number of positive PO instances that are correctly classified while FP is the number of negative PO instances misclassified as positive. FN represents the tally of positive PO instances misclassified as negative whereas the negative instances that are correctly classified are defined by TN. These parameters are represented as percentages; TPR, FPR, true negative rate (TNR), and false negative rate (FNR) and defined in Equations 3 – 6 respectively, as follows.

$$TPR(sensitivity) = \frac{TP}{TP + FN} \tag{3}$$

$$TNR\ (specificity) = \frac{TN}{FP + TN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

$$FNR = \frac{FN}{TP + FN} \tag{6}$$

TPR (or sensitivity) gives a measure of the proportion of actual positive examples which are correctly classified while FPR is the proportion of actual negative examples of PO which are incorrectly identified as positive PO instances. FNR is the percentage of positive PO instances which are wrongly classified as negative POs while the TNR is the fraction of actual negative PO examples which are correctly classified. Accuracy (ACC) has been the widely used metric [8], [40]. It quantifies the predictive capability of elements in the test dataset. Although, it is easy to implement and interpret, it ignores class distribution and frequently skews in the direction of the majority class. It is therefore not suitable for DIP scenario [35]. Apart from ACC (Eq. 7), there are other derivable measures that consider class inequality in their design — precision, recall and F$_m$ given in Equations 8 - 10 respectively and are suitable when the positive class label is the key issue whereas the ROC and area under the curve (AUC) capture performances of minority and majority classes. Precision is a fraction of the predicted positive POs that are actually positive while F$_m$ defines the harmonic mean between precision and sensitivity. The F$_m$ is a more complete measure because it combines precision and recall.

$$ACC = \frac{TN + TP}{FP + TN + TP + TP} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F_m = \frac{2 \times precision \ \times recall}{precision + recall} \tag{10}$$

The ROC curve is a graph of TPR or sensitivity on the *y*-axis against FPR on the *x*-axis while the extreme values are 0 and 1. The total area enclosed by the ROC curve is described by AUC value and is given in Eq. 11. An AUC value of 100% depicts a perfect classification, the one close to 100% depicts a very good performance, while values lower than 50% depicts performance by chance or luck. Another widely used metric of interest, is the root mean squared error (RMSE) which measures the deviation between the classifier's output and actual values. It is defined in Eq. 12 [40].

$$AUC = \frac{1 + TPR - FPR}{2} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(output_c(i) - actual_c(i))^2} \tag{12}$$

where $output_c(i)$ denotes the prediction probability of instance i, which belongs to class c, and $actual_c(i)$ depicts the actual probability.

III. Experimental Settings

A. Dataset Source and Preprocessing

Data was acquired from secondary health facilities in Uyo, Nigeria. A total of one thousand six hundred and thirty-two (1,632) records were obtained from archives of retrospective observations of pregnant women recorded while they enrolled for antenatal care, with an input feature space of forty-two (42) attributes in excluding the target variable. Some of the attributes include; average maternal age, number of children delivered, previous medical history, abortion, miscarriage, prematurity, previous illness, number of attendances to antenatal care, antenatal registration, and mode of delivery, amongst other features. Attribute cleaning, aggregation and elimination of attributes with only a single domain value was performed. The resultant dataset which had thirty-five (35) attributes were subjected to feature ranking [41] and selection via PCA in WEKA software. Attributes with eigenvalue (EV) scores greater than or equal to unity [42] were thirteen (13) and together accounted for 67.13% variation of the target feature. Table II gives a description of attribute description and rank.

As shown in Table II, the average maternal blood pressure topped the list with EV of 3.86 (11.7% proportion of variance), followed by average maternal weight (EV = 2.77, proportion = 8.39%). The thirteenth rank attribute, average ascorbic acid level accounted for 3.17% variation with eigenvalue score of 1.05.

TABLE II. Rank And Description Of Significant Attributes

Rank	Attribute	Description	EV	Proportion (%)	Cumulative (%)
1	Maternal BP	Average maternal blood pressure	3.86	11.69	11.69
2	Maternal Weight	Average maternal weight	2.77	8.39	20.29
3	Hemoglobin Level	Average number of red blood cells count	2.37	7.18	27.47
4	PCV level	Average Packed Cell Volume count	1.92	5.82	33.29
5	Pulse Rate	Average number of heart beats per minute	1.54	4.67	37.67
6	Mode of Delivery	Delivery method vaginal delivery =1; caesarean section = 2	1.42	4.30	42.26
7	Malaria Frequency	Number of times maternal malaria Diagnosis	1.39	4.21	46.47
8	Hepatitis C	Indicates history of hepatitis C disease; presence=1, absence=2	1.26	3.82	50.29
9	Diabetes Status	Maternal Diabetic status non-diabetic=0 type1=1; type2=2, others=3	1.18	3.60	53.89
10	Herbal Ingestion	Use of herbal medicinal products during pregnancy	1.15	3.48	57.37
11	Respiratory disorder	Maternal respiratory disease status; presence=1, absence=2	1.12	3.39	60.76
12	Age	Maternal age during pregnancy	1.06	3.20	63.96
13	Ascorbic acid Level	Average amount of ascorbic acid in the body during pregnancy	1.05	3.17	67.13
14	Pregnancy outcome	Maternal delivery outcome miscarriage = 0; pre-term =1; full-term=2, stillbirth=3	-	-	-

Table II also reveals that PO consists of four distinct classes of instances — miscarriage, preterm, term and stillbirth. Out of the 1,632 records, 198 (12.1%) examples belong to miscarriage class, 65 (4.0%) are representatives of preterm births while 114 instances (7.0%) were stillbirths. Term births had majority of observations with a frequency of 1255 (76.9%) (while observations of preterm, still-births and miscarriage classes together have a tally of 23.1%). The distribution of class examples depicts a severe imbalanced situation where term births is the majority class whereas the other three classes (miscarriage, preterm, term and stillbirth) are in the minority with high IR values as follows; miscarriage (6.3), preterm (19.1) and stillbirth (11.0).

B. Resampling Methodology

The final stage of preprocessing implements three data resampling techniques based on oversampling and undersampling — SMOTE, resample with replacement (RRW) and resample without replacement (RRN). The implementation was performed in WEKA version 3.8.4 using default values of “weka.filters.supervised.instance.resample” function and R library. The distribution of class labels in the resultant datasets (Table III) depicts a substantial reduction in the severity of imbalance in the resampled datasets than the original datset (ORD). There is a uniform spread in the RRW method and a near balance distribution in the SMOTE dataset. The RRN approach randomly eliminated 847 (67.5%) instances of the majority PO class (term births) while other PO classes remained unchanged. The IR of the resampled dataset, given in Table IV and Fig. 1, reveals a maximum inter class IR deviation of 0.43 for SMOTE while zero (0) deviation is observed for RRW dataset.

TABLE III. DISTRIBUTION OF TARGET LABELS DATASET

Dataset Code	**Resample Method**	**Misca-rriage (%)**	**Pret-erm (%)**	**Term (%)**	**Still Birth (%)**	**Total**
ORD	Original data	198 (12.1)	65 (4.0)	1255 (76.9)	114 (7.0)	1632
SMOTE	Oversamp-ling (SMOTE)	1188 (25.5)	1300 (27.9)	1255 (27.0)	912 (25.2)	4655
RRN	Random resampling without replacement	198 (25.2)	65 (8.3)	408 (53.0)	114 (14.5)	785
RRW	Random resample with replacement	408 (25)	408 (25)	408 (25)	408 (25)	1632

TABLE IV. ANALYSIS OF IR IN THE ORIGINAL AND RESAMPLED DATASETS

Dataset Code	Miscarriage (%)	Preterm (%)	Term (%)	Still Birth (%)
ORD	6.3	19.3	1	11.0
SMOTE	1.09	1	1.03	1.43
RRN	2.06	6.27	1	3.58
RRW	1	1	1	1

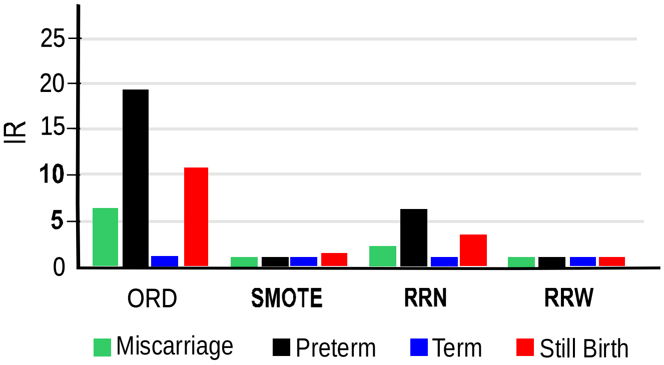

Fig. 1. Visualization of IR for ORD and Resampled Datasets

The RRN dataset has a maximum IR value of 6.27 for preterm class which was hitherto 19.3, while stillbirth drifted to 3.58 from 11.0. This produces a significant balance effect when compared with the ORD dataset.

C. Predictive Modeling and Performance Comparison

The input features correspond to the significant attributes selected during preprocessing with PCA while PO is the target variable. The predictive modeling was performed in WEKA 3.8.4 platform using six classifiers; DT, SVM, KNN, RF, NB and MLP. The default WEKA parameters of each classifier were used for model building and testing processes as follows;

- DT was implemented with C4.5 algorithm with 0.25 as the confidence level, the minimum number of item-sets per leaf was set to 2 while leaf pruning was utilized to get the final tree.
- SVM was trained with John Platt's SMO, Polykernel function, an internal parameter of 1.0 for the exponent of each kernel function and a penalty parameter at 1.0. The model adopted a batch processing mode with a bag-size of 100.
- kNN was invoked through Instance based learning (Ibl) function with one neighbour for returning the output class. Brute force search algorithm was used for nearest neighbours selection based on euclidean distance. The process was iterated with a batch processing size of 100.
- NB parameters were based on weight learning without kernel estimator and supervised discretization functions.
- MLP used backpropagation to learn a multi-layered perceptron. It used a learning rate of 0.3 and momentum of 0.2.
- RF constructed a forest of random trees with an unlimited depth and 100 as the maximum number of iterations.

The models were built and executed with each resampled dataset by adopting four test modes— two based on k-fold cross validation while the other two relied on percentage splitting ratio namely; 10-FCV, 5-fold cross validation (5-FCV), 80% split for training and 20% for testing (80-20) and 70% split for training and 30% for testing (70-30). Since all the classes of PO are of interest in this work, the performances of the classifiers were evaluated based on derivatives from ROC curve — sensitivity, specificity, recall, precision, AUC and other performance measures including kapa statistic (KS), RMSE and CM parameters [43]. Generally, for DIP treatment, measures such as F_m and AUC are recommended rather than traditional classification ACC. Since F_m, combines precision and recall (it eliminates the limitations of other single metric) and also imposes an enhanced inter-class performance equilibrium [44] while RMSE is widely used error measure for classifier evaluation, both measures are more suitable for real-world applications and therefore adopted for the comparative analysis. The overall results (Table V) depict F_m and RMSE values across dataset and classifiers for all the test modes while weighted averages across dataset and classifiers are presented in Tables VI and VII.

The results in Table V, show that the F_m and RMSE values were moderately high for ORD dataset. However due to the imbalance effect, predictions based on the ORD dataset will be biased towards the term births class. The performance of RF on SMOTE dataset ($F_m \geq 0.92$) was the best followed by kNN in RRW ($F_m \geq 0.81$) dataset. In terms of RMSE, the least error value was recorded by RF in SMOTE dataset (RMSE= 0.18) with 10-FCV test mode. The weighted averages in Tables VI and VII, which also appear graphically in Fig. 2 and 3, clearly exposed the performances of the resampled datasets across test modes and classifiers respectively. The average classification result from the resampled dataset is highest with SMOTE dataset ($F_m = 0.73$) in 10-FCV and 80-20 datasets while the least value ($F_m = 0.53$) was attributed to NB in 10-FCV and 80-20 datasets. The performances in terms of both F_m and RMSE for RF and DT are almost the same as evidenced in overlapped trajectories in Fig. 3 and are the topmost performing classifiers while NB is the least performing algorithm.

All F_m and RMSE differences across classifiers and datasets are significant with 95% confidence using a two-way analysis of variance (ANOVA) test in R programming environment. The interaction effect [41] between test modes, resampled datasets and classifiers provided evidence of the existence of a significant interaction between the effects of datasets and classifiers, (F=117.94; p=0.000), while interaction between test mode and other factors yielded no significant effect (F=0.07,0.01;p=1.000 respectively) at 95% confidence level.

A similar result was observed with RMSE as the response variable — classifier and dataset interaction produced (F=17.24; p= 0.00) while interaction involving test modes produced insignificant effects (F = 0.17, 0.14; p= 0.00). This implies that prediction accuracy varies significantly across classifiers and dataset only. Turkey's multiple comparison test [45] showed that the difference between means of dataset pairs and classifier pairs are significantly different. The confidence intervals for SMOTE (mean = 0.73), RF (mean=0.78) are different from others in their respective groups with similar trend significantly observed with RMSE as the dependent variable. However, test modes do not affect the quality of predictions significantly.

TABLE V. PERFORMANCE COMPARISON CLASSIFIERS ON RESAMPLED DATASETS AND TEST MODES

Classifier	Resampling Method	**Test Modes**							
		5-FCV		**10-FCV**		**Train (70%)**		**Train (80%)**	
		F_m	RMSE	F_m	RMSE	F_m	RMSE	F_m	RMSE
NB	ORD	.73	.32	.73	0.32	.70	.33	.71	.32
	SMOTE	.32	.53	.29	.53	.33	.53	.30	.54
	RRN	.59	.38	.60	.38	.61	.38	.59	.41
	RRW	.53	.40	.52	.41	.52	.41	.52	.41
MLP	ORD	.77	.26	.76	.26	.80	.30	.79	.30
	SMOTE	.66	.34	.66	.34	.63	.35	.67	.33
	RRN	.56	.67	.58	.36	.63	.37	.55	.38
	RRW	.58	.37	.60	.37	.60	.36	.58	.36
KNN	ORD	.74	.34	.74	.34	.71	.35	.69	.36
	SMOTE	.89	.23	.89	.23	.88	.24	.88	.23
	RRN	.52	.47	.52	.47	.53	.47	.51	.47
	RRW	**.82**	**.27**	**.84**	**.26**	**.81**	**.29**	**.84**	**.27**
SVM	ORD	.76	.35	.75	.35	.68	.35	.69	.36
	SMOTE	.71	.40	.71	.40	.72	.39	.71	.40
	RRN	.63	.38	.64	.38	.66	.38	.65	.38
	RRW	.50	.40	.50	.39	.49	.40	.49	.39
RF	ORD	.79	.29	.79	.29	.80	.30	.76	.33
	SMOTE	**.94**	**.20**	**.94**	**.18**	**.92**	**.20**	**.93**	**.19**
	RRN	.56	.39	.56	.39	.59	.39	.52	.41
	RRW	**.84**	**.24**	**.85**	**.23**	**.82**	**.26**	**.83**	**.24**
DT	ORD	.76	.33	.81	.26	.78	.31	.80	.30
	SMOTE	.88	.23	.88	.23	.86	.25	.87	.23
	RRN	.58	.36	.58	.37	.70	.35	.55	.37
	RRW	.79	.29	.80	.28	.73	.32	.79	.28

TABLE VI. WEIGHTED AVERAGE OF F_M AND RMSE ACROSS DATASETS

Dataset	**5-FCV**		**10-FCV**		**Train (70%)**		**Train (80%)**	
	F_m	RMSE	F_m	RMSE	F_m	RMSE	F_m	RMSE
ORD	0.76	0.32	0.76	0.31	0.73	0.33	0.74	0.33
SMOTE	0.73	0.32	0.73	0.32	0.72	0.33	0.73	0.32
RRN	0.57	0.45	0.58	0.39	0.63	0.39	0.57	0.40
RRW	0.68	0.35	0.65	0.34	0.63	0.36	0.64	0.34

TABLE VII. WEIGHTED AVERAGE ACROSS CLASSIFIERS

Dataset	**5-FCV**		**10-FCV**		**Train (70%)**		**Train (80%)**	
	F_m	RMSE	F_m	RMSE	F_m	RMSE	F_m	RMSE
NB	0.54	0.41	0.53	0.41	0.54	0.41	0.53	0.42
MLP	0.64	0.41	0.65	0.33	0.67	0.35	0.65	0.34
KNN	0.74	0.33	0.75	0.33	0.73	0.34	0.73	0.33
SVM	0.65	0.38	0.65	0.38	0.64	0.38	0.64	0.38
RF	0.78	0.28	0.79	0.27	0.78	0.29	0.76	0.29
DT	0.77	0.29	0.78	0.27	0.78	0.30	0.77	0.28

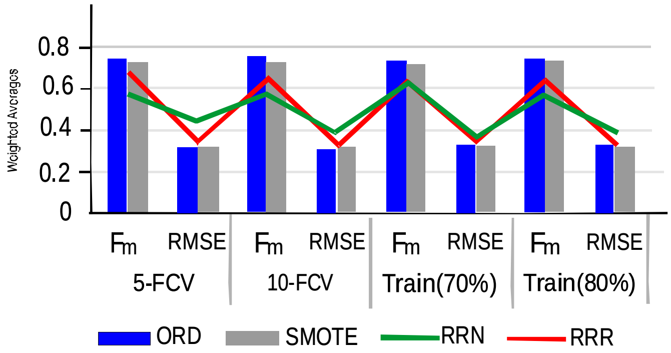

Fig. 2. Graph of Weighted Averages of F_m and RMSE Across Datasets.

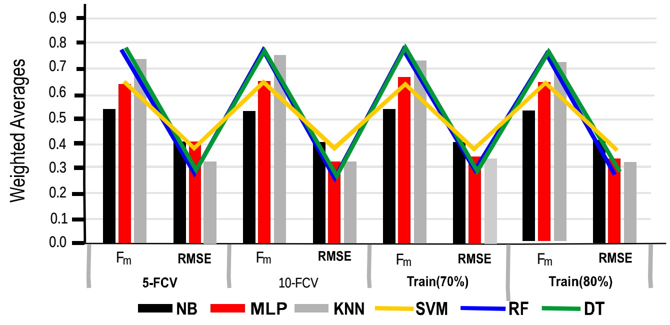

Fig. 3. Graph of Weighted Averages of Fm and RMSE Across Classifiers.

IV. EVALUATION OF RF CLASSIFICATION AND DISCUSSION

The results obtained from PO predictions on all classes using RF classifier and SMOTE dataset in all test mode are reported in Tables VIII and IX.

TABLE VIII. RF CLASS PREDICTIONS WITH SMOTE ACROSS TEST MODES

Test mode	KS	Coverage (%)	ACC (%)	Time (secs)
5-FCV	.89	98.1	92	1.57
10-FCV	.90	97.8	93.1	1.5
Train (70%)	.88	98.3	91.9	0.16
Train (80%)	.89	98.0	93.0	.07

TABLE IX. RF CLASS PREDICTIONS EVALUATION WITH SMOTE

Test mode	Class	TPR	FPR	PR	RE	FM	AUC
5-FCV	Stillbirth	.90	.02	.92	.90	.91	.97
	Term	.87	.06	.84	.87	.86	.96
	Preterm	.96	.01	.98	.96	.97	.99
	Miscarriage	.92	.02	.93	.92	.93	.99
10-FCV	Stillbirth	.94	.02	.93	.91	.92	.96
	Term	.89	.06	.85	.89	.87	.96
	Preterm	.96	.01	.98	.96	.97	.99
	Miscarriage	.93	.02	.94	.93	.94	.99
Train (70%)	Stillbirth	.88	.03	.88	.88	.88	.97
	Term	.88	.07	.82	.88	.85	.95
	Preterm	.96	.01	.98	.96	.97	1.0
	Miscarriage	.91	.01	.96	.91	.93	.99
Train (80%)	Stillbirth	.86	.02	.92	.86	.89	.96
	Term	.89	.07	.82	.89	.85	.95
	Preterm	.96	.01	.97	.96	.96	.99
	Miscarriage	.93	.013	.96	.93	.95	.99

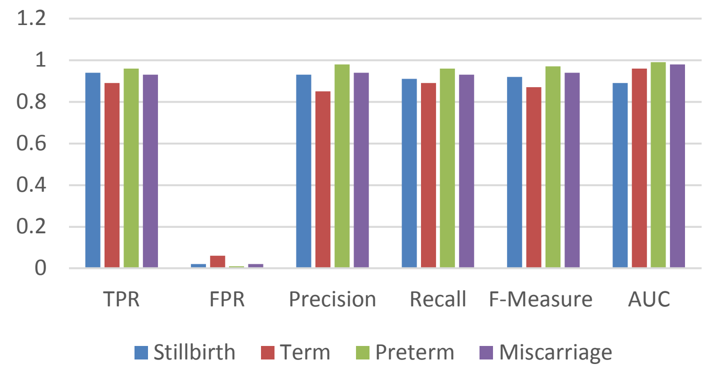

Fig. 4. Graph of RF 10-FCV Performance of PO Classes with SMOTE.

Excellent coverage of instances is expressed in the results with (ACC ≥ 91.9% and coverage ≥ 97.8%) across the test modes. The time used ranges from 0.07 seconds to 1.57 seconds with 80-20 dataset split having the least time due to the number of testing instances used. Average class predictions were greater than 91.8% with 10-FCV having the highest average ACC of 93.1% although computationally expensive. As shown in Table IX, preterm class has the highest sensitivity of 96% in all test modes while the least score is observed for Term class in all test modes except 10-FCV (89%). A similar trend is observed for F_m where term birth earned the least score of 85% in both Train (70%) and Train (80%) test modes. All the performance measures reported in this work depict very good results therefore confirming the suitability of the approach.

The summary of RF predictions using 10-FCV test mode — since it had the highest ACC and KS values (Table VIII) and least classification error (RMSE=0.18) as shown in Table V, is given in Fig. 4 while associated ROC curves for all the PO classes are presented in Fig. 5 to 8.

The sensitivity, precision, recall and F_m for all classes are excellent (ACC ≥ 89). The RMSE=0.01 is least for preterm births in almost all the test modes while term-birth was the least performing class — RMSE = 0.07. The AUC for each class of PO (still birth = 96.36%, term birth = 95.84%, preterm = 99.12% and miscarriage = 98.76%) depict an enhancement in the results as compared to the ORD dataset. The sensitivity also recorded very good results in all categories of PO (still =94%, Term =89%, preterm =96%, miscarriage =93%) with insignificant FPR in each class (0.01≤ FPR ≥0.06).

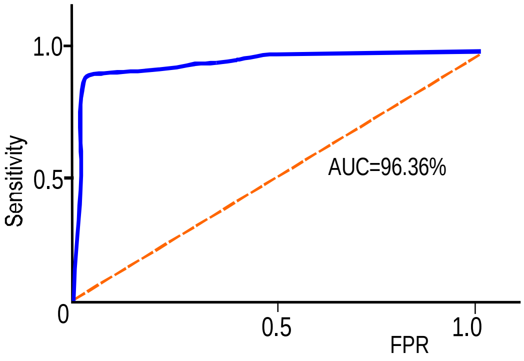

Fig. 5. ROC Curve for Still Birth Prediction.

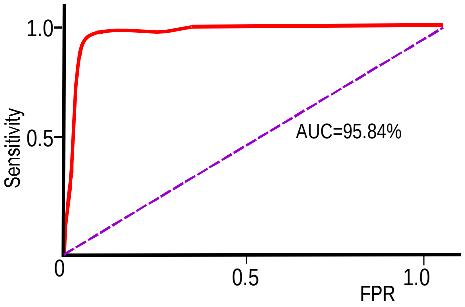

Fig. 6. ROC Curve for Term Birth Class Prediction.

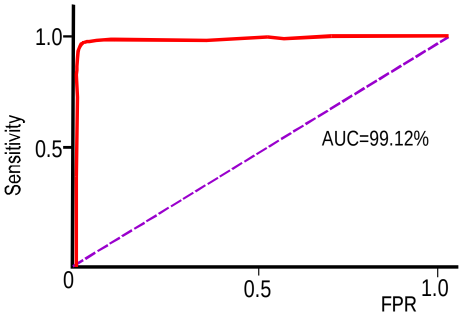

Fig. 7. ROC Curve for Preterm Birth Class Pprediction.

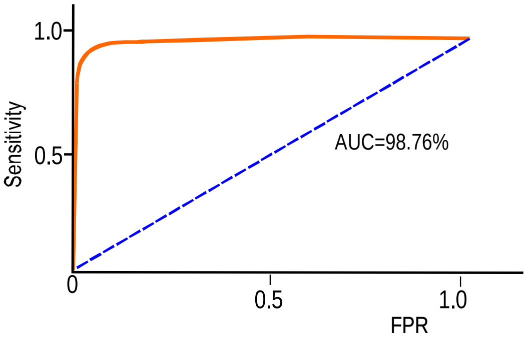

Fig. 8. ROC Curve for Miscarriage Class Prediction.

V. Conclusion

The work reported in this paper implemented a comparative predictive analytics on six machine learning algorithms (RF, DT, NB, SVM, MLP, kNN) and three data imbalance treatment approaches (RRW,RRN, SMOTE) for the prediction of POs using four test dataset modes (10-FCV, 5-FCV, Train (70%), Train (80%)). The aim was to identify the best classifier for PO classification using pregnancy risk factors. The process commenced with data collection and preprocessing — data cleaning, integration, feature selection and imbalance treatment. Feature rank analysis identified 13 principal attributes based on EV scores from PCA, which the other analytic stages depended on. SMOTE, RRN and RRW datasets drastically reduced the IR when compared to ORD dataset and were used for classification and prediction by the six ML algorithms. The experiment was conducted on four different test modes while derivatives of CM and other standard metrics were used to evaluate the performances of the different classifiers.

The results of ANOVA performed separately using F_m and RMSE showed that mean performance of classifiers across the datasets varied significantly (F=117.94; p=0.00) at 95% confidence interval, while turkey multi-comparison test revealed RF (mean=0.78) and SMOTE (mean=0.73) as having outstandingly significant means. In addition, RF model on SMOTE dataset produced ACC ≥ 0.89, AUC ≥ 0.96 and coverage of 97.8% for each PO class which depict a very good performance and was the best performing classifier. However, there was no significant difference (F=0.07, 0.01; p=1.000) in the mean performance of classifiers and datasets across test data modes respectively. The results significantly enhance the predictive accuracy of all the classes (especially adverse PO class) and confirmed the importance of data-imbalance treatment and the suitability of RF for PO classification. In terms of the adopted resampling methods, SMOTE produced the least IR among the various classes while RF and DT were the two most performing classifiers. This implies that oversampling is better than random unsdersampling methodology in the treatment of DIP maternal health domian. The results further proved that train/test data modes insignificantly affect classification accuracy in a balanced data setting, although there are noticeable variations in computational cost. The results of preprocessing identified 13 pregnancy risk factors that significantly impact on PO, therefore provide the right information for the early diagnosis and treatment of the adverse POs thereby reducing MM. The performance of these models on binary classification problems and discovery of optimal classifiers' parameters for improved performance are directions for future work.

References

[1] Goldenberg, Robert L., Elizabeth M. McClure, and Sarah Saleem. "Improving pregnancy outcomes in low-and middle-income countries." Reproductive health 15(1), (2018): 88.

[2] Romero, Roberto, Sudhansu K. Dey, and Susan J. Fisher. "Preterm labor: one syndrome, many causes." Science 345, no. 6198 (2014): 760-765.

[3] Mu, Yu, Kai Feng, Ying Yang, and Jingyuan Wang. "Applying deep learning for adverse pregnancy outcome detection with pre-pregnancy health data." In MATEC Web of Conferences, vol. 189, p. 10014. EDP Sciences, 2018.

[4] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," Int. J. Pattern Recognition Artificial. Intelligence., vol. 23, no. 4, pp. 687_719, 2009.

[5] Ebenuwa, Solomon H., Mhd Saeed Sharif, Mamoun Alazab, and Ameer Al-Nemrat. "Variance ranking attributes selection techniques for binary classification problem in imbalance data." IEEE Access 7 (2019): 24649-24666..

[6] B. A. G. Nguyen Hoang and S. Phung, ``Learning pattern classi_cation tasks with imbalanced data sets," in Pattern Recognition, P.-Y. Yin, Ed., Vukovar, Croatia: InTech, 2009.

[7] Yıldırım, Pınar. "Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes." Procedia Computer Science 83 (2016): 1013-1018.

[8] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, ``An insight into classi_cation with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf. Sci., vol. 250, pp. 113_141, Nov. 2013.

[9] Kim, Sangwook, Zhibin Yu, Rhee Man Kil, and Minho Lee. "Deep learning of support vector machines with class probability output networks." Neural Networks 64 (2015): 19-28.

[10] Moreira, Mario WL, Joel JPC Rodrigues, Guilherme AB Marcondes, Augusto J. Venancio Neto, Neeraj Kumar, and Isabel de la Torre Diez. "A Preterm Birth Risk Prediction System for Mobile Health Applications Based on the Support Vector Machine Algorithm." In 2018 IEEE International Conference on Communications (ICC), pp. 1-5. IEEE, 2018.

[11] Sulistiyanti, A., Farida, S., & Widodo, S. "Decision Support System To Monitoring Maternity Process Using Support Vector Machine Method". International Journal of Research in Engineering and Science, 6(8), (2018). 45-49.

[12] Lunghi, F., G. Magenes, L. Pedrinazzi, and MARIA GABRIELLA Signorini. "Detection of fetal distress though a support vector machine based on fetal heart rate parameters." In Computers in Cardiology, 2005, pp. 247-250. IEEE, 2005..

[13] Prema, N. S., and M. P. Pushpalatha. "Machine learning approach for Preterm Birth Prediction Based on Maternal Chronic Conditions." In Emerging Research in Electronics, Computer Science and Technology, pp. 581-588. Springer, Singapore, 2019.

[14] Sahin, Hakan, and Abdulhamit Subasi. "Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques." Applied Soft Computing 33 (2015): 231-238.

[15] Gilchrist, Jeff, Colleen M. Ennett, Monique Frize, and Erika Bariciak. "Neonatal mortality prediction using real-time medical measurements." In 2011 IEEE International Symposium on Medical Measurements and Applications, pp. 65-70. IEEE, 2011.

[16] Lakshmi, B. N., T. S. Indumathi, and Nandini Ravi. "A Study on C. 5 decision tree classification algorithm for risk predictions during pregnancy." Procedia Technology 24 (2016): 1542-1549..

[17] Hill, Jacquelyn L., M. Karen Campbell, Guang Yong Zou, John RG Challis, Gregor Reid, Hiroshi Chisaka, and Alan D. Bocking. "Prediction of preterm birth in symptomatic women using decision tree modeling for biomarkers." American journal of obstetrics and gynecology 198, no. 4 (2008): 468-e1.

[18] Mehta, Rutvij, Nikita Bhatt, and Amit Ganatra. "A survey on data mining technologies for decision support system of maternal care domain." International Journal of Computers and Applications 138, no. 10 (2016): 20-4.

[19] Ferreira, Duarte, Abílio Oliveira, and Alberto Freitas. "Applying data mining techniques to improve diagnosis in neonatal jaundice." BMC medical informatics and decision making 12, no. 1 (2012): 143.

[20] Akbulut, Akhan, Egemen Ertugrul, and Varol Topcu. "Fetal health status prediction based on maternal clinical history using machine learning techniques." Computer methods and programs in biomedicine 163 (2018): 87-100.

[21] Azar, Ahmad Taher, H. Hannah Inbarani, S. Udhaya Kumar, and Hala Shawky Own. "Hybrid system based on bijective soft and neural network for Egyptian neonatal jaundice diagnosis." International Journal of Intelligent Engineering Informatics 4(1), (2016): 71-90.

[22] Moreira, Mário WL, Joel JPC Rodrigues, Neeraj Kumar, Jianwei Niu, and Arun Kumar Sangaiah. "Multilayer Perceptron Application for Diabetes Mellitus Prediction in Pregnancy Care." In International Conference on Frontier Computing, pp. 200-209. Springer, Singapore, 2017.

[23] Moreira, Mario WL, Joel JPC Rodrigues, Guilherme AB Marcondes, Augusto J. Venancio Neto, Neeraj Kumar, and Isabel de la Torre Diez. "A Preterm Birth Risk Prediction System for Mobile Health Applications Based on the Support Vector Machine Algorithm." In 2018 IEEE International Conference on Communications (ICC), pp. 1-5. IEEE, 2018.

[24] Kuppermann, Miriam, Anjali J. Kaimal, Cinthia Blat, Juan Gonzalez, Mari-Paule Thiet, Yamilee Bermingham, Anna L. Altshuler, Allison S. Bryant, Peter Bacchetti, and William A. Grobman. "Effect of a Patient-Centered Decision Support Tool on Rates of Trial of Labor After Previous Cesarean Delivery: The PROCEED Randomized Clinical Trial." Jama 323, no. 21 (2020): 2151-2159.

[25] Vinks, Alexander A., Nieko C. Punt, Frank Menke, Eric Kirkendall, Dawn Butler, Thomas J. Duggan, DonnaMaria E. Cortezzo et al. "Electronic Health Record–Embedded Decision Support Platform for Morphine Precision Dosing in Neonates." Clinical Pharmacology & Therapeutics 107, no. 1 (2020): 186-194.

[26] López-Martínez, Fernando, Edward Rolando Núñez-Valdez, Vicente García-Díaz, and Zoran Bursac. "A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management." Algorithms 13, no. 4 (2020): 102.

[27] Løhre, Erik Torbjørn, Morten Thronæs, Cinzia Brunelli, Stein Kaasa, and Pål Klepstad. "An in-hospital clinical care pathway with integrated decision support for cancer pain management reduced pain intensity and needs for hospital stay." Supportive Care in Cancer 28, no. 2 (2020): 671-682.

[28] Pick, R. A.: Benefits of decision support systems." Handbook on Decision Support Systems 1. Springer, Berlin, Heidelberg, 719-730. (2008).

[29] Ekong, V. , Inyang, U. G., and Onibere, E. A.: Intelligent decision support system for depression diagnosis based on neuro-fuzzy-CBR hybrid." Modern Applied Science 6.7. (2012)

[30] Venkatesh KK, Strauss RA, Grotegut CA, Heine RP, Chescheir NC, Stringer JS, Stamilio DM, Menard KM, Jelovsek JE. Machine Learning and Statistical Models to Predict Postpartum Hemorrhage. Obstetrics & Gynecology. 2020 Apr 1;135(4):935-44.

[31] Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. JAMA network open. 2020 Jan 3;3(1):e1919396-.

[32] Bahado-Singh RO, Vishweswaraiah S, Aydas B, Yilmaz A, Saiyed NM, Mishra NK, Guda C, Radhakrishna U. Precision cardiovascular medicine: artificial intelligence and epigenetics for the pathogenesis and prediction of coarctation in neonates. The Journal of Maternal-Fetal & Neonatal Medicine. 2020 Feb 4:1-8

[33] García, Vicente, José Salvador Sánchez, and Ramón Alberto Mollineda. "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance." Knowledge-Based Systems 25(1) (2012): 13-21.

[34] Burnaev, Evgeny, Pavel Erofeev, and Artem Papanov. "Influence of resampling on accuracy of imbalanced classification." In Eighth International Conference on Machine Vision (ICMV 2015), vol. 9875, p. 987521. International Society for Optics and Photonics, 2015.

[35] Beckmann, Marcelo, Nelson FF Ebecken, and Beatriz SL Pires de Lima. "A KNN undersampling approach for data balancing." Journal of Intelligent Learning Systems and Applications 7(4), (2015): 104.

[36] [Fahrudin, Tora, Joko Lianto Buliali, and Chastine Fatichah. "Enhancing the Performance of SMOTE Algorithm by Using Attribute Weighting Scheme and New Selective Sampling Method for Imbalanced Data Set." Int. J. Innov. Comput. Inf. Control 15(2) (2018).

[37] Kaur, Prabhjot, and Anjana Gosain. "Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise." In ICT Based Innovations, pp. 23-30. Springer, Singapore, 2018.

[38] Q. Gu, Z. Cai, L. Zhu, and B. Huang, "Data mining on imbalanced data sets," in International Conference on Advanced Computer Theory and Engineering (ICACTE '08), 2008, pp. 1020-1024.

[39] Jeatrakul, Piyasak, and Kok Wai Wong. "Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm." In The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2012.

[40] Liu, Yangguang, Yangming Zhou, Shiting Wen, and Chaogang Tang. "A strategy on selecting performance metrics for classifier evaluation." International Journal of Mobile Computing and Multimedia Communications (IJMCMC) 6, no. 4 (2014): 20-35.

[41] Inyang, Udoinyang G., and Oluwole Charles Akinyokun. "A hybrid knowledge discovery system for oil spillage risks pattern classification." Artificial intelligence Research 3(4), (2014): 77-86.

[42] Akinyokun, Oluwole Charles, and Udoinyang G. Inyang. "Experimental study of neuro-fuzzy-genetic framework for oil spillage risk management." Artif. Intell. Research 2(4), (2013): 13-36.

[43] Bin Othman, Mohd Fauzi, and Thomas Moh Shan Yau. "Comparison of different classification techniques using WEKA for breast cancer." In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, pp. 520-523. Springer, Berlin, Heidelberg, 2007.

[44] Busa-Fekete, Róbert, Balázs Szörényi, Krzysztof Dembczynski, and Eyke Hüllermeier. "Online F-measure optimization." In Advances in Neural Information Processing Systems, pp. 595-603. 2015.

[45] Kim, Hae-Young. "Statistical notes for clinical researchers: post-hoc multiple comparisons." Restorative dentistry & endodontics 40, no. 2 (2015): 172-176.