

A novel approach for imbalanced instance handling toward better preterm birth classification

Himani Deshpande¹, Leena Ragha²

¹Department of Artificial Intelligence and Data Science, Thadomal Shahani Engineering College, Mumbai, India

²Department of Computer Science, BLDEA's V.P. Dr. P.G. Halakatti College of Engineering and Technology, Vijayapur, India

Article Info

Article history:

Received Apr 17, 2024

Revised Jul 14, 2024

Accepted Aug 6, 2024

Keywords:

Class imbalance

Clustering

Majority class

Minority class

Oversampling

Penalty

ABSTRACT

Preterm birth (PTB) is a major cause of child and mother mortality, a PTB classification model can assist in assessing the health condition ahead of time and help avoid complications during childbirth. Mother's significant feature (MSF) dataset created for this study has features derived from mother's physical, lifestyle, social and stress attributes. MSF dataset consists of 119 features of 1,000 mothers with 172 preterm and 828 full-term deliveries, resulting in issues of dataset imbalance namely class inseparability and classification bias. To overcome the imbalance issue, a novel algorithm named majority penalizing minority upsampling (MPMU) is proposed. MPMU forms clusters looking into the degree of dataset imbalance, it analyses the composition of each cluster individually and computes the varied penalty for majority class instances. It further balances dataset composition by oversampling minority class instances. MPMU processed dataset is further used to train the proposed 6L-ANN network which finds the probability of occurrence of PTB. The proposed model has shown efficient results on MSF sub-datasets with precision values ranging from 0.90 to 0.97, area under the curve (AUC) between 0.86 to 0.99, and prediction accuracy ranging from 93.04% to 99.47%. Experiment results show that a mother's lifestyle and stress features have a strong influence on the childbirth outcome.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Himani Deshpande

Department of Artificial Intelligence and Data Science, Thadomal Shahani Engineering College

Bandra (w), Mumbai, Maharashtra, India

Email: himani.deshpande@thadomal.org

1. INTRODUCTION

Maternal health and childbirth experience have always been a priority concern across the globe. There are many complications that mother and child can go through during childbirth, one such complication is when the childbirth takes place before 37 weeks of pregnancy which is termed preterm birth (PTB). PTB is a leading cause of underdeveloped baby organs, maternal mortality and child mortality [1]. In today's era with rapid societal and lifestyle changes, the PTB cases are increasing rapidly. A PTB classification model can help to assess the health condition of a woman ahead of time to prevent any future health complications. Pregnancy outcomes are closely associated with the health of the mother [2]. A large number of health issues that women face is associated with their lifestyle, relationship with family, social surroundings, stress level, and physical health. All these factors influence their reproductive health as well [3] and hence information associated with the mother's mental and physical health can be used for the prognosis of PTB. Exploring machine learning methods and tools on the mother's health dataset can be of great help towards finding useful patterns and facts for a better understanding of PTB. Considering the size of the population, medical

domain datasets are found to be highly imbalanced with fewer instances of diseased patients [4]. Lack of density in imbalanced datasets frequently results in learning models failing to discover unusual patterns and having classification bias towards the majority class, thus resulting in minority class misclassification [5], [6]. This makes imbalanced data handling a high-priority issue that needs to be addressed for efficient classification of health conditions.

Observing the recent work in the maternal, childbirth domain and specifically in the PTB domain [7], firstly it was observed that majorly the researchers and experts are from the medical domain and they have focused on PTB classification models by using existing statistical methodologies and have not focused on dataset enhancement techniques and modern heuristic classification methods. Secondly, there is no standard dataset and researchers are relying on small local datasets with majority data consisting of obstetric features and test results and have not worked on the influencing factors that affect these test results. To fill this gap a dataset titled mother's significant feature (MSF) is created for this study, through rigorous interaction with gynecologists with the aim to derive the influencing features depicting mother's physical health, lifestyle, social surroundings and stress.

Looking into the imbalanced dataset handling methods across different domains, it is observed that researchers have come up with different techniques which are broadly divided into three categories namely pre-processing, cost-sensitive learning and ensemble learning methods [8], [9]. Pre-processing methods work on strengthening the minority and suppressing the majority samples before classification. The dataset can be pre-processed based on instances or features. Instance-based methods either over-sample minority instances or under-sample majority instances or both [10]–[12]. Many researchers have used cluster-based fuzzy weighing [13] and approaches where resampling of clusters is proposed [14], [15], these researchers have focused on changing the size of clusters by over-sampling and under-sampling and have not worked on understanding the impact of instances falling under different clusters towards class separability. Feature selection methods pre-process datasets by observing features, these methods are broadly divided into three categories, namely filter methods, wrapper methods and embedded methods [16], [17]. These methods look into the strength of a feature individually or in combination with other features towards efficient classification. The issue with pre-processing methods is that it is difficult to decide on the change in the distribution of the database to suit the given problem [8].

Cost-sensitive methods are used to solve categorization problems when the costs of incorrect classification differ depending on the instances and features. These techniques panelize to reduce losses faced due to misclassification [18]. As the cost is assessed considering the imbalance in the dataset, these methods are found to be efficient in handling imbalanced datasets. Cost can be incorporated into the classifier's training process as a cost matrix or threshold. To create a cost-sensitive classifier, researchers have followed methodologies like modifying the goal function using a weighing strategy [19]–[21], changing the decision thresholds, the learning process [22], the encoding network [23], and one-class approach [24], [25]. Although these techniques are simple to use, overfitting problems could arise [8]. The intricacy of cost calculation may be one of the main reasons why cost-sensitive approaches are less popular than resampling methods, even though these methods are effective [26]. When a single classifier fails to work efficiently on an imbalanced dataset, multiple classifiers are considered while deciding the final outcome [27]. These types of methods are called ensemble methods. Bagging and Boosting are two different kinds of ensemble approaches [28]. The majority of ensemble models have taken cost-sensitive and resampling techniques into account [26]. Literature suggests that dataset imbalance can be handled by processing datasets or customizing prediction algorithms, the former being a more common choice among researchers. This research aims to overcome issues in the maternal domain by creating a dataset depicting mother's physical, lifestyle, social and stress-related features. Further, a novel imbalance handling method is proposed to overcome the bias and inseparability issues found in the created dataset.

2. METHOD

Researches in the PTB domain are based on limited observations from medical records and have not focused on the lifestyle, stress and social aspects of a woman's life that may also contribute to her maternal health. For the same reason, this research delves into the examination of both the mental and physical health of mothers, leading to the creation of a corresponding dataset. Observing the current population, it is realized that the dataset majorly consists of full-term birth (FTB) cases and few PTB cases, thus resulting in imbalanced dataset formation. Towards this, a novel clustering-based algorithm named majority penalizing minority upscaling (MPMU) is proposed to handle the data imbalance by focusing on minimizing the impact of majority dominance that may influence the classification results. Further, a neural network named 6L-ANN is proposed that classifies the instance outcome being preterm or full-term. Figure 1 shows the architecture of the proposed model.

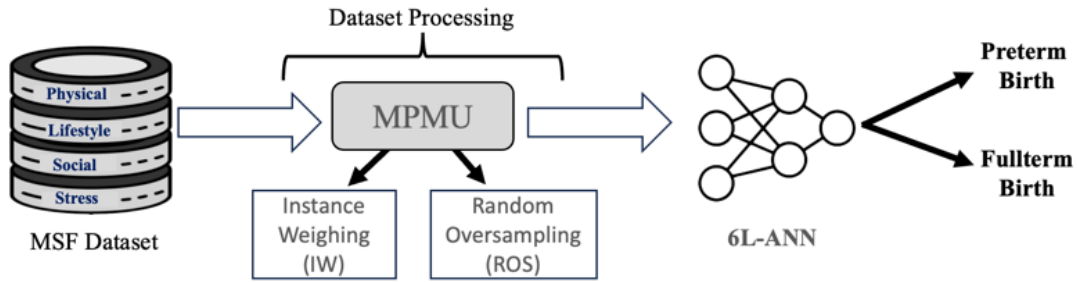


Figure 1. Architecture of proposed model

2.1. Mother's significant feature dataset

With the aim to work towards the betterment of mother and child, a dataset titled MSF dataset is created. MSF dataset gives an insight into the mother's mental and physical health conditions across three phases (teenage, after marriage, during pregnancy) of her reproductive age. The majority of the women who delivered babies between February 2018 and September 2019 at D. Y. Patil Hospital, located in the Mumbai Metropolitan Region of India, were interviewed. MSF dataset consists of 119 features of 1,000 women, which are further divided into four categories namely physical (22 features), social (19 features), lifestyle (68 features), and stress (10 features). Out of all the records 172 women had Preterm delivery while 828 had full-term delivery. Details about the planning, challenges and procedures followed to create the MSF dataset are mentioned in our previous work [29]. MSF dataset is partially available online [30]. Our previously proposed methods on the MSF dataset include models like Variation and Information based random forest (VIBRF) [31], [32] which focuses on relevant feature selection and random forest based fuzzy feature weighing for class imbalance (RFFWCI) [33] which weighs features for imbalance handling. This study, proposes MPMU method which realizes the contribution of instances of the dataset towards imbalance issue. To validate the performance of the proposed MPMU method, seven standard imbalanced datasets from the Keel repository [34] are used for experimentation.

2.2. Proposed imbalance handling method

As discussed in the literature under section 2, existing cluster-based approaches have relied on resampling and have looked upon the clusters individually and not in relation to the dataset as a whole. The proposed imbalance handling method MPMU provides a novel complete solution where clusters are analyzed individually as well as in relation to the overall dataset and instances are weighted by investigating the cluster, they fall in. MPMU consists of two major components namely instance weighing (IW) and random over sampling (ROS), which are explained further in this section.

2.2.1. Majority penalizing minority upscaling

The proposed MPMU algorithm aims at understanding the mixing up of majority class instances with minority class instances. MPMU algorithm works on majority instances with an IW method and on minority instances by randomly upscaling them. MPMU algorithm uses k-means clustering to group similar instances together, considering intra-cluster variance as expressed in (1) for k clusters and n instances with c_j as the centroid of the cluster.

$$j = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad (1)$$

Weights of majority instances are decided based on the composition of the cluster they fall in, thus the number of clusters M to be formed reflects the proportion of majority class instances in the whole dataset i.e. probability of occurrence of majority class as shown in (3) of Table 1, where $P(Min)$ is the probability of occurrence of minority instance and $P(Maj)$ is the probability of occurrence of majority instance in the dataset. M is directly proportional to the dataset imbalance, with more imbalance in the dataset greater number of clusters will be formed, same can be seen in Table 2. A perfectly balanced dataset will have two clusters with very minimal mix. Figure 2 shows the flowchart for the proposed architecture using MSF dataset, MPMU and 6L-ANN. Input to the model is the MSF dataset and output is a fuzzy value depicting the probability of occurrence of PTB and full-term birth. In Figure 2, the blue color boxes show the steps involved for majority instance handling i.e. IW, the green color box depicts the work done for the minority class and the grey color boxes denote the processing of the dataset as a whole. MPMU algorithm is mentioned in Table 1.

Table 1. MPMU algorithm

Input: Dataset Y with 'X' instances and 'F' features	
Output: Processed majority and minority instances	
1.	Calculate dataset imbalance 'DI'
	$DI = \frac{A}{B}$ (2)
2.	Calculate the number of clusters 'M', to be formed.
3.	Split dataset Y into 'M' clusters.
	$M = P(\text{Min U Maj}) - P(\text{Min})$ (3)
4.	Create vector V_CMC[] of size X.
5.	for N=1 to M (each of the M clusters) calculate within Cluster Imbalance 'CI _N '
	$CI_N = \frac{C_{NA}}{C_{NB}}$ (4)
6.	for i=1 to X (each of the X instances) <ul style="list-style-type: none"> • identify cluster 'N' within which instance 'i' falls. • identify CI_N for cluster 'N' <ul style="list-style-type: none"> if (CI_N < β or CI_N < DI) CMC_N=0 else CMC_N= CI_N-DI
	(5)
	• V_CMC[i]=CMC _N
7.	for i=1 to X (across X instances) for j=1 to F (across F features) if (Minority instance) Y[i][j]=Y[i][j]
	else Y[i][j]=Y[i][j]-(α*V_CMC[i])
	(6)
8.	Randomly oversample minority instances such that A=B

Table 2. Datasets analysis

	Dataset	No. of features	No. of instances	Dataset imbalance	No. of clusters
1	Physical	22	1,000	4.82	7
2	Lifestyle	68	1,000	4.82	7
3	Stress	10	1,000	4.82	7
4	Social	19	633	5.95	7
5	Physical+Lifestyle+Stress	100	1,000	4.82	7
6	Pima India	7	768	1.86	4
7	vehicle2	13	988	2.88	6
8	segment0	10	5,472	6.02	7
9	ecoli3	7	336	8.6	8
10	Page-blocks0	8	514	8.79	8
11	Yeast2_vs_4	8	768	9.08	8
12	vowel0	19	2,308	9.98	8

2.2.2. Majority instance weighing

A major contribution of the MPMU algorithm is to penalize the majority class instances through the IW method. Formed clusters are analyzed to identify the majority class instances which are close to minority class instances, thus difficult to identify and causing misclassification. The IW method subdues the identified majority class instances by penalizing them. The penalty for each majority class instance is decided by looking into the composition of the cluster it falls in. The penalty is not the same for all the majority class instances as the aim is to subside only those majority class instances which are inseparable from minority class instances. Cluster imbalance is calculated for each of the 'M' clusters. Cluster imbalance CI_N of the N^{th} cluster is calculated as shown in (4) in Table 1, where C_{NA} is the number of minority class instances in the N^{th} cluster and C_{NB} is the number of majority class instances in the N^{th} cluster respectively. dataset imbalance (DI) is calculated using (2) in Table 1, where A is the number of minority class instances and B is the number of majority class instances in the dataset. The imbalance of the dataset as a whole plays a significant role in deciding the penalty for the majority class instances. If the imbalance within the cluster is greater than the imbalance of the dataset, it is considered to be defying the dataset composition and thus, a penalty will be applied to the majority instances of such clusters. The proposed algorithm penalizes the majority class instances only if the 'CI' value of the cluster into which the instance is falling, is greater than

the value of ‘ DI ’. Cluster misclassification cost (CMC) for each of the ‘ M ’ clusters are different and is decided based on the variation between the cluster imbalance and dataset imbalance. CMC_N value for the N^{th} cluster is calculated using (5) in Table 1. If the cluster imbalance is very low i.e. below a certain threshold i.e. ‘ β ’, it shows that the cluster constitutes mainly of the majority class and the presence of the minority class can be due to noise or outliers. In this case, majority class instances will not be penalized. For a dataset with ‘ X ’ instances, vector $V_CMC[]$ of size X , is used to store the CMC values corresponding to each of the dataset instances looking into the cluster it falls into. The CMC value of N^{th} cluster is calculated using (6) of Table 1, considering the imbalance of the cluster and dataset. As shown in (6), while penalising the j^{th} feature of the i^{th} instance, α value is used to normalize the difference between the value of features and CMC value. For the MSF dataset, optimal value of α and β is found to be 0.7 and 0.3 after empirical analysis.

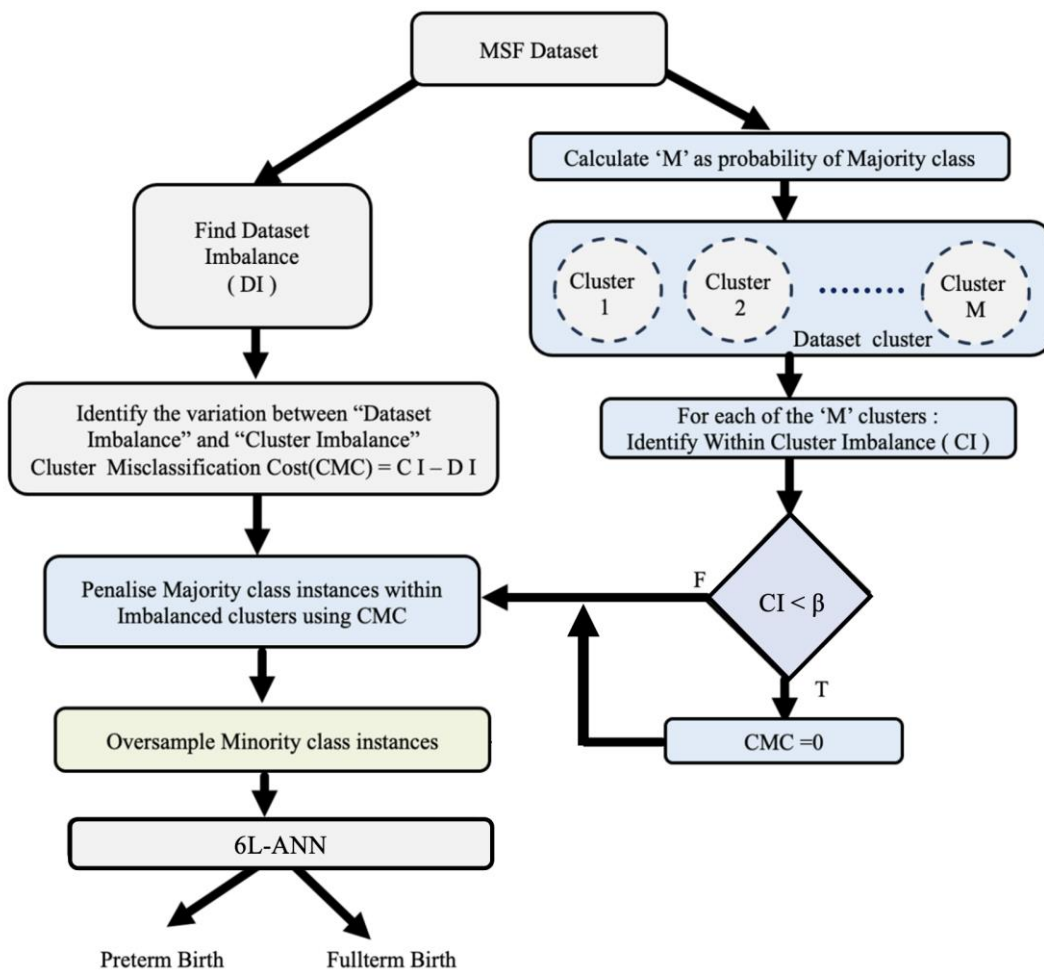


Figure 2. Flowchart of proposed imbalanced instance handling method

2.2.3. Cluster formation

Figure 3, shows an example for the formation of MPMU clusters and how the positive class instances are penalized. In Figure 3, red circles denote minority class instances and the black + (plus) sign denotes majority instances. In Figure 3, cluster 1 and cluster 4 have instances only from one class, thus there is no penalty on these instances. Cluster 2 has a very low imbalance i.e. below threshold β , thus no penalty on majority class instances. Clusters 3 and 5 have more imbalance as compared to the overall dataset, the majority class instances of these clusters are penalized by looking at the composition of the cluster and calculating the CMC value as expressed in (6). Different shades of grey for majority class instances (+) in Figure 3, depicts the different intensity of the penalty, lighter shade means a higher penalty. MPMU algorithm does not penalize minority instances.

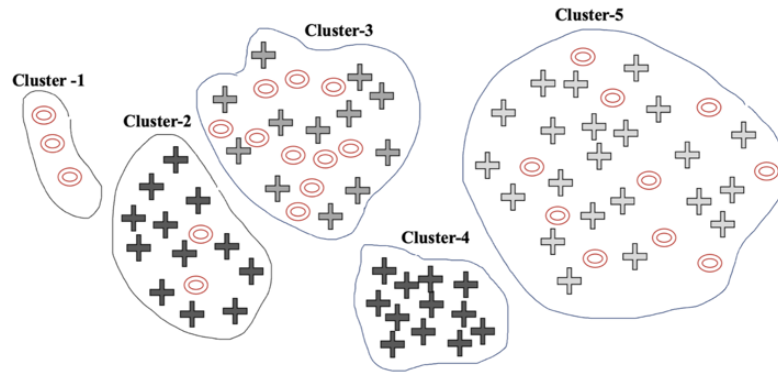


Figure 3. Cluster formation

2.2.4. Minority instance handling

Even after penalizing the majority class instances, the composition of the dataset remains inclined towards the majority class. To balance the composition of the dataset, the MPMU algorithm oversamples the minority class instances in the dataset, such that the number of minority class instances gets equal to the number of majority class instances as shown in step 8 of Table 1. For the MSF dataset, samples are collected by interviewing women from similar geographic and demographic locations. Keeping the same in mind, for over-sampling, no synthetic instances are generated instead it is done by random duplication of minority samples, thus even after over-sampling the instance, the feature value remains true to the original population of the MSF dataset.

2.3. 6L_ANN

Variations across different hyperparameters of artificial neural network (ANN) models using multi-layer perceptron (MLP) with back propagation are used in the experiments. Based on the empirical analysis, 6L_ANN, a six-layered ANN architecture is selected for classification which consists of two drop-out layers and two activation layers along with an output and an input layer. Rectified linear activation unit (ReLU) activation function is used for hidden layers and the output layer uses the sigmoid activation function. MSF dataset processed using the proposed MPMU method is further used to train the 6L_ANN network to classify PTB and FTB outcomes.

3. RESULTS AND DISCUSSION

This section of the paper presents the classification results using the proposed MPMU method and 6L-ANN network. Multiple experiments are performed to understand the nature of the proposed methods with different classifiers and datasets. To check the imbalance handling efficiency of the MPMU algorithm, along with the proposed 6L-ANN, five popular classification algorithms, namely, random forest (RF), decision tree (DT), gaussian NB (GNB), k-nearest neighbors (KNN) ($neighbor = 3$), and MLP are used for experimentation. Researchers suggest that while working on an imbalanced dataset area under the curve (AUC) curve, precision value and prediction accuracy together need to be considered to evaluate the performance of the classification task [35], [36]. Prediction accuracy evaluates the overall performance of a classifier, AUC value assesses the capability to distinguish between the classes, and precision measures the PTB class. Thus, this study considers precision, AUC and prediction accuracy for evaluating the performance of the proposed method.

3.1. Datasets analysis

For experimentation, along with MSF dataset, 7 standard datasets are used. Table 2, shows the dataset details in terms of dataset size and imbalance. The first five rows of Table 2, shows MSF datasets information and the remaining 7 rows have information about the other standard datasets used. There are missing records under social features of the MSF dataset, for the same reason only 633 instances of it are considered during experimentation. As shown in Table 2, standard datasets are selected with variations in terms of imbalance ratio, number of features and number of instances so that the experimental results validate the versatility of the proposed MPMU method. The last column of Table 2 shows the count of clusters formed as expressed as 'M' in (3), of the MPMU algorithm. For experimentation, all the datasets are preprocessed for handling missing values and have been normalized, such that for each record, the sum of the

square of normalized values is equal to one. A total of 70% of the whole dataset is taken for training while 30% is considered for testing.

3.2. Experimental results on MSF dataset

Multiple experiments are performed on the MSF dataset to understand the effectiveness of penalizing (IW) and oversampling steps of the MPMU algorithm. Figure 4 shows the classification results on five different MSF sub-datasets as represented in the first five rows of Table 2. Experiments are conducted with seven combinations of data processing methods and classifiers. Figure 4 shows the results of instance weight (IW) and MPMU method in combination with MLP, GNB, KNN, RF, DT and the proposed 6L-ANN classifier. In Figure 4, the MPMU model is evaluated with 6L-ANN classifier. There were many missing values in the social sub-dataset for the same reason it was not included while conducting combined MSF dataset experimentation. Figures 4(a) to 4(c) show the precision value, AUC and prediction accuracy using the proposed IW step and MPMU method on MSF sub-datasets.

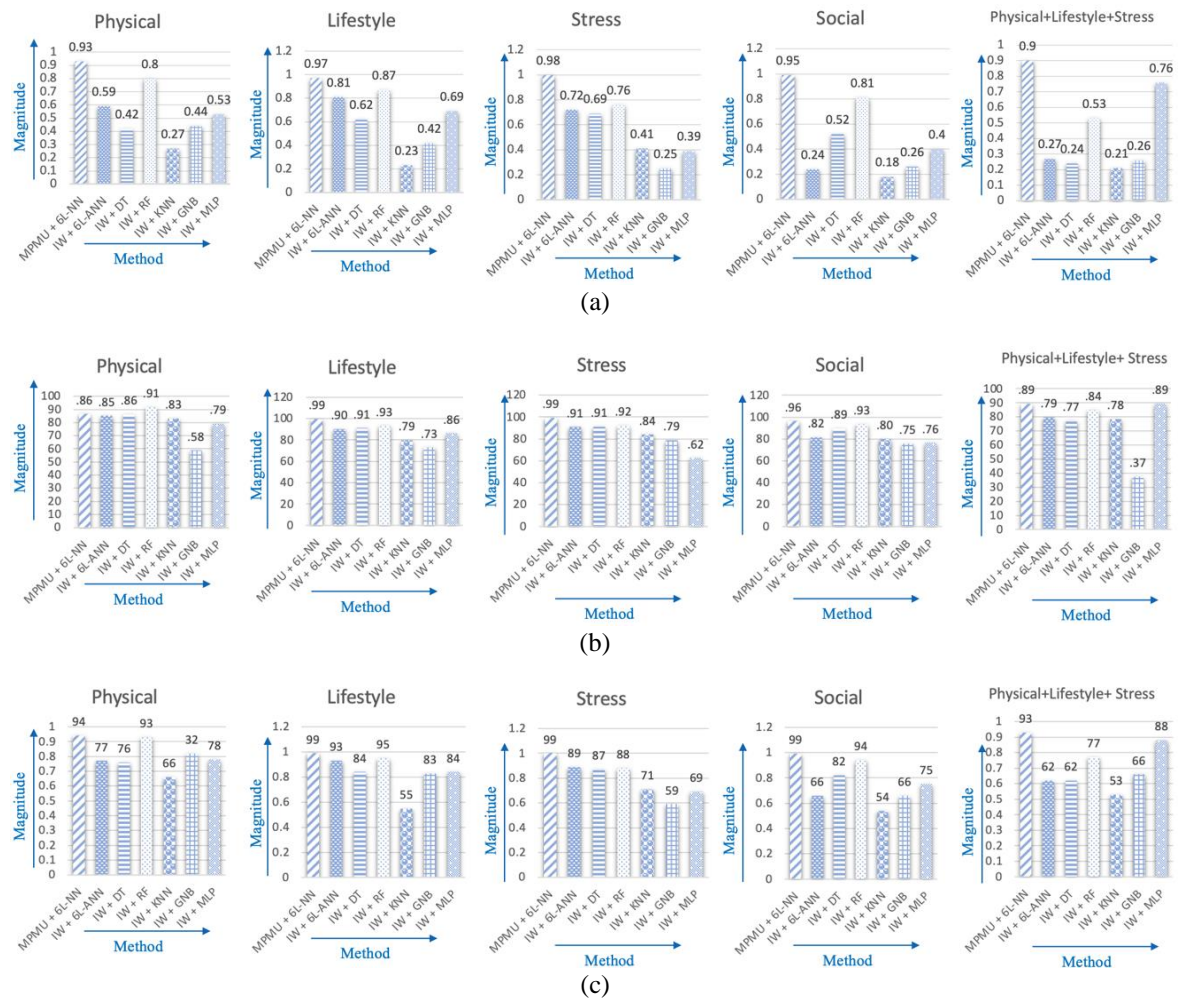


Figure 4. Classification results on MSF Sub-datasets (physical, lifestyle, stress, social and combined MSF features (physical, lifestyle and stress features)) using the proposed IW step and proposed MPMU method with classifiers namely MLP, GNB, KNN, RF, DT and proposed 6L-ANN (a) demonstrate classification results in terms of precision, (b) prediction results in terms of AUC value, and (c) prediction accuracy of classification

MSF dataset is an imbalanced dataset and results reflect classification bias, same can be justified by the results in Figure 4. While pre-processing the datasets using IW, despite higher prediction accuracy and a decent AUC value, the precision value is very low for classifiers like GNB, KNN and DT for all MSF datasets, this shows that the results are inclined toward the minority class. Considering the prediction of PTB

as a priority concern over the prediction of FTB, thus counting precision as the most crucial evaluation parameter, it could be concluded that RF is giving the best results for all the IW pre-processed MSF datasets, and the second-best results are achieved by 6L-ANN network for all the MSF sub-datasets. It is further observed that the IW step alone is not as effective as the MPMU method which combines majority instance weighing and minority oversampling, as there is a significant improvement in results for all MSF sub-datasets while pre-processing using the proposed MPMU method and further using 6L-ANN network as the classifier.

Experimental results in Figure 4, validate the strength of the proposed model i.e. MPMU with 6L-ANN classifier, precision between 0.90 to 0.97 is achieved on MST sub-datasets, an AUC value ranging between 0.86 to 0.99 is observed, and prediction accuracy of 93.00% to 99.47% is achieved. Looking into the results in Figure 4, it is observed that mother's lifestyle and stress features are achieving better classification results for PTB prediction. Third-best classification results are achieved using social features and physical features classification results are found to be the least accurate ones. One of the probable reasons for inferior results for social features would be a smaller dataset and higher-class imbalance. This signifies that the mother's mental health features are closely related to the pregnancy outcome in terms of the classification of PTB and FTB. Looking into the results of the proposed MPMU method in Table 3, with standard classifiers and with the results in Figure 4 with 6L-ANN classifier shows that among all the classifiers MPMU is giving the best results using 6L-ANN classifier. The experimental results on different sub-datasets of MPMU as shown in Figure 4, support the hypothesis that a mother's mental health plays a crucial role in predicting the PTB birth outcome. As seen in Figure 4, most of the classifiers are giving better predictions using the lifestyle and stress feature sub-dataset. Using lifestyle features precision is achieved in the range of 0.23 to 0.97, AUC value is achieved between 0.72 to 0.99 and prediction accuracy is in the range of 55% to 99%. Using stress features precision is achieved in the range 0.25 to 0.98, AUC value is achieved between 0.62 to 0.99 and prediction accuracy in the range 59% to 100%.

To understand the performance of MPMU method over existing data imbalance solutions, experiments are conducted using algorithms proposed by researchers in the past. Table 3 demonstrates the results for the same. The data processing methods used for implementation are scaled values (SV), clustered SMOTE (CS) [37], random over sampling (ROS) [38] and the proposed MPMU algorithm.

Being imbalanced MSF dataset is found to be performing extremely poorly in terms of precision and AUC value, even after scaling the values using SV method, classifiers fail to perform the same can be seen in the results of Table 3. The cluster-based instance weighing and oversampling approach of MPMU is capable of handling majority bias and inseparability issues of MSF datasets the same can be seen in the results of Table 3. There is a significant improvement in the results of MPMU method with all the used classifiers when compared to SV, CS, and ROS imbalance handling methods. On average precision value is improved by 48%, AUC by 26% and prediction accuracy by 16%, when using MPMU against SV. Precision value is improved by 52%, AUC is improved by 32% and prediction accuracy is improved by 26%, comparing the results of MPMU against CS. Against ROS method MPMU results are improved by 54%, 31% and 16% for precision, AUC and prediction accuracy. Greater improvement in precision values using MPMU validates the success of MPMU in identifying PTB instances correctly. Among the used standard classifiers RF is showing the best results for all the selected imbalance handling methods on MSF sub-datasets.

Table 3. Comparative results of different imbalance handling methods on MST sub-datasets

		SV	CS	ROS	MPMU	SV	CS	ROS	MPMU	SV	CS	ROS	MPMU	SV	CS	ROS	MPMU	SV	CS	ROS	MPMU
Physical	Precision	0.18	0.19	0.19	0.42	0.22	0.21	0.2	0.8	0.2	0.2	0.2	0.27	0.29	0.29	0.21	0.44	0.23	0.23	0.23	0.53
	AUC	0.53	0.47	0.55	0.76	0.6	0.45	0.59	0.93	0.52	0.51	0.52	0.66	0.64	0.52	0.49	0.82	0.52	0.5	0.52	0.78
	Accuracy	71.7	74.5	72.8	86.2	83.9	80.9	82.8	91.6	77.6	65.1	77	83.2	18.4	18.3	31.3	58.7	66.5	66.5	66.5	79.2
Lifestyle	Precision	0.17	0.19	0.17	0.62	0.21	0.2	0.21	0.87	0.17	0.17	0.17	0.23	0.16	0.16	0.15	0.42	0.21	0.21	0.21	0.69
	AUC	0.49	0.51	0.5	0.84	0.54	0.51	0.54	0.95	0.49	0.51	0.49	0.55	0.5	0.51	0.49	0.83	0.52	0.48	0.52	0.84
	Accuracy	69.3	69.3	71.2	91	82.2	75.3	82.4	93.2	79.2	63.7	79	79.8	16.8	16.8	26.5	72.8	70	70	70	86
Stress	Precision	0.2	0.2	0.19	0.69	0.21	0.21	0.21	0.76	0.2	0.19	0.2	0.41	0.23	0.23	0.2	0.25	0.24	0.24	0.24	0.39
	AUC	0.53	0.52	0.53	0.87	0.51	0.49	0.51	0.88	0.54	0.5	0.53	0.71	0.58	0.52	0.51	0.59	0.56	0.52	0.56	0.69
	Accuracy	81.1	58.6	80.3	91.1	79.5	58.8	79.1	92.1	77.3	62.1	74.7	83.8	80.4	59.9	68.7	78.7	59.5	59.5	59.5	62.5
Social	Precision	0.17	0.17	0.16	0.52	0.18	0.18	0.17	0.81	0.15	0.14	0.15	0.18	0.17	0.18	0.15	0.26	0.17	0.17	0.17	0.4
	AUC	0.54	0.48	0.52	0.82	0.51	0.48	0.51	0.94	0.53	0.53	0.53	0.54	0.51	0.48	0.43	0.66	0.51	0.54	0.51	0.75
	Accuracy	75.6	73.5	74.7	89.7	81.9	76.7	80.9	93.3	82.7	69	80.8	80.1	59.7	34.2	62.3	75.6	66.8	66.8	66.8	76.4
Lifestyle+	Precision	0.2	0.2	0.19	0.24	0.25	0.22	0.26	0.53	0.17	0.16	0.17	0.21	0.24	0.24	0.18	0.26	0.21	0.21	0.21	0.76
Physical+	AUC	0.53	0.5	0.53	0.62	0.58	0.52	0.59	0.77	0.5	0.48	0.5	0.53	0.58	0.47	0.46	0.66	0.53	0.49	0.53	0.88
Stress	Accuracy	72.2	72.2	70.3	77.3	81.8	80.1	81.7	84.6	79	62.9	78.4	78.6	24.7	25	52.6	37.6	72.3	72.3	72.3	89.5

3.3. Experimental results on standard datasets

For validation of the efficiency of the proposed MPMU algorithm, experiments are conducted on seven standard datasets using five different classifiers. These datasets are commonly used by researchers in the imbalanced handling literature. The number of clusters formed using MPMU, for each of the datasets are as shown in Table 2. It is observed that the MPMU algorithm is overshadowing other algorithms for most of the datasets and classifier combinations. Results in Table 4, depict that the proposed MPMU method works well with large size datasets like vowel0 and segment0. For smaller datasets like Pima India and ecoli3 the final results are not very good but there is a remarkable improvement in results when compared with SV, CS and ROS imbalance handling methods. While comparing the results of SV over the results of the MPMU algorithm, improvement in results has been observed between 0% to 21% for precision, between 0% to 10% for AUC and between 0% to 17% for prediction accuracy. The results in the table convey that along with the MSF dataset, the proposed methods work equally well for other imbalanced datasets with varied sizes and imbalances.

Table 4. Comparative results of different imbalance handling methods on standard imbalanced datasets

Datasets	Decision Tree				Random Forest				KNN				Gaussian NB				MLP				
	SV	CS	ROS	MPMU	SV	CS	ROS	MPMU	SV	CS	ROS	MPMU	SV	CS	ROS	MPMU	SV	CS	ROS	MPMU	
Precision	ecoli3	0.47	0.32	0.39	0.68	0.72	0.68	0.65	0.86	0.46	0.4	0.37	0.66	0.63	0.6	0.65	0.8	0.66	0.66	0.66	0.81
	Pima India	0.47	0.49	0.46	0.47	0.71	0.7	0.7	0.73	0.56	0.57	0.56	0.58	0.67	0.66	0.67	0.72	0.7	0.7	0.7	0.78
	yeast-2_vs_4	0.65	0.57	0.56	0.67	0.92	0.93	0.92	0.94	0.63	0.64	0.59	0.88	0.72	0.76	0.72	0.83	0.84	0.84	0.84	0.88
	page-blocks0	0.72	0.71	0.7	0.92	0.94	0.94	0.93	0.99	0.78	0.78	0.72	0.8	0.58	0.54	0.58	0.55	0.89	0.89	0.89	0.92
	segment0	0.95	0.94	0.95	0.95	1	1	1	1	0.99	0.98	0.97	1	0.92	0.8	0.92	0.99	1	1	1	1
	vowel0	0.85	0.79	0.82	0.9	0.99	1	0.99	1	1	1	0.99	1	0.88	0.58	0.87	0.92	1	1	1	1
	vehicle2	0.85	0.86	0.85	0.89	1	0.99	1	1	0.93	0.94	0.88	0.96	0.62	0.54	0.61	0.76	0.99	0.99	0.99	1
	ecoli3	0.79	0.71	0.49	0.89	0.95	0.94	0.53	0.95	0.87	0.86	0.5	0.87	0.9	0.9	0.5	0.91	0.95	0.95	0.51	0.98
	Pima India	0.67	0.68	0.5	0.68	0.82	0.83	0.5	0.85	0.74	0.74	0.49	0.77	0.81	0.8	0.51	0.84	0.83	0.83	0.49	0.87
	yeast-2_vs_4	0.88	0.86	0.52	0.92	1	1	0.51	0.99	0.86	0.86	0.51	0.95	0.9	0.88	0.5	0.9	0.94	0.94	0.43	0.96
AUC	page-blocks0	0.91	0.91	0.49	0.98	0.99	0.99	0.49	1	0.93	0.93	0.51	0.93	0.93	0.89	0.5	0.91	0.98	0.98	0.49	0.99
	segment0	0.98	0.98	0.5	0.99	1	1	0.5	1	0.99	0.99	0.5	1	0.98	0.96	0.48	1	1	1	0.51	1
	vowel0	0.97	0.95	0.51	0.97	1	1	0.5	1	1	1	0.51	1	0.98	0.9	0.47	0.99	1	1	0.49	1
	vehicle2	0.94	0.95	0.49	0.96	1	1	0.53	1	0.98	0.98	0.53	0.99	0.82	0.71	0.47	0.9	1	1	0.5	1
	ecoli3	91.1	87.3	89.94	95.88	94.1	92.9	91.8	95.88	91.8	89.8	88.8	93.24	75.1	75.7	73.4	81.69	89.8	89.8	89.8	91.59
	Pima India	69.5	70.6	68.55	69.84	75.3	76.1	75.5	78.21	71.1	71	69.2	74.46	74.4	73.3	72.8	77.42	74.8	74.8	74.8	79.44
	yeast-2_vs_4	95.5	94.1	94.09	95.91	95.9	96.1	95.9	97.53	95.3	95.2	94	97.74	31.3	49	27.7	33.98	93.1	93.1	93.1	91.61
	page-blocks0	96.6	96.6	96.48	99.11	97.6	97.6	97.6	99.41	96.1	96.1	95.3	96.02	89.7	90	89.1	89.31	94.8	94.8	94.8	94.81
	segment0	99.2	99.1	99.21	99.26	99.7	99.8	99.7	99.76	99.4	99.3	99.3	99.88	82.2	88.8	79.9	98.92	99.7	99.8	99.7	98.57
	vowel0	98.5	97.8	98.15	98.99	99.4	99.6	99.6	99.61	99.9	99.9	99.8	99.72	92.7	89.2	86.8	95.74	99.8	99.8	99.8	99.77

4. CONCLUSION

This work proposes a novel imbalance handling model towards the prediction of pre-term birth using the created MSF dataset, which consists of lifestyle, social, physical and stress features of 1,000 mothers. The proposed MPMU imbalance handling method works on class inseparability and classification bias issues by subduing the majority class instances following a cluster-based penalty approach and strengthening the minority instances by random oversampling. It has been observed that the proposed model consisting of MPMU and 6L-ANN classifier, is highly efficient in classifying PTB and FTB outcomes on MSF datasets with precision ranging from 0.90 to 0.97, AUC between 0.86 to 0.99, and prediction accuracy falling under 93% to 99.47%. MPMU imbalance handling method has been validated on other imbalanced datasets too. MPMU shows a superior performance when compared to other Imbalance handling methods on MSF as well as other datasets. Thus, this research work succeeds in proposing an efficient cluster-based imbalanced instance handling method. Experimental results on the mental and physical health-related features of the mother, show that the lifestyle of the mother and Stress-related features are major contributors towards analyzing her maternal health which defines her mental health. Thus, this study, concludes that a mother’s mental health is a major contributor towards understanding the pre-term birth possibilities.

REFERENCES




- [1] T. Strunk, A. Currie, P. Richmond, K. Simmer, and D. Burgner, “Innate immunity in human newborn infants: prematurity means more than immaturity,” *The Journal of Maternal-Fetal and Neonatal Medicine*, vol. 24, no. 1, pp. 25–31, Jan. 2011, doi: 10.3109/14767058.2010.482605.
- [2] M. A. Arfandi, A. Ansariadi, R. Amiruddin, W. Wahiduddin, A. U. Salmah, and A. Salam, “Preterm birth risk in mother with hypertensive disorders of pregnancy,” *International Journal of Public Health Science*, vol. 12, no. 2, pp. 590–597, 2023, doi: 10.11591/ijphs.v12i2.22599.
- [3] G. Homan, J. Litt, and R. J. Norman, “The FAST study: Fertility assessment and advice targeting lifestyle choices and behaviours: A pilot study,” *Human Reproduction*, vol. 27, no. 8, pp. 2396–2404, 2012, doi: 10.1093/humrep/des176.
- [4] S. S. Kaddi and M. M. Patil, “Ensemble learning based health care claim fraud detection in an imbalance data environment,”

- Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 3, pp. 1686–1694, 2023, doi: 10.11591/ijeecs.v32.i3.pp1686-1694.
- [5] G. Naik, D. Siroya, M. Nisar, B. Shah, and H. Deshpande, “Analysis on the efficacy of ANN on small imbalanced datasets,” *Springer Proceedings in Mathematics and Statistics*, vol. 403, pp. 129–137, 2023, doi: 10.1007/978-3-031-16178-0_11.
 - [6] M. H. B. M. Razali, R. Bin Saian, Y. B. Wah, and K. R. Ku-Mahamud, “A class skew-insensitive ACO-based decision tree algorithm for imbalanced data sets,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 412–419, 2021, doi: 10.11591/ijeecs.v21.i1.pp412-419.
 - [7] M. Son and E. S. Miller, “Predicting preterm birth: Cervical length and fetal fibronectin,” *Seminars in Perinatology*, vol. 41, no. 8, pp. 445–451, 2017, doi: 10.1053/j.semperi.2017.08.002.
 - [8] P. Branco, L. Torgo, and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” *ACM Computing Surveys*, vol. 49, no. 2, 2016, doi: 10.1145/2907070.
 - [9] F. Nieto-Del-amor *et al.*, “Combination of feature selection and resampling methods to predict preterm birth based on electrohysterographic signals from imbalance data,” *Sensors*, vol. 22, no. 14, 2022, doi: 10.3390/s22145098.
 - [10] M. H. Saeed and J. I. Hama, “Cardiac disease prediction using AI algorithms with SelectKBest,” *Medical and Biological Engineering and Computing*, vol. 61, no. 12, pp. 3397–3408, 2023, doi: 10.1007/s11517-023-02918-8.
 - [11] W. A. Kusuma, N. Noviana, L. S. Hasibuan, and M. Nurilmala, “Improving DNA barcode-based fish identification system on imbalanced data using SMOTE,” *Telkonnika (Telecommunication Computing Electronics and Control)*, vol. 15, no. 3, pp. 1230–1238, 2017, doi: 10.12928/TELKOMNIKA.v15i3.5011.
 - [12] B. Zhu, X. Pan, S. vanden Broecke, and J. Xiao, “A GAN-based hybrid sampling method for imbalanced customer classification,” *Information Sciences*, vol. 609, pp. 1397–1411, 2022, doi: 10.1016/j.ins.2022.07.145.
 - [13] S. Ali, P. Chourasia, Z. Tayebi, B. Bello, and M. Patterson, “ViralVectors: compact and scalable alignment-free virome feature generation,” *arXiv preprint arXiv: 2304.02891*, Apr. 2023.
 - [14] G. Andresini, A. Appice, and D. Malerba, “Dealing with class imbalance in android malware detection by cascading clustering and classification,” *Studies in Computational Intelligence*, vol. 880, pp. 173–187, 2020, doi: 10.1007/978-3-030-36617-9_11.
 - [15] R. Choudhary and S. Shukla, “A clustering based ensemble of weighted kernelized extreme learning machine for class imbalance learning,” *Expert Systems with Applications*, vol. 164, 2021, doi: 10.1016/j.eswa.2020.114041.
 - [16] H. H. Htun, M. Biehl, and N. Petkov, “Survey of feature selection and extraction techniques for stock market prediction,” *Financial Innovation*, vol. 9, no. 1, 2023, doi: 10.1186/s40854-022-00441-7.
 - [17] P. Dhal and C. Azad, “A comprehensive survey on feature selection in the various fields of machine learning,” *Applied Intelligence*, vol. 52, no. 4, pp. 4543–4581, 2022, doi: 10.1007/s10489-021-02550-9.
 - [18] J. W. Wan and M. Yang, “Survey on cost-sensitive learning method,” *Ruan Jian Xue Bao/Journal of Software*, vol. 31, no. 1, pp. 113–136, 2020, doi: 10.13328/j.cnki.jos.005871.
 - [19] A. Telikani, A. H. Gandomi, K. K. R. Choo, and J. Shen, “A cost-sensitive deep learning-based approach for network traffic classification,” *IEEE Transactions on Network and Service Management*, vol. 19, no. 1, pp. 661–670, 2022, doi: 10.1109/TNSM.2021.3112283.
 - [20] F. Cheng, J. Zhang, and C. Wen, “Cost-sensitive large margin distribution machine for classification of imbalanced data,” *Pattern Recognition Letters*, vol. 80, pp. 107–112, 2016, doi: 10.1016/j.patrec.2016.06.009.
 - [21] R. Gunawan, H. A. Ghani, N. Khamis, J. Al Amien, and E. Ismanto, “Deep learning approach to DDoS attack with imbalanced data at the application layer,” *Telkonnika (Telecommunication Computing Electronics and Control)*, vol. 21, no. 5, pp. 1060–1067, 2023, doi: 10.12928/TELKOMNIKA.v21i5.24857.
 - [22] G. Casañola-Martin *et al.*, “Exploring different strategies for imbalanced ADME data problem: case study on Caco-2 permeability modeling,” *Molecular diversity*, vol. 20, no. 1, pp. 93–109, 2016.
 - [23] J. Chen, Z. Wu, and J. Zhang, “Driving safety risk prediction using cost-sensitive with nonnegativity-constrained autoencoders based on imbalanced naturalistic driving data,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4450–4465, Dec. 2019, doi: 10.1109/TITS.2018.2886280.
 - [24] M. G. Madden and S. S. Khan, “One-class classification: taxonomy of study and review of techniques,” *Knowledge Engineering Review*, vol. 20, no. 2, pp. 117–125, 2004.
 - [25] H. J. Lee and S. Cho, “The novelty detection approach for different degrees of class imbalance,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 21–30, 2006, doi: 10.1007/11893257_3.
 - [26] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” *Expert Systems with Applications*, vol. 73, pp. 220–239, May 2017, doi: 10.1016/j.eswa.2016.12.035.
 - [27] L. Rokach, “Ensemble-based classifiers,” *Artificial Intelligence Review*, vol. 33, no. 1–2, pp. 1–39, 2010, doi: 10.1007/s10462-009-9124-7.
 - [28] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. Shaadan, and S. W. Indratno, “Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 598–608, 2023, doi: 10.11591/ijeecs.v29.i1.pp598-608.
 - [29] H. Deshpande and L. Ragha, “Realizing mother’s features influential on childbirth experience, towards creation of a dataset,” *Data Science*, pp. 143–167, 2022, doi: 10.1201/9781003283249-10.
 - [30] H. Deshpande and L. Ragha, “Mother’s significant feature (MSF) dataset,” *IEEE DataPort*, 2021.
 - [31] H. Deshpande and L. Ragha, “Mother’s lifestyle feature relevance for NICU and preterm birth prediction,” *ITM Web of Conferences*, vol. 40, 2021, doi: 10.1051/itmconf/20214003039.
 - [32] L. Ragha and H. S. Deshpande, “A hybrid random forest-based feature selection model using mutual information and F-score for preterm birth classification,” *International Journal of Medical Engineering and Informatics*, vol. 15, no. 1, 2023, doi: 10.1504/ijmei.2023.10051207.
 - [33] H. Deshpande and L. Ragha, “Random forest based fuzzy feature weighing model for imbalance class distribution towards preterm-birth classification,” *SSRN Electronic Journal*, 2021, doi: 10.2139/ssrn.3867354.
 - [34] L. S. Ramos *et al.*, “KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,” *Journal of Multiple-Valued Logic and Soft Computing*, 2011.
 - [35] H. He and E. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, 2009.
 - [36] B. Mohamed, D. H. Khelouane, and T. A. Alitouche, “Evaluation measures for models assessment over imbalanced data sets,” *Journal of Information Engineering and Applications*, vol. 3, no. 10, pp. 27–38, 2013.
 - [37] F. Last, G. Douzas, and F. Bacao, “Oversampling for imbalanced learning based on k-means and SMOTE,” *arXiv preprint arXiv: 1711.00837*, Nov. 2017, doi: 10.1016/j.ins.2018.06.056.




- [38] M. Hayaty, S. Muthmainah, and S. M. Ghufuran, "Random and synthetic over-sampling approach to resolve data imbalance in classification," *International Journal of Artificial Intelligence Research*, vol. 4, no. 2, 2021, doi: 10.29099/ijair.v4i2.152.

BIOGRAPHIES OF AUTHORS



Himani Deshpande    a dedicated scholar, holds a Ph.D. degree in computer science from the University of Mumbai in the year 2022. With more than a decade of experience in academics, she serves as an assistant professor in the Department of Artificial Intelligence and Data Science at Thadomal Shahani Engineering College, Mumbai, India. She has an industry experience of 3 years as software engineer. Her expertise spans diverse domains, including machine learning, deep learning, and generative AI. Beyond her academic pursuits, she channels her expertise towards addressing crucial societal issues, notably in women's health. She can be contacted at email: himani.deshpande@thadomal.org.



Leena Ragha    a distinguished Professor at BLDEA's V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapur, India, brings 32 years of enriching teaching experience. Renowned for her expertise in image processing, data science, and deep learning, she actively contributes to pioneering research endeavors. She is author of 50 plus research articles in renowned journals and conferences. She is equally adept at administrative duties, ensuring a holistic contribution to academia. She can be contacted at email: cse.leenaragha@bldeacet.ac.in.