Leveraging sampling schemes on skewed class distribution to enhance male fertility detection with ensemble AI learners

Debasmita GhoshRoy
Banasthali Vidyapith
P. A. Alvi
Banasthali Vidyapith
KC Santosh
santosh.kc@usd.edu

University of South Dakota

Research Article

Keywords: Skewed dataset, Re-sampling, Male fertility, and Ensemble AI learners

Posted Date: September 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3311423/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at International Journal of Pattern Recognition and Artificial Intelligence on March 7th, 2024. See the published version at https://doi.org/10.1142/S0218001424510030.

Abstract

Designing effective AI models becomes a challenge when dealing with imbalanced/skewed class distributions in datasets. Addressing this, re-sampling techniques often come into play as potential solutions. In this investigation, we delve into the male fertility dataset, exploring fifteen re-sampling approaches to understand their impact on enhancing predictive model performance. The research employs conventional AI learners to gauge male fertility potential. Notably, five ensemble AI learners are studied, their performances compared, and their results are evaluated using four measurement indices. Through comprehensive comparative analysis, we identify substantial enhancement in model effectiveness. Our findings showcase that the LightGBM model with SMOTE-ENN re-sampling stands out, achieving an efficacy of 96.66% and an F1-score of 95.60% through 5-fold cross-validation. Interestingly, the CatBoost model, without re-sampling, exhibits strong performance, achieving an efficacy of 86.99% and an F1-score of 93.02%. Furthermore, we benchmark our approach against state-of-the-art methods in male fertility prediction, particularly highlighting the use of re-sampling techniques like SMOTE and ESLSMOTE. Consequently, our proposed model emerges as a robust and efficient computational framework, promising accurate male fertility prediction.

1. Introduction

In the era of artificial intelligence, data is an essential component. It can assist in extracting useful information from enormous, heterogeneous, and hierarchical data. Regarding classification, several methods, including support vector machines, random forests, decision trees, extreme gradient boosting, and K-nearest neighbors, among others, have demonstrated superior predictive performance [1, 2]. A classifier can achieve high classification accuracy and not correctly predict a single minority class in the event of an imbalanced dataset because these cutting-edge models place a greater emphasis on classification accuracy than on the imbalanced nature of the input data. For instance, in a dataset having 0.2% negative cases, a primary classifier that predicts all information focuses on being positive will score a classification accuracy of 99.8%. However, none of the negative issues have been discovered in this instance. As a result, classification will continue to be biased towards the majority class, and the decision boundary line will be biased toward samples from the minority class when the input data distribution is uneven [3, 4]. In classification, the minority observations are frequently dismissed as noise. Now and again, most test information tests are arranged into the majority group. Consequently, minority class classification accuracy is significantly lower than that of the majority class. In this situation, most ML models will likely produce unsatisfactory results, which may not be suitable for real-world domain applications. In the present day, most of the datasets have unequal distribution, especially in the medical sector. To overcome this issue, sampling is a probable solution to handle an imbalanced class dataset. A better learning model can be developed using the sampling process to differentiate between majority and minority classes effectively.

There are three possible ways to handle the skewness distribution of data complexity: data level, algorithm level, and cost-sensitive technique. The data level scheme is most considerable and is

implemented in the pre-processing stage. In literature, over-sampling and under-sampling procedures are mainly used. The prime focus is reducing or increasing the dataset size [5] to balance the sample ratios. The purpose of oversampling schemes is to create extra synthetic minority observations by interpolating a few instances that nearly lie together in the feature space [6]. In contrast, the under-sampling approach eliminates the majority class samples by randomly selecting a fixed number of classes equal to those minority classes. However, both applications have merits and demerits; for this reason, some hybrid approaches (under-and over-sampling methods in parallel) are utilized to alleviate these problems [7]. Many researchers have considered all the sampling schemes mentioned above for disease diagnosis. Hence, class imbalance is a vital research issue in ML and AI.

On the other side, many public health issues are a significant concern for society due to environmental causes, changing lifestyles, and the increasing influence of media and advertising. For example, curing diseases like heart disease, obesity, type 2 diabetes, lung cancer, asthma, and infertility after progressing to a critical stage is challenging. According to the World Health Organization (WHO), a large population has suffered worldwide. In the last two decades, the number of people diagnosed with cancer has nearly doubled, whereas obesity and diabetes have become common problems for the younger generation. Additionally, asthma affects 262 million people and is the most common chronic disease in children today [8]. Apart from all these diseases, the rate of infertility is increasing, and its impact is widespread. As per global estimation, 186 million individuals are infertile, and about 48.5 million couples experience infertility [9]. As a result, several emotional stressors have developed in men and women, bringing many difficulties in marital life and identity problems. In general population, major depression is twice as common among all. Additionally, it is underrecognized and underrepresented as a disease. Still, no one can opt for a preventive strategy initially, especially in case of males [10]. Conversely, women's infertility has been an important research topic for the last 40 years, whereas half of the patients involve male factors. Worldwide, sperm counts have decreased by half, and 1 in 20 men have frighteningly poor sperm quality. Moreover, the rising rate of male infertility necessitates the development of new prevention, diagnosis, and treatment strategies.

A diagnostic procedure in the health sector is both time-consuming and costly. In addition, during the testing and therapy process, many emotional affective reactions came into the picture (anxiety, distress, fear, worry, loneliness, and depression) [11]. Prevention is better than cure, and it can be possible with AI. ML is extremely helpful in detecting disease faster and more accurately. The presented research is one of the solutions for choosing a model and a class balancing strategy to expand the model presentation on the binary class male fertility dataset. The proposed research outcome would serve as a tool to predict male infertility status in the early stage, which can provide an effective solution to decrease the global male infertility scenario. The following are the contributions of this research article:

1. Five ensemble AI learners are implemented on an imbalanced dataset to predict male fertility;
2. To tackle the male fertility imbalanced dataset, four oversampling approaches are deployed;
3. A state-of-the-art comparative analysis of seven under-sampling schemes; and
4. Deployment of hybrid re-sampling strategies for optimal model performance.

The document's layout remains as follows: Section 2 explains the background work and a few significant published works in this domain. Section 3 provides brief integrated information related to research topics (data balancing, skewed dataset, male fertility, and AI learners). Section 4 explains the computational experiments, and Section 5 defines the results and discussion. Finally, section 6 concludes and offers suggestions for future research.

2. Literature review

The effect of re-sampling techniques on disease diagnosis many experts and researchers have conducted extensive research on this topic. Most experts have put their interest in individual sampling processes, mainly oversampling. On the other hand, some have applied an under-sampling or combinational approach. Hence, the process of generating new data is very much typical in healthcare applications. Albert et al. [12] recently used smote and adasyn approach to detect heart disease using ML algorithms. The result has shown that oversampling improved model performance by 11% compared to the original dataset. Nishat et al. [13] performed a comprehensive investigation using six different ML algorithms via smote-enn technique. The result portrayed that the application of SMOTE-ENN increased model performance, and 90% accuracy was achieved by RFC. Yang et al. [14] used a hybrid sampling method to identify missed abortion diagnoses via ensemble AI learners. The result was compared with 11 sampling algorithms, and finally, maximum efficacy was reported via RFC. Naz et al. [15] used SMOTE technique to detect people with diabetes two by the deployment of the SMO classifier. They noted that the proposed model achieved an accuracy of 99.07%. Kumar et al. [16] applied several data balancing approaches: random oversampling, random under-sampling, smote, adasyn, svm-smote, borderline smote, smotenc, smoteen, and smote-tomek to diagnose the acute disease to prevent covid 19 on the original dataset. A total of 6 classifiers are used, and performance has been compared. They reported DT with smoteen based model performed best. Gupta and Gupta [17] conducted a comprehensive data-level investigation of cancer diagnosis using ensemble AI learners. Eight data-handling methods were employed and designed 14 classification models. They reported smote-enn achieved the best accuracy via ensemble stacking. GhoshRoy et al. [18] used smote oversampling technique to increase dataset size, and XGB classifier was used to predict male fertility. The accuracy improvement is noticed after the application of oversampling. GhoshRoy et al. [19] used industry-standard ML algorithms to predict male fertility where smote technique is deployed on the original dataset. Each model's accuracy was compared, and model performance was enhanced due to sampling. Yibre et al. [20] used smote technique along with the adaboost classifier to predict male fertility. They reported that the model achieved high performance with a data-balancing approach. Lin et al. [21] performed an experimental study to check the effect of sampling techniques using 44 datasets. They reported that the hybrid sampling application becomes more impactful than over and under-sampling individually. Islam et al. [22] studied imbalanced image datasets using different samplers. They reported KNNOR oversampling approach performed better as compared to 9 samplers. Ma et al. [23] used an evolutionary safe level synthetic minority oversampling approach with BPNN to predict the seminal quality. They reported model achieved the highest AUC of 97.2% after data balancing. Feng et al. [24] used 15 datasets with the C4.5 classifier and applied a hybrid

sampling approach (CUS and SMOTE). They found that hybrid methods performed well and provided higher classification accuracy. Fujiwara et al. [25] conducted a study based on a hybrid sampling approach using a CART classifier via eight datasets. Xu et al. [26] used a hybrid sampling strategy with ten datasets and an RF classifier chosen. Table 1 lists recent related works focusing on the different sampling approaches to diagnose disease serially. Remarkably, the order to combine other samplers, datasets used, and constructed classifiers are compared. Moreover, the developed model performs better after applying these sampling methods to a dataset.

Moreover, data balancing approaches are a considerable framework before building an effective model. Re-sampling approaches help researchers to manage a class imbalance problem in a simple, easy, and understandable way. In this article, we follow and investigate the effect of data-balancing approaches which are widely used in various disease diagnoses. In other words, it is unknown which data balancing approaches perform better in the male fertility dataset; it is the critical question of this research work. We are mainly focusing on binary classification and multivariate data.

Table 1
sampling approach related work

Authors [Ref]	Samplers	Classifier	Disease
Albert et al. [12]	SMOTE, and ADASYN	DT	Heart disease
Nishat et al. [13]	SMOTE-ENN	RFC	Heart failure
Yang et al. [14]	SMOTE-ENN	RFC	Missed abortion
Naz et al. [15]	SMOTE	SMO	
Kumar et al. [16]	ROS, RUS, SMOTE, ADASYN, SVM -SMOTE, Borderline SMOTE, SMOTENC, SMOTEEN, and SMOTE-Tomek	LDA, SVM, GNB, ANN, k-NN, DT	Covid-19
Gupta and Gupta [17]	SMOTE, ADAYSN, TL, RUS, Borderline SMOTE, SMOTE-SVM	Stacked Ensemble	Cancer

3. Integrated Information of Research

3.1 Sampling Schemes

Different types of data balancing schemes are already proposed in the literature. These methods greatly uplift the data size and improve the classifier's performance. The data-balancing strategies and classifier-level approach are two different aspects of dealing with class imbalance. Therefore, the classifier

approach mainly focuses on changing present learning methods to favor the minority group. The data balancing schemes seek to rebalance class dispersion by re-sampling data space, which entails oversampling instances of minority classes and under-sampling instances of majority classes. We have implemented three strategies for this work: over, under, and hybrid sampling. Figure 1. exhibits the sampling task (oversampling vs. under-sampling).

3.1.1 Oversampling Schemes

In this work, four well-known methods are employed: random oversampling (ROS), adaptive synthetic (ADASYN), synthetic minority oversampling (SMOTE), and borderline synthetic minority oversampling (Borderline-SMOTE).

- *ROS*: In this method, observations from minority classes are randomly selected and added to the training dataset. It indicates that the minority classes in the training dataset are duplicated, which leads to overfitting in ML algorithms. [27].
- *ADASYN*: It is a generic framework, and more observations from minority observations are produced along the borderline in ADASYN. The ratio of majority samples in the k-nearest neighbours of a minority sample (t_i) is $h\,[i]$. It establishes the probability of being near the border. It is then normalized to calculate $h\,[i]$ and $F\,(i)$, the number of samples to synthesize from $(t_i\,)[\,28]$.
- *Borderline-SMOTE*: It begins to function by classifying the minority class observations. The instances close to the borderline are more important for classification than those far from it. All K neighbours of a borderline sample belong to the majority class. This borderline method, using SMOTE, pre-samples the minority class [29].
- *SMOTE*: By interpolating a collinear point, it produces synthetic samples from the minority class based on the KNN method. This observation or point creates a gap between a minority class's observation and that of its closest neighbour. [30].

3.1.2 Under sampling schemes

Eight schemes such as random under-sampling (RUS), Near miss, cluster-based under-sampling (CUS), condensed nearest neighbour rule (CNN), tomek-links (TL), one-sided selection (OSS), and edited nearest neighbour (ENN) are deployed.

- *RUS*: It is one of the most basic approaches where samples are chosen randomly, and the training dataset is removed. This technique's fundamental concept of achieving class balance is eliminating enormous samples, with or without replacement [31].
- *Near-miss*: It works based on the distance of majority class examples to minority class examples. In this technique, we prevent the problem of information loss and find n closest instances in the majority class; three variations, version 1, version 2, and version 3, can be applied [32]. This study uses version 1.

- *CUS*: This technique's motivation is to prevent a loss of information that may use a downsizing dataset. This approach first clusters all the training samples into the K cluster and then chooses appropriate training samples from derived clusters. The main idea is that different clusters seem to have distinct characteristics depending on the number of majority and minority samples present in the cluster [33].
- *CNN*: It operates on the NNR principle, and motivation is derived from statistical concerns related to NNR. Before that, be familiar with the meaning of a consistent subset of a sample set. A consistent subset is a subset that classifies all the remaining points in the sample set, or the dataset is used as a stored reference set for NNR and classifies all remaining dataset samples correctly; then, this subset is known as the consistent subset of a sample set. CNN aims to find a compatible subset [34].
- *TL*: The fundamental goal is to locate pair (a,b) such that either belongs to the minority class and the others belong to the majority class. Further, a and b need to be the nearest neighbour, and such instances are known as TL, in which majority class instances are eliminated. This technique clarifies the boundary between two classes, leading to more distinct minority regions [35].
- *OSS*: This method is the outcome of the implementation of TL followed by the use of US-*CNN*. TL is utilized to remove noisy and borderline majority class observation, whereas CNN aims to eliminate observations from the majority class distant from the decision border [36].
- *ENN*: It is a workable strategy, and the nearest neighbour rule states that the closest samples of each majority sample are determined based on the distance between two samples and that the majority samples can be identified as noise samples by assessing whether or not their labels are consistent. Eliminating observations whose class differs from that of its k-nearest neighbours is the cornerstone of ENN. This technique's primary goal is to eliminate most noise observations. [37].

3.1.3 Hybrid Sampling Schemes

Three hybrid sampling schemes, namely smote-tomek, somte-rus, and smote-enn are used simultaneously for upsizing and downsizing the dataset.

- *SMOTE-Tomek*: In this method, two sampling approaches are deployed. Smote is an oversampling scheme in which minority class observations are oversampled. In other words, Tomek is an under-sampling approach that removes observations from the majority class with overlapping values. Hence, the ratio of observations becomes 1:1 [38].
- *SMOTE-RUS*: This hybrid scheme is based on smote, which synthesizes samples of the minority class based on their nearest neighbor and, in the next rus, randomly reduces the majority samples to match the size of the minority class [39].
- *SMOTE-ENN*: In this scheme, smote is developed by enn to search noisy observations. Instead of disregarding observations from one class, ENN removes observations from both categories. To eliminate the misclassified observations, it uses its three nearby neighbors [40].

3.2 Skewed Dataset

Medical data are often not symmetric enough to be adequately modeled through usual normal distributions, mostly skewed patterns. The maximum samples are either positive or negative in a two-class target label. The demonstration of skewness is represented by a bell curve when the right and left side data points are distributed equally; it is said to be a normal or symmetric distribution. In the ML model, a skewed dataset degrades overall model performance, and a sampling-based pre-processing approach is used to overcome this problem. After sampling approaches are implemented, uneven data distribution issues are reduced [41]. As a result, precise model performance is vital in the medical sector.

3.3 Male Fertility

Male fertility is associated with poor sperm quality. Still, we cannot define the precise etiology of infertility due to a lack of proper diagnosis and treatment. Many clinical factors (hormonal, immunological, genetic/chromosomal, behavioral, and environmental) are relevant causes, but somehow it isn't easy to quantify or relate to male fertility. Hence, much effort is placed into comprehending the association between semen traits and male fertility [42]. The possible bottom line merely affects male reproductive ability; fortunately, the improvement is within our grasp. The significant role of lifestyle and environmental factors is an alarming disruptor that can impact the reproductive system.

3.4 Ensemble AI learners

An incisive overview of the AI learners of each classification technique has been discussed here to present vital insight into five ensemble learners.

1) RF: RF employs bootstrapping, averaging, and bagging to train several decision trees. By utilizing various combinations of the given attributes, numerous distinct decision trees can be constructed concurrently on various subsets of the training data. Bootstrapping guarantees that each decision tree within the RF is distinct, which lowers the variance of the RF. The RF classifier's ability to combine the outcomes of various tress judgments into a single conclusion allows for good generalization. The RF classifier seeks to continuously surpass nearly all existing classifier algorithms in terms of precision without the issues of unbalanced datasets and overfitting. [43]. The mean square error for RF can be defined as Eq. (1)

$$MSE = \frac{1}{N}\sum_{k=0}^{n}\binom{n}{k}\left(F_i - y_i\right)b^2$$

1

where N represents the number of distinct data points, and F_i replicates the outcome returned by the model y_i and the precise value of the point value is i.

2) CatBoost: It is a novel variant of gradient boosting trees that deals with categorical and ordered features. The permutation method solves the categorical attributes, providing a gradient-boosting framework. As a result, the modified target-based statistics offer a more efficient implementation with reduced computational complexity, and model overfitting was overcome via Bayesian optimization. Catboost employs greedy search to generate a robust competitive system by the combination of many

weak systems sequentially. With the help of ordered boosting, decision trees are fitted one after the other to minimize errors. CatBoost, in contrast to gradient boosting models, uses the oblivious tree approach, which produces a straightforward fitting scheme and excellent computational efficiency and uses loss function change to rank the feature importance of the built model. [44].

3) LightGBM: LightGBM distinguishes itself by gradient-based one-sided sampling (GOSS), mutually exclusive feature bundling (EFB), and differential acceleration via a histogram technique. The basic idea behind GOSS is to reserve large gradient samples and randomly select tiny gradient samples in proportion to their sizes. Similarly, EFB combines two non-exclusive features. This binding minimizes the number of features and the temporal complexity, boosting the model's computational efficacy. Finally, the fundamental of a histogram is to continuously construct a histogram with width K by discretizing successive floating-point eigenvalues into k integers. After traversing the data, the accumulated statistics in the histogram are selected as an index based on these values. Finally, the computational segmentation score is determined [45]. This classifier performed well in classification challenges, with improved performance due to faster training, decreased memory use, and parallel processing.
4) ADA: It is a supervised algorithm for binary classification problems. ADA combines several weak classifiers into a single classifier known as a robust classifier. The most common ADA algorithm is decision trees with only one level or a single split. These trees are often referred to as decision stumps. This approach produces a system by giving all data points equal weights. It then assigns a higher weight to improperly classified points. In the following model, all higher-weight points are given more weight. It keeps training systems until a lower error is obtained. To begin the ADA, the weight of the training set is used [46].
5) XGB: XGB is a tree integration model that uses the cumulative sum of anticipated values of a sample in each tree as the sample prediction in the XGB system. It is an extensible and cutting-edge use of gradient-boosting machines that have been shown to push the limits of computational power for boosted tree algorithms. Adding fresh models to an ensemble technique known as boosting allows for correcting faults generated by older models. Models are introduced repeatedly until no discernible improvements can be found. The primary elements for its design are model efficiency and computing speed [47].

4. Experimental Design

In this segment, we have reported on the experimental setting that aims to appraise the performance of binary classification models with distinct sampling schemes as accurately as possible. The pictorial representation of data re-sampling schemes is given in Figure 2.

4.1 Dataset

This research is conducted on a male fertility dataset, and the description is given below:

The dataset was divided into two classes [49], consisting of 100 samples with ten attributes each. Six of the ten attributes have categorical values, while the others have numerical values. There are 12 occurrences of infertile. Fertile, with 88 examples, makes up the second class. The challenge in the

research is figuring out if a male is considered fertile or not. This study converts all variables into the normalized range with different rules. For example, age, year of analysis, the number of exposures, or the average number of cigarettes per day are normalized onto the interval (0,1). Similarly, the variable with only two independent variables is assigned binary values. We have six variables in which we used this representation. The variable with three independent attributes, we represent it in ternary values (-1,0,1). Finally, for the variables with four independent attributes, we used four equal and different values (-1, -0.33, 0.33,1).

4.2 Performance evaluation criteria

There are several supervised models for classification. However, side-by-side comparisons can quickly determine a problem's most influential and reliable model. Here are the most popular classifier performance evaluation metrics: (1) accuracy, (2) precision, (3) recall, and (4) f1-score.

1) Accuracy (ACC) is defined by . It calculates the classifier's capacity to identify only accurate observations for each class.

2) Precision (PREC) measures a classifier's capacity to locate genuine cases inside a class.

3) Recall (REC) is defined by as a classifier's ability to find all truthful cases in a class.

4) F1-score replicates the perfect balance due to PREC and REC are inversely related. PREC and REC are both vital; a high F1 score is beneficial.

In this research, we have used four evaluation protocols to assess model performance.

5. Results and Discussion

In this research, we have investigated and compared the applicability of five ensemble AI-learners and fourteen significant sampling schemes (1. ROS, 2. ADASYN, 3. Borderline SMOTE, 4. SMOTE, 5. RUS, 6. Near-miss, 7. CNN, 8. TL, 9. ENN, 10. OSS, 11. CUS, 12. SMOTE-Tomek, 13. SMOTE-RUS, 14. SMOTE-ENN) to predict male fertility based on lifestyle and environmental factors. The performance of five different classifiers is assessed using the ACC, PREC, REC, and F1-score metrics for both the original and re-sampled datasets. The data used to create the model is split into 30% for testing and 70% for training observations. Moreover, the K-fold cross-validation protocol enhances model performance and robustness. K is a number that represents the number of iterations used to train and validate a model, with each iteration using a new fold of data for validation. Besides, the K-5 fold is often utilized for learning. We also considered the s.d.(σ) value during performance analysis to understand the robustness of the model.

5.1 Performance comparison of different AI learners with the original dataset

In this paragraph, we have represented the experiments performed over the original dataset results. Table 2, Among all classifiers, Catboost performed best. The catboost achieved a mean efficacy of 93.02% with s.d. 0.13. The classifier's performance without balancing schemes in decreasing order came as Catboost > LightGBM > XGB > RF > ADA.

Table 2
Male fertility prediction using the original dataset

Models	Test Set Performance (in %)			
	ACC±σ	PREC±σ	REC±σ	F1-Score±σ
RF	83.99$\pm$.086	88.28 ± .038	94.37 ± .080	91.03 ± .051
CatBoost	86.99 ± .024	87.89 ± .023	98.88 ± .022	93.02 ± .013
LightGBM	86.00 ± .058	95.49 ± .064	89.31 ± .006	92.22 ± .034
ADA	80.00 ± .083	87.31 ± .038	90.25 ± .095	88.50 ± .053
XGB	85.71 ± 0.0	89.41 ± .030	95.12 ± .039	92.05 ± .003

5.2 Performance comparison of different AI learners with an over-sampling variant dataset

This experiment demonstrates the performance of different classifiers with an oversampled dataset. The classification performance of five classifiers using the ROS, ADASYN, borderline SMOTE, and SMOTE methods are displayed in Table 3. After analysis, we found that every classifier worked well with different combinations of re-sampling techniques. For the CatBoost classifier, borderline SMOTE provides the highest accuracy and F1-score of 89.30% and 88.0%, respectively. Similarly, other classifiers like RF-ROS obtained an accuracy of 95.06% with an s.d. of 0.40. LightGBM-SMOTE provides a maximum accuracy of 90.19% and s.d. of 0.05. Finally, ADA-ROS shows better performance among all combinations and the obtained accuracy of 92.56% with s.d. 0.04. XGB classifier performed best with ROS sampler and the reported accuracy of 95.05% where the s.d. value is 0.01. In this table, we found some classifier provides the same level of classification accuracy; in that case, we have compared their f1-score value to reach the final decision. We also found each classifier's contributions ranking from the seven under-sampling techniques, from best to worst: 1. Catboost: Borderline SMOTE > ROS > ADASYN > SMOTE, 2. RF-ROS > ADASYN > Borderline SMOTE > SMOTE, 3. Light GBM: SMOTE > ROS > Borderline SMOTE > ADASYN, 4. ADA- ROS > Borderline SMOTE > SMOTE > ADASYN, 5. XGB- ROS > ADASYN > SMOTE > Borderline SMOTE.

Table 3
Over-sampling on dataset

Models	Samplers	Test Set Performance (in %)			
		ACC±σ	PREC±σ	REC±σ	F1-Score±σ
CatBoost	ROS	89.3 ± .050	1.0 ± .00	78.5 ± .101	87.6 ± .069
	ADASYN	86.9 ± .03	96.0 ± .07	78.4 ± .087	85.6 ± .034
	Borderline SMOTE	89.3 ± 0.02	97.0 ± .05	81.0 ± .064	88.0 ± .023
	SMOTE	86.06 ± .042	96.00 ± .079	76.79 ± .112	84.34 ± .056
RF	ROS	95.06 ± .040	1.0 ± 0.0	90.1 ± .081	94.0 ± .047
	ADASYN	93.5 ± .089	93.6 ± .126	96.6 ± .040	94.5 ± .069
	Borderline SMOTE	93.4 ± .074	95.0 ± .099	94.4 ± .062	93.8 ± .064
	SMOTE	91.89 ± 0.104	93.33 ± 0.133	93.46 ± 0.062	92.79 ± 0.08
Light GBM	ROS	90.9 ± .067	1.0 ± 0.0	81.9 ± 0.133	81.6 ± .081
	ADASYN	89.3 ± 0.06	94.4 ± 0.11	86.6 ± 0.113	89.2 ± 0.063
	Borderline SMOTE	90.0 ± 0.66	95.2 ± 0.09	86.6 ± 0.135	89.6 ± 0.072
	SMOTE	90.19 ± 0.059	95.0 ± 0.099	86.79 ± 0.067	90.13 ± 0.052
ADA	ROS	92.56 ± 0.049	1.0 ± 0.0	85.12 ± 0.098	91.64 ± 0.061
	ADASYN	87.73 ± 0.068	92.77 ± 0.098	83.46 ± 0.092	87.28 ± 0.069
	Borderline SMOTE	88.56 ± 0.079	92.77 ± 0.098	85.12 ± 0.111	88.15 ± .080
	SMOTE	87.76 ± 0.88	93.68 ± 0.126	85.00 ± 0.152	87.35 ± 0.090
XGB	ROS	95.06 ± 0.017	1.0 ± 0.0	90.12 ± 0.034	94.77 ± 0.019
	ADASYN	92.69 ± 0.068	95.00 ± 0.099	91.79 ± 0.052	92.99 ± 0.0587
	Borderline SMOTE	91.06 ± .085	94.11 ± 0.117	90.12 ± 0.081	91.40 ± 0.073
	SMOTE	92.66 ±	95.29 ±	91.66 ±	92.90 ± 0.036

	0.046	0.094	0.0527

5.3 Performance comparison of different AI learners with under-sampling variant dataset

In this section, we evaluated the classifier performance and different under samplers. All results are well documented in Table 4. Firstly, for catboost classifier OSS sampler performance is outstanding, with accuracy and F1score of 95.63% and 97.77%, respectively. In both cases, the reported s.d. is 0.00. Conversely, RF also provides superior performance with the combination of the OSS scheme and the achieved accuracy, and F1-score is precisely the same as catboost. In case of LightGBM classifier, the highest accuracy was obtained by OSS. The reported accuracy is 96.12% with s.d. 0.01. The ADA and XGB classification models perform best with OSS; the reported accuracies are 95.63% and 96.12%. Furthermore, when we compared the contributions of each of the seven under-sampling techniques, we discovered the following ranking from best to worst: OSS > Tomek-Links > ENN > CUS > RUS > CNN > Near Miss.

Table 4
Under-sampling on dataset

Models	Samplers	Test Set Performance (in %)			
		ACC±σ	PREC±σ	REC±σ	F1-Score±σ
CatBoost	RUS	80.00 ± .2915	80.00 ± 0.4	70.00 ± 0.4	73.33 ± 0.388
	Near Miss	48.33 ± 0.2198	33.33 ± 0.298	50.00 ± 0.44	40.0 ± 0.357
	CNN	62.00 ± 0.0979	68.30 ± 0.033	73.33 ± 0.133	70.40 ± 0.076
	Tomek Links	86.59 ± 0.0289	86.59 ± 0.0289	100.00 ± 0.0	92.78 ± 0.016
	ENN	86.36 ± 0.077	86.14 ± 0.076	100 ± 0.0	92.37 ± 0.042
	OSS	95.63 ± 0.0096	95.63 ± 0.0096	100 ± 0.0	97.77 ± 0.005
	CUS	80.00 ± 0.244	74.32 ± 0.39	70.00 ± 0.44	70.00 ± 0.49
RF	RUS	80.00 ± 0.1870	83.33 ± 0.2108	80.00 ± 0.244	79.33 ± 0.193
	Near Miss	65.0 ± 0.22	43.33 ± 0.388	50.00 ± 0.44	45.99 ± 0.407
	CNN	57.00 ± 0.2181	62.00 ± 0.1122	73.33 ± 0.326	65.28 ± 0.210
	Tomek Links	86.59 ± 0.028	86.59 ± 0.028	100.00 ± 0.0	92.78 ± 0.016
	ENN	86.54 ± 0.0936	87.55 ± 0.0783	97.77 ± 0.044	92.26 ± 0.053
	OSS	95.63 ± 0.0096	95.63 ± 0.0096	100.00 ± 0.0	97.77 ± 0.005
	CUS	63.33 ± 0.194	60.00 ± 0.374	50.00 ± 0.31	53.33 ± 0.32
Light GBM	RUS	43.33 ± 0.081	36.66 ± .1943	80.00 ± 0.4	50.00 ± 0.258
	Near Miss	43.33 ± 0.0816	36.66 ± 0.19	80.00 ± 0.4	50.00 ± .2581
	CNN	63.00 ± 0.060	63.00 ± 0.060	100.00 ± 0.0	77.14 ± 0.042
	Tomek Links	85.16 ± 0.043	86.37 ± 0.0299	98.33 ± 0.033	91.92 ± 0.025
	ENN	82.36 ± 0.0388	82.36 ± 0.0388	100.00 ± 0.0	92.03 ± 0.022
	OSS	96.12 ± 0.0117	96.57 ± 0.0113	99.48 ± 0.0102	97.99 ± 0.006
	CUS	43.33 ± 0.0816	36.66 ± 0.194	80.00 ± 0.4	50.00 ± 0.258
ADA	RUS	63.33 ± 0.2505	50.00 ± 0.33	60.00 ± 0.4	54.0 ± 0.344
	Near Miss	58.33 ± 0.2818	36.66 ± .3055	50.00 ± .4472	42.00 ± .4079
	CNN	58.00 ± 0.039	70.66 ± 0.149	73.33 ± 0.2494	66.66 ± 0.091
	Tomek Links	74.83 ± 0.118	85.67 ± 0.0497	84.39 ± 0.1011	84.93 ± 0.075

	ENN	78.54 ± 0.0696	85.33 ± 0.0826	90.83 ± 0.0840	87.41 ± 0.040
	OSS	95.63 ± 0.018	96.57 ± 0.019	98.98 ± 0.012	97.77 ± 0.009
	CUS	65.00 ± 0.374	70.00 ± 0.4	60.00 ± 0.37	63.33 ± 0.371
XGB	RUS	56.66 ± .2758	50.00 ± 0.333	60.00 ± 0.3741	54.00 ± 0.344
	Near Miss	71.60 ± 0.16	50.00 ± 0.447	63.33 ± 0.3711	59.33 ± 0.338
	CNN	53.00 ± 0.289	52.00 ± 0.3357	60.00 ± 0.3887	55.00 ± 0.348
	Tomek Links	85.27 ± 0.0615	88.82 ± 0.0346	94.84 ± 0.0421	91.71 ± 0.035
	ENN	86.54 ± 0.0936	89.05 ± 0.0639	95.27 ± 0.0580	91.98 ± 0.055
	OSS	96.12 ± 0.0117	96.57 ± 0.0113	99.48 ± 0.010	97.99 ± 0.006
	CUS	90.00 ± 0.200	87.30 ± 0.24	91.20 ± 0.230	90.01 ± 0.200

5.4 Performance comparison of different AI-learners with hybrid sampling variant dataset

In identifying male fertility, we used four hybrid samplers on the dataset and compared the classifier's performance individually. In Table 5, experimental results are presented. The Catboost with SMOTE-ENN is the best among the three samplers. This model can predict male fertility with a classification accuracy of 94.37%, and the s.d. is 0.03. RF classifier worked best with the SMOTE-Tomek sampler and reported an accuracy of 94.96% with an s.d. of 0.06. For LightGBM, SMOTE-ENN performed best, with an accuracy of 96.66% with an s.d. of 0.044. ADA performs better with SMOTE-Tomek, with the highest F1 score of 90.73% and reported accuracy of 90.72%. XGB with SMOTE-ENN attained a maximum accuracy of 96.66% and the highest F1 score of 95.38%. When we analyzed how well each classifier predicted male fertility, we obtained the following ranking from best to worst: SMOTE-ENN > SMOTE-Tomek > SMOTE-RUS.

Table 5
Effect of Hybrid Sampling

Models	Samplers	Test Set Performance (in %)			
		ACC±σ	PREC±σ	REC±σ	F1-Score±σ
Cat Boost	SMOTE-Tomek	91.52 ± .0264	98.33 ± .033	84.69 ± .063	90.78 ± .0318
	SMOTE-RUS	81.28 ± .1127	87.11 ± .1268	88.88 ± 1.11	87.50 ± .067
	SMOTE-ENN	94.37 ± .0351	100 ± 0.0	85.71 ± .090	92.05 ± .052
RF	SMOTE-Tomek	94.96 ± .0611	95.00 ± .099	96.51 ± .0427	95.32 ± .0521
	SMOTE-RUS	84.35 ± .1175	87.69 ± .125	93.33 ± .0544	89.83 ± .070
	SMOTE-ENN	93.33 ± .0544	94.00 ± .120	91.42 ± .0699	91.85 ± .056
Light GBM	SMOTE-Tomek	90.65 ± .049	95.71 ± .085	86.36 ± .0684	90.32 ± .047
	SMOTE-RUS	73.20 ± .0736	76.95 ± .080	91.11 ± .044	83.06 ± .041
	SMOTE-ENN	96.66 ± .044	97.14 ± .057	94.28 ± .0699	95.60 ± .0577
ADA	SMOTE-Tomek	90.72 ± .0722	91.80 ± .107	91.21 ± .099	90.73 ± .069
	SMOTE-RUS	71.53 ± .109	81.50 ± .1152	82.22 ± .1805	79.88 ± .0904
	SMOTE-ENN	91.04 ± .066	92.72 ± .145	88.57 ± .057	89.40 ± .0581
XGB	SMOTE-Tomek	92.39 ± .0484	95.00 ± .099	91.51 ± .052	92.63 ± .0389
	SMOTE-RUS	81.15 ± .0758	87.32 ± .1173	88.88 ± .0702	87.24 ± .0457
	SMOTE-ENN	96.66 ± .0272	100.00 ± 0.0	91.42 ± .0699	95.38 ± .0376

5.5 Performance comparison of all best classifiers with over-sampling schemes

Previously, we have conducted a comparison study on different classifiers with 15 data-balancing approaches employed to identify male fertility. After analysis, we found five classifiers where three over-samplers provided the best performance regarding ACC, PREC, REC, and F1-score. The results are documented in Table 6. Catboost with borderline-SMOTE provides the least accuracy compared to other models, whereas XGB outperformed with an accuracy of 95.06% and s.d. is 0.01. Moreover, Table 6 shows that the XGB-ROS scheme is the best-performing model for male fertility detection.

Table 6
Performance comparison of the best classifier with over-sampler

Classifier	Samplers	$\|S_M\|/\|S_N\|$	Ratio	Test Set Performance (in %)			
				ACC±σ	PREC±σ	REC±σ	F1-Score±σ
CatBoost	Borderline - SMOTE	88/64	1.38	89.3 ± 0.02	97.0 ± 0.05	81.0 ± 0.06	88.0 ± 0.02
RF	ROS			95.06 ± .04	1.0 ± 0.0	90.1 ± 0.08	94.0 ± 0.04
LightGBM	SMOTE			90.19 ± 0.05	95.0 ± 0.09	86.79 ± 0.06	90.13 ± 0.05
ADA	ROS			92.56 ± 0.04	1.0 ± 0.0	85.12 ± 0.09	91.64 ± 0.06
XGB	ROS			95.06 ± 0.01	1.0 ± 0.0	90.12 ± 0.03	94.77 ± 0.01

5.6 Performance comparison of all best classifiers with under-sampling schemes

A similar strategy is applied for the comparison of under-sampling schemes. We used five different classifiers followed by eight under-sampling procedures. Table 7 displays the model evaluation report. In this way, we found that maximum accuracy is obtained by two classifiers such as LightGBM and XGB. The application of OSS delivered excellent performance. Moreover, we can say OSS is a suitable under-sampler, among others.

Table 7
Performance comparison of the best classifier with under sampler

Classifier	Samplers	$\|S_M\|/\|S_N\|$	Ratio	Test Set Performance (in %)			
				ACC±σ	PREC±σ	REC±σ	F1-Score±σ
CatBoost	OSS	227/12	18.91	95.63 ± 0.009	95.63 ± 0.0096	100 ± 0.0	97.77 ± 0.005
RF				95.63 ± 0.009	95.63 ± 0.0096	100.00 ± 0.0	97.77 ± 0.005
LightGBM				96.12 ± 0.011	96.57 ± 0.011	99.48 ± 0.010	97.99 ± 0.006
ADA				95.63 ± 0.018	96.57 ± 0.019	98.98 ± 0.012	97.77 ± 0.009
XGB				96.12 ± 0.011	96.57 ± 0.0113	99.48 ± 0.010	97.99 ± 0.006

5.7 Performance comparison of all best classifiers with hybrid-sampling schemes

In case of hybrid samplers, we have listed all the best classifier's performances in Table 8. We compared all classifier performances and found that five individual classifiers performed well with different sampling methods. For example, LightGBM provides maximum efficacy and F1-score of 96.66% and 95.60%, respectively. However, we also observed that XGB delivers the same accuracy, but the value of the F1-score is lesser than LightGBM.

Table 8
Performance comparison of the best classifier with a hybrid sampler

Classifier	Samplers	$\|S_M\|/\|S_N\|$	Ratio	Test Set Performance (in %)			
				ACC±σ	PREC±σ	REC±σ	F1-Score±σ
CatBoost	SMOTE-ENN	81/38	2.12	94.37 ± 0.0351	100 ± 0.0	85.71 ± 0.0903	92.05 ± 0.052
RF	SMOTE-Tomek	87/63	1.38	94.96 ± 0.0611	95.00 ± 0.099	96.51 ± 0.0427	95.32 ± 0.052
LightGBM	SMOTE-ENN	81/38	2.12	96.66 ± 0.044	97.14 ± 0.057	94.28 ± 0.0699	95.60 ± 0.057
ADA	SMOTE-Tomek	87/63	1.38	90.72 ± 0.0722	91.80 ± 0.107	91.21 ± 0.099	90.73 ± 0.069
XGB	SMOTE-ENN	81/38	2.12	96.66 ± 0.0272	100.00 ± 0.0	91.42 ± 0.0699	95.38 ± 0.037

5.8 Performance comparison of best classifiers with all samplers

Although, we have performed rigorous analysis to determine the best classification model for predicting male fertility (from Table 2 to Table 8). To reach the research goal, again, we investigate the final observations from the best classifiers. In this way, we discovered the top classifier (LightGBM), which provides promising results for predicting male fertility. The reported model (LightGBM-SMOTE-ENN) accuracy is 96.66%, the best among all ensemble AI -learners.

Table 9
Performance comparison of the best classifier with all samplers

Classifier	Samplers	Test Set Performance (in %)			
		ACC±σ	PREC±σ	REC±σ	F1-Score±σ
XGB	ROS	95.06 ± 0.01	1.0 ± 0.0	90.12 ± 0.03	94.77 ± 0.01
XGB	OSS	96.12 ± 0.011	96.57 ± 0.0113	99.48 ± 0.010	97.99 ± 0.00
LightGBM	OSS	96.12 ± 0.011	96.57 ± 0.011	99.48 ± 0.010	97.99 ± 0.00
LightGBM	SMOTE-ENN	96.66 ± 0.044	97.14 ± 0.057	94.28 ± 0.0699	95.60 ± 0.057

5.9 Performance comparison between sampling vs. no sampling model

Understanding how sampling schemes improved the model performance compared to the non-sampled dataset, a quick view and detailed analysis are presented in Table 10. The total result is documented in Tables 2 to 9. In Table 2, we have observed that the Catboost classifier provides the highest accuracy with the imbalanced dataset. Similarly, LightGBM is our proposed model, which offers outstanding performance with an accuracy of 96.66%. Therefore, it is visible that sampling helps to enhance the model accuracy, and a 10% accuracy upliftment is reported. The depicted framework in Fig. 3 utilizes the SMOTE-ENN algorithm to detect male fertility.

Table 10
Performance comparison between no sample vs. sample dataset

Classifier	Samplers	Test Set Performance (in %)			
		ACC±σ	PREC±σ	REC±σ	F1-Score±σ
LightGBM	SMOTE-ENN	96.66 ± 0.044	97.14 ± 0.057	94.28 ± 0.0699	95.60 ± 0.057
CatBoost	None	86.99 ± .0244	87.89 ± .023	98.88 ± .022	93.02 ± .013

5.10 Performance comparison between proposed model vs. existing model

In addition to the abovementioned analysis, we have focused on a comparative study between the proposed and existing models. Few researchers have used sampling methods in the area of male prediction. In this context, it was necessary because less observation is presented in the dataset. More data is required to build model intelligence; here, the need for samplers arises (over, under, or hybrid). In Table 11, we have compared our best model performance with existing model performance where samplers are used. Four articles are documented where scientists mainly worked on oversampling methods and mostly used SMOTE.

Table 11
Performance comparison between proposed vs. existing models

Authors [Ref]	Classifiers	Samplers	Test Set Performance (in %)			
			ACC	PREC	REC	F1-score
GhoshRoy et.al. [19]	RF	SMOTE	90.47	-	-	91.99
GhoshRoy et.al. [18]	XGB	SMOTE	93.22	-	-	-
Yibre et al. [20]	FFNN	SMOTE	97.5	-	-	96.66
MA et al. [23]	ADA	ELSMOTE	95.1	95.5	97.2	-
Our study	LightGBM	SMOTE-ENN	96.66 ± 0.044	97.14 ± 0.057	94.28 ± 0.0699	95.60 ± 0.057

The uniqueness of our study is not only designing an AI model but also exploring and investigating the effect of data re-sample on the prediction of male fertility. In literature, oversampling is the most opted scheme, whereas, in this study, we deployed fifteen re-sampling procedures with an advanced AI framework. In the final stage of this experiment, we discovered which combination of data balancing and classification approach performs well with a small sample size (100 samples). SMOTE-ENN paired with the LightGBM model ranked best regarding sampling strategy performance on male fertility prediction. In addition, we have provided a visual representation of all classifier performance across 15 sampling strategies (see Figs. 4–8). It helps to comprehend how sampling influences overall model performance.

6. Conclusion and Future Works

This study delves into three prominent data balancing strategies: over-sampling, under-sampling, and hybrid sampling. These techniques are widely recognized for their ability to mitigate the imbalanced class distribution often encountered in datasets. The focus of this research is not limited to the conventional SMOTE variant oversampling methods commonly found in the literature. Instead, the study explores fifteen distinct re-sampling approaches to assess their impact on the performance of AI models. Five ensemble AI learners, specifically Catboost, XGB, ADA, RF, and Lightgbm, are employed in conjunction with both the original and re-sampled datasets. After analyzing the distribution of the original sample, Catboost emerges as the optimal performer in terms of accuracy. Notably, a substantial enhancement in the overall model performance is observed upon applying re-sampling techniques. This study introduces a novel intelligent model for detecting male fertility, utilizing Lightgbm learners alongside the SMOTE-ENN re-sampling approach. The suggested model's performance is then benchmarked against various state-of-the-art resampling techniques. This recommended approach holds the potential to support computer-aided decision-making and contribute to preventing male infertility. The research findings of this paper contribute to a deeper understanding of the impact of re-sampling techniques on predictive performance. Furthermore, they highlight that combining under-sampling and over-sampling strategies yields superior results. It's important to note that this paper solely focuses on male fertility, using two-class target labels, a limited set of features, and constrained data. Numerous unexplored research avenues remain in this domain. Besides the ensemble classifiers studied here, considering other conventional classifiers for performance comparison could further optimize model parameters.

Declarations

Ethics Declarations

Data Availability: https://archive.ics.uci.edu/ml/datasets/Fertility (Last accessed: 12 Jan, 2023)

Ethical approval and consent to participate: This article does not include any human participant studies conducted by any of the authors. Since this is a review, consent is not required.

Human and animal ethics: This study did not include any human subjects or animals.

Consent for publication: This article contains no identifying information, so it is inapplicable.
Competing interests: There are no potential conflicts of interest reported by any of the authors.

Funding: Not applicable.

Contributions: DGR and KC conceptualized the study and its methodology. KC analyzed the study. DGR wrote the original draft that was reviewed by PA and KC. KC finalized/corrected the manuscript.

References

1. El Bouchefry, K., & de Souza, R. S. (2020). Learning in big data: Introduction to machine learning. In *Knowledge discovery in big data from astronomy and earth observation* (pp. 225-249). Elsevier.
2. Hosni, M., Abnane, I., Idri, A., de Gea, J. M. C., & Alemán, J. L. F. (2019). Reviewing ensemble classification methods in breast cancer. *Computer methods and programs in biomedicine*, *177*, 89-112.
3. Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429-441.
4. Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1560-1571.
5. Geetha, R., Sivasubramanian, S., Kaliappan, M., Vimal, S., & Annamalai, S. (2019). Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier. *Journal of medical systems*, *43*, 1-19.
6. Zhu, T., Lin, Y., & Liu, Y. (2020). Improving interpolation-based oversampling for imbalanced data learning. *Knowledge-Based Systems*, *187*, 104826.
7. Desuky, A. S., & Hussain, S. (2021). An improved hybrid approach for handling class imbalance problem. *Arabian Journal for Science and Engineering*, *46*, 3853-3864.
8. Gupta, S., & Thériault, G. (2023). Do not diagnose or routinely treat asthma or chronic obstructive pulmonary disease without pulmonary function testing. *bmj*, *380*.
9. Zehra, A. C. A. R., & SATILMIŞ, İ. G. CULTURAL PERSPECTIVE ON INFERTILITY IN TURKISH SOCIETY: THE ISTANBUL SAMPLE. *Izmir Democracy University Health Sciences Journal*, *5*(3), 635-650.
10. Hazlina, N. H. N., Norhayati, M. N., Bahari, I. S., & Arif, N. A. N. M. (2022). Worldwide prevalence, risk factors and psychological impact of infertility among women: a systematic review and meta-analysis. *BMJ open*, *12*(3), e057132.
11. Ghazal, T. M., Rehman, A. U., Saleem, M., Ahmad, M., Ahmad, S., & Mehmood, F. (2022, February). Intelligent Model to Predict Early Liver Disease using Machine Learning Technique. In *2022 International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-5). IEEE.
12. Albert, A. J., Murugan, R., & Sripriya, T. (2023). Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology. *Research on Biomedical Engineering*, *39*(1), 99-

113

13. Muntasir Nishat, M., Faisal, F., Jahan Ratul, I., Al-Monsur, A., Ar-Rafi, A. M., Nasrullah, S. M., ... & Khan, M. R. H. (2022). A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. *Scientific Programming*, *2022*, 1-17.
14. Yang, F., Wang, K., Sun, L., Zhai, M., Song, J., & Wang, H. (2022). A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis. *BMC Medical Informatics and Decision Making*, *22*(1), 344.
15. Naz, H., & Ahuja, S. (2022). SMOTE-SMO-based expert system for type II diabetes detection using PIMA dataset. *International Journal of Diabetes in Developing Countries*, *42*(2), 245-253.
16. Kumar, V., Lalotra, G. S., & Kumar, R. K. (2022). Improving performance of classifiers for diagnosis of critical diseases to prevent COVID risk. *Computers and Electrical Engineering*, *102*, 108236
17. Gupta, S., & Gupta, M. K. (2022). A comprehensive data-level investigation of cancer diagnosis on imbalanced data. *Computational Intelligence*, *38*(1), 156-186.
18. GhoshRoy, D., Alvi, P. A., & Santosh, K. C. (2022). Explainable AI to Predict Male Fertility Using Extreme Gradient Boosting Algorithm with SMOTE. *Electronics*, *12*(1), 15.
19. GhoshRoy, D., Alvi, P. A., & Santosh, K. C. (2023, March). Unboxing Industry-Standard AI Models for Male Fertility Prediction with SHAP. In *Healthcare* (Vol. 11, No. 7, p. 929). MDPI.
20. Yibre, A. M., & Koçer, B. (2021). Semen quality predictive model using feed forwarded neural network trained by learning-based artificial algae algorithm. *Engineering Science and Technology, an International Journal*, *24*(2), 310-318.
21. Lin, C., Tsai, C. F., & Lin, W. C. (2023). Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: an experimental study. *Artificial Intelligence Review*, *56*(2), 845-863.
22. Islam, A., Belhaouari, S. B., Rehman, A. U., & Bensmail, H. (2022). KNNOR: An oversampling technique for imbalanced datasets. *Applied Soft Computing*, *115*, 108288.
23. Ma, J., Afolabi, D. O., Ren, J., & Zhen, A. (2021). Predicting seminal quality via imbalanced learning with evolutionary safe-level synthetic minority over-sampling technique. *Cognitive Computation*, *13*, 833-844.
24. Feng, S., Zhao, C., & Fu, P. (2020). A cluster-based hybrid sampling approach for imbalanced data classification. *Review of Scientific Instruments*, *91*(5), 055101.
25. Fujiwara, K., Huang, Y., Hori, K., Nishioji, K., Kobayashi, M., Kamaguchi, M., & Kano, M. (2020). Over-and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. *Frontiers in public health*, *8*, 178.
26. Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data. *Journal of Biomedical Informatics*, *107*, 103465.

27. Viloria, A., Lezama, O. B. P., & Mercado-Caruzo, N. (2020). Unbalanced data processing using oversampling: Machine Learning. *Procedia Computer Science*, *175*, 108-113.
28. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE.
29. Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1* (pp. 878-887). Springer Berlin Heidelberg.
30. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
31. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29.
32. Shilaskar, S., & Ghatol, A. (2019). Diagnosis system for imbalanced multi-minority medical dataset. *Soft Computing*, *23*(13), 4789-4799.
33. Hoyos-Osorio, J., Alvarez-Meza, A., Daza-Santacoloma, G., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2021). Relevant information undersampling to support imbalanced data classification. *Neurocomputing*, *436*, 136-146.
34. Bansal, A., & Jain, A. (2021, June). Analysis of Focussed Under-Sampling Techniques with Machine Learning Classifiers. In *2021 IEEE/ACIS 19th International Conference on Software Engineering Research, Management and Applications (SERA)* (pp. 91-96). IEEE.
35. Zhang, H., Zhang, H., Pirbhulal, S., Wu, W., & Albuquerque, V. H. C. D. (2020). Active balancing mechanism for imbalanced medical data in deep learning–based classification models. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(1s), 1-15.
36. Batista, G. E., Carvalho, A. C., & Monard, M. C. (2000, April). Applying one-sided selection to unbalanced datasets. In *MICAI* (Vol. 2000, pp. 315-325).
37. Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3), 408-421.
38. Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003, December). Balancing training data for automated annotation of keywords: a case study. In *WOB* (pp. 10-18).
39. Sui, Y., Wei, Y., & Zhao, D. (2015). Computer-aided lung nodule recognition by SVM classifier based on combination of random undersampling and SMOTE. *Computational and mathematical methods in medicine*, *2015*.
40. Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003, December). Balancing training data for automated annotation of keywords: a case study. In *WOB* (pp. 10-18).
41. Sen, S., Singh, K. P., & Chakraborty, P. (2023). Dealing with imbalanced regression problem for large dataset using scalable Artificial Neural Network. *New Astronomy*, *99*, 101959.

42. Jorgensen, A., Svingen, T., Miles, H., Chetty, T., Stukenborg, J. B., & Mitchell, R. T. (2023). Environmental impacts on male reproductive development: lessons from experimental models. *Hormone research in paediatrics*, *96*(2), 190-206.
43. Mishra, S., Mallick, P. K., Jena, L., & Chae, G. S. (2020). Optimization of skewed data using sampling-based pre-processing approach. *Frontiers in Public Health*, *8*, 274.
44. Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31-39.
45. Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of big data*, *7*(1), 1-45.
46. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.
47. Schapire, R. E. (2013). Explaining adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 37-52.
48. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.
49. https://archive.ics.uci.edu/ml/datasets/Fertility

Figures

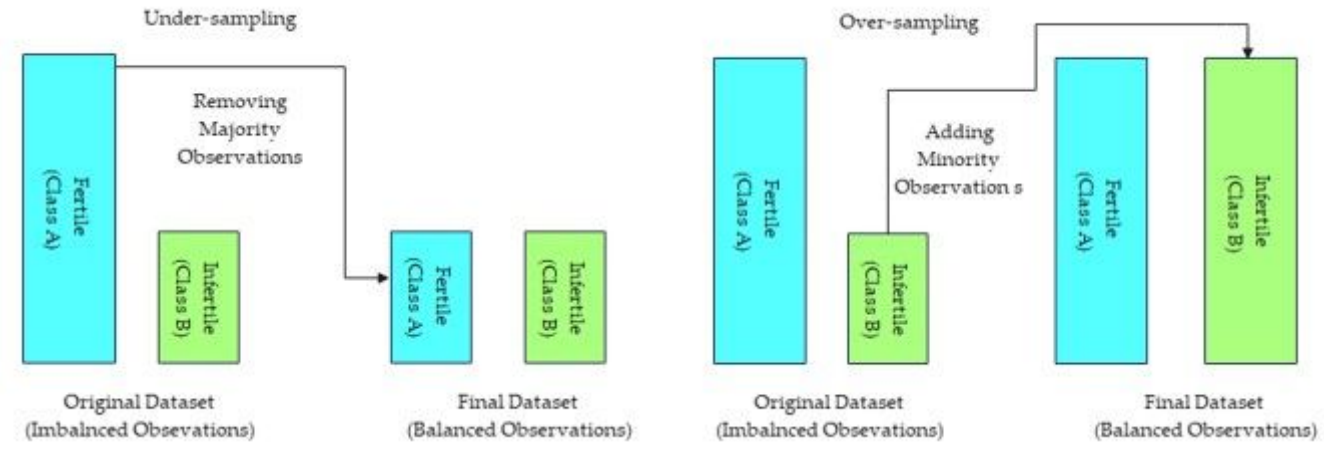

Figure 1

The role of samplers in data balancing

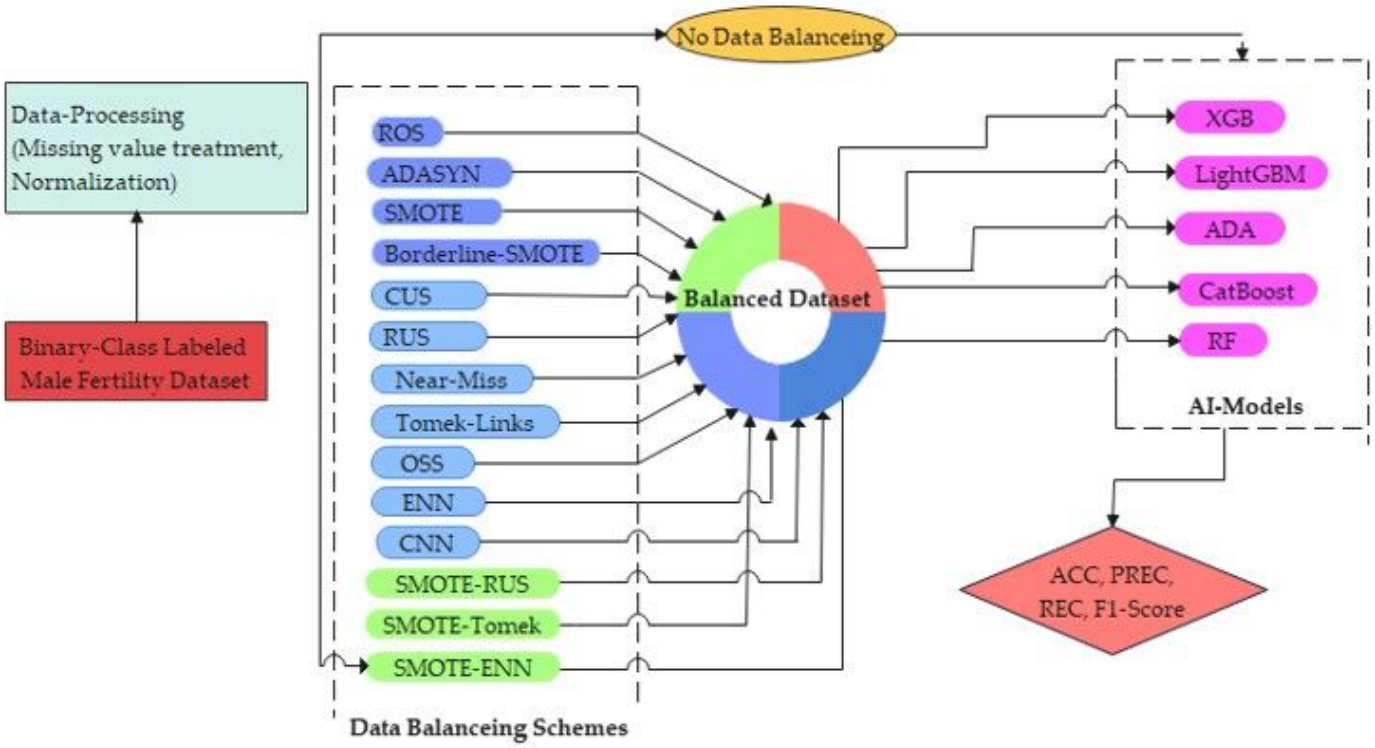

Figure 2

The framework of data re-sampling using ensemble learners

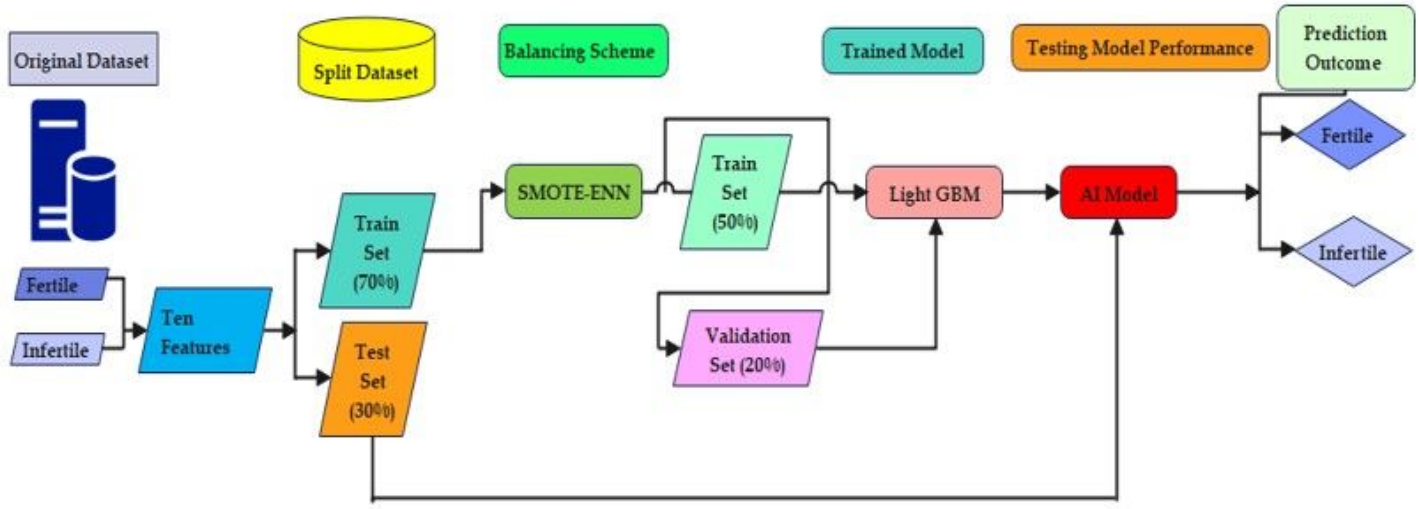

Figure 3

Predictive model for male fertility using SMOTE-ENN

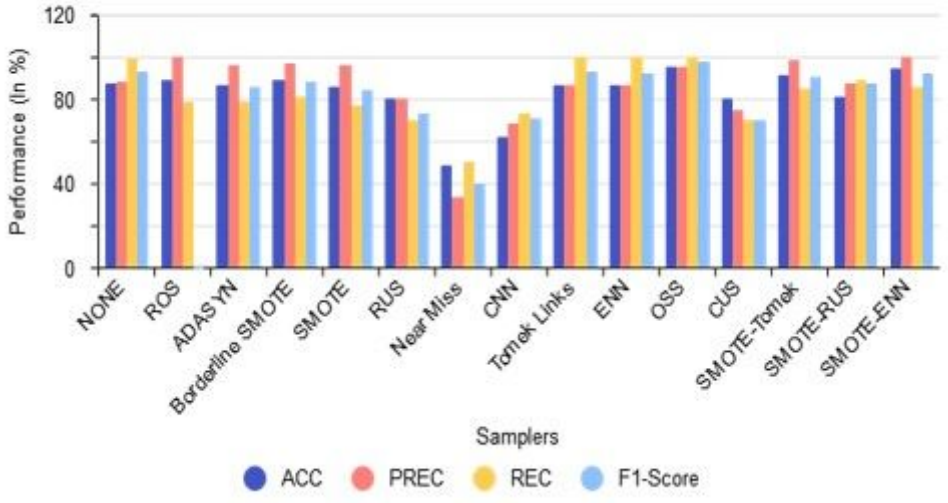

Figure 4

CatBoost Classifier Performance Evaluation

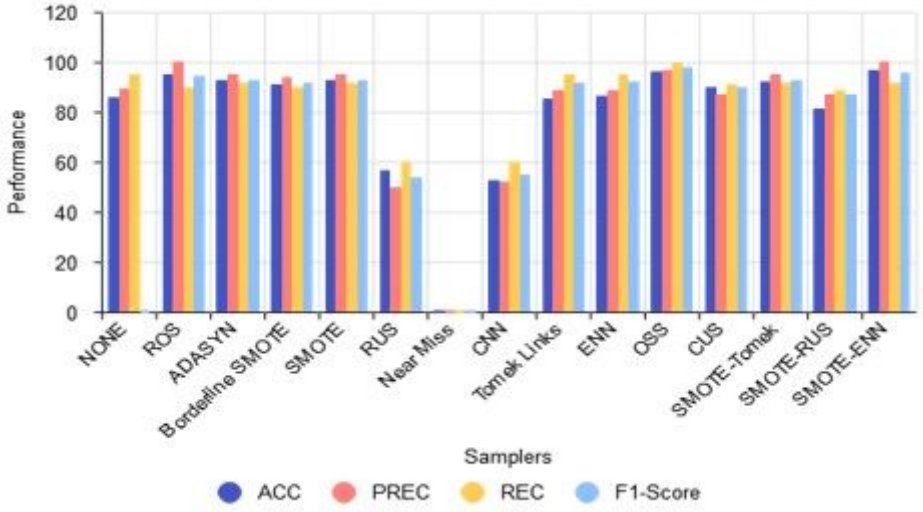

Figure 5

XGB Classifier Performance Evaluation

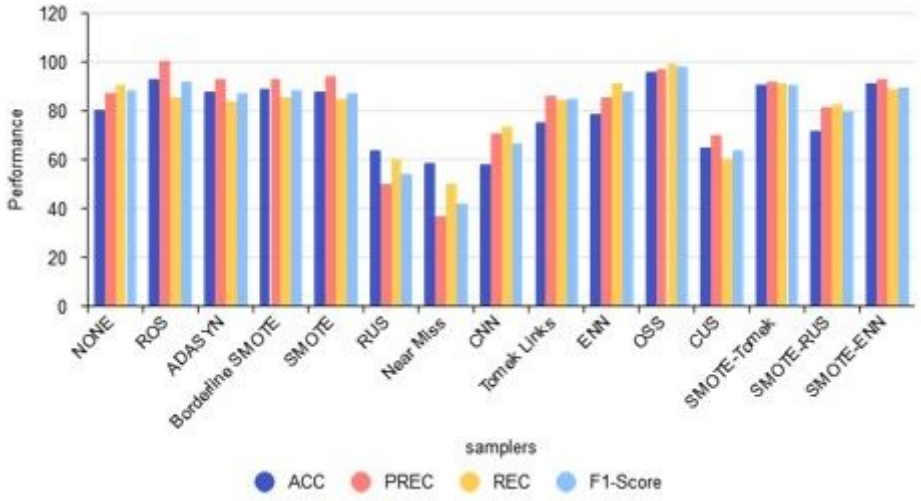

Figure 6

ADA Classifier Performance Evaluation

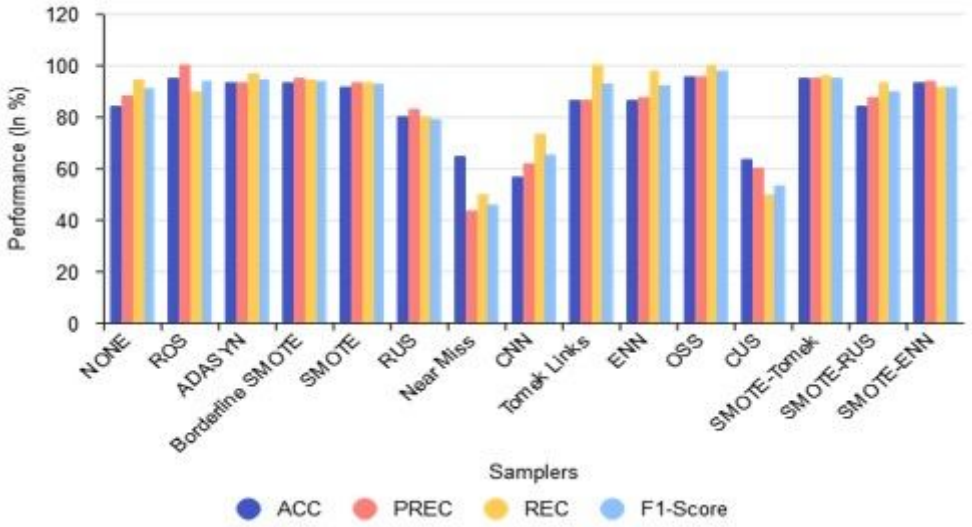

Figure 7

RF Classifier Performance Evaluation

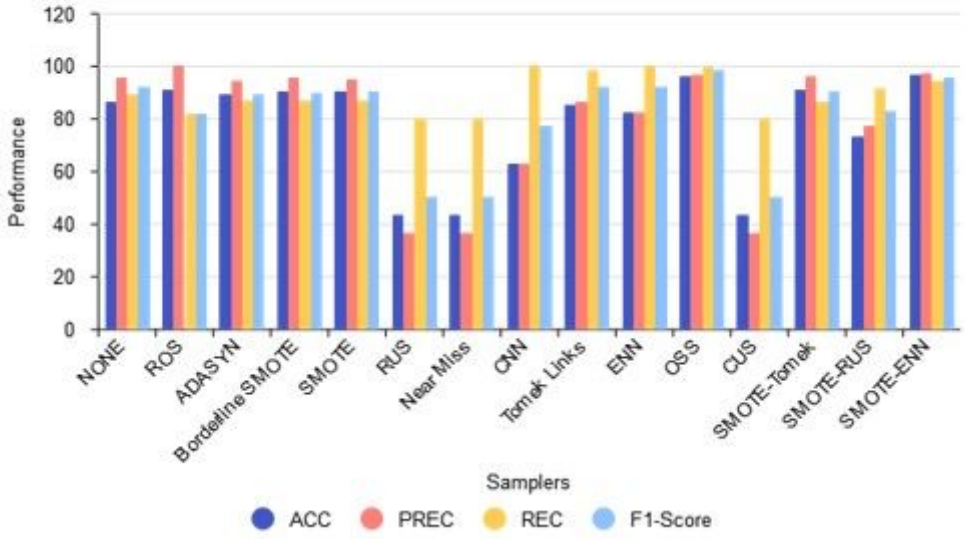

Figure 8

LightGBMClassifier Performance Evaluation