VOL. 16, NO. 21, NOVEMBER 2021 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A RESAMPLE-SMOTE BALANCE WITH RANDOM FOREST FOR IMPROVING SEMINAL QUALITY PREDICTION IN HEALTHCARE INFORMATICS

Raihani Mohamed[1], Abdul Rafiez Abdul Raziff[2] and Sabri Mohd. Nasir[3]
[1]Department of Computer Science, FCSIT, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[2]Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Gombak, Malaysia
[3]Intel Technology Sdn. Bhd. (PG12), Bayan Lepas, Penang, Malaysia
E-Mail: raihanimohamed@upm.edu.my

ABSTRACT

Previous research has shown that men seminal fertility rates significantly decreased in the last twenty years due to health status, life habits, and environment. Thus, seminal quality prediction in men's fertility has become a very demanding decision support system in the biomedical engineering field. Existing solutions focused on producing the accuracy of the prediction model. They also acknowledged that there were problems of imbalance class, ambiguity, and noise of the sample dataset. However, the real issues of the domain were still vague. A framework consisting of Resample-Smote Balance with the state-of-art Random Forest (RF) method is proposed to overcome the real-world problem and tested on the Fertility dataset. This method is introduced to alleviate the ambiguity and noise of the sample set. Subsequently, the Smote method is applied to balance the size of the dataset before the classification phase using RF is taken place. The performance of the proposed model is compared with other state-of-art classifiers such as MLP, SVM, and DT. Consequently, this work manages to produce the best accuracy model with 98.2% that can improve the ambiguity and noise from the sample set at the same time competent to handle the imbalance class of the dataset. The RF also boosts the accuracy model compared with other classifiers due to its capability to produce the most probable class from its majority-voting task as output.

Keywords: seminal quality men fertility, fertility dataset, imbalance class, resample, smote, random forest.

INTRODUCTION

Study in healthcare informatics involves acquiring data from patients under examination and executing several techniques on the data. Accurate patients' diagnosis and effective treatments control are standards of the quality services that should doctors, and caretakers provide to the patients. Inaccurate diagnosis and poor services can lead the patients in jeopardy, and thus these reasons are intolerable [1]. There are millions of patients' treatments, and records are stored in many databases. By utilizing data mining techniques, important healthcare-related questions can be answered. Furthermore, information integration is crucial to provide proper analysis of raw data in delivering the right decision. The data are extracted to form useful information and discover hidden details to make effective decision-making for the user from massive databases. An enormous amount of data are gathered including lung cancer [2], asthma, tuberculosis, diabetes, heart disease data [1], men and women fertility data [3], [4]. Seminal quality is referred to male reproductive fluid, containing spermatozoa in suspension [5]. Seminal quality prediction in men fertility has become a very demanding decision support system in the biomedical engineering field. Its fertility is essential to measure potential male fertility due to the notable decline in fertility rates [6]–[8]. It also involves the incorporation of women into the labor force and consequent delay in the age at which one decides to have descendants, as well as the widespread use of contraceptives [9]. Although semen analysis alone cannot determine whether a male can have descendants, however, it is a good predictor of male fertility potential [10]-[12]. Hence, a semen analysis is also essential to evaluate the quality of the semen donors [11]. Applications with a decision support system should be able to grant adequate information from the perspective of the possible influence of environmental factors and life habits on semen quality [3]. There are several challenges that need to be tackled to produce a sound decision support system for a medical solution. To predict the accuracy of potential male fertility, the problems of the implementation need to be investigated. It is either from the data level or at the implementation level. One of the challenges in bioinformatics and medical diagnosis is a class imbalance. It is a common problem in biometrics and medical datasets [13], [14]. Hence, the standard learning algorithms cannot perform well in this case due to ambiguity in the data boundary, and data are noisy. Thus, many efforts are proposed in tackling the mentioned problem such as Clustering-Based Decision Forests (CBDF) [15], genetic algorithm (GA) [16], and Particle swarm optimization (PSO) [17], that is carried out before the classification phase. However, the chosen methods might not consider the sample size that could be further explored mainly in promising an outstanding model performance. This condition occurs when the number of examples that represent one class is much lower than the ones of the other classes [18]. Thus, this may cause inadequate accuracy of the model. In this context, the prediction model should consider the alleviation of class imbalance problems from the testing data. Regarding the complexity of the problem domain, the model is required to handle the class imbalance problem.

There are several contributions to this study. A framework consisting of proposed methods with a Resample-Smote balance is proposed at the pre-processing phase to cater to the unbalanced class in the real-world dataset from the UCI Machine learning repository [3]. The class prediction is proposed using the state-of-art ensemble decision tree of Random Forest (RF) to improve the model accuracy. The proposed method can automatically detect and resolve the abovementioned problems to attain overall satisfaction in the decision support system in healthcare applications. Finally, we compare the results with the several state-of-art classifiers such as SVM, DT, and MLP to measure its performance. The rest of this paper is organized as follows. Section 2 explores the previous work related to healthcare informatics. Meanwhile, Section 3 explains the materials and methods proposed in this study. Section 4 presents the results and discussions. Section 5 clarifies the conclusion of the overall conducted experiments.

LITERATURE REVIEW

Health informatics can also be addressed as healthcare informatics, biomedical informatics, medical informatics, nursing informatics, clinical informatics as well as biomedical informatics is considered as information engineering that is applied to the field of healthcare for patient healthcare information. Informatics in healthcare is an extensive area that involves achieving the data of the patient under test and performing various techniques on the data. Patient data will vary in terms of the data that are obtained from patients that are clinically validated, including data that are recorded by a medical transcriptionist. Nowadays, the need for disease prediction is considered a necessity in healthcare informatics. Many patients undergo a series of checks in-order to know the possibilities of getting serious diseases such as heart, kidney, cancer, etc. This functionality falls into the category of predictive analysis. One of the technologies that can be employed, such as data mining can be very useful in transforming raw data from a massive database into the right decision of predicting a possible disease. The sufficient quality of service resembles the ability of correct diagnostic together with proper treatment and vice versa [19]. Millions of patients' records can be stored and processed using data mining methods. With the availability of integrated data, disease prediction or diagnostic can be made at higher accuracy that can increase a better treatment.

According to Cornwallis & O'Connor (2009), male fitness can be measured by its efficiency by looking at its ejaculation productivity due to males apportioning a different number of sperm to ejaculates [5]. Semen emission is the formed of greyish, white bodily fluid that is secreted by the gonads of males. It consists of sperm or the spermatozoa and fructose and other enzymes that help the sperm to survive to facilitate successful fertilization. Nowadays, prediction on seminal quality of men's fertility is essential for decision support system applications and tools. However, there are limitations and issues that need to consider and rectify for further improvement on the services based on the users' needs. Useful model in prediction men fertility, class imbalance complexity and dimension of sample size are the main challenges in determining the men fertility prediction. These issues required a high commitment to producing the best quality of application and services. The standard issue in biometrics and healthcare data is the class imbalance problem. When the samples are imbalanced, this scenario will influence model accuracy. Hence, to tackle the issue, there is a need to acknowledge the level of its complexity, whether it is under easy or difficult imbalanced class. The unbalanced class scenario that lead to the low accuracy of the model showed clearly that the complexity of the data might exist due to the majority class samples overlapping with the samples of minority class [15]. This condition is also addressed as noisy in samples data. The most significant concern regarding class imbalance is the dimensionality of sample size. There are issues of majority and minority class. A dataset can be considered imbalanced when the class proportions are highly skewed. It occurs when there are many more examples of some classes than others [20]. Datasets that are imbalanced from the real-world setting are standard in biometrics, gene recognition, and medical datasets [13], [14], [21]. In binary-two class of seminal quality of men fertility scenario, majority class occurs when most of the available data has the negative number referred as majority observations outnumbers of positive (minority) observations and otherwise. These scenarios prevent the classification model from performing in high accuracy without prior treatment of the input data before the classification stage. There are reasons for data treatment for the imbalanced class problem. According to Visa & Ralescu (2005), the reasons are as follows: 1) Standard classifiers are driven by accuracy, so the minority class may be ignored. 2) Standard classification methods operation under the assumption that the data sample is a faithful representation of the population of interest, which is not always the case with imbalanced problems. 3) The classification methods for imbalance issues ems should allow for errors coming from different classes to have different costs [22]. One of the recent techniques and approaches to tackle the class imbalance problem is at data level approaches [18]. This category considers the data treatment at the pre-processing stage to transform the imbalanced problem into balanced data by influencing the class distribution [23].

This work presents the approaches used to tackle the imbalance class problem using the Resample-Smote method. Resampling is a sequence of methods used to reconstruct the sample datasets, including training and validation sets. This method is referred to economically using a collected dataset to improve the estimation of the population parameter and quantify the uncertainty of the evaluation [24]. The Synthetic Minority Over-Sampling Technique (Smote) is the approach that has "oversampling" in its name. Thus, it does not add copies of existing instances but creates new artificial examples using the procedures [23]. It has been proved as a handy tool to deal appropriately with imbalanced datasets [23].

This method tries to avoid over-fitting using a random procedure to create new samples, but this can introduce noise or nonsensical samples. However, Smote remains highly regarded because of its simplicity. Additionally, this work also highlighted the significance of the supervised learning method using Random Forest (RF) as well as Decision Tree (DT), Multi-layer Perceptron (MLP), and Support Vector Machine (SVM). In case of predicting seminal quality in men's fertility, other methods that have been reported in previous works include Radial Basis Function Neural Network (RBFNN) [25], and Artificial Neural Network (ANN) [26]. The Clustering-Based Decision Forest (CBDF) is introduced as unsupervised learning to tackle the class imbalance problems in seminal quality prediction [15].

MATERIAL AND METHOD

The details of work regarding this study are explained in this section. The datasets, pre-processing stage, and classification with detailed explanation are discussed in the following subsections.

A. Dataset

The dataset is publicly available, namely the Fertility dataset that is retrievable from the UCI Repository dataset [3]. There are consisted of 100 volunteers that provide the semen sample analyses according to the World Health Organization based on 2010 criteria [27]. The samples were obtained from anonymous as well as young and healthy university students' ages between 18 and 36 years old. Based on the criteria, the sperm concentration is related to socio-demographic data, environmental factors, health status, and sample life habits. The features for each attribute are consisting of individual lifestyles such as alcohol consumption, smoking habit, and the number of hours sitting per day. Other data such as the environmental factors (including but not limited to season of the data collection performed), high fevers of the last year, and accidents were also collected.

B. Pre-Processing Phase

The dataset includes two classes as semen quality: "normal" indicated as (N) and "altered" as (O). There are 12 samples as altered and 88 samples as normal-the altered class regarded as sperm concentration parameter. The objective of this work is to predict the unknown sperm concentration result y on new testing data x' with the learning model $Y = f(X)$ built upon training dataset D. To predict the seminal quality, such as sperm concentration can be regarded as a binary classification problem. Given the training dataset of n labelled observations $D = \{(x_i, y_i)|i = 1, 2, \cdots, n\}$, where $x = \left(x_1, x_2, \cdots, x_p\right)^T \in X$ is a vector of attributes including life habits, environmental factors, etc., $y \in Y = \{0, 1\}$ is the class label (0 for "Normal" sperm concentrations and 1 for "Altered" from the Fertility dataset D). The dataset is so unbalanced and exposed to ambiguity and data noisy. Therefore, the data is required for treatment at the pre-processing phase before the classification takes place.

This pre-processing phase involves two main stages: Resample and Smote method. The detailed explanations of the proposed methods will be discussed in the following sections.

a) Resample method: This method is proposed to cater to the ambiguity and noise in the samples of the Fertility dataset. The procedure is to perform random sampling with replacement and maintain the distribution of class values seen in the input data. The original dataset must fit entirely in the memory. There is an option –v for the no replacement option to create a disjoint test set then the difference in performance probably indicates that the original separate test set has a different distribution than that of the training data.

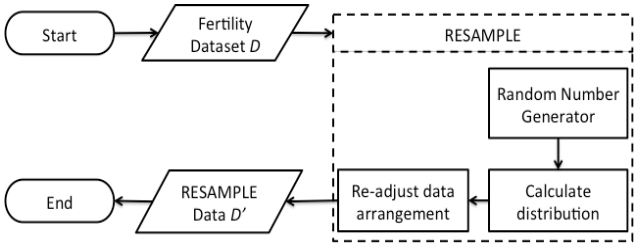

Figure-1. Resample method process flow.

Figure-1 presents the new subsample to create dataset D' to tackle the problem mentioned. The Resample method caters to the random number generator process, then calculates distribution before re-adjust the data arrangement. The Resample dataset D' is then ready for the next stage.

b) Synthetic minority over-sampling (Smote) method: In this research, Smote method is proposed to tackle the imbalance class from the Fertility dataset. It has "oversampling" in its name. Thus, it does not add copies of existing instances but creates new artificial examples using the procedures [23]. The procedures are as follows: 1) A member of the minority class is selected, and its k nearest neighbors (from the minority class) are identified. One of them is chosen randomly. 2) Then, the new example added to the set is a random point in the line segment defined by the member and its neighbor. A value of $k = 5$ has been recommended and is the one used in many studies. This method is considering the process of generating a new dataset from the minority class and is widely used in the dataset with binary classes. The procedure is processed at the training set, which the number of instances for a minority class will be increased [28]. The pre-processed dataset with Resample method is denoted as dataset D'. Then D' will continue its second phase as in Figure-2 illustrates the Smote process flow that took place after the first stage with the amount of Smote percentage $p = 100$ is proposed. This Smote method is applied three times iteratively.

Mainly, three inputs involve in generating the synthetic samples: the number of minority class samples, amount of Smote percentage, p and number of nearest neighbors to be used. Algorithm 1 indicates the pseudo-code of Smote method [29]].

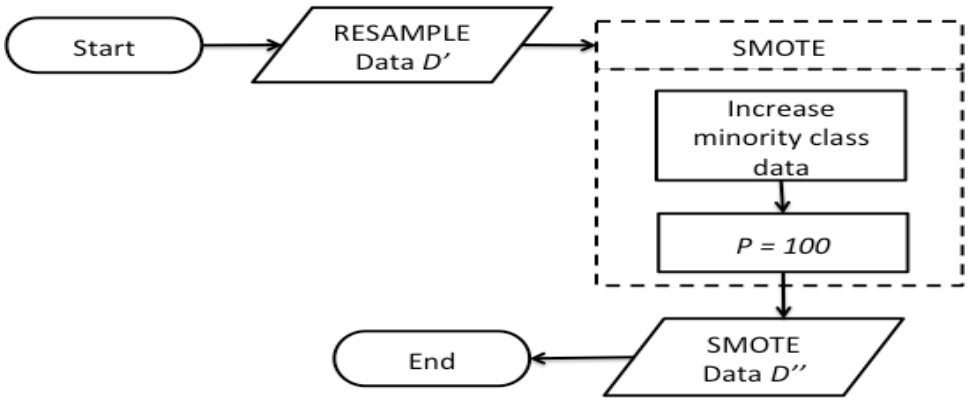

Figure-2. Smote method process flow.

Algorithm 1: Smote	
Input: Z, samples of Fertility Dataset D' Output: New synthetic instances generated from Z_+ and formed new dataset D''	
1.	Procedure generating synthetic samples from Z_+
2.	While $n < N$ do, Iterate through all N in Z_+
3. 4.	Find its k-nearest neighbor Randomly select some of the neighbor
5.	for a line that is connected between neighbors, generate new synthetic samples.
6.	*end* while
7.	*end* procedure

Both methods: Resample and Smote are applied at the pre-processing stage to the data. After this procedure, the dataset D'' simultaneously needs to continue to the next classification stage.

C. Classification Phase

After this stage, the dataset D'' is the pre-processed data using the Resample-Smote balance method that are proposed in this study. The classification method that is proposed is RF. Then the other classifiers are also selected to compare its results such as MLP, SVM, and DT. Details explanations on each classifier are as follows:

a) Random forest: This research proposed RF as a classifier due to RF method being capable of ensembling many of the individual decision trees to improve model prediction accuracy. A decision tree is a statistical model that relates the response of the input data as features X, seminal quality attributes to binary class label Y, and makes recursive for binary partitioning. The data is partitioned at each node into two similar branches using simple rules (greater or less than) and repeated at each new node and stop with specific criteria reached [30], [31]. The n-trees node is created randomly according to the chosen number of trees. Each instance is predicted by using each decision tree that has been created, and the class that receives the highest vote is predicted as a final class. RF changes dynamically based on how the classification trees are constructed.

Hence, each of the trees is generated using different bootstrap from the sample data. Meanwhile, each node is separated using the best randomly chosen node among the subset of predictors. Unlike RF, bagging uses a standard decision trees split mechanism where each node uses the best split among all the variables. This ensemble classifier model has proven their its effectiveness as compared to other ordinal classifiers such as SVM and ANN.

In a simple explanation, let M is the total number of features x, and then the number of sample features can be close to $\sqrt{M}$, $½\sqrt{M}$ or $2\sqrt{M}$. RF can be presented as follows: $\{h(x, \emptyset_k), k = 1, \ldots\}$, where $\emptyset_k$ are independent and identically distributed random vectors. Hence, each tree casts a unit vote for the most popular class at input x. The algorithm is illustrated in Algorithm 2 [32].

It creates the final tree model that has J terminal nodes that correspond to the disjoint regions R_j that cover the space of all values of the features X. The response $f(X)$ is predicted with a constant b_j for $X \in R_j$. The formation is represented in Equation 1 [33].

$$f(x) = \sum_{j=1}^{J} b_j I\{X \in R_j\} \quad (1)$$

Where: $I = \begin{cases} 1 & true \\ 0 & otherwise \end{cases}$

b_j = the average over all the response values of the training data in the leaf R_j.

Algorithm 2: Random Forest (RF) creations in the training stage	
Input: Training dataset $\boldsymbol{D''} = (\boldsymbol{x_i}, \boldsymbol{y_i}), \ldots, (\boldsymbol{x_n}, \boldsymbol{y_n})$, features $\boldsymbol{X}$, number of trees in forest $\boldsymbol{B}$ Output: Final tree model	
1	Function$\boldsymbol{RandomForest(D'', X)}$
2	$\boldsymbol{H \leftarrow 0}$
3	For $\boldsymbol{i \in 1, \ldots B}$ do
4	$\boldsymbol{D^{(i)}}$ ←a bootstrap sample from $\boldsymbol{D}$
5	$\boldsymbol{h_i \leftarrow RandomizedTreeLearn(D^{(i)}, X)}$
6	$\boldsymbol{H \leftarrow H \cup \{h_i\}}$
7	end for
8	Return $\boldsymbol{H}$
9	end function
10	Function $\boldsymbol{RandomizedTreeLearn\ (D'', X)}$
11	at each node:
12	$\boldsymbol{x}$ ←very small subset of $\boldsymbol{X}$
13	Split on best feature in $\boldsymbol{x}$
14	Return $\boldsymbol{The\ learned\ tree}$
15	end function

The wide range of relationships among the nodes can be captured from simple decision rules. Hence its hierarchical structure provides the capacity to account for interactions among the independent variables. With boosting, pruning and bagging methods, the optimal tree and idiosyncratic distinctions can be generalized when the new data is observed. At the testing stage, new data arrives is then classified based on the sign of $f(x)$ to produce a final result of the classification problem.

b) Multilayer perceptron (MLP): MLP contains nodes that are arranged in layers. Each node is built with an activation function and connected with links. Each link has its own weights that will determine the value of the output of the nodes. Algorithm 3 presents the algorithm on how it works. In MLP, the dataset is then passed into the network line-by-line. At the hidden layer, the weight is adjusted according to the initial weight and bias. Then the output of the hidden layer will be calculated using a sigmoid function. This process will happen again at the output layer. Then the error will be calculated, adjust weight, and will be repeated until the global error can achieve 0. This process is called backpropagation.

Algorithm 3: Multilayer Perceptron (MLP)	
Input: Training dataset $\boldsymbol{D''} = (\boldsymbol{x_i}, \boldsymbol{y_i}), \ldots, (\boldsymbol{x_n}, \boldsymbol{y_n})$, features $\boldsymbol{X}$ Output: MLP model with adjusted weights	
1	Function $\boldsymbol{MLP(D'', X)}$
2	Init Values
3	Pass input into the network
4	Compute summation in hidden layer: $\boldsymbol{V = w_i x_i + bias}$
5	Compute output using activation function: $\boldsymbol{Y = sigmoidFunction(v)}$
6	Compute summation in hidden layer: $\boldsymbol{V = w_i x_i + bias}$
7	Compute output using activation function: $\boldsymbol{Y = sigmoidFunction(v)}$
8	Compute error: $\boldsymbol{error = target - output}$
9	Adjust weight
10	If global error != 0, go to step 2
11	end function

c) Decision tree (DT): In this work, the algorithm used is J48 that is adopted by [34]. The working algorithm can be seen in Algorithm 4. From the code above, the first stage is to initialize the tree structure. The process of creating a tree is in a recursive structure that firstly, will create nodes. Then it will create branches and finally will form into a tree. Calculation on the information-theoretic will be performed on all the attributes if it is split to a. The best attribute will be initialized, and a tree will be created. Sub-dataset will be created from *D* based on the best attribute. This creating of

tree process will be stopped when the dataset is "pure" which means that a definite value can be predicted from D.

d) Support vector machine (SVM): SVM is one of the most popular machine-learning algorithms in solving a binary classification problems. It works by finding an optimal solution in separating hyper-plane between the classes. In this work, LibSVM is used with the radial basis function (RBF) as the kernel type. It maps the samples to a higher dimensional space.

D. Modeling Phase

In the process of defining and analyzing data requirements to support the accuracy of the model performance, three main data models evolve as this study progress from the initial datasets. It contained three models to represent the different levels of abstraction, including:

a) Raw data model, where raw fertility data to be represented as a stream features a sample,

b) Resample model, which features are alleviated with Resample method to overcome the data ambiguity and noise,

c) Smote model, where the features are balanced for prediction of seminal quality of men fertility domain.

Resample-Smote balance model, where the features are alleviated then balance for prediction.

EXPERIMENTAL RESULT AND DISCUSSIONS

Generally, the complete dataset is divided into two subsets: the training set and the test set. Then, the ten-fold cross-validation has been applied and used to assess the performance of every method proposed as other authors also applied the same cross-validation approach [3], [16].

The data has been divided into ten sets $(S_1, S_2, .., S_{10})$ and the ten experiments performed are shown in every section as follows. The test data are chosen randomly from the initial data and the remaining data from the training data. The method is called ten-fold cross-validation since this process has been performed ten times. Subsequently, the classification model is tested under four different types of classifiers, namely Random Forest (RF), Decision Tree (DT), Multilayer Perceptron Neural Network (MLP) and Support Vector Machine (SVM) are applied at the training stage. The parameter is set up under RF with 100-batch tree size, libSVM with kernel type Radial Basis Function (RBF), DT namely addresses as J48 with three numbers of folds, and MLP with learning rate at 0.3 and one hidden layer.

A. Experimental Result using Different Classifiers

This experiment is implemented to present the performance of the proposed classifier methods such as MLP, J48, RF, and SVM. Firstly, we will show the result in the confusion matrix so that each prediction against the actual value can be reviewed and analyzed. Then the model performance based on the accuracy, precision, recall, and f-measure is also given. These results are presented in a separate section as in the following sections.

Table-1 shows the values of normal (N) and altered (O) class values of the predicted dataset under different classifiers. All the classifiers results are tabulated its TP and FP on Normal (N) with TN and FN for Altered (O).

Table-1. Confusion matrix result of different classifiers.

ACTUAL	PREDICTION							
	MLP		J48		SVM		RF	
	N	O	N	O	N	O	N	O
N	84	4	85	3	88	0	84	4
O	6	6	12	0	12	0	10	2

Results based on different classifier methods are displayed in Table-2. Accuracy using MLP indicates that the highest achievement model with 0.9 followed by J48, SVM, and RF with 0.89, 0.88, and 0.86, respectively. Hence, recall results are as same as accuracy results. The results on precision are 0.893, 0.864, 0.878 and 0.826 for MLP, J48, SVM, and RF respectively. Meanwhile, F-measure indicates 0.896, 0.873, 0.879, and 0.839 respectively.

Table-2. Classification accuracy performance using different classifier.

Classifier	Accuracy	Precision	Recall	F-Measure
MLP	0.900	0.893	0.900	0.896
J48	0.890	0.864	0.890	0.873
SVM	0.880	0.878	0.880	0.879
RF	0.860	0.826	0.860	0.839

B. Experimental Result with Resample Method

This method is proposed to alleviate the imbalance of data due to the ambiguity of data boundary and noise in the dataset. Table-3 indicates the result of TP, TF, FP and FN on the actual and prediction results of the Fertility dataset. The proposed Resample method changes the result from before. With classifier RF, the result is showing the highest compare with other classifiers. The MLP falls into third place after RF and SVM, while J48 obtained the lowest result among others.

Table-3. Confusion matrix result of different classifiers using the Resample method.

ACTUAL	PREDICTION							
	RF		SVM		MLP		J48	
	N	O	N	O	N	O	N	O
N	88	2	90	0	83	7	87	3
O	3	7	10	0	4	6	8	2

The performance model of the Fertility dataset using the Resample method is measured with accuracy, precision, recall, and f-measure. Table-4 indicates the result based on accuracy, precision, recall, and f-measure when the proposed method utilizes the Resample method. Results under RF show the significant result with 0.95, 0.948, 0.95, and 0.949 under accuracy, precision, recall and f-measure respectively. Hence, the results are followed by SVM, MLP, and J48.

Table-4. Performance model using the Resample method under different classifiers.

Classifier	Accuracy	Precision	Recall	F-Measure
RF	0.950	0.948	0.950	0.949
SVM	0.900	0.880	0.900	0.890
MLP	0.890	0.905	0.890	0.896
J48	0.890	0.860	0.890	0.941

C. Experimental Result with Smote Method

This experiment is carried out to show the proposed method using Smote method. This method is proposed to balance the dataset that is an imbalance between the minority and majority data at the pre-processing phase. With the amount of Smote percentage $p = 100$ that is applied three times, the new total examples of the dataset are 184 from the original of 100 examples only. The results are tabulated in the confusion matrix and its performance model in the sections as follows.

Table-5 shows the results in the confusion matrix table under different classifiers when imposed with Smote method at the pre-processing stage.

Table-5. Confusion matrix result of different classifiers using Smote method in the Fertility dataset.

ACTUAL	PREDICTION							
	RF		J48		MLP		SVM	
	N	O	N	O	N	O	N	O
N	83	5	83	5	69	19	61	27
O	8	88	5	87	4	92	12	84

From the experiment, the results are changing when different classifiers are tested. Yet, RF remains the highest compared with others. However, J48 is becoming the second higher, followed by MLP, and SVM. The results under different classifiers are presented in Table-6. Smote with RF shows the significant result with 0.929, 0.930, 0.929, and 0.929 under accuracy, precision, recall, and f-measure respectively. The recall results for J48, MLP, and SVM classifiers were obtained as same as from the accuracy result. Meanwhile, precision result is 0 0.905, 0.884 and 0.794 for J48, MLP and SVM correspondingly. F-measure results are 0.905, 0.874 and 0.786 for J48, MLP and SVM respectively. Hence the results are still lacking compared with the result of Resample method before.

Table-6. Result using Smote method under different classifiers.

Classifier	Accuracy	Precision	Recall	F-Measure
RF	0.929	0.930	0.929	0.929
J48	0.924	0.925	0.924	0.924
MLP	0.875	0.884	0.875	0.874
SVM	0.788	0.794	0.788	0.786

D. Experimental Result with Resample-Smote Balance Method

This experiment is carried out to present the proposed Two-Stage Resample-Smote Balance Method using the Fertility dataset to predict the seminal quality of men's fertility. The proposed method is taken place at the pre-processing stage. Resample method is applied at the first phase then followed by Smote method with the amount of Smote percentage, $p = 100$ with three times iteratively at the second phase. The result of the proposed Resample-Smote method under different classifiers is presented in Table-7 in the form of a confusion matrix.

Table-7. Confusion matrix result of different classifiers using the proposed Resample-Smote balance method.

ACTUAL	PREDICTION							
	RF		J48		MLP		SVM	
	N	O	N	O	N	O	N	O
N	89	1	84	6	83	7	79	16
O	2	78	1	79	1	79	17	63

Table-8 presents the results under different classifiers. The highest result was obtained for the proposed Resample-Smote with RF. Its result shows the most significant result with 0.982 under accuracy, while precision, recall, and f-measure also indicate the same result. The true positive (TP) result is 0.982 and the false positive (FP) 0.018 correspondingly. Meanwhile, J48 falls after RF that is followed by MLP, and SVM with the accuracy of 0.959, 0.953 and 0.806.

Table-8. Results using A Resample-Smote Balance method.

Classifier	Accuracy	TP	FP	Precision	Recall	F-Measure
RF	0.982	0.982	0.018	0.982	0.982	0.982
J48	0.959	0.959	0.038	0.959	0.959	0.919
MLP	0.953	0.953	0.043	0.955	0.953	0.953
SVM	0.806	0.806	0.196	0.806	0.806	0.806

E. Overall Results and Comparisons

This section is explicitly highlighting the comparison of the best results using an RF classifier with different methods is compared due to its significant result with other classifiers such as MLP, SVM, and DT. Figure-3 presents the results of different methods experimented to furnish the most excellent model for the prediction of seminal quality in men's fertility.

We can observe that from the results using Smote, and the accuracy result is 0.92. However, when we impose the Resample method, the result is higher than using Smote. Consequently, we proposed to combine Resample-Smote to balance the data before the classification phase, the result of the proposed model achieves the highest among other methods before. As mentioned, the Fertility dataset consists of an imbalanced class problem. With this consideration, the improvement of classification performance in this work with the exploration of the data structure is regarded. Subsequently, the Resample is proposed to alleviate the imbalance sample from ambiguity and noise. Thus, Smote is introduced to balance the class in the dataset at the same time RF as a classifier that is initiated and capable of producing the most probable class from its majority-voting task as output.

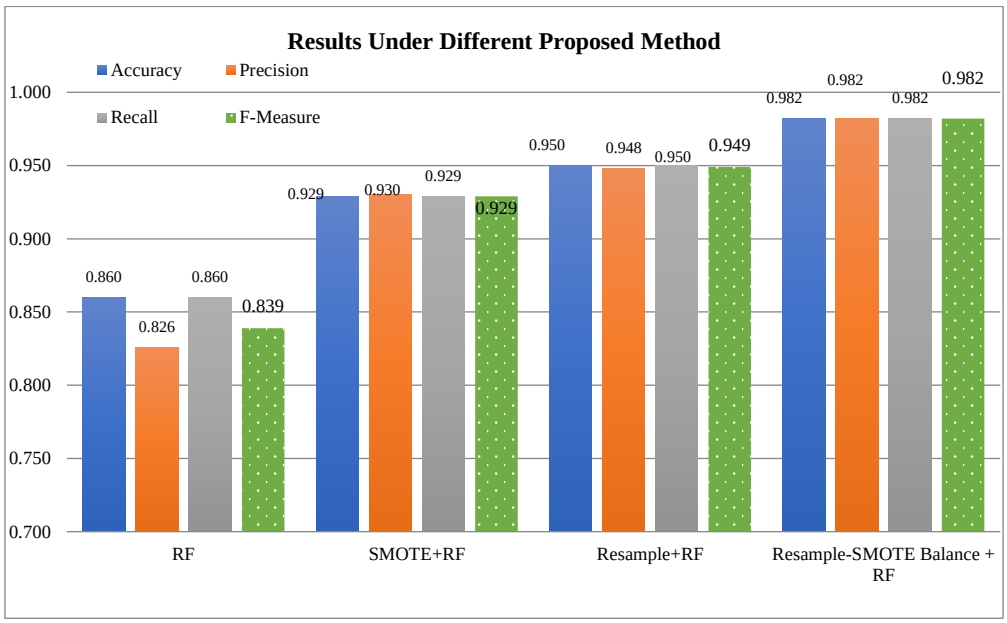

Figure-3. Comparison of results under different methods using RF classifier.

F. Comparison with Benchmarked Works

This section discusses the comparison between previous works and the proposed method. The significant difference of the proposed method with the earlier works is the methods that are proposed different based on the objective of each study on the same dataset; the Fertility dataset. We are comparing with the same supervised learning classification approach that they are proposed previously [3], [16], [17] to achieve an accuracy model based on the method proposed respectively. The achievement based on the accuracy model is illustrated in Figure-4.

As mentioned, the comparisons are based on the accuracy results achieved. However, based on the objective, Gil *et al*. (2012) introduce the MLP method to make prediction accuracy values based on AI technique. Thus, their proposed method successfully achieved 0.86 accuracies of the performance model [3]. Meanwhile, Sahoo & Kumar (2014) propose the SVM+PSO method using the feature selection technique on 100 samples of the Fertility dataset with nine attributes with two classes and obtained 0.94 [17]. Bidgoli et al. (2016) produce a similar objective with our proposed method, a Resample-Smote balance method that is to tackle the class imbalance of the Fertility dataset. However, the method proposed is slightly different. They propose to use the number of neurons, and the learning rate should be optimized in the learning phase to reach the best performance [16]. Hence, the Genetic Algorithm (GA) is applied to find the appropriate artificial neural network structure and parameters.

Consequently, MLP that is combined with GA obtained 0.9386 accuracies gives better performance compared with the previous works. Thus, the proposed Resample-Smote balance method with RF classifier successfully achieved 0.982 accuracy superseding the others. With the Resample-Smote balance method using RF classifier, we manage to produce the best accuracy model that is capable to alleviate the ambiguity and noise from the sample set at the same time competent to handle the imbalance class of the dataset. The RF also boosts the accuracy model compared with MLP, SVM, and DT due to its capability to produce the most probable class from its majority-voting task as output.

CONCLUSIONS

Seminal quality prediction in men's fertility is motivated by the demand to improve the decision support systems and application in healthcare informatics. Apparently, the proposed framework is differentiated from previous work in the domain area due to techniques in problem-solving and the limitation of the model in terms of problems tackling that are relevant to the methods proposed. The design and development of the framework are in a practical implementation with the RF method is proposed. It is tested in a real-world UCI Repository dataset, namely the Fertility dataset. This dataset is collected over 100 participants with 100 examples, nine attributes, and a binary class dataset. The Resample-Smote balance method to cater to the problem mentioned is well defined for the domain in ensuring the decision support system is deliverables and beyond the expectation of its end-user. In the last segment of this research, the significance of the framework offered over the previous works in solving the imbalance class and ambiguity and noise in the sample set of the Fertility dataset are accounted for and investigated. Even though previous works are slightly different and constructed the method based on issues they want to tackle, thus, this work is compared against the accuracy results obtained from the proposed methods. Consequently, the result improves and is applicable for accurate decision-making in healthcare applications.

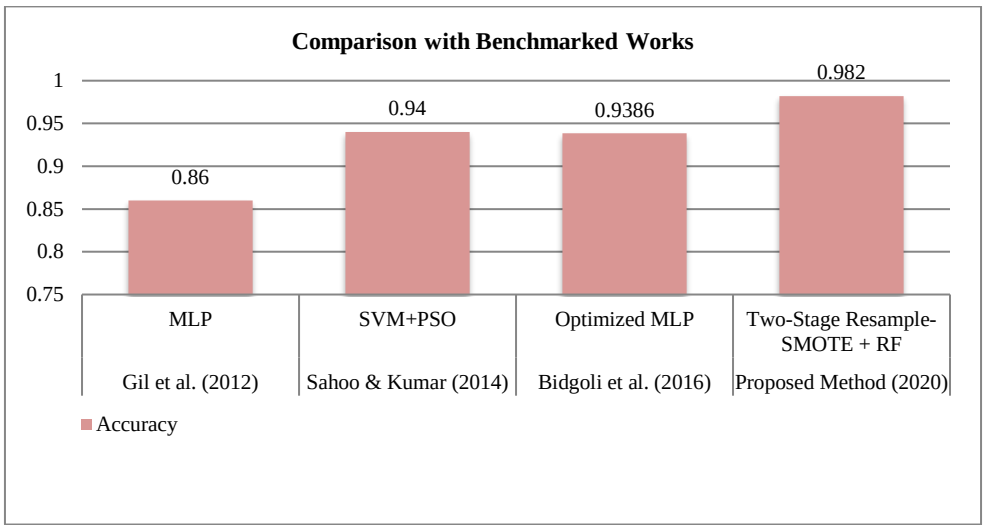

Figure-4. Comparison results based on benchmarked works.

REFERENCES

[1] S. Goyal. 2015. Graphical User Interface developed for Diseases Prediction by Mean of Clustering and Apriori Algorithm. 119(12): 1-7.

[2] H. Vundavilli, S. Member, A. Datta and C. Sima. 2020. Cryptotanshinone Induces Cell Death in Lung Cancer by Targeting Aberrant Feedback Loops. 24(8): 2430-2438.

[3] D. Gil, J. L. Girela, J. De Juan, M. J. Gomez-Torres and M. Johnsson. 2012. Predicting seminal quality with artificial intelligence methods. Expert Syst. Appl. 39(16): 12564-12573.

[4] R. Kiran P. and N. N. C. 2017. A Thorough Study on Machine Learning. Int. J. Innov. Res. Comput. Commun. Eng. 5(2): 65-71.

[5] C. K. Cornwallis and E. A. O'Connor. 2009. Sperm: Seminal fluid interactions and the adjustment of sperm quality in relation to female attractiveness. Proc. R. Soc. B Biol. Sci.

[6] M. C. Inhorn. 2003. Global infertility and the globalization of new reproductive technologies: Illustrations from Egypt. In Social Science and Medicine.

[7] M. C. Inhorn and P. Patrizio. 2014. Infertility around the globe: New thinking on gender, reproductive technologies and global movements in the 21st century. Hum. Reprod. Update. 21(4): 411-426.

[8] S. C. Richards. 2003. Infertility around the Globe: New Thinking on Childlessness, Gender, and Reproductive Technologies. Med. Anthropol. Q.

[9] J. E. Darroch. 2013. Trends in contraceptive use. In Contraception. 87(3): 259-263.

[10] J. P. E. Bonde *et al.* 1998. Relation between semen quality and fertility: A population-based study of 430 first-pregnancy planners. Lancet.

[11] M. Gomendio, A. F. Malo, J. Garde and E. R. S. Roldan. 2007. Sperm traits and male fertility in natural populations. Reproduction. 134(1): 19-29.

[12] S. A. Kidd, B. Eskenazi, and A. J. Wyrobek. 2001. Effects of male age on semen quality and fertility: A review of the literature. Fertil. Steril. 75(2): 237-248.

[13] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker and G. D. Tourassi. 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks. 21(2-3): 427-436.

[14] P. Radivojac, N. V. Chawla, A. K. Dunker and Z. Obradovic. 2004. Classification and knowledge discovery in protein databases. J. Biomed. Inform. 37(4): 224–239.

[15] H. Wang, Q. Xu and L. Zhou. 2014. Seminal quality prediction using clustering-based decision forests. Algorithms. 7(3): 405-417.

[16] A. A. Bidgoli, H. E. Komleh and S. J. Mousavirad. 2016. Seminal quality prediction using optimized

artificial neural network with genetic algorithm. ELECO 2015 - 9th Int. Conf. Electr. Electron. Eng. pp. 695-699.

[17] A. J. Sahoo and Y. Kumar. 2014. Seminal quality prediction using data mining methods. Technol. Heal. Care.

[18] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera. 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews. 42(4): 463-484.

[19] S. Muhammad, N. Admodisastro, N. Mohd Ali and H. Osman. 2021. The Correctness of Service in Runtime Adaptation for Context-Aware Mobile Cloud Learning. Turkish J. Comput. Math. Educ. 12(3): 2236-2241.

[20] A. Estabrooks, T. Jo and N. Japkowicz. 2004. A multiple resampling method for learning from imbalanced data sets. Comput. Intell. 20(1): 18-36.

[21] X. Liang, D. Li, M. Songi, A. Madden, Y. Ding, and Y. Bu. 2019. Predicting biomedical relationships using the knowledge and graph embedding cascade model. PLoS One. 14(6): 1-23.

[22] S. Visa and A. Ralescu. 2005. Issues in mining imbalanced data sets-a review paper. In Proceedings of the sixteen midwest artificial intelligence and cognitive science conference. pp. 67-73.

[23] J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio and L. I. Kuncheva. 2015. Random Balance: Ensembles of variable priors' classifiers for imbalanced data. Knowledge-Based Syst. 85: 96-111.

[24] F. Charte, A. J. Rivera, M. J. Del Jesus and F. Herrera. 2015. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. Knowledge-Based Syst. 89: 385-397.

[25] A. Helwan, A. Khashman, E. O. Olaniyi, O. K. Oyedotun and O. A. Oyedotun. 2016. Seminal quality evaluation with RBF neural network. Bull. Transilv. Univ. Brasov, Ser. III Math. Informatics, Phys. 9(2): 137-146.

[26] J. L. Girela, D. Gil, M. Johnsson, M. J. Gomez-Torres and J. De Juan. 2013. Semen Parameters Can Be Predicted from Environmental Factors and Lifestyle Using Artificial Intelligence Methods. Biol. Reprod. 88(4): 1-8.

[27] J. C. Lu, Y. F. Huang and N. Q. Lü. 2010. [WHO Laboratory Manual for the Examination and Processing of Human Semen: its applicability to andrology laboratories in China]. Zhonghua Nan ke xue= Natl. J. Androl.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16: 321-357.

[29] A. R. Abdul Raziff, N. Sulaiman, N. Mustapha and T. Perumal. 2017. Smote and OVO Multiclass Method for Multiple Handheld Placement Gait Identification on Smartphone's Accelerometer. J. Eng. Appl. Sci. 12(2): 204-207.

[30] L. Breiman. 2001. Random forests. Mach. Learn. 45(1): 5-32.

[31] A. Liaw and M. Wiener. 2002. Classification and Regression by Random Forest. R news. 2(December): 18-22.

[32] W.-Y. Loh. 2011. Classification and regression trees. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1(1): 14-23.

[33] D. Mennitt, K. Sherrill, and K. Fristrup. 2014. A geospatial model of ambient sound pressure levels in the contiguous United States. J. Acoust. Soc. Am. 135(5): 2746-2764.

[34] J. Quinlan Ross. 1993. C4. 5: Programs for Machine Learning. Machine Learning.