Male Fertility Classification using Machine Learning and Oversampling Techniques

AGP Sidhawara[1]
Department of Informatics, Faculty of Industrial Technology, Universitas Atma Jaya Yogyakarta,
Sleman 55282, Indonesia
Email: [1]aloysius.gonzaga@uajy.ac.id

Abstrak. Klasifikasi Kesuburan Pria Menggunakan Machine Learning dan Teknik Oversampling. *Metode pembelajaran mesin telah diterapkan pada diagnosis kesuburan pria dalam beberapa tahun terakhir. Melalui deteksi dini kasus infertilitas, penerapan teknologi ini menawarkan potensi manfaat di bidang medis. Studi ini menyajikan penyelidikan eksperimental yang mengkaji prospek penggunaan teknik oversampling dan pemilihan fitur untuk meningkatkan kinerja pengklasifikasi sederhana untuk mengklasifikasikan kesuburan pria pada Fertility Dataset. Dua teknik oversampling (SMOTE dan ADASYN), dua scaler berbeda (MinMax dan Standard), dan dua metode pemilihan fitur berbeda (SelectKBest dan SelectFromModel) digunakan untuk meningkatkan performa pengklasifikasi. Hasil menunjukkan bahwa performa model pembelajaran mesin lebih baik pada dataset hasil oversampling dibandingkan dataset asli. Random Forest mencapai kinerja terbaik pada set tes SMOTE dengan akurasi 90%, Recall 89% dan 100% masing-masing di kelas Normal dan Altered. Fitur Kecelakaan atau Trauma, Usia, dan Demam Tinggi dipilih oleh SelectKBest, dan dianggap sebagai faktor yang berkontribusi terhadap kesuburan pria dalam penelitian-penelitian sebelumnya.*

Kata Kunci: *kesuburan laki-laki, klasifikasi, pembelajaran mesin, SMOTE, ADASYN*

Abstract. *Machine learning methods have been applied to male fertility diagnosis in recent years. Through early infertility case detection, this technology application offers potential benefits to the medical field. This study presents an experimental investigation that examines the prospect of using the oversampling technique and feature selection to enhance the performance of shallow classifiers to classify male fertility on the Fertility Dataset. Two oversampling techniques (SMOTE and ADASYN), two different scalers (MinMax and Standard), and two different feature selection methods (SelectKBest and SelectFromModel) were used to improve the performance of the classifier. The results show that the performance of machine learning models is better on the oversampled dataset than the original dataset. Random Forest performed best on the SMOTE test set with 90% accuracy, 89% and 100% Recall in Normal and Altered classes, respectively. Accidents or trauma, Age, and High Fevers features are selected by SelectKBest, and considered as factors that contribute to male fertility in prior studies.*

Keywords: *male fertility, classification, machine learning, SMOTE, ADASYN*

1. Introduction

The capacity of a man to bear children or the ideal function of his reproductive organs for conception is known as fertility. Because of the complexity of the human reproductive system, appropriate ovulation and fertilization are essential for a healthy pregnancy. Infertility affects around 17.5% of adult people globally, or 1 in 6 of them, as reported by the World Health Organization (WHO) [1]. About 50% of infertility problems globally are caused by males [2]. Meanwhile, in Indonesia, 22.3% of couples experience infertility [3], with men between the ages of 30 and 40 having the highest rate [4]. The quality of semen, or sperm, is influenced by several triggering variables, including lifestyle choices, medical history, and physical trauma [5], [6].

In recent years, male reproduction health topics have utilized machine learning methods to diagnose male fertility [7]. A public dataset named Fertility Dataset was donated on 16 January 2013 at the UCI Machine Learning Repository [8] and has been widely used in machine learning research [9], [10], [11], [12], [13], [14]. Several shallow machine learning models such as the Support Vector Machine (SVM) [12], K-Nearest Neighbors (KNN) [9], and tree-based models

[10], [13] have been employed to predict the male's seminal quality. This technology application can benefit the medical sector by providing early detection of infertility case.

However, the baseline study mentioned the need to investigate the effect of imbalanced classes [15]. Several previous studies have not yet employed a specific method to deal with the imbalanced dataset and boosting classifier performance [9], [11], [13]. Numerous methods can be utilized to overcome the effect of imbalanced datasets. One of the popular methods is the oversampling technique. Synthetic Minority Oversampling Technique or SMOTE has been a popular and effective technique for learning from unbalanced data since 2002 [16]. SMOTE for oversampling efficiently balances training data while managing outlier effects, improving machine learning prediction performance [17]. Another oversampling technique that has recently gained popularity is Adaptive Synthetic (ADASYN). For the minority class, ADASYN can generate synthetic data samples adaptively to mitigate the bias resulting from the unequal distribution of the data [18]. ADASYN lessens the impact of class imbalance on prediction performance [19].

On the other side, improving machine learning models' performance can be done by utilizing feature selection methods [20]. By lowering dimensionality and enhancing computational efficiency, the SelectKBest feature selection technique enhances model performance [21]. A model-based feature selection method also potentially improved the classifier performance [22]. Therefore, this work presents an experimental investigation that examines the prospect of using the oversampling technique and feature selection to enhance the performance of shallow classifiers toward datasets related to male fertility.

2. Literature Review

The study by [9] developed an application that aids in the early detection of sperm fertility. This study uses the UCI Machine Learning Repository's Fertility Dataset with 60 records used as the train set and 20 records as the test set. The k-Nearest Neighbors method was implemented in the web-based application in the form of the KNN method calculation. The web-based application consists of an input page and a classification result page. Twenty records of test data were tested on the web application and yielded results of 85% accuracy.

A study by [14] uses the sperm sample data of 100 volunteers (Fertility Dataset) to create a classification model in the form of rules. The degree of male fertility is divided into classes normal (N) and altered fertility (O). It can be predicted using the classification model using Classification and Regression Trees (CART). Research methods employed in this work include data preprocessing by checking on duplicates, missing values handling, and checking on inconsistent data. The next steps are data transformation and classification, followed by model evaluation using a confusion matrix. The result is CART algorithm was able to provide a decision tree with 84% accuracy and 12 rules to predict male fertility.

A similar study by [11] predict male fertility levels using the UCI Machine Learning Fertility Dataset and utilized the CART algorithm. Age, childish diseases, accidents or severe trauma, surgical intervention, high fevers within the previous year, frequency of alcohol consumption, smoking habit, number of hours spent sitting per day, and diagnosis were the features used in the dataset. To prevent the model from being either overfit or underfit, K-Fold Cross Validation was implemented on CART to evaluate the model's performance on several sets of data. The average accuracy for training and testing data is 98.70% and 81.16%, respectively. The study demonstrates that the CART algorithm is a valuable tool for identifying men's reproductive levels. To make it even easier to use, the developed model is integrated into an Android mobile application.

In the study by [12], the Fertility Dataset's semen quality is predicted using SVM methods and Genetic Algorithm—experiments with 10 iterations. The result shows that SVM+GA (dot kernel) has the greatest accuracy at 89%, followed by SVM at 88%, Decision Tree at 84%, Neural Network at 82%, and Naïve Bayes at 82%. It has been demonstrated that GA raises the accuracy value of SVM with kernel dot, which displays a considerable difference among the other methods.

Another work on tree-based algorithms by [13] aims to determine the accuracy of Fertility Dataset classification success by comparing the Random Forest and C4.5 Decision Tree. Surgical procedures, age, childhood illnesses, trauma or accidents, alcohol use, and smoking habits are the dataset's attributes used. This study used 10-fold cross-validation to assess how the two models performed. The C4.5 decision tree has an accuracy rate of 85.90%, whereas the random forest has an average accuracy of 87.20%. Based on the results, Random Forest outperforms Decision Tree C4.5 by 1.3% in terms of accuracy.

A study by [23] aims to predict the male fertility dataset through the application of feature selection techniques and several classifier algorithms. The SMOTE method was used to improve the accuracy and representativeness of classifier results. A combination of feature selection and classification techniques predicts male fertility. MLP, Naïve Bayes, Random Forest, KNN, and SVM classifiers were employed in this investigation. Based on comparison data, the Naive Bayes classifier outperforms the others with a classification accuracy of 90.65%.

These previous studies show the potential use of machine learning methods in the classification of male fertility. In addition, feature selection and oversampling techniques have been used to improve the performance of the classifiers. Therefore, this study will explore the shallow classifiers' performance on the Fertility Dataset with combinations of oversampling techniques and feature selection methods.

3. Methodology

In this section, the dataset and research method will be explained. Figure 1 shows the stages of data preprocessing (including data loading, data encoding, train-test split, and oversampling), machine learning model training, and model evaluation.

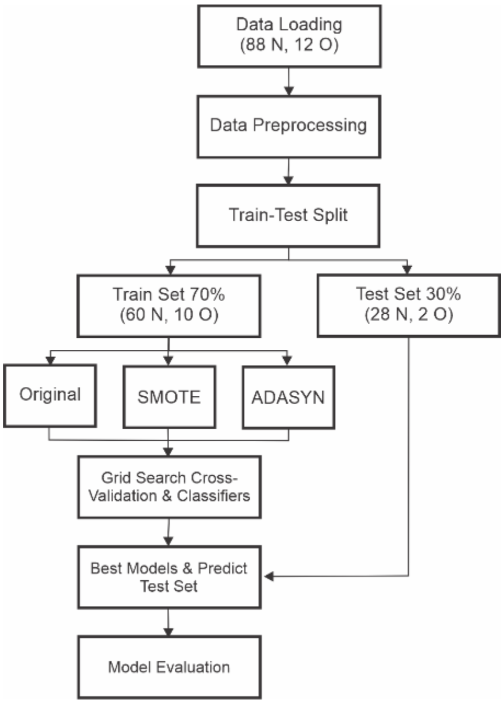
Data Loading (88 N, 12 O)
Data Preprocessing
Train-Test Split
Train Set 70% (60 N, 10 O)
Test Set 30% (28 N, 2 O)
Original
SMOTE
ADASYN
Grid Search Cross-Validation & Classifiers
Best Models & Predict Test Set
Model Evaluation

Figure 1. Research Methodology Flowchart

3.1. Dataset

The dataset used in this research (https://archive.ics.uci.edu/dataset/244/fertility) was acquired from a study by Gil et al. [15] One hundred healthy volunteers from the University of Alicante, age between 18 and 36 years, participated in the study. They were requested to provide a semen sample following 3 to 6 days of sexual abstinence. The semen analysis was conducted according to the 2010 WHO guidelines. There are nine attributes supplied in the dataset such as

(1) season in which the sample was acquired (categorical); (2) age at the time the study was conducted (continuous); (3) `childish diseases` shows if the participants ever experienced chicken pox, measles, mumps, or polio (binary); (4) if the participants ever experienced accident or trauma (binary); (5) surgical intervention (binary); (6) high fevers in the last year (ordinal); (7) frequency of alcohol consumption (ordinal); (8) smoking habit (ordinal); (9) number of hours spent sitting per day (continuous). One column `diagnosis` is designated as the label with values N (Normal) and O (Altered). The dataset provided was in normalized value.

3.1. Data Loading and Preprocessing

The dataset was acquired from the UCI Machine Le Jupyter Notebook. Because the dataset file is in Txt format and has no column names, the attributes were added to the file according to the repository information such as `season`, `age`, `child_diseases`, `accident_or_trauma`, `surgical_intervention`, `high_fevers_last_year`, `alcohol_consumption`, `smoking_habit`, `number_of_hours_sitting`, and `diagnosis`. The `season` column was dropped to focus on more general attributes. Binary values (0 and 1) in th `child_diseases`, `accident_or_trauma`, and `surgical_intervention` columns were replaced with ‘no’ and ‘yes’, respectively. Then, the dataset was split into a train set and a test set with a 70:30 ratio. The label is also divided into y_train and y_test variables. The categorical features (`child_diseases`, `accident_or_trauma`, and `surgical_intervention`) in the train and test set were encoded using OneHotEncoder, while ordinal and numerical columns remained the same. Machine learning models trained using the original dataset will be compared with machine learning models using a dataset transformed using oversampling techniques. The encoded dataset was oversampled using Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN). After that, original and oversampled datasets were assigned to the machine learning pipelines.

3.3. Machine Learning Model Pipelines

Table 1. Machine Learning Models Pipeline Setting

Pipeline	***Data Scaler***	***Feature Selection***	***Classifiers***	***Hyperparameter Setting***
KNN Pipeline	MinMaxScaler, StandardScaler	SelectKBest (K=3-10 features) SelectFromModel (Decision Tree, max_features=5)	KNeighborsClassifier	n_neighbors=[3,5,7,9], weights=[uniform,distance], p=[1,2]
DT Pipeline	None		DecisionTreeClassifier	max_depth=[3,4,5], criterion=[gini,entropy]
GBT Pipeline	None		GradientBoostingClassifier	n_estimators=[100,150], learning_rate=[0.01,0.1,1], max_depth=[3,4,5]
RF Pipeline	None		RandomForestClassifier	n_estimators=[50,100,150], max_depth=[3,4,5], criterion=[gini,entropy]
LogReg Pipeline	MinMaxScaler, StandardScaler		LogisticRegressionClassifier	C= [0.001, 0.01, 0.1, 1, 10], penalty=['l1', 'l2']
SVC Pipeline	MinMaxScaler, StandardScaler		SVC	kernel=['poly','rbf'], C=[0.001, 0.01, 0.1, 1, 10], gamma=[0.001, 0.01, 0.1, 1, 10]

The machine learning pipeline in this study consists of data scaling, feature selection, and classifier. There were six classifiers compared such as K-Nearest Neighbors (KNN), Decision Tree (DT), Gradient Boosting Classifiers (GBT), Random Forest (RF), Logistic Regression (LogReg), and kernelized Support Vector Machine (SVC). Each classifier is assigned to one pipeline with KNN, LogReg, and SVC using a data scaler. SelectKBest and SelectFromModel feature selection methods were employed. Using the `f-classif} function, SelectKBest selected the features with the highest score based on their ANOVA F-score and P-value. Meanwhile, SelectFromModel uses a Decision Tree estimator to choose features based on their corresponding importance values according to specified threshold parameters. Hyperparameters of classifiers were empirically tuned using the parameters grid (see Table 1).

3.3. Machine Learning Model Evaluation

Stratified K-fold with 5 splits was used in the cross-validation process. The Grid Search Cross Validation assessed each model's performance and determined the best parameters. Performance metrics for machine learning models were Accuracy, Precision, Recall, and F1 Score. The ratio of True Positives to the entirety of test set samples is called Accuracy. The ratio of True Positives to the total of True Positives and False Positives is known as Precision. The Precision of a classifier is an indication of its capability to prevent incorrectly identifying a negative sample as a positive class. The ratio of True Positives to the total of True Positives and False Negatives is recognized as Recall. The classifier's Recall measures its capacity to acknowledge every positive sample.

4. Results and Discussion

Table 2. Machine Learning Models Evaluation on Original Dataset

Classifier	*Best Selected Features*	*CV Accuracy*	*Test Accuracy*	*Precision*		*Recall*		*F1 Score*	
				N	O	N	O	N	O
KNN	accident_or_trauma_no, accident_or_trauma_yes, high_fevers_last_year, alcohol_consumption	0.87	0.97	0.97	1.00	1.00	0.50	0.98	0.67
Decision Tree	child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption	0.83	0.90	0.93	0.00	0.96	0.00	0.95	0.00
Gradient Boosting Classifier	accident_or_trauma_no, accident_or_trauma_yes, high_fevers_last_year	0.84	0.93	0.93	0.00	1.00	0.00	0.97	0.00
Random Forest	child_diseases_no, child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.86	0.93	0.93	0.00	1.00	0.00	0.97	0.00
Logistic Regression	accident_or_trauma_no, accident_or_trauma_yes, high_fevers_last_year	0.86	0.93	0.93	0.00	1.00	0.00	0.97	0.00
SVC	child_diseases_no, child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.87	0.83	0.93	0.00	0.89	0.00	0.91	0.00
	Average	**0.86**	**0.92**	**0.94**	**0.17**	**0.98**	**0.08**	**0.96**	**0.11**

The dataset was split into 70 instances, with the train-validation set consisting of 60 Normal (N) and 10 Altered (O), and the test set consisting of 30 cases (28 Normal and 2 Altered). Table 2 compares evaluation metrics from the best model of each combination of scaler, feature selection, and classifier pipeline on the original dataset. K-Nearest Neighbors gained the highest test accuracy with 97% value. However, the Recall value on the Altered class only increased 50%. This shows the effect of imbalanced data on the performance of the model to recall the characteristics of minority samples. The other classifiers also achieved a high Accuracy rate, however, they failed to classify the Altered class. Most classifiers only learn the characteristics of the majority class (Normal) and fail in the minority class.

SMOTE balanced the number between Normal and Altered data from 60 and 10 to 60 and 60, respectively. Table 3 compares evaluation metrics on the SMOTE dataset from the best models of each pipeline combination. The Random Forest model yields a higher Cross-Validation Accuracy (90%) than the Gradient Boosting Classifier (89%) while on par with the Gradient Boosting Classifier in Test Accuracy. Random Forest uses five features from SelectKBest with four levels of maximum depth, 150 estimators, and using gini criterion. The Random Forest model acquired a 94% F1 Score in the Normal class and a 57% F1 Score in the Altered class. On the Recall metric, Random Forest gained 89% in the Normal class and 100% in the Altered class. These metrics show that Random Forest has a good recall in both classes. The average evaluation metrics from six shallow classifiers employed on the SMOTE dataset have higher values than the original dataset.

Table 3. Machine Learning Models Evaluation on SMOTE Dataset

Classifier	*Best Selected Features*	*CV Accuracy*	*Test Accuracy*	*Precision*		*Recall*		*F1 Score*	
				N	O	N	O	N	O
KNN	child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.90	0.77	1.00	0.22	0.75	1.00	0.86	0.36
Decision Tree	child_diseases_no, accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.88	0.67	1.00	0.17	0.64	1.00	0.78	0.29
Gradient Boosting Classifier	accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption	0.89	0.90	1.00	0.40	0.89	1.00	0.94	0.57
Random Forest	accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption	0.90	0.90	1.00	0.40	0.89	1.00	0.94	0.57
Logistic Regression	Accident_or_trauma_no, accident_or_trauma_yes, age, alcohol_consumption	0.73	0.73	1.00	0.20	0.71	1.00	0.83	0.33
SVC	accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.94	0.77	1.00	0.22	0.75	1.00	0.86	0.36
	Average	**0.87**	**0.79**	**1.00**	**0.27**	**0.77**	**1.00**	**0.87**	**0.41**

ADASYN generates synthetic data to increase the number of Altered from 10 to 62 instances, meanwhile, Normal data remained at 60 cases. Table 4 shows the comparison of test accuracy on the ADASYN dataset from the best models of each pipeline combination. The results show that Random Forest yields 87% Test Accuracy. Random Forest uses nine features from SelectKBest with five levels of maximum depth, 150 estimators, and the gini criterion. The Random Forest model acquired a 92% F1 Score in the Normal class and a 50% F1 Score in the

Altered class. On the Recall metric, Random Forest gained 86% in the Normal class and 100% in the Altered class. However, the average value of evaluation metrics from the classifiers employed on the ADASYN dataset, lower than the SMOTE dataset, but higher than the original dataset.

Table 4. Machine Learning Models Evaluation on ADASYN Dataset

Classifier	*Best Selected Features*	*CV Accuracy*	*Test Accuracy*	*Precision*		*Recall*		*F1 Score*	
				N	O	N	O	N	O
KNN	child_diseases_no, child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, surgical_intervention_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.85	0.87	1.00	0.33	0.86	1.00	0.92	0.50
Decision Tree	child_diseases_no, child_diseases_yes accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.83	0.70	1.00	0.18	0.68	1.00	0.81	0.31
Gradient Boosting Classifier	child_diseases_no, child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, surgical_intervention_no, surgical_intervention_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.89	0.83	0.96	0.20	0.86	0.50	0.91	0.29
Random Forest	child_diseases_no, child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, surgical_intervention_yes, age, high_fevers_last_year, alcohol_consumption, smoking_habit	0.86	0.87	1.00	0.33	0.86	1.00	0.92	0.50
Logistic Regression	child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption	0.71	0.70	1.00	0.18	0.68	1.00	0.81	0.31
SVC	child_diseases_yes, accident_or_trauma_no, accident_or_trauma_yes, age, high_fevers_last_year, alcohol_consumption	0.87	0.83	0.96	0.20	0.86	0.50	0.91	0.29
	Average	**0.84**	**0.80**	**0.99**	**0.24**	**0.80**	**0.83**	**0.88**	**0.37**

The findings indicate that, when comparing the original dataset to the modified dataset with oversampling techniques, the average performance of machine learning models is better on the altered dataset. Most classifiers fail to classify the Altered class. In this case SMOTE as an oversampling technique has been proven in improving classification accuracy [24], [25], [26]. Although the average performance of ADASYN is lower than SMOTE, ADASYN performs better than the original dataset [27]. The best model on the SMOTE and the ADASYN dataset is Random Forest. Prior studies have also discovered that Random Forest outperforms other machine learning classifiers in several fields, particularly in medical applications [28], [29], [30].

Random Forest offers significant benefits in its resistance to overfitting, capacity to manage highly non-linear data, detect complex boundaries, and stability even when outliers are present [31], [32] On the SMOTE dataset Random Forest using a smaller number of features, compared with Random Forest on ADASYN which utilized nine features. The Random Forest model uses five features selected by SelectKBest which are \`accident_or_trauma_no\`, \`accident_or_trauma_yes\`, \`age\`, \`high_fevers_last_year\`, and \`alcohol_consumption\`. Accidents or trauma in the form of testicular trauma or long-term pain was a common cause of male infertility, and one to be wary of [33]. On the age factor, male seminal quality is lower for men above 29 years old. Higher male age was linked to poorer embryo development and a decreased chance of fertilization [34], [35]. On its own, a high fever can result in considerable alterations to sperm parameters, sperm DNA damage, sperm apoptosis, and germ cell death [5], [36]. Severe drinking and smoking were linked to worse seminal parameters. Chronic alcohol use has been linked to low semen quality, primarily because of oxidative stress. These oxidative stress also showed correlations with the motility, morphology, and concentration of sperm [37], [38].

5. Conclusion and Future Works

In this investigation, the category of male fertility on the Fertility Dataset was classified using machine learning classifiers and oversampling techniques. We used two oversampling techniques (SMOTE and ADASYN), two different scalers on distance-based and linear classifiers (MinMax and Standard) and none on tree-based classifiers, two different feature selection (SelectKBest and SelectFromModel), and six different classifiers (KNN, Decision Tree, Gradient Boosting Tree, Random Forest, Logistic Regression, and SVC). The evaluation metrics of each model on three datasets were compared. Random Forest performed best on the SMOTE test set with 90% accuracy, 89%, and 100% Recall in Normal and Altered classes, respectively. \`Accident or trauma\`, \`Age\`, \`High Fevers\`, and \`Alcohol Consumption\` features are selected by SelectKBest and considered factors contributing to male seminal quality in prior studies. Despite our results, oversampling techniques such as Deep-SMOTE, BI-BMCSMOTE have not yet been used in this work. Therefore, additional research is required to investigate the effects of additional oversampling techniques on the Fertility Dataset and machine learning classifiers' performance.

References

[1] Sexual and Reproductive Health and Research (SRH), "Infertility Prevalence Estimates, 1990–2021," World Health Organization, Apr. 2023.

[2] J. Fainberg and J. Kashanian, "Recent advances in understanding and managing male infertility," *F1000Research*, vol. 8, 2019, doi: 10.12688/f1000research.17076.1.

[3] F. Damayanti, M. Hakimi, M. Anwar, and D. A. Puspandari, "Psychometric properties of the Indonesian online version of fertility quality of life tool: a cross-sectional study," *International Journal of Community Medicine and Public Health*, vol. 8, p. 2768, 2021, doi: 10.18203/2394-6040.IJCMPH20211944.

[4] A. Guyansyah *et al.*, "Primary infertility of male and female factors, polycystic ovary syndrome and oligoasthenoteratozoospermia dominate the infertile population in agricultural and industrial areas in Karawang Regency, West Java Province, Indonesia," *Bali Medical Journal*, 2021, doi: 10.15562/bmj.v10i1.2281.

[5] M. Bendayán and F. Boitrelle, "COVID-19: semen impairment may not be related to the virus," *Human Reproduction (Oxford, England)*, 2021, doi: 10.1093/humrep/deab082.

[6] M. Nistal, R. Paniagua, P. Gónzalez-Peramato, and M. Reyes-Múgica, "Perspectives in Pediatric Pathology, Chapter 23. Testicular Pathology Secondary to Physical and Chemical Injury," *Pediatric and Developmental Pathology*, vol. 19, pp. 452–459, 2016, doi: 10.2350/16-04-1811-PB.1.

[7] A. Syarif and F. R. Lumbanraja, "SYSTEMATIC REVIEW: PERKEMBANGAN MACHINE LEARNING PADA SPERMA MANUSIA," 2023. [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

[8] D. Gil and J. Girela, “Fertility.” UCI Machine Learning Repository, 2013. doi: https://doi.org/10.24432/C5Z01Z.

[9] R. Yepriyanto and Y. Retno Wahyu Utami, “SISTEM DIAGNOSA KESUBURAN SPERMA DENGAN METODE K-NEAREST NEIGHBOR (K-NN),” *Jurnal Ilmiah SINUS*, vol. 13, no. 2, pp. 33–44, 2015.

[10] T. W. Pratiwi and T. Arifin, “Optimasi Decision Tree Menggunakan Particle Swarm Optimization untuk Klasifikasi Kesuburan pada Pria,” *SISTEMASI: Jurnal Sistem Informasi*, vol. 10, no. 1, pp. 1–12, 2021.

[11] A. Rahman Hakim, D. Marini Umi Atmaja, A. Basri, and A. Ariyanto, “Performance Analysis of Classification and Regression Tree (CART) Algorithm in Classifying Male Fertility Levels with Mobile-Based,” *JOURNAL OF TECH-E*, vol. 7, no. 1, pp. 10–20, 2023.

[12] H. Harafani and A. Maulana, “Penerapan Algoritma Genetika pada Support Vector Machine Sebagai Pengoptimasi Parameter untuk Memprediksi Kesuburan,” *Jurnal Teknik Informatika*, vol. 5, no. 1, pp. 51–59, 2019.

[13] A. Prabowo, S. Wardani, R. Wijaya Dewantoro, and W. Wesly, “Komparasi Tingkat Akurasi Random Forest dan Decision Tree C4.5 Pada Klasifikasi Data Penyakit Infertilitas,” *Media Online)*, vol. 4, no. 1, pp. 218–224, 2023, doi: 10.30865/klik.v4i1.1115.

[14] U. Khaira, N. Syarief, and I. Hayati, “Prediksi Tingkat Fertilitas Pria Dengan Algoritma Pohon Keputusan Cart,” *Jurnal Ilmiah Umum dan Kesehatan Aisyiyah*, vol. 5, no. 1, 2020, [Online]. Available: https://download.garuda.kemdikbud.go.id/article.php?article=3026906&val=27399&title=Prediksi%20Tingkat%20Fertilitas%20Pria%20Dengan%20Algoritma%20Pohon%20Keputusan%20Cart

[15] D. Gil, J. L. Girela, J. De Juan, M. J. Gomez-Torres, and M. Johnsson, “Predicting seminal quality with artificial intelligence methods,” *Expert Systems with Applications*, vol. 39, no. 16, pp. 12564–12573, Nov. 2012, doi: 10.1016/j.eswa.2012.05.028.

[16] A. Fernández, S. García, F. Herrera, and N. Chawla, “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary,” *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[17] N. Nnamoko and I. Korkontzelos, “Efficient treatment of outliers and class imbalance for diabetes prediction,” *Artificial intelligence in medicine*, vol. 104, p. 101815, 2020, doi: 10.1016/j.artmed.2020.101815.

[18] C. Liu and L. Zhu, “A two-stage approach for predicting the remaining useful life of tools using bidirectional long short-term memory,” *Measurement*, vol. 164, p. 108029, 2020, doi: 10.1016/j.measurement.2020.108029.

[19] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. Davis, “DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values,” *IEEE Access*, vol. 7, pp. 102232–102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[20] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, 2018, doi: 10.1016/j.neucom.2017.11.077.

[21] D. A. Otchere, T. Ganat, J. Ojero, B. N. Tackie-Otoo, and M. Y. Taki, “Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions,” *Journal of Petroleum Science and Engineering*, 2021, doi: 10.1016/J.PETROL.2021.109244.

[22] A. K. Srivastava, D. Singh, A. S. Pandey, and T. Maini, “A Novel Feature Selection and Short-Term Price Forecasting Based on a Decision Tree (J48) Model,” *Energies*, 2019, doi: 10.3390/en12193665.

[23] B. Karlik, A. M. Yibre, and B. Koçer, “Comprising Feature Selection and Classifier Methods with SMOTE for Prediction of Male Infertility,” 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:204826876

[24] N. Cahyana, S. Khomsah, and A. Aribowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," *2019 5th International Conference on Science in Information Technology (ICSITech)*, pp. 217–222, 2019, doi: 10.1109/ICSITech46713.2019.8987499.

[25] X. Tan *et al.*, "Wireless Sensor Networks Intrusion Detection Based on SMOTE and the Random Forest Algorithm," *Sensors (Basel, Switzerland)*, vol. 19, 2019, doi: 10.3390/s19010203.

[26] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pp. 1–8, 2021, doi: 10.1109/ICIC54025.2021.9632912.

[27] K. Davagdorj, J. S. Lee, V.-H. Pham, and K. Ryu, "A Comparative Analysis of Machine Learning Methods for Class Imbalance in a Smoking Cessation Intervention," *Applied Sciences*, vol. 10, p. 3307, 2020, doi: 10.3390/app10093307.

[28] S. Benbelkacem and B. Atmani, "Random Forests for Diabetes Diagnosis," *2019 International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–4, 2019, doi: 10.1109/ICCISCI.2019.8716405.

[29] R. Buettner, M. Hirschmiller, K. Schlosser, M. Rössle, M. Fernandes, and I. Timm, "High-performance exclusion of schizophrenia using a novel machine learning method on EEG data," *2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, pp. 1–6, 2019, doi: 10.1109/HealthCom46333.2019.9009437.

[30] N. M. Abdulkareem and A. Abdulazeez, "Machine Learning Classification Based on Radom Forest Algorithm: A Review," vol. 5, pp. 128–142, 2021, doi: 10.5281/ZENODO.4471118.

[31] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review," *Frontiers in Aging Neuroscience*, vol. 9, 2017, doi: 10.3389/fnagi.2017.00329.

[32] M. Shojaeizadeh, S. Djamasbi, R. C. Paffenroth, and A. C. Trapp, "Detecting task demand via an eye tracking machine learning system," *Decision Support Systems*, vol. 116, no. June 2018, pp. 91–101, 2019, doi: 10.1016/j.dss.2018.10.012.

[33] R. Mora, J. Nabhani, T. Bakare, R. Khouri, and M. Samplaski, "The effect of testicular trauma on male infertility.," *Human fertility*, pp. 1–6, 2022, doi: 10.1080/14647273.2022.2135464.

[34] K. Gill, J. Jakubik-Uljasz, A. Rosiak-Gill, M. Grabowska, M. Matuszewski, and M. Piasecka, "Male aging as a causative factor of detrimental changes in human conventional semen parameters and sperm DNA integrity," *The Aging Male*, vol. 23, pp. 1321–1332, 2020, doi: 10.1080/13685538.2020.1765330.

[35] O. A. Oluwayiose *et al.*, "Sperm DNA methylation mediates the association of male age on reproductive outcomes among couples undergoing infertility treatment," *Scientific Reports*, vol. 11, 2021, doi: 10.1038/s41598-020-80857-2.

[36] M. Bendayan and F. Boitrelle, "What could cause the long-term effects of COVID-19 on sperm parameters and male fertility?," *QJM*, vol. 114, no. 4, p. 287, Jul. 2021, doi: 10.1093/qjmed/hcab028.

[37] L. Boeri *et al.*, "Heavy cigarette smoking and alcohol consumption are associated with impaired sperm parameters in primary infertile men," *Asian Journal of Andrology*, vol. 21, pp. 478–485, 2019, doi: 10.4103/aja.aja_110_18.

[38] S. S. Ramgir and V. Abilash, "Impact of Smoking and Alcohol Consumption on Oxidative Status in Male Infertility and Sperm Quality," *Indian Journal of Pharmaceutical Sciences*, 2019, doi: 10.36468/pharmaceutical-sciences.588.