2024 7th International Conference on Informatics and Computational Sciences (ICICoS) | 979-8-3503-7588-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICICoS62600.2024.10636933

Diabetes Prediction in Machine Learning using Feature Selection: A Comparative Analysis based on Sampling Techniques

Zulchair Asy'ari
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6025222005@student.its.ac.id

Shintami Chusnul Hidayati
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
shintami@its.ac.id

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@its.ac.id

***Abstract*—Diabetes poses a global threat, impacting both patient well-being and healthcare resources. Preventing diabetes has become a key focus in mitigating its impact in the medical field. This research leverages machine learning to predict diabetes in patients by utilizing a dataset of diabetic and non-diabetic individuals to identify patterns indicative of diabetes. The study uses the Diabetes 130-US hospitals dataset, covering the years 1999-2008, and analyzes its features, applies data preprocessing, selects relevant features, and addresses data imbalance through various sampling techniques to enhance prediction accuracy. The machine learning models employed in this research include Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM). The findings highlight the importance of feature selection in optimizing model performance, identifying key features to improve accuracy, and assessing the effectiveness of each model. A comparative analysis of the models using different sampling techniques demonstrates the exceptional performance of the Random Forest model, achieving 99.8% accuracy with the raw dataset. However, Logistic Regression shows potential for improvement, as its performance increases with the combination of various techniques, indicating its value for future enhancement of predictive capabilities.**

***Index Terms*—machine learning, diabetes prediction, feature selection, sampling technique, comparative analysis**

I. Introduction

Diabetes presents a global threat that has been growing over the years [1]. This condition risks complicating other diseases and injuries due to elevated glucose levels, significantly worsening the patient's quality of life if left untreated [2]. Prevention is a more practical approach than medication in managing the condition [3], [4]. Diabetes not only affects patients but also strains healthcare resources and systems due to the increasing number of cases worldwide. The urgency to mitigate the impact of diabetes on society underscores the importance of effective predictive models in addressing this global problem.

Diabetes prediction research has evolved by harnessing the potential of machine learning in healthcare. A notable gap in past research lies in the limitation of medical datasets, which often contain a relatively small amount of data and features. The PIMA Indian Diabetic Dataset (PIDD) [5], the most commonly used dataset for diabetes prediction, includes 768 instances and eight features. Recent research has utilized the Early-Stage Diabetes Risk Prediction Dataset (ESDRPD) [6], comprising 520 instances and 17 features, to explore the analysis of richer features in this field. Meanwhile, the commonly used models for diabetes prediction are Random Forest, Support Vector Machine, and Gradient Boosting [7], [8], [9]. These models are favored for their unique traits and effectiveness in diabetes prediction, delivering exceptional results while offering room for improvement [10].

Feature selection in diabetes prediction is crucial to data preprocessing, helping models learn the data structure more effectively. This process identifies and selects the most important features based on their relevance to the label [11], [12], [13]. Effective feature selection allows models to focus on significant data, bypassing redundant and irrelevant features, thereby providing more robust predictive models [14], [15].

Datasets often need more data distribution between classes, where each label does not have an equal amount of representative data [16], [17]. Imbalanced datasets can lead to overfitting, as predictive models receive less training on the minority class. Therefore, sampling techniques such as undersampling and oversampling are necessary to improve diabetes prediction.

This research aims to enhance the robustness of diabetes prediction by achieving the best accuracy across different models using larger, feature-rich datasets. The preprocessing steps before model testing include data cleaning, feature selection, and sampling techniques. The performance of each model will be compared to determine the most effective approach.

II. Related Works

In a study by Jiang *et al.* [9], diabetes prediction was conducted using a dataset obtained from public healthcare services in the Haizhu District, Guangzhou City, China. This dataset includes follow-up records of diabetic patients from 2016 to 2023, consisting of 252,176 instances with nine features. Compared to PIDD, this dataset is more extensive regarding the total amount of data and number of features but contains fewer features than ESDRPD. The study employed

comprehensive steps, starting with data preprocessing to handle missing values, variable conversion, and outlier management. Oversampling techniques were applied to balance the training data and address data imbalance. Feature selection was performed to identify features with optimal values for diabetes prediction. The dataset was split into a 70:30 ratio for training and testing purposes. The best predictive model in this research was Random Forest, achieving an overall accuracy of 91.24%.

Ahmed *et al.* [8] conducted diabetes prediction using two datasets simultaneously. The first dataset was PIDD, and the second, referred to as dataset 2, contained 950 instances with 19 features. The study employed a comprehensive approach, including data preprocessing to handle outlier removal, missing values, and label encoding. The datasets were divided into an 80:20 ratio for training and testing. Seven machine learning models were used for prediction and analysis. The comparison showed that the Support Vector Machine (SVM) was the best predictive model for PIDD with 80.26% accuracy. In contrast, Decision Tree and Random Forest provided the best accuracy of 96.81% for dataset 2.

Rajendra *et al.* [18] proposed an improved Logistic Regression model to predict diabetes using PIDD and the Vanderbilt Diabetes Dataset. This improved model utilized a feature selection method developed by Scikit-learn to identify the most influential features for diabetes prediction. The results after feature selection showed a slight accuracy improvement of 0.98-1.74%. Despite the modest increase in accuracy, the execution time required for the model to perform predictions was reduced by 80-86%.

Later, Kuriakose *et al.* [7] conducted diabetes prediction research using PIDD and the Diabetes 130-US hospitals dataset from 1999-2008. The research involved data cleaning as part of data preprocessing and splitting the dataset into training and testing data. The study reported a lower accuracy for PIDD, with the best predictive model being Random Forest, achieving 79% accuracy compared to Taz *et al.* [19], who reported an ensemble model accuracy of 95%. For the Diabetes 130-US hospitals dataset, the best predictive model was Random Forest with 80% accuracy.

III. Methodology

This research involves several steps to predict diabetes using the Diabetes 130-US hospitals dataset for the years 1999-2008. Fig. 1 shows the proposed approach.

A. Data Preprocessing

1) Data Cleaning: Each column is examined for its basic structure, presence of missing values, and features that require transformation. Rows with missing values are deleted to enhance the model's performance. Features with unusual representations are transformed into a format that is readable and interpretable by the predictive model.

2) Feature Selection (FS): Feature selection enhances the accuracy of the predictive model by calculating the correlation between each feature and the diabetes class using the chi-square method [20], as shown in (1).

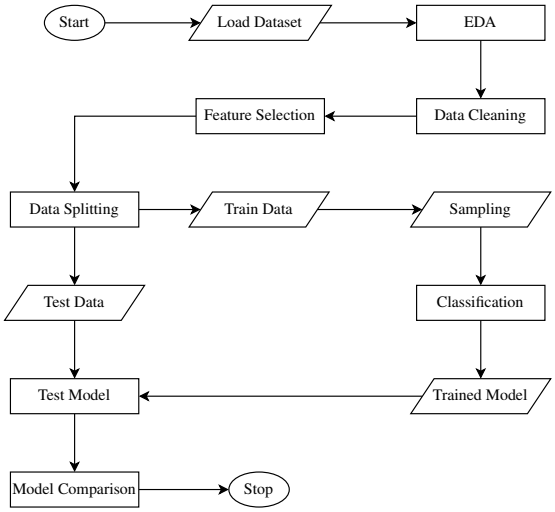

Fig. 1. Flowchart of the proposed approach.

$$\chi^2 = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{.} E_i \tag{1}$$

Here, χ^2 denotes the chi-squared score, O_i is the observed value, E_i is the expected value, and I is the total number of rows in the feature.

B. Data Imbalance

Diabetic patients in the dataset might not have equal data compared to non-diabetic patients. Handling imbalanced data between classes is crucial for giving the predictive model a higher chance of learning the minority class, thereby providing better prediction accuracy. Undersampling and oversampling methods are used to compare the performance of each predictive model.

Undersampling reduces the number of samples in the majority class to match the minority class, as shown in Fig. 2. Oversampling increases the number of samples in the minority class to match the majority class, as shown in Fig. 3. These methods provide the models with balanced learning data to prevent bias in diabetes prediction.

1) NearMiss: NearMiss undersampling [21] reduces the number of majority class samples by selecting those "near" the minority class instances, improving model performance in handling imbalanced data.

2) Synthetic Minority Over-sampling Technique (SMOTE): SMOTE [22] balances imbalanced datasets by creating synthetic data for the minority class without duplicating existing data points, generating new instances based on the characteristics of existing minority class samples.

3) Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors (SMOTEENN): The combination of SMOTE and Edited Nearest Neighbors (ENN) [23] handles imbalanced datasets by oversampling the minority class using

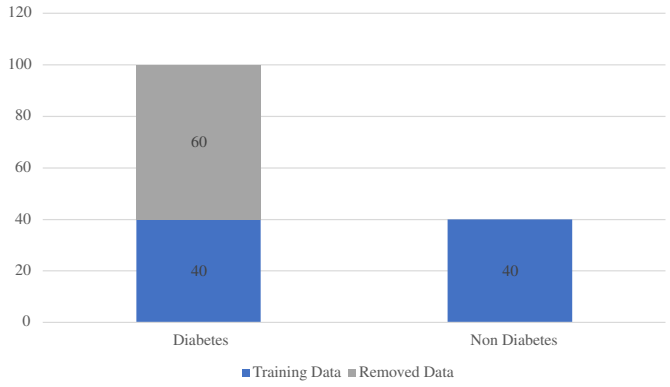

Fig. 2. Undersampling.

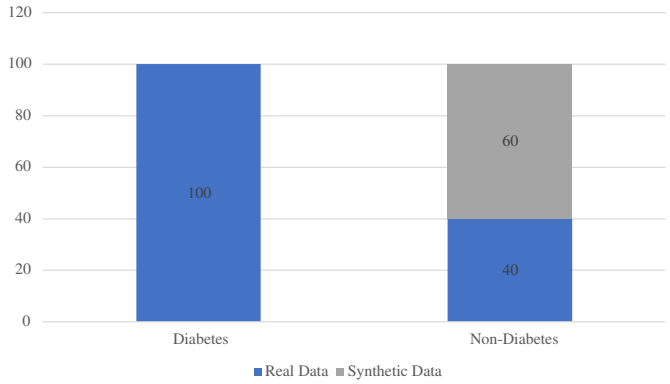

Fig. 3. Oversampling.

SMOTE and cleaning noisy samples with ENN, resulting in a more balanced and refined dataset.

4) SMOTETomek: SMOTETomek [24] combines SMOTE and Tomek links to address imbalanced datasets by oversampling the minority class using SMOTE and removing overlapping instances with Tomek links.

C. Training and Testing

This research employs some popular low-cost machine learning classifiers [25], [26] to predict diabetes and aims to identify the model that provides the highest accuracy.

1) Logistic Regression: Logistic Regression (LR) [27] estimates the probability of an event by modeling the relationship between features and a binary outcome. It uses the logistic function, or sigmoid function, to transform the output of a linear combination of predictors into a probability score between 0 and 1. With a prediction threshold set at 0.5, the model decisively determines the event's occurrence based on whether the probability estimate surpasses this threshold. The formulation of LR can be defined as

$$\sigma(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{x}}}, \tag{2}$$

where $\sigma(\mathbf{x})$ represents the binary probability between 0 and 1, e is the mathematical constant approximately equal to 2.718, and $\mathbf{x}$ represents the data input.

2) Random Forest: Random Forest (RF) [28] is an ensemble learning method that constructs multiple decision trees during training and merges their predictions for more robust and accurate results. Each decision tree is built using a random subset of the training data and features to prevent overfitting. During prediction, each tree "votes" on the outcome, and the final result is determined by the majority vote. Formally, this process is formulated as:

$$\hat{y} = \text{MajorityVote}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_B), \tag{3}$$

where $\hat{y}$ is the predicted class, $\hat{y}_i$ = represents the predicted class by each decision tree, and MajorityVote is the function that selects the class with the most votes.

3) K-Nearest Neighbors: The K-Nearest Neighbors (KNN) [29] algorithm makes predictions based on the majority class or the average of the nearest neighbors' values. KNN identifies the closest points in the training set using a distance metric and assigns the class label by majority voting among these neighbors. The predicted class is formulated as:

$$\hat{y} = \arg\max \left(\sum_{i=1}^{k} I(y_i = c_j) \right). \tag{4}$$

Here, $\hat{y}$ is the predicted class, y_i represents the class of the k-nearest neighbors, c_j represents the classes, and I is the indicator function returning a boolean value.

4) Support Vector Machine: Support Vector Machine (SVM) [30] operates by finding an optimal hyperplane that separates different classes in the input features. By maximizing the margin, or distance, between the hyperplane and the nearest data points from each class, SVM ensures precise and effective classification.

$$\hat{y} = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b), \tag{5}$$

where $\hat{y}$ represents the predicted class for a new input $\mathbf{x}$, $\mathbf{w}$ is the weight vector perpendicular to the hyperplane, $\mathbf{x}$ represents the input features, and b is the bias term.

IV. Experiments

This section reports the performance of each model, starting from the prediction of the raw dataset, feature selection, and prediction after sampling. The predictive model that provides the highest accuracy will be chosen as the best model for diabetes prediction.

A. Dataset

The Diabetes 130-US hospitals dataset for 1999-2008 contains 101,766 records, comprising 78,363 diabetic and 23,403 non-diabetic patients. Fig. 4 illustrates the data distribution for each class. The dataset includes 51 features, consisting of 14 numerical and 37 categorical features. These features encompass patient administration details, number of medications, types of medicines, and diagnoses.

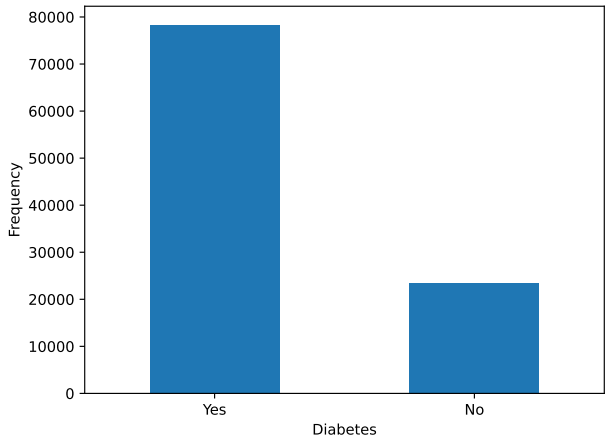

Fig. 4. Total number of diabetic patients.

TABLE I
TRANSFORMATION OF AGE FEATURES.

Age	Transformed Age
[0-18)	Child
[18-30)	Young Adult
[30-50)	Middle-Aged
[50-100)	Senior

The dataset has missing values, which will be processed later in the experiment. For our study, the dataset is split into an 80:20 ratio for training and testing purposes. This ensures that the predictive models can accurately predict unseen data, separate from the training data.

B. Results and Discussion

Rows containing missing values were removed, and features with approximately 80% of null values were also excluded. Feature transformation was performed to facilitate model interpretation, as shown in Table I. Data labeling using Scikit-learn's Label Encoder converted all categorical data into numerical data, improving the models' ability to understand the dataset's structure.

After data cleaning and splitting the dataset with an 80:20 ratio, initial testing was conducted on the proposed models to evaluate their performance on the raw dataset, as shown in Table II. The results indicated that the Random Forest model achieved an accuracy of 99.8%, while the other models' accuracies ranged from 70.0-77.0%.

TABLE II
DIABETES PREDICTION USING RAW DATASET.

Performance Metrics	Model (%)			
	LR	KNN	RF	SVM
Accuracy	77.0	70.0	99.8	77.0
Precision	77.0	77.4	100.0	77.0
Recall	100.0	86.7	99.7	100.0
F1-Score	87.0	81.8	99.8	87.0

TABLE III
SELECTED FEATURES.

Features	Score
num_medications	14334
change	11704
diag_1	5903
insulin	3391
metformin	944
num_lab_procedures	900
time_in_hospital	772
diag_2	541
diag_3	509
glipizide	391
number_emergency	298
glyburide	278
number_inpatient	176
pioglitazone	144
number_outpatient	127
rosiglitazone	110
readmitted	108
glimepiride	71

TABLE IV
DIABETES PREDICTION WITH FEATURE SELECTION.

Performance Metrics	Model (%)			
	LR	KNN	RF	SVM
Accuracy	90.7	73.0	99.3	76.9
Precision	94.0	77.9	100.0	76.9
Recall	93.9	90.5	99.1	100.0
F1-Score	93.9	83.8	99.5	86.9

Chi-square analysis was employed to select features that strongly correlate with the diabetes label. Using a critical value of 0.05 [20], the features with the highest chi-square scores were identified. These selected features, which are shown in Table III, play a significant role in predicting diabetes and enhance the performance of the predictive models.

The results of diabetes prediction using the selected features are shown in Table IV. The performance of Random Forest decreased by 0.5%, while Logistic Regression showed an improvement of 13.7%. KNN improved slightly by 3.0%, and SVM showed a marginal improvement of 0.1%. The selected features were applied as a new dataset, and sampling techniques were compared using this refined dataset.

Table VI compares each sampling technique applied to the new dataset. SMOTE showed a significant decrease in accuracy, with a 15.2% drop for KNN and a 17.3% drop for SVM, while Logistic Regression and Random Forest maintained their performance compared to the raw dataset prediction shown in Table II. Among the sampling techniques, Logistic Regression performed the best with SMOTE, achieving an accuracy of 90.7%. SMOTETOMEK resulted in the second-best accuracy with 90.3%, followed by SMOTEENN at 87.3%, and NearMiss had the lowest performance with 86.7%.

Logistic Regression initially showed a lower accuracy of 77% in the raw prediction, as shown in Table II. However, after data preprocessing and feature selection, the accuracy improved significantly to 90.7%, marking the best performance

TABLE V
MEAN EXECUTION TIME FOR EACH MODEL.

	LR	KNN	RF	SVM
Execution Time (seconds)	8.03	57.27	94.69	6,307.53

TABLE VI
ACCURACY COMPARISON BETWEEN SAMPLING TECHNIQUES.

Sampling Technique	Model Accuracy (%)			
	LR	KNN	RF	SVM
SMOTE	90.7	57.8	99.3	59.6
Near Miss	86.4	46.9	99.3	51.3
SMOTEENN	87.3	49.8	99.3	40.7
SMOTETOMEK	90.3	57.9	99.3	59.5

for this model in this study. Additionally, Logistic Regression demonstrated the fastest execution time among all the proposed models in every test, as detailed in Table V.

+KNN achieved the highest accuracy, 57.9%, using SMOTETOMEK, followed by SMOTE at 57.8%, SMOTEENN at 49.8%, and NearMiss at 46.9%. The raw dataset prediction showed 70% accuracy for KNN, and feature selection significantly improved the accuracy to 73% compared to other sampling methods.

Random Forest demonstrated the best accuracy in raw dataset prediction with 99.8% accuracy, as shown in Table II. However, feature selection and all sampling techniques slightly reduced the model's accuracy to 99.3%.

SVM performed similarly to Logistic Regression, achieving 77.0% accuracy in raw dataset prediction. Feature selection reduced the accuracy by 0.1%, and applying sampling techniques significantly decreased the accuracy to 40.7-59.6%, as shown in Table VI.

Performance comparison between SMOTE, raw prediction, and feature selection prediction can be seen in Fig. 5. The results indicate that Logistic Regression and KNN perform better when the dataset is combined with feature selection. In contrast, Random Forest and SVM performed best with the raw dataset.

The Random Forest model in this study outperformed the other proposed models, achieving the best performance with an accuracy of 99.8% without any feature selection or sampling techniques applied, as shown in Table II. This performance surpasses the models proposed in previous research, as detailed in Table VII. Additionally, Random Forest demonstrated a 20.8% faster execution time in this study.

TABLE VII
ACCURACY COMPARISON WITH PREVIOUS RESEARCH.

Reference	Model	Accuracy
Kuriakose *et al.* [7]	LR	76%
	KNN	69%
	RF	79%
	SVM	77%
Highest Accuracy Proposed Models	LR	90,7%
	KNN	73%
	RF	99,8%
	SVM	77%

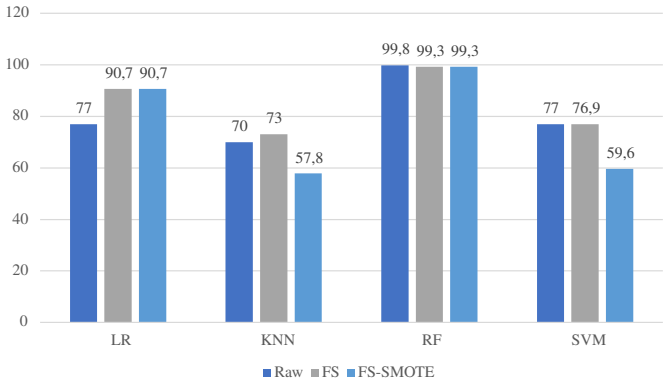

Fig. 5. Comparison of sampling and no sampling techniques.

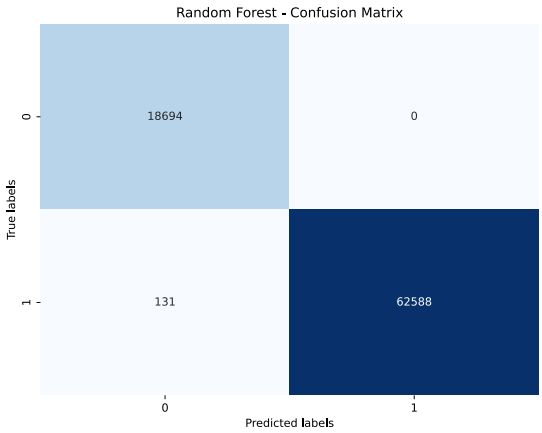

Fig. 6. Confusion matrix for random forest on raw dataset prediction.

In the raw dataset prediction, Random Forest accurately predicted 18,694 negative diabetes cases with no errors, as illustrated in Fig. 6. However, the model misclassified 131 positive diabetes cases as negatives out of 62,588 data points.

Several strategies could be implemented to address false negative errors. Integrating advanced feature engineering techniques and exploring more sophisticated machine learning algorithms can enhance diabetes prediction accuracy. Specifically, incorporating deep learning models, such as neural networks [31], [32], and integrating local features of multimodality [33], [34], could provide significant improvements. Expanding the dataset by including more diverse patient demographics and health parameters, as demonstrated in [35], will allow for a more comprehensive analysis.

V. CONCLUSION

After conducting several diabetes prediction tests using various sampling techniques to handle data imbalance and improve accuracy, the results indicated that SMOTE provides

the best performance compared to other sampling techniques. However, proper data preprocessing is essential to achieve optimal results for all predictive models.

Diabetes prediction with this dataset was best performed by Random Forest, which demonstrated remarkable performance, surpassing other proposed models. Despite having lower accuracy than Random Forest, Logistic Regression showed promising growth and still leaves room for improvement. Future research will involve further exploration of Logistic Regression, Random Forest, and other robust models and techniques to enhance the quality of diabetes prediction.

REFERENCES

[1] M. Laakso, "Epidemiology of type 2 diabetes," in *Type 2 Diabetes*. CRC Press, 2016, pp. 15–26.

[2] R. Ramtahal, C. Khan, K. Maharaj-Khan, S. Nallamothu, A. Hinds, A. Dhanoo, H.-C. Yeh, F. Hill-Briggs, and M. Lazo, "Prevalence of self-reported sleep duration and sleep habits in type 2 diabetes patients in south trinidad," *Journal of epidemiology and global health*, vol. 5, no. 4, pp. S35–S43, 2015.

[3] C. Fatichah, M. F. M. Robby, S. C. Hidayati, and T. Mustaqim, "Detection of covid-19 based on lung ultrasound image using convolutional neural network architectures," in *Proc. International Seminar on Research of Information Technology and Intelligent Systems*, 2021, pp. 155–160.

[4] D. Aulia, R. Sarno, S. C. Hidayati, and M. Rivai, "Optimization of the electronic nose sensor array for asthma detection based on genetic algorithm," *IEEE Access*, vol. 11, pp. 74 924–74 935, 2023.

[5] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the adap learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Symposium on Computer Applications and Medical Care*, 1988, pp. 261–265.

[6] "Early Stage Diabetes Risk Prediction," UCI Machine Learning Repository, 2020, DOI: https://doi.org/10.24432/C5VG8H.

[7] S. M. Kuriakose, P. B. Pati, and T. Singh, "Prediction of diabetes using machine learning: Analysis of 70,000 clinical database patient record," in *Proc. International Conference on Computing Communication and Networking Technologies*. IEEE, 2022, pp. 1–5.

[8] N. Ahmed, R. Ahammed, M. M. Islam, M. A. Uddin, A. Akhter, M. A. Talukder, and B. K. Paul, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021.

[9] L. Jiang, Z. Xia, R. Zhu, H. Gong, J. Wang, J. Li, and L. Wang, "Diabetes risk prediction model based on community follow-up data using machine learning," *Preventive Medicine Reports*, vol. 35, p. 102358, 2023.

[10] S. K. Kalagotla, S. V. Gangashetty, and K. Giridhar, "A novel stacking technique for prediction of diabetes," *Computers in Biology and Medicine*, vol. 135, p. 104554, 2021.

[11] E. Massip, S. C. Hidayati, W.-H. Cheng, and K.-L. Hua, "Exploiting category-specific information for image popularity prediction in social media," in *Proc. IEEE International Conference on Multimedia Expo Workshops*, 2018, pp. 45–46.

[12] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digital health*, vol. 6, p. 2055207620914777, 2020.

[13] S. C. Hidayati, M. R. Fiqih Thalib, and A. Munif, "The influence of user profile and post metadata on the popularity of image-based social media: A data perspective," in *Proc. International Conference on Artificial Intelligence in Information and Communication*, 2024, pp. 806–811.

[14] G. Marzani, N. Suciati, and S. C. Hidayati, "Morphological preprocessing for low-resolution face recognition using common space," in *Proc. International Conference on ICT for Smart Society*, 2021, pp. 1–5.

[15] A. Widyadhana, S. C. Hidayati, D. A. Navastara, and Y. Anistyasari, "ASF-LLRDA: Locality-regularized linear regression discriminant analysis with approximately symmetrical face preprocessing for face recognition," in *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2023, pp. 2031–2036.

[16] N. Junsomboon and T. Phienthrakul, "Combining over-sampling and under-sampling techniques for imbalance dataset," in *Proc. international conference on machine learning and computing*, 2017, pp. 243–247.

[17] S. C. Hidayati and Y. Anistyasari, "Body shape calculator: Understanding the type of body shapes from anthropometric measurements," in *Proc. ACM International Conference on Multimedia Retrieval*, 2021, pp. 461–465.

[18] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100032, 2021.

[19] N. H. Taz, A. Islam, and I. Mahmud, "A comparative analysis of ensemble based machine learning techniques for diabetes identification," in *Proc. International Conference on Robotics, Electrical and Signal Processing Techniques*. IEEE, 2021, pp. 1–6.

[20] O. S. Bachri, M. H. Kusnadi, and O. D. Nurhayati, "Feature selection based on chi square in artificial neural network to predict the accuracy of student study period," *International Journal of Civil Engineering and Technology*, vol. 8, no. 8, 2017.

[21] N. M. Mqadi, N. Naicker, and T. Adeliyi, "Solving misclassification of the credit card imbalance problem using near miss," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–16, 2021.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[23] M. Lamari, N. Azizi, N. E. Hammami, A. Boukhamla, S. Cheriguene, N. Dendani, and N. E. Benzebouchi, "Smote–enn-based data sampling and improved dynamic ensemble selection for imbalanced medical data classification," in *Proc. International Conference of Advanced Computing and Informatics*, 2021, pp. 37–49.

[24] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement performance of the random forest method on unbalanced diabetes data classification using smote-tomek link," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 1, pp. 258–264, 2023.

[25] S. C. Hidayati, Y.-L. Chen, C.-L. Yang, and K.-L. Hua, "Popularity meter: An influence- and aesthetics-aware social media popularity predictor," in *Proc. ACM International Conference on Multimedia*, 2017, pp. 1918–1923.

[26] Z. D. Meilani, I. M. Nizar, M. F. Sunandar, and S. C. Hidayati, "Lawyering social: Navigating legal issues on social media posts with a low-cost data algorithm," in *Proc. International Conference on Informatics and Computational Sciences*, 2021, pp. 266–271.

[27] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.

[28] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[29] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[30] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, pp. 389–422, 2002.

[31] F. Aofa, P. S. Sasongko, Sutikno, Suhartono, and W. A. Adzani, "Early detection system of diabetes mellitus disease using artificial neural network backpropagation with adaptive learning rate and particle swarm optimization," in *Proc. International Conference on Informatics and Computational Sciences*, 2018, pp. 1–5.

[32] R. Sofiana and S. Sutikno, "Optimization of backpropagation for early detection of diabetes mellitu," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 5, pp. 3232–3237, 2018.

[33] D. D. Thang, S. C. Hidayati, Y.-Y. Chen, W.-H. Cheng, S.-W. Sun, and K.-L. Hua, "A spatial-pyramid scene categorization algorithm based on locality-aware sparse coding," in *Proc. IEEE Second International Conference on Multimedia Big Data*, 2016, pp. 342–345.

[34] S. C. Hidayati, K.-L. Hua, Y. Tsao, H.-H. Shuai, J. Liu, and W.-H. Cheng, "Garment detectives: Discovering clothes and its genre in consumer photos," in *Proc. IEEE Conference on Multimedia Information Processing and Retrieval*, 2019, pp. 471–474.

[35] N. Dewilza, S. Suryono, and Y. Kartikasari, "Digital image processing for pancreatic size detection usingmagnetic resonance imaging (mri) in diabetes mellitus case," in *Proc. International Conference on Informatics and Computing*, 2019, pp. 1–5.