Machine Learning Techniques for Predicting Brain Stroke Risk: Addressing Data Imbalance

L B H C Pathmakumara
University of Plymouth
United Kingdom
hchandeepap@gmail.com

R W K T Rajapaksha
Faculty of Computing
NSBM Green University
Sri Lanka
kavishka.r@nsbm.ac.lk

Abstract—**Stroke represents a significant global health concern, markedly influencing morbidity and mortality rates across the world. This research investigates the utilization of various machine learning algorithms to forecast stroke risk, with a specific emphasis on tackling the widespread issue of class imbalance found in stroke prediction datasets. We conduct a thorough assessment of the performance of several models, including Random Forest, K-Nearest Neighbors (K-NN), Naive Bayes, Decision Trees, Support Vector Machines (SVM), and Logistic Regression, analyzing both the original imbalanced dataset and a balanced dataset created using the Synthetic Minority Over-sampling Technique (SMOTE). Our results indicate that although the initial accuracy of the models on the imbalanced dataset was substantial, the implementation of SMOTE provided a more accurate evaluation of model efficacy, with Random Forest attaining the highest accuracy at 92%. This study highlights the critical role of class balancing methods such as SMOTE in improving the predictive capabilities of machine learning models within medical applications. The findings of this research carry considerable implications, as enhanced stroke prediction models can facilitate better patient outcomes and optimize healthcare resource allocation through timely interventions.**

Keywords— stroke prediction, machine learning, class imbalance, SMOTE, healthcare optimization

I. INTRODUCTION

A stroke is a serious disease with a high chance of death that arises when the brain's blood flow is partially interrupted. It ranks as the second leading cause of mortality and disability globally. Annually,15 million stroke victims, 5 million fatalities, and an additional 5 million irreversibly damaged individuals occur globally according to the World Health Organization (WHO) [1]. A stroke has an impact on the sufferer as well as their family, job, and social surroundings. Furthermore, it can affect anyone of any age, gender, or physical condition, despite what the general public believes[2]. A stroke is an abrupt and serious neurological issue that impacts the blood vessels in the brain. It happens when there is a disturbance in the blood flow to a specific area of the brain, leading to a lack of oxygen for the brain's cells. Hemorrhagic stroke and ischemic stroke are the two forms of stroke. The former can cause mild to severe permanent damage, while the latter is rare and results from a blood artery burst that can cause brain hemorrhage. The most common type of stroke is ischemic, occurring when an artery narrows or becomes blocked, halting the flow of blood to a specific area of the brain. [3]. Particularly machine learning (ML) and artificial intelligence (AI) are quickly becoming crucial for the early diagnosis of a variety of illnesses, including diabetes.(as a classification task [4] or regression task for continuous glucose prediction [5]), hypertension [6], cholesterol [7], COVID-19 [8], sleep disorders, hepatitis C [9], CKD [10], etc. During this investigation, we will pay special attention to the stroke. Many research papers have employed machine learning algorithms to study this specific medical condition.

The techniques for developing effective binary classification models to predict the frequency of strokes are presented in this research. The synthetic minority over-sampling (SMOTE) technique was used. [11] to assess the impacts of class balance. The source of the dataset was Kaggle. Initially, the imbalanced dataset was used to construct and assess models. The effect of class balancing on model performance was subsequently demonstrated by reevaluating these models on a balanced dataset generated using SMOTE. Despite extensive research on stroke prediction, there is still a gap in analyzing model performance before and after class balancing procedures. In order to address this, this study uses SMOTE to create binary classification models on both balanced and imbalanced stroke datasets. We highlight the significance of resolving class imbalance for successful stroke prediction models by

analyzing these models in both circumstances and offering insights into how well class balancing improves model performance. We conducted an assessment of Naive Bayes, K-NN, Support Vector Machine, Random Forests, Decision Trees, and Logistic Regression. In our evaluation of both imbalanced and balanced datasets, we illustrated the efficacy of the models by achieving high AUC, precision, recall, F-measure, and accuracy. The comparative analysis emphasizes the importance of class balance for creating effective stroke prediction models.

II. LITRATURE REVIEW

Recent studies have focused on applying machine learning approaches to predict stroke risk, as part of the research community's efforts to build tools and tactics to detect and anticipate serious illnesses. Numerous research have shown how well machine learning models can predict the occurrence of strokes. For example, [12] highlighted machine learning (ML) 's crucial role in determining stroke risk variables and predicting stroke occurrences. By analyzing patient data and predicting stroke using a range of machine learning algorithms, they demonstrated the value of these models in healthcare settings.

The class imbalance in datasets, where non-stroke cases greatly outnumber stroke cases, is one of the main obstacles to stroke prediction. This disparity may cause models to become skewed in favor of the majority class, which would impair the model's predictive power for stroke incidence. Several class balancing methods have been suggested to address this issue, with the Synthetic Minority Over-sampling Technique (SMOTE) being one of them.

[13] introduced SMOTE, a technique that generates synthetic samples for the minority class by creating interpolations based on existing examples of the minority class. This method has been widely used in medical research to enhance machine learning models' performance on unbalanced datasets.

[7] investigated the use of SMOTE in COVID-19 and cholesterol prediction models. According to their findings, SMOTE successfully reduced the issue of class imbalance, improving performance indicators like accuracy, precision, recall, and F-measure.

Several machine learning models have been assessed for their efficacy in the context of stroke prediction. The techniques commonly used are naive bayes, k-nearest neighbors (KNN), support vector machines (SVM), decision trees, random forests, and logistic regression. The advantages and disadvantages of each model vary depending on the dataset and its intended use.

For example, [9] demonstrated the interpretability and simplicity of logistic regression in contrast to the complexity of decision trees by using both logistic regression and decision trees to predict hepatitis C. In a similar vein, Wang et al. (2021) showed the great accuracy and resilience of ensemble approaches like random forests by using random forests and SVMs for the prediction of chronic kidney disease.

Despite advancements, challenges remain in optimizing model performance and ensuring robustness in diverse clinical settings. Addressing class imbalance in medical datasets, which impacts predictive accuracy, is crucial. Techniques like SMOTE are used to mitigate this, but their effectiveness varies across different contexts and datasets. A key research gap is the detailed comparison of model performance before and after applying SMOTE for stroke prediction. Closing this gap is essential to validate advanced class balancing methods.

III. METHODOLOGY

A. Dataset Description

Our study was developed upon a Kaggle dataset (“Stroke Prediction Dataset,” n.d.). There were 5110 participants in total, and the attributes are all stated in Table 1 (One attribute for the target class and ten attributes as input to ML models):

TABLE I. DATASET ATTRIBUTES

Attribute name	Attribute Description
id	A distinct number assigned to every patient
gender	"Male", "Female" or "Other"
age	Patients' ages (1–82)
hypertension	0-no hypertension, 1-hypertension
heart_disease	0-no heart diseases, 1-heart disease
ever_married	"No" or "Yes"
work_type	"Children","Govt_job","Never_worked", "Private","Self-employed"
residence_type	"Rural" or "Urban"
avg_glucose_level	Blood glucose level on average
bmi(Kg/m2)	Body mass index
smoking_status	"Formerly smoked", "never smoked", "smokes", "Unknown"
stroke	1-Stroke, 0- no stroke

(*Source: Author's compilation)*

There are far more non-stroke cases in the original dataset than stroke patients, resulting in an imbalance. It's given in Figure 1.

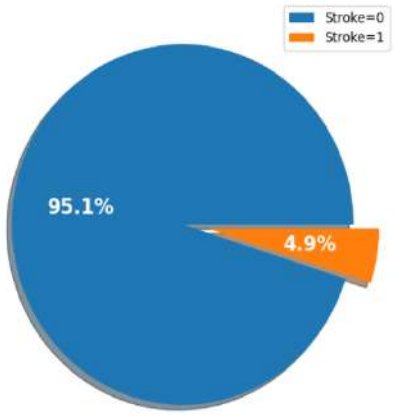

Fig 1-Distribution of Stroke Before apply SMOTE

(Source: Author's compilation)

B. Data Preprocessing

There are 5,110 items in the collection. To preserve data integrity and avoid analytical distortion, careful data preprocessing was essential, including filling in missing values and eliminating outliers. These procedures guaranteed a dependable and clean dataset for additional research and model training. We employed the Synthetic Minority Over-sampling Technique (SMOTE) to tackle the imbalance in the class, generating synthetic samples for the minority class consisting of stroke sufferers. By balancing the dataset, improves the models' capacity to learn and provide precise stroke predictions. Figure 2 - A displays a balanced distribution of the target variable, stroke, after applying SMOTE. This illustrates how SMOTE can effectively solve class imbalance by guaranteeing that machine learning models can predict stroke occurrences properly and impartially toward the majority class. A balanced dataset's age distribution is depicted in Figure 2 - B, with the majority of stroke victims falling into the 70–79 age range, then the 50–59 age group. This emphasizes the necessity of providing older persons with specialized healthcare resources and preventative measures. The gender distribution in the balanced dataset is displayed in Figure 2 - C. Males are more likely than females to experience strokes, with most ladies never experiencing one. This distribution draws attention to gender disparities in the incidence of stroke and implies that these variations may need to be taken into account in targeted interventions.

Figures 2 - D and 3 - A below show heart disease and hypertension prevalence in stroke participants. We note that a significant proportion of stroke survivors in both figures do not have a diagnosis of heart disease or hypertension.As seen in Figure 3 - B, the majority of stroke patients are overweight. This graphic highlights the significance of tracking and controlling weight as a preventive strategy against stroke by suggesting a possible link between increased body weight and the risk of stroke.Figure 3 - C shows that most stroke victims work in private employment. This shows that those in the private sector may have a greater incidence of stroke,

potentially as a result of lifestyle choices connected to job and stress. This realization highlights the necessity of workplace prevention and focused health interventions.

To sum up, the visual aids provide a comprehensive view of stroke distribution based on various characteristics. Key findings include a higher incidence of stroke among older age groups, particularly those aged 70-79, and a noticeable gender disparity, with more women affected than men. Additionally, individuals with private employment and those who are overweight show higher stroke risk. These insights underscore the need for targeted preventive measures and health interventions for at-risk populations, crucial for effective stroke prevention and management strategies.

C. Machine Learning Models

The classifiers utilized in this portion include Naive Bayes, Logistic Regression, Decision Trees, Random Forests, SVM, and KNN for the purpose of predicting strokes.

In our system, the first model that is used is logistic regression (LR) [14]. Logistic regression, initially created for binary tasks, is a statistical technique used for data classification and has since been expanded to handle multi-class issues. The outcome from the model is a binary variable denoted as p=P(Y=1), representing the probability of an instance being in the "Stroke" class. Conversely, 1−p=P(Y=0) represents the likelihood of an instance belonging to the "Non-Stroke" class.The following is the expression for the relationship between the log-odds and the model parameters β_i:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Next we used Decision Trees, which logically and clearly classify data by splitting it into branches based on feature values. Then we employed an ensemble of decision trees called a random forest classifier after the decision tree. Random forests lower the chance of overfitting and increase accuracy by averaging the predictions of several trees. Then we used naive Bayes classifier assumes predictor independence based on Bayes' theorem. Despite its simplicity, it processes huge datasets efficiently and produces accurate results for classification tasks. When classifying a new subject i with characteristics vector f_i the maximum value of $P(c|\ f_{i1},\ldots,\ f_{in})$ corresponds to class c. The conditional probability is defined in the following way:

$$P(c|\ fi1,\ldots,fin)\ = \frac{P(c|\ fi1,\ldots,fin)\ Pc()}{P(\ fi1,\ldots,fin)}$$

Next, the K-nearest neighbors (KNN) classifier uses distance-based methods, such as Manhattan or Euclidean distance, to determine how similar or different two examples are from one another. For a new sample, the KNN finds the nearest K vectors and classifies the sample based on the majority class of its neighbors, aiding in stroke prediction [15]. A support vector machine (SVM) model was also used, which finds the optimal hyperplane with the biggest margin for splitting the data points into different classes. SVMs are flexible with various kernel functions and efficient in high-dimensional spaces.

D. Evaluation

Several performance measures were noted during the ML models under consideration's evaluation procedure. We shall take into account the most widely utilized in the pertinent literature in the current analysis [16].
Recall is the percentage of stroke survivors who, in relation to all positive participants, were correctly classified as positive. Sometimes it's referred to as true positive rate or sensitivity. Precision and recall are better suitable for detecting a model's flaws when dealing with unbalanced data. The accuracy shows how many actual stroke sufferers fit into this group. The recall rate is the percentage of stroke patients who can be anticipated with accuracy. The F1-Score, which offers a summary of a model's prediction ability, is the harmonic mean of precision and recall.

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}'} \qquad \text{Precision} = \frac{\text{TP}}{\text{TP + FP}}$$

$$\text{F1-Score} = 2.\frac{\text{Precision . Recall}}{\text{Precision + Recall}'} \qquad \text{Accuracy} = \frac{\text{TN + TP}}{\text{TN+TP+FN+FP}}$$

Please bear in mind that FP represents false positive, FN represents false negative, TP represents true positive, and TN represents true negative.

IV. RESULTS AND DISCUSSION

In this segment, we compare the results of six different machine learning models—Naive Bayes, Logistic Regression, Random Forests, Decision Trees, SVM, and KNN—both prior to and after implementing SMOTE for predicting stroke risk. The

performance of each model was assessed using accuracy, precision, recall, and F1-score to obtain a complete overview, accounting for class imbalance in the analysis. Table 2 summarizes the performance measures on the unbalanced dataset, while Table 3 presents the metrics after balancing the dataset with SMOTE.

TABLE II. MODEL PERFORMANCE BEFORE SOMTE

ML Model	Precision	Recall	F1-Score	Accuracy
LR	0.92	0.96	0.94	0.96
DT	0.93	0.93	0.93	0.93
RF	0.92	0.96	0.94	0.96
NB	0.94	0.89	0.91	0.89
K-NN	0.96	0.96	0.94	0.96
SVM	0.92	0.96	0.94	0.96

TABLE III. MODEL PERFORMANCE AFTER SMOTE

ML Model	Precision	Recall	F1-Score	Accuracy
LR	0.94	0.74	0.82	0.74
DT	0.93	0.88	0.91	0.88
RF	0.93	0.92	0.92	0.92
NB	0.95	0.91	0.73	0.61
K-NN	0.93	0.8	0.85	0.8
SVM	0.96	0.61	0.72	0.61

(*Source: Author's compilation)*

A. Model Performance Visualization

The effect of correcting class imbalance on accuracy, precision, recall, and F1-score is illustrated through visual depictions of model performance before and after SMOTE. Figure 4-A shows the accuracy of each model, highlighting performance enhancement with balanced data. Figure 4-B compares precision before and after SMOTE, indicating increased precision in stroke prediction. Figure 4-C illustrates recall metrics, showing improved ability to correctly identify stroke patients after addressing class imbalance. Figure 4-D presents the F1-score, combining recall and precision, offering an overview of overall model performance improvement with SMOTE. Models with the highest accuracy before SMOTE are Random Forest, KNN, SVM, and Logistic Regression, each achieving about 96%, while Naive Bayes has the lowest at 89%. After SMOTE, Random Forest retains the highest accuracy at roughly 92%, while Naive Bayes and SVM drop to around 61%. These visualizations and insights highlight the impact of correcting class imbalance on model performance, showing that balancing the dataset with SMOTE enhances performance to varying degrees, with some models benefiting more significantly than others. This demonstrates the importance of addressing class imbalance in improving predictive accuracy and reliability.

V. CONCLUSION

In this work, we examined how class balance affects the predictive power of many machine learning models in stroke incidence using data from Kaggle. The study consisted of two phases: modeling the original imbalanced dataset and modeling a balanced dataset created using the Synthetic Minority Over-sampling Technique (SMOTE). Initial evaluations on the imbalanced dataset showed high accuracy (around 96%) for models like Logistic Regression, Random Forest, KNN, and SVM, while Naive Bayes lagged at 89%. The problem of class imbalance has been highlighted by this difference, where the overwhelming presence of the majority class can give the impression of models being highly accurate. The models' performance metrics saw a substantial alteration following the use of SMOTE. The Random Forest model demonstrated the highest accuracy at 92%, indicating its robustness. Conversely, the accuracy of models like SVM and Naive Bayes dropped to around 61%, showing their sensitivity to changes in data distribution. The dataset balancing generally results in more realistic performance evaluations, as indicated by the comparison of precision, recall, F1-score, and accuracy before and after SMOTE. This is crucial for medical diagnostics,

where false negatives can have serious consequences, such as failing to predict a stroke. The analysis reveals that no single model excels across all metrics. The choice of a model should take into account the particular needs and preferences of the application. For example, following the deployment of SMOTE, Random Forest continued to perform well across a number of metrics. Future research should explore hybrid approaches that combine several models and include more feature engineering and selection strategies to improve predictive performance. Furthermore, putting these models to the test in actual clinical situations would shed light on how reliable and useful they are. To sum up, resolving class imbalance through methods such as SMOTE is essential to building strong and effective machine learning models for stroke prediction. This research underscores the importance of this step in ensuring that models provide accurate and balanced predictions, contributing to early and accurate stroke detection and timely intervention.

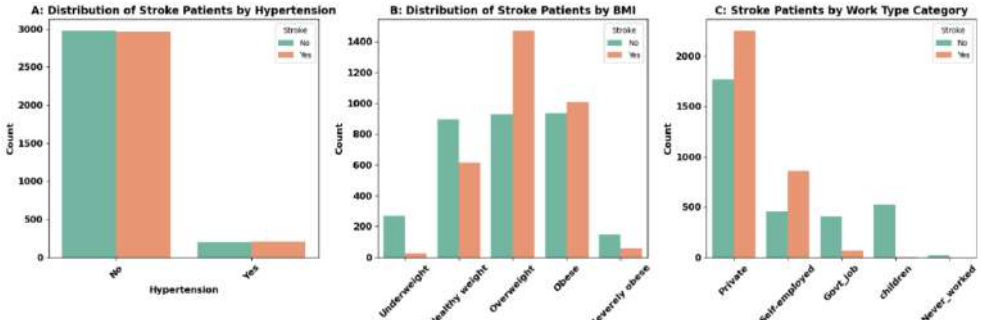

Fig 3 -Hypertension, BMI, Work Type distributions *(Source: Author's compilation)*

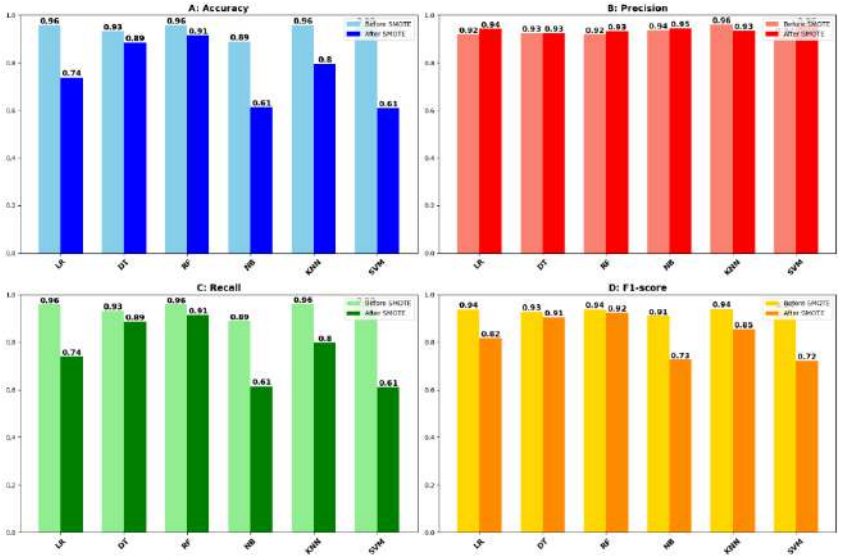

Fig 4-Accuracy, Precision, Recall, F1-Score comparison (*Source: Author's compilation*

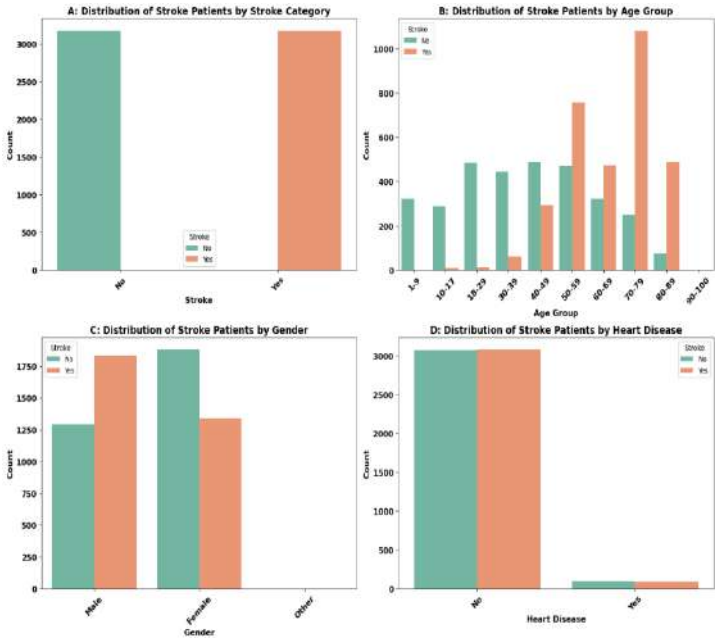

Fig 2 - Balanced dataset, Age, Gender, Heart distributions (Source: Author's compilation)

REFERENCES

[1] "WHO EMRO | Stroke, Cerebrovascular accident | Health topics," World Health Organization - Regional Office for the Eastern Mediterranean. Accessed: Jun. 02, 2024. [Online]. Available: http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

[2] T. Elloker and A. J. Rhoda, "The relationship between social support and participation in stroke: A systematic review," *Afr J Disabil*, vol. 7, p. 357, 2018, doi: 10.4102/ajod.v7i0.357.

[3] A. Bustamante *et al.*, "Blood Biomarkers to Differentiate Ischemic and Hemorrhagic Strokes," *Neurology*, vol. 96, no. 15, pp. e1928–e1939, Apr. 2021, doi: 10.1212/WNL.0000000000011742.

[4] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," *IEEE Access*, vol. 9, pp. 103737–103757, 2021, doi: 10.1109/ACCESS.2021.3098691.

[5] S. Alexiou, E. Dritsas, O. Kocsis, K. Moustakas, and N. Fakotakis, *An approach for Personalized Continuous Glucose Prediction with Regression Trees*. 2021, p. 6. doi: 10.1109/SEEDA-CECNSM53056.2021.9566278.

[6] E. Dritsas, N. Fazakis, O. Kocsis, N. Fakotakis, and K. Moustakas, "Long-Term Hypertension Risk Prediction with ML Techniques in ELSA Database," 2021, pp. 113–120. doi: 10.1007/978-3-030-92121-7_9.

[7] N. Fazakis, E. Dritsas, O. Kocsis, N. Fakotakis, and K. Moustakas, *Long-term Cholesterol Risk Prediction using Machine Learning Techniques in ELSA Database*. 2021, p. 450. doi: 10.5220/0010727200003063.

[8] A. S. Kwekha-Rashid, H. N. Abduljabbar, and B. Alhayani, "Coronavirus disease (COVID-19) cases analysis using machine-learning applications," *Appl Nanosci*, vol. 13, no. 3, pp. 2013–2025, Mar. 2023, doi: 10.1007/s13204-021-01868-7.

[9] M. A. Konerman *et al.*, "Machine learning models to predict disease progression among veterans with hepatitis C virus," *PLOS ONE*, vol. 14, no. 1, p. e0208141, Jan. 2019, doi: 10.1371/journal.pone.0208141.

[10] W. Wang, G. Chakraborty, and B. Chakraborty, "Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm," *Applied Sciences*, vol. 11, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/app11010202.

[11] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Applied Soft Computing*, vol. 76, pp. 380–389, Mar. 2019, doi: 10.1016/j.asoc.2018.12.024.

[12] E. M. Alanazi, A. Abdou, and J. Luo, "Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models," *JMIR Form Res*, vol. 5, no. 12, p. e23440, Dec. 2021, doi: 10.2196/23440.

[13] 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, Jun. 2002, doi: 10.1613/jair.953.

[14] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of Clinical Epidemiology*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/j.jclinepi.2020.03.002.

[15] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, Jul. 2022, doi: 10.1145/3459665.

[16] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *IJDKP*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.