Modified SMOTE and Ensemble Learning Based on Expert Judgment for Chronic Diseases Prediction

Nur Ghaniaviyanto Ramadhan [a], Warih Maharani [b], Alfian Akbar Gozali [c] and Adiwijaya Adiwijaya [d,*]

[a] *School of Computing, Telkom University, Bandung, West Java, Indonesia*
E-mail: nuruer@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0003-0304-516X
[b] *School of Computing, Telkom University, Bandung, West Java, Indonesia*
E-mail: wmaharani@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0003-1574-921X
[c] *School of Applied Science, Telkom University , Bandung, West Java, Indonesia*
E-mail: alfian@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0001-9377-9532
[d] *School of Computing, Telkom University, Bandung, West Java, Indonesia*
E-mail: adiwijaya@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0002-3518-7587

Abstract. Chronic non-communicable diseases such as cancer, stroke, diabetes mellitus (DM), hypertension (HT), chronic kidney failure (CKF), and cardiovascular disease (CVD) have become major health issues worldwide. Another challenge arises when predicting these diseases using datasets from general checkup (GCU) examinations. One of the problems is the imbalance in the number of positive and negative classes in the data. In addition, doctors need additional information from GCU data to provide preventive therapy to people at risk of developing chronic diseases in the future. This can be achieved by integrating expert knowledge with machine learning models. This research aims to predict chronic diseases using a single type of GCU data. Another objective is to modify the synthetic minority oversampling technique (SMOTE) to handle imbalanced data and implement voting ensemble learning based on expert judgment. The results show that the proposed model improves the prediction performance by 10% to 47% compared to traditional models. This system provides guidance to medical professionals to perform preventive interventions more accurately and efficiently, helping to improve the quality of life of patients.

Keywords: Chronic disease prediction, GCU dataset, Weighted SMOTE, Tree-based ensemble learning

1. Introduction

Non-communicable diseases such as cancer, stroke, DM, HT, CKF, and CVD have become major health concerns worldwide [34]. These diseases have a significant impact on the quality of life of individuals and communities and impose a substantial economic burden on healthcare systems [30]. Effective management and comprehensive prevention strategies are critically important [1],[21],[35]. This includes health education [12], managing risk factors [20], [25], [43], and predicting disease likelihood [2]. These efforts are key to reducing the prevalence and impact of these diseases. In addressing this challenge, innovations in medical and health technology are highly needed.

*Corresponding author. E-mail: adiwijaya@telkomuniversity.ac.id.

Several studies on chronic disease prediction are continually improving to develop more accurate prediction methods. Almadani in their research, predictions were made using data mining techniques to identify patients with the highest likelihood of experiencing a stroke [3]. However, the model applied in the study did not include any modifications or the addition of variables. In their research, Latha applied an ensemble strategy to improve the accuracy of CVD risk prediction based on existing risk factors [24]. This strategy achieved a maximum improvement of 7% in the precision of the prediction. However, the ensemble strategy used only existing machine learning models combined without incorporating additional variables.

Fitriyani *et al*. proposed an early prediction model for diabetes mellitus and hypertension based on individual risk factor data [15]. The study also developed a mobile application to provide a practical tool. However, the data used was derived from four different secondary datasets to predict these two diseases. Ren et al. studied the problem of predicting chronic kidney disease in hypertensive patients using a hybrid model combining Bidirectional Long Short-Term Memory (BiLSTM) and an autoencoder network [38]. Howlader et al. conducted an identification of significant attributes and a prediction of diabetes mellitus [19]. The feature identification techniques used included various methods, such as information gain and analysis of variance (ANOVA). However, these studies did not incorporate expert judgment in identifying features and risk factors.

Su *et al*. identified the main issue in their research as the low generalizability of the prediction model, caused by an imbalanced dataset [42]. The study addressed this by grouping data based on age categories using a feature compensation technique. However, the adaptation technique did not incorporate expert judgment in the synthetic data generation process, nor did it include weighted variables in the SMOTE algorithm. Castellanos et al. addressed issues related to the maximum rule and intersection rule in datasets for DM, cancer, and CVD [9]. The rules were generated through their classification model. However, the depth of the rules produced by the algorithm remained fixed (unable to increase or decrease), and the generated rules were not derived from healthcare experts.

Based on the limitations of previous studies, such as reliance on secondary (public) datasets, the lack of integration between expert judgment and machine learning models, and the absence of weighted variables in the SMOTE algorithm, these constraints have led to prediction accuracy that could still be improved and limited generalization ability, especially on imbalanced datasets. Therefore, this study proposes a new model for predicting several chronic diseases based on expert judgment. Specifically, the main contributions of our research are as follows: First, using a single type of primary dataset (general checkup dataset) to predict multiple chronic diseases. Second, adding weighted variables to the SMOTE algorithm. Third, enhancing prediction performance using tree-based ensemble learning integrated with expert judgment. This study also aligns with SDG 3 (Good Health and Well-Being) by promoting better health outcomes using advanced machine learning techniques and expert knowledge.

This paper is structured as follows: Section 1 discusses the background of the research conducted. Section 2 reviews related works on chronic disease prediction using machine learning and approaches to addressing data issues. Section 3 describes the research methodology applied in this study. Section 4 and 5 presents the research findings and discusses the results. Finally, Section 6 concludes the research findings and highlights potential future works.

2. Related Works

2.1. Handling Imbalanced Data

Lopez Martinez *et al*. conducted research on HT prediction using a dataset derived from questionnaires in the US region. Imbalanced data handling was applied using the SMOTE technique, which successfully improved the F1-score by 29.6%. The F1-score increased from 47.4% before applying SMOTE to 77% after its application [26]. However, details such as the number of samples after applying SMOTE were not provided, and the validity of the questionnaire data used was unclear.

Ramezankhani *et al*. specifically examined the impact of the SMOTE oversampling technique on the performance of three classifiers for predicting diabetes mellitus. The study also analyzed the percentage of synthetic data generated, applying values ranging from 100% to 700%. The best F1-score was achieved by generating synthetic data equivalent to 700% of the minority class in the training data. The F1-score increased from 33.6% before applying SMOTE to 43.6% after its application, indicating that SMOTE improved performance by 10% [37].

Azad *et al*. discussed the application of SMOTE, genetic algorithm, and decision tree models for disease prediction. The study also examined the impact of different training-testing data proportions on prediction results. The dataset used was obtained from the National Diabetes and Kidney Disease Institute. The training-testing proportions applied were 60-40, 65-35, 70-30, 75-25, and 80-20. The best prediction results were achieved with an 80-20 dataset split, yielding an F1-score of 78.38% and an AUC-ROC of 78.62% [6]. However, the study did not report the size of the dataset after applying SMOTE.

Sreejith *et al*. proposed a framework to address class imbalance and feature selection issues. An enhanced SMOTE technique using the Orchard algorithm was applied to handle imbalance, while feature subset selection was used for feature selection. Three public datasets from the UCI repository were utilized, including the Pima Indian Diabetes (PID) dataset. The F1-score achieved on the PID dataset was 89% [41]. However, the dataset balancing process was applied to the entire dataset rather than just the training data, which is not ideal. As stated by, Ramadhan et al., dataset balancing should be performed specifically on the training data [32].

Maldonado *et al*. proposed an enhancement to the SMOTE algorithm by introducing feature weighting, named Feature-Weight SMOTE (FW-SMOTE). This approach replaces the Euclidean distance with the Induced Minkowski OWA Distance (IMOWAD). Additionally, the method integrates feature selection techniques, such as direct feature ranking, into the oversampling process [29]. However, feature ranking is often specific to particular datasets and may not generalize well across diverse domains. Another limitation is that FW-SMOTE relies on filter-based methods, such as mutual information and correlation scores, for feature ranking. These methods might miss opportunities for better feature selection, which could be achieved through the integration of expert judgment tailored to the problem domain.

Wang *et al*. introduced an adaptive weighted oversampling method that combines the Support Vector Machine (SVM) algorithm with the SMOTE technique, called Adaptive Weighting SMOTE (AWSMOTE). This approach addresses a key limitation of traditional SMOTE, specifically the collinearity problem between synthetic and original samples. The variable weights are determined based on estimation vectors from SVM [44]. However, the method heavily relies on the SVM model to distinguish between support vectors and non-support vectors, which limits its applicability to SVM-based models and leaves its potential unexplored for other methods, such as ensemble or decision tree-based approaches. Another limitation is the absence of datasets with extreme imbalance ratios, such as 1:100, in the evaluation, which restricts the validation of the method in more challenging scenarios.

Fahrudin *et al*. proposed an approach called Attribute Weighted and KNN Hub on SMOTE (AWH-SMOTE). Attribute weighting was implemented using four methods: Wojna1, Wojna2, Scaled Misclassification Ratio (SMR) Weight, and Information Gain [13]. However, the selection of the attribute weighting method was performed randomly. A key limitation of this study is the lack of evaluation on datasets with extremely imbalanced ratios, such as 1:100. Additionally, the approach has not been tested with other machine learning algorithms, such as ensemble learning, to explore its broader applicability.

2.2. Chronic Diseases Prediction

Sorayaie Azar *et al*. conducted cancer prediction using six machine learning models: K-Nearest Neighbours (KNN), SVM, Decision Tree (DT), Random Forest (RF), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost). The dataset faced challenges such as class imbalance and an excessive number of features. To address these issues, SMOTE was employed for imbalanced data handling, and feature selection was applied to identify relevant features. The best prediction performance was achieved with the RF model, yielding an F1-score of 71.78% and an AUC of 82.38% [40].

Kibria *et al*. employed a soft voting ensemble approach for predicting diabetes mellitus. The dataset used was the public Pima Indian dataset, which faced the issue of class imbalance. SMOTE-Tomek was utilized to handle the imbalance. The voting ensemble model combined XGBoost and RF, while several standalone machine learning models, such as AdaBoost, XGBoost, RF, SVM, and Logistic Regression (LR), were used for comparison. The soft voting ensemble model achieved an F1-score of 89% and an AUC of 95%, outperforming the standalone machine learning models [22]. However, a limitation of the study is that the data used was secondary and widely used by other researchers, with no direct validation by medical experts to ensure the synthetic data's interpretation aligns with clinical realities.

Ashfaq *et al*. analysed the application of several ensemble models, including stacking, bagging, and voting, for diagnosing CVD. The study utilized the Cleveland dataset from the UCI open repository. The best accuracy was achieved with the bagging ensemble model at 86%, while the other ensemble models showed only a 1% difference: 85% for voting and 84% for stacking [5]. However, the study did not specify the individual models used in the voting ensemble. In other disease prediction studies, voting ensembles have been shown to outperform stacking ensembles by a margin of 10% [33]. This is due to the selection of base models being a critical factor in determining prediction outcomes.

Habib predicted CVD by implementing a hard voting ensemble. The base models used for voting were LR, RF, Multi-Layer Perceptron (MLP), and Gaussian Naïve Bayes (GNB). The study also considered several critical factors that increase the risk of CVD, such as the number of cigarettes smoked per day, glucose levels, and blood pressure. Additionally, imbalanced data handling was addressed using random under sampling. The voting model achieved an F1-score of 82% and an AUC of 73% [16].

3. Method

Handling imbalanced datasets in medical data has become crucial as it can lead to inaccuracies in prediction [34]. Additionally, handling imbalanced datasets prior to the machine learning process can improve the quality of prediction models [37]. This study will modify the SMOTE algorithm. The modification was made by adding a weight variable to the algorithm. SMOTE was chosen because, in previous research, it demonstrated superior results compared to other oversampling algorithms such as SMOTE-Tomek and Adaptive Synthetic (ADASYN) [35]. Additionally, SMOTE is independent of data

distribution, so it can be applied to different types of datasets [10]. The traditional SMOTE oversampling algorithm has a significant limitation: the quality of resampled data can be low when minority data points are too far from their nearest neighbours or when neighbouring data points belong to a different class (overlapping) [23].

This issue can be addressed using weighting, which aims to bring data points that are too far apart closer together. The weighting concept can be applied to the attributes of the dataset [13]. Current attribute weighting techniques include information gain [13], correlation score [28], and mutual information [39]. Ramadhan et al. stated in their research that future studies could incorporate medical experts' knowledge into the machine learning prediction process [34]. Therefore, in this study, the attribute weighting technique utilizes weights determined based on expert judgment. Weighting is applied to help the model prioritize relevant variables, enhancing prediction result. Additionally, it also utilizes doctors' knowledge and expertise to assess attribute importance.

Fig. 1 illustrates the expert judgment rules for diagnosing several chronic diseases by identifying the most influential features in the dataset. This expert judgment was obtained through discussions with a team of doctors at the Telkom Health Foundation. The role of expert judgment in this study is to assign weights to the attributes in the dataset. Assigning these weights requires a method to integrate expert judgment with the SMOTE algorithm. In this study, the integration method involves incorporating a weighting formula. Formulas (1) - (3) represent mathematical calculations to generate weight values, which will serve as the weights for each feature in the data.

$$W\alpha = K \times \alpha \tag{1}$$

$$W\beta = K \times \beta \tag{2}$$

$$\left(\sum_{SFE} \times W\alpha\right) + \left(\sum_{NSFE} \times W\beta\right) = 1 \tag{3}$$

Where:

- $\sum SFE$ represents the number of features in the GCU data that are significant according to expert judgment.
- $\sum NSFE$ represents the number of features in the GCU data that are non-significant according to expert judgment.
- K is a base constant used to ensure the total weight equals 1. The value of K in this formula must be determined first before calculating the values of $W\alpha$ and $W\beta$.
- α is a constant used to determine the relative weight difference for significant features, while β is a constant for determining the relative weight difference for non-significant features.
- $W\alpha$ represents the weight value for significant features, while $W\beta$ represents the weight value for non-significant features.
- The value of α is always set to be 10 times greater than β, as features deemed significant by experts are considered to have 10 times more importance than non-significant features.

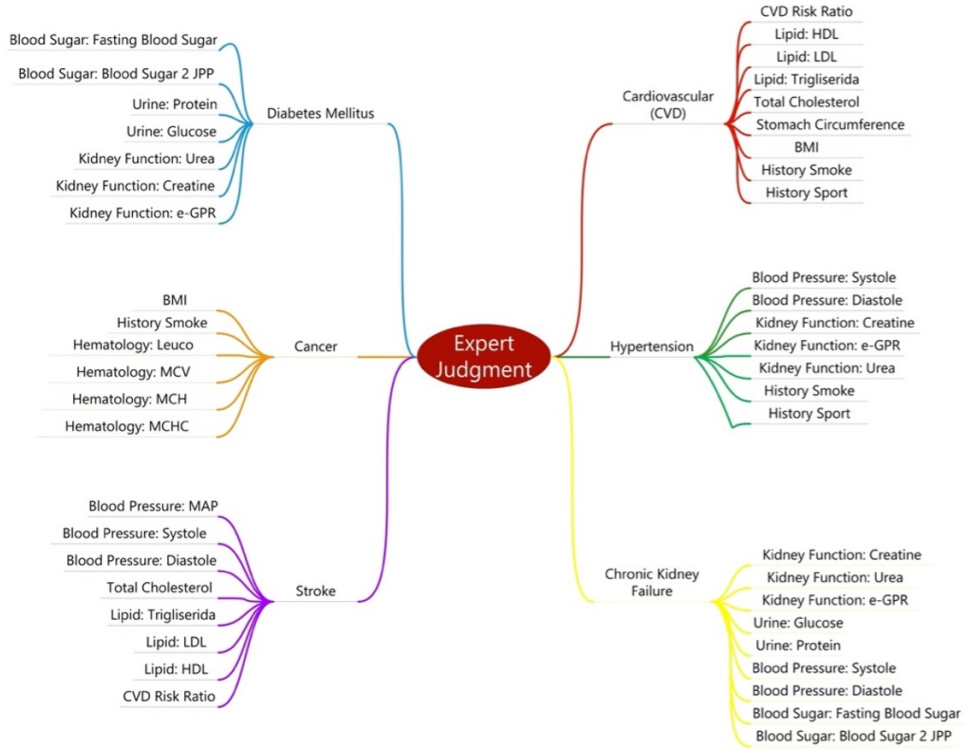

Fig. 1. Expert Judgment Knowledge

The relationship between Fig. 1 and the methodology in this study serves as a conceptual framework that illustrates how Expert Judgment is integrated with machine learning models to enhance prediction outcomes for various medical conditions. Each branch in the figure represents a specific disease or medical condition (e.g., Diabetes Mellitus, cardiovascular diseases, stroke), while the sub-branches denote clinical features or variables identified as significant by experts. These features are selected based on their proven relevance in clinical practice by the experts.

In the methodology, this framework guides the weighting process for features, ensuring that: relevant features are prioritized, meaning variables identified by experts (e.g., blood glucose levels for diabetes or blood pressure for hypertension) are given higher weights compared to non-significant features according to the experts. By basing the feature weighting process on this expert judgment-driven methodology, the approach ensures that the machine learning model is not only data-driven but also clinically informed.

Fig. 2 presents the flow diagram of the proposed research methodology. In this proposed diagram, the process is divided into three stages. The first stage begins with the availability of the GCU dataset for chronic diseases. Exploratory Data Analytics (EDA) is conducted on the GCU dataset to examine its characteristics, structure, and existing issues. Details about the GCU dataset are presented in Section 3.3. The results of the EDA indicate that the GCU data has issues with missing values and outliers. Feature encoding is performed to convert string-type data into integer or numeric formats to facilitate machine learning models in processing the data efficiently [32].

The second stage begins with checking whether the GCU data for each disease is imbalanced. If the data is not imbalanced, the process directly proceeds to the third stage. However, if the data is imbalanced, handling is performed specifically on the training data. The imbalanced data handling is

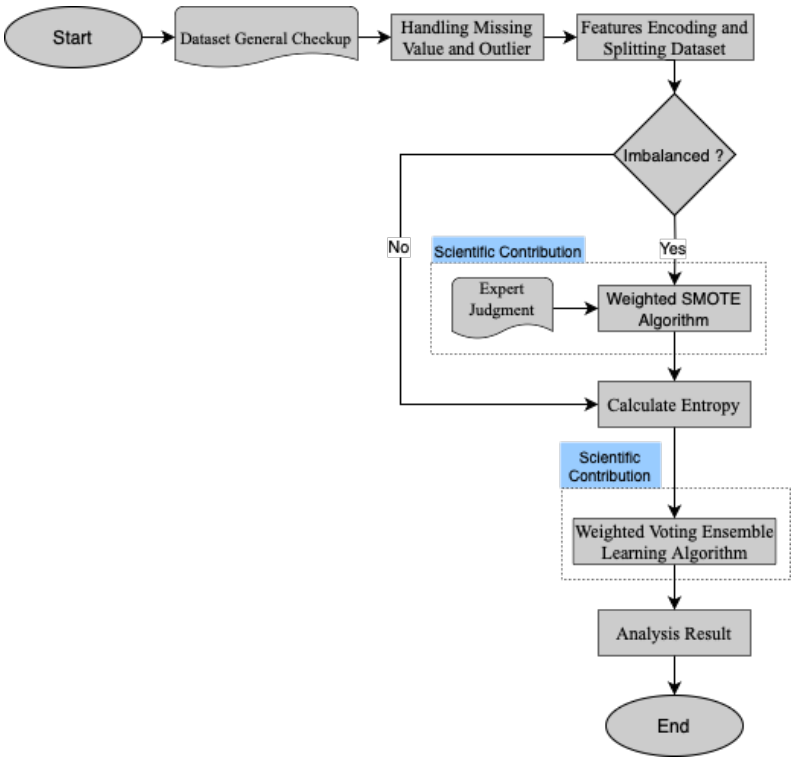

Fig. 2. Proposed Research Diagram

carried out by integrating expert judgment and the SMOTE algorithm through weighting. The result of this handling is that the number of minority class data becomes balanced with the number of majority class data.

The third stage begins with calculating the entropy values for each GCU dataset. Entropy calculation is performed to evaluate the relevance between features [8] and to assist in decision-making within machine learning algorithms [7]. During the entropy calculation process, weighting is also applied to each data attribute. The purpose of this weighting is to ensure that attributes given higher weights are prioritized during the prediction process, enhancing the focus on critical features in the decision-making of the ML model. Consequently, the prediction process utilizes data with entropy values that have already been weighted. This study employs a tree-based voting ensemble learning prediction model.

3.1. Dataset GCU

This study uses a single type of GCU dataset for multiple chronic diseases, obtained from the Telkom Health Foundation in Bandung, with sample collection spanning 2019–2021. The dataset comprises 26 features (including the class label) for six types of chronic diseases: DM, cancer, CVD, stroke, CKF, and HT. The dataset characteristics includes 5 categorical features and 20 numerical features [36]. Details of the GCU dataset used are available in the Zenodo data repository: https://doi.org/10.5281/zenodo.14725457. The EDA results indicate a skewed data distribution, suggesting the presence of noise or outliers. Therefore, outlier removal is necessary to achieve a cleaner and more normal data distribution. Additionally, the dataset faces an imbalanced class issue, where the number of class 0 (negative)

instances significantly outweighs class 1 (positive) instances. Addressing this imbalance is crucial to ensure it does not adversely affect prediction results.

3.2. Handling Missing Value and Outlier

Based on the detection of missing values in the GCU dataset, it was observed that the features with missing values are consistent across all diseases. The feature "history sport" has the highest missing rate: 7.4% in the DM dataset, 7% in the cancer dataset, 6.9% in the CVD and stroke datasets, 8.2% in the CKF dataset, and 8% in the HT dataset, while other features have a lower average missing rate. In this study, missing values will be replaced using the mean value. Outliers in the GCU data will be removed to ensure that the predictions are free from noise and outliers. However, categorical features such as "history smoke," "history sport," "urine protein," and "urine glucose" will not undergo outlier removal, as most values in these features represent general categories, and removing outliers could result in the elimination of data that is valid.

3.3. Weighting of SMOTE

Data is considered highly imbalanced when the imbalance ratio (IR) approaches 0, whereas an IR value close to 1 indicates balanced data [35]. The IR values for the GCU dataset used in this study are presented in Table 1. The formula for calculating the IR can be seen in Formula (4) [31]. In this research, the majority class refers to the negative label, while the minority class refers to the positive label.

$$IR = \frac{\text{Number of Data Minority}}{\text{Number of Data Majority}} \tag{4}$$

Table 1
Imbalanced Ratio GCU Dataset

Dataset	**Label Negative**			**Label Positive**			**Imbalanced Ratio**		
	2019	2020	2021	2019	2020	2021	2019	2020	2021
DM	932	926	911	0	6	21	0	0.00647	0.02305
Cancer	1146	1149	1148	5	2	3	0.00436	0.00174	0.00261
CVD	1207	1202	1198	0	5	9	0	0.00415	0.00783
Stroke	1333	1332	1325	0	1	8	0	0.00075	0.00603
CKF	1200	1199	1196	0	1	4	0	0.00083	0.00334
HT	1221	1217	1188	0	4	33	0	0.00328	0.02777

During the distance calculation in the SMOTE algorithm, Euclidean distance is used, as it is considered the most effective distance metric for determining K [13]. Here is the formula to calculate the euclidean distance using weights.

$$\text{distance}_{i,j} = \sqrt{\sum \left(\frac{x_j - x_i}{\text{weights}} \right)^2} \tag{5}$$

Each variable in the equations (5) is defined as follows: x_j and x_i represent the values or coordinates of two data points whose distance is being calculated. The term *weights* refers to the weight value of the variable $W\alpha$ or $W\beta$. Algorithm 1 is the pseudocode for the traditional SMOTE algorithm, enhanced with weight variables for each data feature. Several steps in the weighted SMOTE process are as follows:

Step 1: Initialize SMOTE Object

In this step, the initial setup for the SMOTE algorithm variables is performed. The variable N is used to determine the number of synthetic samples to be generated for each sample in the minority class. This is typically expressed as a percentage of the minority class samples. The variable K specifies the number of nearest neighbors to be used to find other similar minority class samples. The variable distance defines the distance metric (e.g., Euclidean distance) to measure similarity between samples. The variable weights assigns a weight to each data dimension, allowing specific dimensions to have a greater influence on distance calculations. A blank list named synthetic arr is created to store the synthetic data samples generated by SMOTE. The variable newindex is initialized to 0, serving to track the index of new synthetic samples added to syntheticArr. This variable ensures the new synthetic data is added sequentially to the list as the algorithm runs.

Step 2: Generate Synthetic Points

This step generates synthetic samples. However, before proceeding, it validates the input parameters. If the value of N (percentage) is less than 100, an error is raised. It verifies that the distance metric used is either Euclidean or Ball Tree. If neither is used, an error is raised. It ensures that K does not exceed the number of minority samples. After validation, the algorithm computes the number of synthetic samples to generate: N = N/100. T = the total number of minority samples, is also calculated.

Step 3: Find K Nearest Neighbors

In this step, the algorithm identifies the nearest neighbors for each minority sample to generate synthetic samples. During this process, weights are applied to distance calculations. For each minority sample i is the algorithm calculates the weighted distance to all other samples. These distances are stored in a matrix, sorted, and the K nearest neighbors are selected.

Step 4: Populate Synthetic Samples

This step generates new synthetic samples based on the nearest neighbors. For each sample i in the minority class: Randomly select one neighbor from the K-Nearest Neighbors. For each feature of the sample, calculate the difference between the sample and its neighbor. Generate a new synthetic point along the line connecting the sample and the neighbor using a random gap. The new synthetic sample is added to syntheticarr, and newindex is incremented.

Step 5: Return Synthetic Samples

In this final step, the algorithm finalizes and returns the generated synthetic samples. The syntheticArr list is converted into a NumPy array. The array is returned as output, containing the newly generated synthetic samples. The process continues until the number of minority class samples equals the number of majority class samples.

3.4. Weighting of Ensemble Learning Method

This study employs a machine learning prediction model based on Decision Tree (DT). In addition to DT, the research will implement a voting ensemble using RF (Random Forest), AdaBoost, and XGBoost models. These three models are selected for the voting ensemble because, in several studies on chronic disease prediction, this method has demonstrated robust prediction results. Furthermore, in preliminary experiments conducted by the researchers, this method outperformed other machine learning and deep

Algorithm 1 Weighting of SMOTE

Initialize SMOTE Object
Set N, K, distance metric, and weights
Initialize an empty list `synthetic_arr` to store synthetic samples
Set `newindex` to 0
Generate Synthetic Points
Input Validation:
if $N < 100$ **then**
    Raise an error
end if
if distance metric is not Euclidean or Ball Tree **then**
    Raise an error
end if
if K exceeds the number of minority samples **then**
    Raise an error
end if
Compute the number of synthetic samples to generate:
$N = \text{integer}(N/100)$
$T = \text{len}(\text{min\_samples})$
Find K Nearest Neighbors
if distance metric is Euclidean **then**
    for each sample i in `min_samples` **do**
        Perform weighted calculation using formulas (1)–(3)
        for each sample j in `min_samples` **do**
            Compute the weighted distance between sample i and j
            Calculate the weighted distance using formula (4)
            Store the distance in a distance matrix
        end for
        Sort the distances and select the K nearest neighbors
    end for
else
    Use Ball Tree algorithm to find K for each sample, considering weights
end if
Populate Synthetic Samples
for each sample i in `min_samples` **do**
    Randomly select a neighbor nn from the K nearest neighbors
    for each feature $attr$ **do**
        Compute the difference $diff$ between sample i and neighbor nn
        Generate a synthetic point along the line between i and nn using a random gap
    end for
    Add the new synthetic sample to `synthetic_arr`
    Increment `newindex` by 1
end for
Return Synthetic Samples
Convert `synthetic_arr` to a numpy array and return as synthetic samples

learning methods [32]. Additionally, model ensemble learning has a strong ability to capture non-linear interactions between features [27], which aligns with the characteristics of GCU data that include complex features such as risk factors for disease (lifestyle, age, and genetics). An analysis will be performed to evaluate the differences in prediction results obtained by each of these models.

Attribute weighting is also applied in the voting ensemble model. The weighting process begins by calculating the entropy value for each dataset after data balancing. Additionally, weighting is applied to each attribute using the weighting formula. This ensures that during the voting ensemble process, the model places greater emphasis on attributes with higher entropy values. The entropy calculation is performed using formula (6) [31].

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \tag{6}$$

In the formula:

- $H(X)$ represents the entropy value of the random variable X.
- $p(x_i)$ is the probability of the occurrence of the value x_i in the random variable X.
- n denotes the total number of possible values that the random variable X can take.
- $\log_b$ is the logarithm function, where the base b is typically 2, commonly used in the context of binary classification.

The evaluation metrics used in this study for analyzing the results are F1-score, Receiver Operating Characteristic-Area Under Curve (ROC-AUC), and Balanced Accuracy Score (BAS). The F1-score is utilized as it represents the harmonic mean of precision and recall [17]. ROC-AUC is employed to evaluate the performance of the classification model and to assess how well the model can distinguish between positive and negative classes [14]. BAS ensures that the performance of both classes is weighted equally, providing a more realistic evaluation [18]. The accuracy metric is excluded because it can produce high scores that may cause confusion when analyzing imbalanced datasets, as it tends to focus on the majority class [18]. The algorithm related to the proposed voting ensemble can be seen on Algorithm 2. Several steps in the weighted voting ensemble process are as follows:

Step 1: Define Function to Calculate Entropy
This function is used to measure the uncertainty in a dataset by calculating how diverse the data is. Higher entropy values indicate higher uncertainty or impurity in the dataset. This concept is often used in decision tree algorithms to effectively split nodes.

Step 2: Define Function to Calculate Entropies for All Features
In this step, a function is created to calculate and store the entropy values of all features in the dataset. These entropy values help measure the uncertainty of each feature and can be used for further feature evaluation, such as selecting the most informative feature.

Step 3: Define Feature Weights
In this step, a dictionary called weights is created to map each feature name to a corresponding weight. The dictionary contains the feature names as keys and the weights assigned to those features as values. These weights indicate how important or relevant a feature is in the analysis or predictive model. The weights can be based on factors such as expert consultation, evaluation of the feature's information,

Algorithm 2 Weighting of Voting Ensemble

Define Function to Calculate Entropy:
Compute the frequency counts of values in the column
Convert count to probabilities
Calculate entropy
Return the entropy values
Define Function to Calculate Entropies for All Features:
Create a function `feature_entropies`
Initialize an empty dictionary `entropies`
for each column in the `dataframe` **do**
    Calculate the entropy using `calculate_entropy`
    Store the entropy in the `entropies` dictionary with the column name as the key
end for
Define Feature Weights:
Create a dictionary `weights` mapping each feature name to a corresponding weight
Calculate Entropy for Each Feature:
Calculate entropy using the formula (4)
Calculate Weighted Entropies:
Initialize an empty dictionary `weighted_entropies`
for each feature **do**
    Multiply the feature's entropy by its corresponding weight
    Store the result in the `weighted_entropies` dictionary
end for
Adjust Weighted Entropies:
Add the weighted entropies to the corresponding columns in the training `dataframe` to create a new weighted `dataframe`
Make Predictions:
Use weighted voting ensemble (RF, AdaBoost, and XGBoost)
Evaluate the Model:
Evaluate the model using F1-score, BAS, and AUC

or certain statistical calculations. For example, if a dataset has the features age, blood pressure, and cholesterol, the weights dictionary might look like this:

$$\text{weights} = \begin{cases} \text{'age'} : 0.2, \\ \text{'blood\_pressure'} : 0.5, \\ \text{'cholesterol'} : 0.3 \end{cases}$$

Step 4: Calculate Entropy for Each Feature

In this step, the previously defined feature entropies function is called to calculate the entropy of each feature in the dataset. The function iterates through each column and computes the entropy, which represents the level of uncertainty or diversity in the values within that column.

Step 5: Calculate Weighted Entropies

A blank dictionary called weighted entropies is created to store the weighted entropy calculations for

each feature. This function iterates through each feature in the dataset. Each feature has an entropy value (stored in the entropy dictionary) and a weight (stored in the weights dictionary). The entropy of each feature is multiplied by its assigned weight. This step emphasizes or reduces the uncertainty of a feature based on its importance. For example, if the entropy of the blood pressure feature is 1.2 and its assigned weight is 0.5, the weighted entropy will be calculated as: 1.2 + 0.5 = 1.7

Step 6: Adjust Weighted Entropies

This step helps improve the interpretation and results of the analysis or predictive model. By adjusting the dataset using the weighted entropy values, the model can focus more on features that have a greater impact based on expert assessment or prior calculations. For example, if the blood pressure feature has a weighted entropy of 0.6, its values in the dataframe can be updated to reflect the effect of that weight. This creates a new dataset called the "weighted dataframe," which is used to train the model and better account for the relative impact of each feature.

Step 7 and Step 8: Make Predictions and Evaluate the Model

In this step, predictions are made using a weighted voting ensemble model that combines three models: RF (Random Forest), AdaBoost, and XGBoost. These models make predictions, and the final result is determined based on a weighted voting mechanism across the three models. Once the ensemble model generates predictions, its performance is evaluated by calculating the F1-Score, BAS, and AUC.

4. Result

In this study, four testing scenarios were conducted: Scenario 1 evaluates the extent of differences before and after applying normalization to the weight values. Scenario 2 determines the optimal weight value, ranging from 10 to 10,000. Scenario 3 identifies which model performs the best and assesses the differences before and after applying weighting. Scenario 4 tests the best model for multi-year predictions. The purpose of these four scenarios is to evaluate and optimize the performance of machine learning models under various conditions and to understand the impact of different techniques, such as normalization and weighting.

4.1. Scenario 1

In the first scenario, the objective is to evaluate the differences in applying normalization to the weight values generated from the weighting formula. The normalization criteria for the weight values used in this study are as follows: (1) Significant features identified by experts have a minimum weight of 0.5 and a maximum weight of 0.9. (2) Non-significant features identified by experts have a minimum weight of 0.1 and a maximum weight of 0.4. These minimum and maximum thresholds are determined while ensuring that the total weight of the 26 features used equals 1, and no feature has a weight of 0. The results of this first scenario are presented in Table 2.

Based on Table 2, after the normalization process, there is an improvement in all evaluation metrics for several datasets. Specifically, the F1-Score for the DM dataset increased by 4%, while BAS and ROC-AUC improved by 7%. For the CKF dataset, the F1-Score increased by 5%, whereas BAS and ROC-AUC showed a smaller improvement of 0.3%. These improvements indicate that data normalization combined with weighted SMOTE algorithms and entropy-based voting ensembles can enhance the model's ability to detect positive cases in datasets with class imbalances. Meanwhile, for other datasets such as Cancer, CVD, Stroke, and HT, evaluation metrics remained stable before and after normalization.

Table 2
Result of Scenario 1

Dataset	Before Normalization			After Normalization		
	F1-Score (%)	BAS (%)	ROC-AUC (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)
DM	65	63.5	63.5	69	70.5	70.5
Cancer	83	75	75	83	75	75
CVD	75	66.7	66.7	75	67	67
Stroke	70	66.5	66.5	70	67	67
CKF	60	49.7	49.7	65	50	50
HT	61	61.5	61.5	61	62	62

This stability suggests that the applied method does not degrade the model's performance on these datasets, even though it does not always result in significant improvements. Thus, data normalization can enhance model performance, especially when combined with appropriate oversampling and ensemble techniques. Furthermore, normalization and weighting have proven to contribute positively to the model's ability to capture better patterns, particularly in datasets with significant label imbalances.

4.2. Scenario 2

In the testing of Scenario 2, used normalized weight values, as in several cases, normalization successfully improved prediction performance. The purpose of this phase is to determine the optimal weight value within a specific range (10–10,000). The use of a range with increments in multiples of ten aims to observe whether there are significant jumps in prediction results compared to using shorter increments. The results obtained from this second scenario are presented in Table 3.

Based on the results from Table 3, with a low weight value of 10, the model performance is relatively low, with F1-Score, BAS, and ROC-AUC ranging between 50–75% for most datasets. This indicates that low weights do not give sufficient priority to important features, preventing the model from effectively capturing patterns. At a weight value of 100, a significant improvement is observed in datasets such as DM, where the F1-Score increased by 14%, while BAS and ROC-AUC improved by 4%. This suggests that weighting begins to influence the synthetic distribution in the minority class data. The weight value of 1.000 delivers the best results across almost all datasets, particularly for DM, CKF, and HT. For instance: DM: F1-Score reached 86%, with BAS and ROC-AUC achieving 90%. CKF: F1-Score increased to 65%, with BAS and ROC-AUC both reaching 74%. HT: F1-Score improved from 61% (without weighting) to 75%, while BAS and ROC-AUC rose to 85%. These results demonstrate that a weight value of 1000 allows the model to focus optimally on significant features without causing overfitting.

On the other hand, with a weight value of 10.000, performance decreases for most datasets, such as DM (F1-Score dropping from 86% to 75%) and CKF (F1-Score reverting to 50%). This shows that excessively high weights can lead to overfitting, where the model becomes too focused on the minority class and loses its ability to generalize. Therefore, a weight value of 1.000 provides the best results for most datasets, balancing improvements in the minority class with maintaining the model's generaliza-

tion. However, very high weights, like 10.000, tend to cause overfitting, which lowers performance in most datasets.

Table 3
Result of Scenario 2

Dataset	Weight Value: 10			Weight Value: 100			Weight Value: 1000			Weight Value: 10000		
	F1-Score (%)	BAS (%)	ROC-AUC (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)
DM	68	74	74	80	78	78	86	90	90	75	74	74
Cancer	83	74	75	73	75	75	83	80	80	83	75	75
CVD	75	67	67	75	67	67	78	78	78	75	75	75
Stroke	70	67	67	70	67	67	70	67	67	75	75	75
CKF	50	50	50	60	69	69	65	69	69	65	65	65
HT	61	72	72	61	75	75	75	85	85	71	77	77

4.3. Scenario 3

In this third testing scenario, a weight of 1000 is used to compare the performance of individual models (RF, AdaBoost, and XGBoost) with a tree-based soft voting ensemble combining these three models after applying weighting. The purpose of this testing is to evaluate the impact of applying weights to the SMOTE algorithm and the tree-based voting ensemble, by analyzing the differences in performance before and after the weighting application.

Based on Table 4, the addition of weighting significantly improved model performance across nearly all datasets, particularly in the metrics F1-Score, BAS, and AUC. Before applying weighting, most models recorded low F1-Scores (49–50%) with BAS and AUC stagnating in the 50–65% range. However, after applying weighting, models like Decision Tree (DT) demonstrated drastic improvements in the DM dataset, with the F1-Score increasing from 49% to 86% and BAS/AUC rising from 55% to 88%. A similar pattern was observed in the Cancer and CVD datasets, where the DT model achieved F1-Scores of 83% and 75%, respectively, after weighting was applied. This indicates that weighting effectively enhances the model's ability to capture patterns in datasets with imbalanced classes or complex feature distributions.

Additionally, ensemble models such as XGBoost and Voting Ensemble demonstrated the highest performance after weighting, particularly on the DM dataset, with F1-Score reaching 96% and AUC 97%. Models like RF and AdaBoost also showed significant improvements but remained below the performance of the ensemble models. This improvement highlights that weighting helps ensemble models leverage the combined strengths of individual models, resulting in more accurate and balanced predictions. However, there were cases where the impact of weighting was less pronounced, such as in the Stroke and CKF datasets. This suggests that datasets with less complex data distributions or less informative features may require additional approaches beyond weighting. Overall, these results underline the importance of weighting in enhancing model performance, particularly for datasets with significant class imbalances, such as those with a ratio of 1:1000.

Table 4
Result of Scenario 3

Model	Dataset	Before Weighting			After Weighting		
		F1-Score (%)	BAS (%)	AUC (%)	F1-Score (%)	BAS (%)	AUC (%)
DT	DM	49	55	55	86	88	88
	Cancer	49	49	59	83	85	85
	CVD	49	49	59	75	77	77
	Stroke	49	49	59	70	73	73
	CKF	50	50	50	65	67	67
	HT	53	50	59	65	68	68
RF	DM	49	49	65	83	86	87
	Cancer	49	50	45	90	92	92
	CVD	50	50	74	90	92	92
	Stroke	50	50	64	75	78	78
	CKF	50	50	50	65	66	67
	HT	49	50	63	71	73	73
AdaBoost	DM	49	49	50	77	80	80
	Cancer	50	49	59	55	57	58
	CVD	50	50	55	90	91	91
	Stroke	50	49	41	75	78	78
	CKF	50	50	65	65	67	67
	HT	49	50	56	65	69	69
XGBoost	DM	50	50	45	96	97	97
	Cancer	50	49	45	82	84	84
	CVD	50	50	74	90	92	93
	Stroke	50	50	45	75	78	78
	CKF	50	50	55	65	67	68
	HT	49	50	63	66	69	69
Voting Ensemble	DM	49	49	59	96	97	97
	Cancer	50	49	45	83	85	85
	CVD	50	50	84	90	92	92
	Stroke	50	50	49	75	78	78
	CKF	50	50	68	65	68	75
	HT	49	50	72	73	76	76

4.4. Scenario 4

In this scenario, the voting ensemble model was tested for multi-year predictions. Multi-year predictions refer to forecasting for the next year (Y+1) and for two years ahead (Y+2). The results of this testing are presented in Table 5.

Table 5
Result of Scenario 4

Dataset	Y+1			Y+2		
	F1-Score (%)	BAS (%)	AUC (%)	F1-Score (%)	BAS (%)	AUC (%)
DM	75	78	82	83	86	88
Cancer	83	83	86	83	84	87
CVD	61	66	67	75	73	77
Stroke	60	65	64	75	74	77
CKF	50	51	67	65	67	70
HT	78	75	75	83	81	81

The multi-year prediction results reveal a noticeable difference between the model's performance for short-term predictions (Y+1) and long-term predictions (Y+2). In the DM dataset, there was a significant improvement in F1-Score (from 75% to 83%), BAS (from 78% to 86%), and AUC (from 82% to 88%) when predictions were extended to Y+2. This suggests that long-term trends are more stable, making it easier for the model to identify relevant patterns. A similar pattern was observed in the HT dataset, where scores consistently increased across all metrics for Y+2. This indicates that chronic diseases with clear progression and well-defined risk factors tend to have better predictability over longer timeframes.

Conversely, datasets such as CVD, Stroke, and CKF presented greater challenges, particularly for Y+1, with F1-Scores of 61%, 60%, and 50%, respectively. This may be due to the dynamic nature of these diseases, characterized by sudden complications or high variability in short-term risk factors. However, the scores improved significantly for Y+2, with CKF showing an F1-Score increase from 50% to 65%. This indicates that long-term trends are more predictable, even though CKF remains the most difficult disease to forecast. Overall, the model demonstrates better performance for long-term predictions across most datasets, emphasizing the importance of leveraging stable trends for improved accuracy in chronic disease prediction.

5. Discussion

5.1. Discussion of the results for all scenarios

Weight normalization ensures that significant features have an appropriate influence on the model, allowing it to focus more on relevant patterns. Conversely, less significant features are still considered but with a smaller impact. By normalizing weights, the model reduces the risk of overfitting on minority classes. This enables the model to learn more consistently from synthetic data generated by oversampling

techniques (e.g., SMOTE). Weight normalization helps the model more effectively detect patterns related to minority classes, thus improving evaluation metrics. Without normalization, features with large values can dominate the training process. With normalization, the model can fairly consider each feature based on its significance. On the other hand, the complexity of disease patterns also plays a role. For example, diabetes (DM) data often has more stable risk patterns and clearer predictive features (e.g., blood sugar, BMI), making it easier for the model to identify patterns after significant weights are normalized. In contrast, stroke often involves more dynamic or indirect risk factors (e.g., hypertension, family history), making normalization have a less significant impact on the outcomes.

The optimal weight value was found to be 1,000 because it provides the ideal balance between improving performance for the minority class and maintaining the model's generalization ability, compared to lower values like 10 or higher values like 10,000. At a weight of 1,000, the model can sufficiently prioritize significant features without completely disregarding non-significant ones. This enables the model to capture relevant patterns from both types of features, which is crucial for datasets with complex features. A low weight, such as 10, results in an influence that is too small, making significant features insufficiently helpful for the model in handling the minority class. Conversely, a high weight, such as 10,000, overemphasizes the minority class, which can lead to reduced generalization ability. With a weight of 1,000, synthetic data generated by SMOTE has a more representative distribution for the minority class. At lower weights, the influence of significant features on synthetic data is inadequate. At higher weights, synthetic data may become overly dependent on certain features, leading to unrealistic patterns.

Using excessively high weights risks causing overfitting to the minority class and has implications for reduced model generalization ability. With high weights, such as 10,000, the model becomes overly focused on the minority class and learns patterns specific to the synthetic data. As a result, the model struggles to recognize variations in unseen data during testing. Excessive weights make the model less effective in handling new data, particularly from the majority class, as its primary attention is directed toward the minority class. High weights can also lead significant features to dominate the training process, while non-significant features are completely ignored. This may cause the model to miss important information contained in the non-significant features. Synthetic data generated by SMOTE with overly high weights may not reflect the true distribution of the minority class, reducing the validity of predictions. The implication of overly high weights is that the model may perform well on the training dataset but poorly on the testing dataset, undermining the primary goal of providing reliable predictions. Therefore, it is crucial to maintain weights within an optimal range to maximize the model's ability to capture patterns from both classes effectively.

The tree-based ensemble voting method demonstrates improved performance after weight adjustment because ensemble voting combines the strengths of multiple models (Random Forest, AdaBoost, and XGBoost), enabling it to capture more diverse patterns than individual models. The addition of weights enhances this capability by ensuring a stronger focus on significant features. Individual models, such as Decision Trees, are prone to either underfitting or overfitting [4]. In contrast, ensemble methods benefit from the collective decision-making process, reducing the likelihood of these issues. By leveraging weighted voting, the ensemble becomes better at addressing class imbalances and detecting meaningful patterns, leading to more reliable and robust predictions. Using an ensemble, both bias and variance can be minimized, while the addition of weights helps the model handle class imbalances more effectively. Weights place greater emphasis on significant features relevant to the minority class, enabling the models within the ensemble to leverage this information more optimally compared to individual models.

Tree-based ensembles are more resilient to data imbalances than single models because the voting mechanism ensures that errors from one model can be compensated by others. This collective approach

enhances the model's ability to generalize across diverse patterns in the data while maintaining robustness against the challenges posed by class imbalances [11]. Weights enhance this by making the minority class data more representative during training. Overall, adding weights to tree-based ensemble voting not only improves performance on imbalanced datasets but also enhances the model's ability to identify complex patterns within the data. This ensures that the ensemble leverages its collective strengths to achieve better generalization and more accurate predictions.

The model demonstrates improved performance in long-term predictions (Y+2) compared to short-term predictions (Y+1) because long-term trends tend to be more stable and consistent, allowing the model to identify clearer patterns. The ensemble voting method provides an advantage in recognizing long-term patterns by combining the predictive strengths of multiple algorithms, each capable of detecting different trends. As a result, long-term predictions are generally more stable and reliable for diseases with clear risk patterns, such as diabetes mellitus (DM) and hypertension (HT), as the model can better recognize consistent trends. However, there are challenges in predicting diseases with high variability, such as chronic kidney disease (CKD), compared to diseases with more stable risk patterns like DM or HT. These challenges include dynamic risk factors, limitations in minority class data, the complexity of feature interactions, and the influence of external factors.

5.2. Implications for chronic disease prediction

The implications of this study cover several aspects, including impacts on the healthcare field, technological innovation, and the use of data for decision-making. The modified Weighted SMOTE method allows for a more representative distribution of minority data, enabling the model to learn patterns more effectively from previously underrepresented data. The addition of ensemble learning algorithms such as Random Forest, AdaBoost, and XGBoost strengthens the model, as each algorithm excels in capturing complex patterns. In the context of disease prediction, such as diabetes mellitus (DM) and hypertension (HT), a higher F1-score aids in diagnosing high-risk patients, enabling earlier preventive actions. The improved F1-score also reduces the likelihood of diagnostic errors that could lead to inappropriate treatments.

The use of the GCU dataset sourced from the Telkom Health Foundation provides an advantage in the form of primary data that is more relevant to the local context compared to commonly used secondary datasets. This ensures that the predictions generated are more aligned with real-world conditions. In local clinical settings, primary data can reflect unique disease patterns, such as the prevalence of certain diseases influenced by lifestyle or environmental factors. This also enables the personalization of models, making them more suitable for specific communities or populations.

The addition of weights based on expert assessments enhances the interpretability of the model, which is crucial in clinical decision-making. The model not only provides predictions but also offers insights into which variables are most relevant, such as blood pressure or blood sugar levels. Physicians can use the model's results to design more specific therapies, for example, giving special attention to patients with high blood pressure values in hypertension predictions. The use of an expert knowledge-based model can foster better collaboration between data scientists and medical professionals, resulting in more practical and applicable solutions.

The improvement in chronic disease prediction supports the global agenda to reduce the burden of non-communicable diseases. Conditions such as diabetes and hypertension often go undetected until advanced stages, making early prediction crucial. The implementation of this model can be utilized in mass health screening programs to identify high-risk individuals, who can then be referred for further

care. This model can also be adopted by other healthcare institutions to optimize limited resources, for instance, by focusing on more vulnerable population groups.

The issue of data imbalance, where the majority class (negative class) dominates the minority class (positive class), often causes bias in the model. The modification of Weighted SMOTE by incorporating weights based on relevant attributes provides better data distribution and improves model performance. This algorithm could set a new standard for handling imbalanced medical data, especially for datasets with extreme imbalance ratios such as 1:1000. It also has the potential to extend SMOTE applications beyond disease prediction, for example, in analyzing imbalanced sales data or other rare events.

Multi-year predictions demonstrate that the model can capture long-term trends better than short-term ones. This is crucial because chronic diseases often develop gradually and are influenced by cumulative risk factors. Long-term predictions aid in the formulation of public health strategies, such as targeted health resource allocation or more focused awareness campaigns. Medical institutions can design more measurable population-based prevention programs, emphasizing the prevention of disease progression over several years.

The approach applied can be utilized in other domains facing similar challenges, such as data imbalance or the need to integrate expert domain knowledge. Examples include fraud analysis in financial systems or damage monitoring in the manufacturing industry. In the financial sector, this technique can help detect rare but highly impactful fraudulent activities. In manufacturing, the algorithm can be used for machine failure prediction, where failures are rare but require significant attention.

5.3. Future Directions and Challenges

Improvements to traditional SMOTE often address the issue of generating synthetic data that is less representative when minority data has a distribution significantly different from the majority. By incorporating attribute weights based on domain knowledge, Weighted SMOTE ensures that critical attributes are given higher priority in the generation of synthetic data.

This study also incorporates Expert Knowledge Integration. Typically, algorithms rely solely on data to make decisions, but this research demonstrates that incorporating expert input (e.g., from doctor) can produce results that are more clinically relevant.

This model sets the stage for further research on integrating domain knowledge into other algorithms, such as clustering, neural networks, or reinforcement learning. This concept can be expanded to tackle extreme data imbalance across various fields, including bioinformatics (e.g., genetic prediction), finance (fraud detection), and transportation (rare incident analysis).

Dynamic attribute weighting: The current method utilizes fixed weights based on expert judgment. Future research can explore the impact of adaptive methods, such as genetic algorithms or deep reinforcement learning, to determine weights in real-time. Evaluation of impact on non-SMOTE algorithms: This weight-based approach can also be integrated into other algorithms, such as ADASYN (Adaptive Synthetic Sampling), to compare its effectiveness.

A Weighted SMOTE-based system can automatically analyze general check-up data and provide risk scores for various chronic diseases, such as diabetes, hypertension, or cardiovascular diseases. In the future, with the integration of technologies like the Internet of Things (IoT), wearable devices could collect real-time data and transmit it to the predictive system, enabling early warnings for patients and doctors.

Several challenges arise, such as implementing this model on a large scale, which requires adequate computational infrastructure to process data in real time. In large hospitals or healthcare centers, integrating this technology must account for both hardware and software requirements. Health data is highly

sensitive and requires strict security measures. This research should be followed by the development of encryption and data anonymization protocols to ensure patient privacy. In areas with limited internet access or inadequate hardware, this technology may require adaptation in the form of models optimized for low-power devices.

6. Conclusion

This study presents a novel framework for chronic disease prediction leveraging GCU data, integrating expert judgment into machine learning techniques. The research demonstrated a significant enhancement in prediction accuracy, ranging from 10% to 47% compared to conventional methods, which underscores the effectiveness of the proposed Weighted SMOTE and ensemble learning approach. Specifically, this method addresses the challenges of imbalanced datasets by introducing attribute weighting informed by expert judgment, enabling more accurate identification of minority classes—often the most critical in medical diagnostics.

The integration of expert judgment not only enhanced the interpretability of the model but also ensured the prioritization of clinically significant features, such as blood glucose levels and blood pressure. This approach bridges the gap between data-driven algorithms and clinical expertise, making it a practical solution for real-world healthcare applications.

The experimental scenarios validated the robustness and versatility of the proposed framework. For instance, the weight normalization process improved key metrics, including F1-Score and ROC-AUC, across various datasets. The optimal weighting factor of 1,000 provided the best balance between performance and generalization, avoiding the pitfalls of overfitting or underrepresentation of critical features. Moreover, the tree-based ensemble voting method, combining Random Forest, AdaBoost, and XGBoost, further enhanced prediction accuracy by leveraging the strengths of multiple algorithms.

Long-term prediction scenarios (Y+2) demonstrated better performance than short-term predictions (Y+1), particularly for chronic diseases with stable progression patterns like diabetes mellitus and hypertension. This highlights the model's ability to capture cumulative risk factors and stable trends over time, making it a valuable tool for preventive healthcare planning.

Future research should focus on refining the dynamic weighting process, potentially employing adaptive algorithms like genetic optimization or reinforcement learning, to further enhance the flexibility and applicability of the model. Additionally, expanding the framework to other domains, such as bioinformatics, fraud detection, or rare event analysis, could validate its broader applicability.

This study contributes to the global health agenda by providing a scalable, interpretable, and accurate predictive model that can be integrated into healthcare systems. It not only enhances diagnostic precision but also aids in resource optimization for early intervention programs, ultimately improving patient outcomes and reducing the burden of chronic diseases.

Acknowledgment

This work was supported in part by the Directorate of Research, Technology, and Community Service; in part by the Directorate General of Higher Education, Research, and Technology; in part by the Ministry of Education, Culture, Research, and Technology, Republic of Indonesia under Grant 030/LIT07/PPM-LIT/2024; and in part by Telkom University.

References

[1] A. Akbar Gozali, Hypertension multi-year prediction and risk factors analysis using decision tree, in: *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, IEEE, 2023, pp. 76–82.

[2] A. Akbar Gozali, Multi-years diabetes prediction using machine learning and general check-up dataset, in: *11th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2023, pp. 98–103.

[3] O. Almadani and R. Alshammari, Prediction of stroke using data mining classification techniques, *International Journal of Advanced Computer Science and Applications* **9**(1) (2018), 457–460.

[4] A. Amro, M. Al-Akhras, K.E. Hindi, M. Habib and B.A. Shawar, Instance reduction for avoiding overfitting in decision trees, *J. Intell. Syst.* **30**(1) (2021), 438–459.

[5] A. Ashfaq, A. Imran, I. Ullah, A. Alzahrani, K.M. Ali Alheeti and A. Yasin, Multi-model ensemble based approach for heart disease diagnosis, in: *2022 International Conference on Recent Advances in Electrical Engineering & Computer Sciences (RAEE & CS)*, IEEE, 2022, pp. 1–8.

[6] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K.K. Singh and A. Khamparia, Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus, *Multimedia Systems* **28**(4) (2022), 1289–1307.

[7] P. Biecek and T. Burzykowski, *Explanatory model analysis: explore, explain, and examine predictive models*, taylorfrancis.com, 2021, p. 324.

[8] B. Butcher and B.J. Smith, Feature engineering and selection: A practical approach for predictive models, *Am. Stat.* **74**(3) (2020), 308–309.

[9] J.A. Castellanos-Garzón, E. Costa, J.L. Jaimes S. and J.M. Corchado, An evolutionary framework for machine learning applied to medical data, *Knowledge-Based Systems* **185** (2019), 104982.

[10] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, *jair* **16** (2002), 321–357.

[11] A.W. Dar and S.U. Farooq, An ensemble model for addressing class imbalance and class overlap in software defect prediction, *Int. J. Syst. Assur. Eng. Manag.* **15**(12) (2024), 5584–5603.

[12] M.A. Faghy, J. Yates, A.P. Hills, S. Jayasinghe, C. da Luz Goulart, R. Arena, D. Laddu, R. Gururaj, S.K. Veluswamy, S. Dixit and R.E.M. Ashton, Cardiovascular disease prevention and management in the COVID-19 era and beyond: An international perspective, *Prog. Cardiovasc. Dis.* **76** (2023), 102–111.

[13] T. Fahrudin, J.L. Buliali and C. Fatichah, Enhancing the performance of smote algorithm by using attribute weighting scheme and new selective sampling method for imbalanced data set, *Int. J. Innov. Comput. Inf. Control* **15**(2) (2019), 423–444.

[14] T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.* **27**(8) (2006), 861–874.

[15] N.L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, Development of disease prediction model based on ensemble learning approach for diabetes and hypertension, *IEEE Access* **7** (2019), 144777–144789.

[16] A.-Z.S.B. Habib and T. Tasnim, An ensemble hard voting model for cardiovascular disease prediction, in: *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, IEEE, 2020, pp. 1–6.

[17] D.J. Hand, P. Christen and N. Kirielle, F*: an interpretable transformation of the F-measure, *Mach. Learn.* **110**(3) (2021), 451–456.

[18] H. He and E.A. Garcia, Learning from imbalanced data, *IEEE Trans. Knowl. Data Eng.* **21**(9) (2009), 1263–1284.

[19] K.C. Howlader, M.S. Satu, M.A. Awal, M.R. Islam, S.M.S. Islam, J.M.W. Quinn and M.A. Moni, Machine learning models for classification and identification of significant attributes to detect type 2 diabetes, *Health Inf Sci Syst* **10**(1) (2022), 2.

[20] M.A. Islam, S. Akter, M.S. Hossen, S.A. Keya, S.A. Tisha and S. Hossain, Risk factor prediction of chronic kidney disease based on machine learning algorithms, in: *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, 2020, pp. 952–957.

[21] M. Kaur, S.R. Sakhare, K. Wanjale and F. Akter, Early stroke prediction methods for prevention of strokes, *Behav. Neurol.* **2022** (2022), 7725597.

[22] H.B. Kibria, M. Nahiduzzaman, M.O.F. Goni, M. Ahsan and J. Haider, An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI, *Sensors* **22**(19) (2022).

[23] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen and Y. Lin, A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare, *BioData Min.* **16**(1) (2023), 15.

[24] C.B.C. Latha and S.C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, *Informatics in Medicine Unlocked* **16** (2019), 100203.

[25] X. Li, D. Bian, J. Yu, M. Li and D. Zhao, Using machine learning models to improve stroke risk level classification methods of China national stroke screening, *BMC Med. Inform. Decis. Mak.* **19**(1) (2019), 261.

[26] F. López-Martínez, E.R. Núñez-Valdez, R.G. Crespo and V. García-Díaz, An artificial neural network approach for predicting hypertension using NHANES data, *Sci. Rep.* **10**(1) (2020), 10620.
[27] P. Mahajan, S. Uddin, F. Hajati and M.A. Moni, Ensemble Learning for Disease Prediction: A Review, *Healthcare (Basel)* **11**(12) (2023).
[28] S. Maldonado, J. López and C. Vairetti, An alternative SMOTE oversampling strategy for high-dimensional datasets, *Appl. Soft Comput.* **76** (2019), 380–389.
[29] S. Maldonado, C. Vairetti, A. Fernandez and F. Herrera, FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification, *Pattern Recognit.* **124** (2022), 108511.
[30] R. Nugent, Chronic diseases in developing countries: health and economic burdens, *Ann. N. Y. Acad. Sci.* **1136** (2008), 70–79.
[31] N.G. Ramadhan, Adiwijaya and A. Romadhony, Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest, *International Journal of Advanced Computer Science and Applications* **12**(7) (2021), 223–228.
[32] N.G. Ramadhan, Adiwijaya, W. Maharani and A. Akbar Gozali, Prediction of diabetes mellitus in the upcoming year using SMOTE and random forest, in: *2023 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, 2023, pp. 316–321.
[33] N.G. Ramadhan, Adiwijaya, W. Maharani and A. Akbar Gozali, Prediction of hypertension in the upcoming year: feature correlation analysis and handling imbalanced based on random forest, in: *2023 Eighth International Conference on Informatics and Computing (ICIC)*, IEEE, 2023, pp. 1–6.
[34] N.G. Ramadhan, Adiwijaya, W. Maharani and A. Akbar Gozali, Chronic diseases prediction using machine learning with data preprocessing handling: a critical review, *IEEE Access* **12** (2024), 80698–80730.
[35] N.G. Ramadhan, Adiwijaya, W. Maharani and A. Akbar Gozali, Prediction of cardiovascular disease (CVD) in the upcoming year using tree-based ensemble model, in: *12th International Conference on Information and Communication Technology (ICOICT)*, IEEE, 2024, pp. 210–216.
[36] N.G. Ramadhan, A. Adiwijaya, W. Maharani and A.A. Gozali, Supplementary File for Title Paper: Modified SMOTE and Ensemble Learning Based on Expert Judgment for Chronic Diseases Prediction, Zenodo, 2025. doi:10.5281/zenodo.14725457.
[37] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh and D. Khalili, The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes, *Med. Decis. Making* **36**(1) (2016), 137–144.
[38] Y. Ren, H. Fei, X. Liang, D. Ji and M. Cheng, A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records, *BMC Med. Inform. Decis. Mak.* **19**(Suppl 2) (2019), 51.
[39] X.-F. Song, Y. Zhang, D.-W. Gong and X.-Y. Sun, Feature selection using bare-bones particle swarm optimization with mutual information, *Pattern Recognit.* **112** (2021), 107804.
[40] A. Sorayaie Azar, S. Babaei Rikan, A. Naemi, J. Bagherzadeh Mohasefi, H. Pirnejad, M. Bagherzadeh Mohasefi and U.K. Wiil, Application of machine learning techniques for predicting survival in ovarian cancer, *BMC Med. Inform. Decis. Mak.* **22**(1) (2022), 345.
[41] S. Sreejith, H. Khanna Nehemiah and A. Kannan, Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection, *Comput. Biol. Med.* **126** (2020), 103991.
[42] Y. Su, C. Huang, W. Yin, X. Lyu, L. Ma and Z. Tao, Diabetes mellitus risk prediction using age adaptation models, *Biomed. Signal Process. Control* **80** (2023), 104381.
[43] Z. Ullah, F. Saleem, M. Jamjoom, B. Fakieh, F. Kateb, A.M. Ali and B. Shah, Detecting high-risk factors and early diagnosis of diabetes using machine learning methods, *Comput. Intell. Neurosci.* **2022** (2022), 2557795.
[44] J.-B. Wang, C.-A. Zou and G.-H. Fu, AWSMOTE: An SVM-based adaptive weighted SMOTE for class-imbalance learning, *Sci. Program.* **2021** (2021), 1–18.