Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

A Random Forest and SMOTE-Based Machine Learning Model for Predicting Recurrence in Papillary Thyroid Carcinoma

Edi Jaya Kusuma*[1], Ririn Nurmandhani[2], Ika Pantiawati[3], Yusthin Meriantti Manglapy[4], Evina Widianawati[5]

[1,2,3,4]Faculty of Health Science, Universitas Dian Nuswantoro, Indonesia
[5]Student of Department of Biomedical Engineering, Chung Yuan Christian University, Taoyuan City, Taiwan
Email: [1]edi.jaya.kusuma@dsn.dinus.ac.id

Received : Jun 10, 2025; Revised : Jun 28, 2025; Accepted : Jun 28, 2025; Published : Aug 18, 2025

Abstract

PTC (Papillary Thyroid Carcinoma) is one subtype of thyroid cancer occurred most frequently in thyroid cancer cases. Although the prognosis of this cancer is typically positive, its recurrence remains a key challenge requiring early detection. This study proposes machine learning models to predict PTC recurrence, explicitly addressing the inherent class imbalance in the recurrence data. This study implemented three supervised learning algorithms, namely Random Forest (RF), Extreme Gradient Boost (XGB), and Support Vector Machine (SVM) with the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset. SMOTE was chosen for its capacity to generate synthetic minority class samples while minimizing information loss, thus effectively addressing class imbalance and improving classification outcomes. Model performance was assessed using accuracy, precision, recall (sensitivity), and F1-score. Among all approaches tested, RF with SMOTE demonstrated superior performance, achieving 0.98 accuracy, perfect precision (1.0), high recall (sensitivity) (0.95), and a strong F1-score (0.97), outperforming previous methods including SMOTEENN-based approaches. The result of this study demonstrates SMOTE specifically outperforms SMOTEENN in this clinical context, likely due to better preservation of subtle prognostic indicators with minimal information loss. This improvement suggests SMOTE's effectiveness in preserving valuable decision boundary information while addressing class imbalance in PTC recurrence prediction. These findings establish RF with SMOTE as a robust and well-balanced approach for predicting PTC recurrence, contributing significantly to the development of more precise and responsive AI-driven decision support tools for thyroid cancer.

Keywords : *Class Imbalance, Clinical Decision Support, Machine Learning, Papillary Thyroid Carcinoma, SMOTE.*

1. INTRODUCTION

Among all types of thyroid malignancies, Papillary Thyroid Carcinoma (PTC) is the most frequently diagnosed, comprising approximately 75–85% of reported cases. It is generally slow-growing and often confined to a single lobe of the thyroid gland. Most patients, especially those under 45 years of age, have a favourable prognosis [1][2]. In certain cases, PTC may exhibit aggressive behaviour, with reported recurrence rates ranging from 8% to 28% [3][4]. Contributing factors to an increased risk of recurrence include the follicular variant of PTC, advanced patient age, and the presence of lymph node metastases [5]. This significant recurrence potential makes early identification of recurrence risk factors a major challenge. This step is crucial for improving patient quality of life and mitigating potential complications in the future.

With advancements in technology, particularly in the field of machine learning (ML), new solutions have emerged that enable the automatic learning of complex patterns from multidimensional data. The application of ML in detecting PTC recurrence aims to support faster and more accurate clinical decision-making, allowing for early intervention while reducing the risks of overdiagnosis and

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

overtreatment. A study conducted by Jinhua Yu et al. [6] developed a Transfer Learning Radiomics (TLR) model based on B-mode ultrasound imaging to predict the risk of lymph node metastasis (LNM) in PTC patients. This study was conducted in a multi-centre, multi-machine, and multi-operator setting. TLR demonstrated superior performance, achieving an average Area Under the Curve (AUC) of 0.93, which was higher than that of traditional clinical statistical models and radiomics approaches.

Another study [7] evaluated various ML approaches, including Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), and Artificial Neural Network (ANN), to identify the recurrence of differentiated thyroid cancer based on clinicopathological data. The results showed that the model incorporating all features achieved the best performance, with SVM reaching an AUC of 99.71%, a recall (sensitivity) of 93.33%, and a specificity of 97.14%. A major challenge in the aforementioned studies was dataset imbalance. In study [6], the number of relevant LNM cases was significantly lower than non-LNM cases, while in study [7] only 28% of patients experienced recurrence, whereas the remaining 72% did not. This imbalance may affect the model's reliability in detecting patterns within the minority class, which often holds critical clinical implications.

Addressing similar challenges, Young Min Park and Byung-Joo Lee [8] evaluated five ML models in predicting PTC recurrence using pathology data. This research utilized the Synthetic Minority Oversampling Technique or SMOTE to balance the previously highly imbalanced distribution of recurrence and non-recurrence cases. By applying SMOTE, DT model achieved 95% of accuracy, which was later followed by the LightGBM and stacking models with an accuracy of 93%. Additionally, an AUC of 0.742 for the number of metastatic lymph nodes highlighted the significance of this variable as a risk indicator for recurrence. SMOTE played a crucial role in this study by enhancing the representation of minority cases without losing essential information [9][10]. By generating synthetic data similar to real cases, SMOTE enabled the model to learn risk patterns more effectively, reduced bias toward the majority class, and improved the model's generalization ability. This underscores the importance of oversampling techniques in developing more accurate and reliable clinical prediction models. Another study [11] addressed the issue of class imbalance in clinical datasets using SMOTE-Edited Nearest Neighbours (ENN) and an Explainable Artificial Neural Network (EANN), focusing on the early detection of thyroid cancer. SMOTEENN combines the SMOTE technique with a noise elimination process based on Edited Nearest Neighbours (ENN), effectively balancing the dataset while removing ambiguous or misclassified samples [12]. However, ENN may aggressively remove synthetic or real samples considered misclassified or noisy or inconsistent with its local neighbourhood, potentially leading to the loss of valuable information[13][14]. Given SMOTEENN's potential to aggressively remove samples, in certain rare cases or medical datasets where minority class samples hold critical information, relying solely on SMOTE can better preserve these valuable data points [15].

Therefore, this study proposes the use of the SMOTE technique to address dataset imbalance in PTC recurrence cases, particularly within cohort datasets. This approach aims to enhance data quality, enabling machine learning models to classify more accurately and fairly across all classes. The evaluation is conducted using various machine learning models, including RF, Extreme Gradient Boost (XGB), and SVM, which widely acknowledged for excelling in classification tasks. By leveraging an innovative combination of resampling methods and diverse classification models, this study is expected to provide an effective solution for handling imbalanced datasets while making a significant contribution to improving the accuracy of clinical diagnosis in PTC recurrence cases. To provide broader context, a comparative discussion of SMOTE and SMOTEENN as prior methods is included at the end of this paper, outlining their relevance to class imbalance issues in clinical datasets.

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

2. METHOD

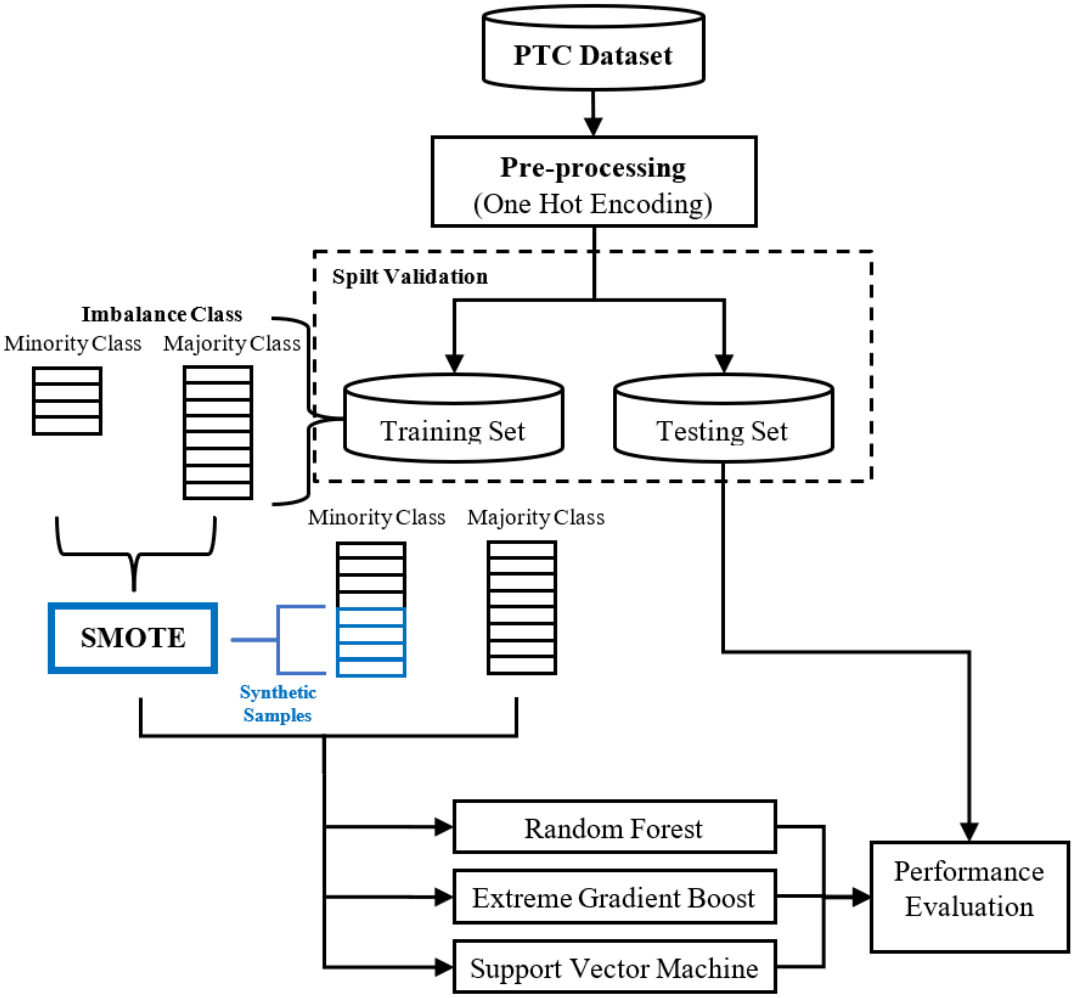

Figure 1. Proposed Research Workflow

This study is experimental research aimed at evaluating multiple machine learning models using a publicly available dataset. The proposed research framework is illustrated in Figure 1. In this framework, the Papillary Thyroid Carcinoma (PTC) dataset undergoes an initial pre-processing stage, specifically data transformation using One-Hot Encoding (OHE), to ensure compatibility with subsequent processing steps. The dataset is then partitioned using split validation into training set and testing set. Afterwards, SMOTE technique is employed on the training data to handle class imbalance. As shown in Figure 1, the synthetic samples generated through SMOTE is then utilized to train several ML models, including RF, XGB, and SVM. Upon completion of the training process, the models are evaluated using the test data to assess their classification performance. A comprehensive discussion of the experimental setup and the methodologies employed in this study is provided in the following sections.

2.1. Papillary Thyroid Carcinoma (PTC) Dataset

This study utilizes a publicly available dataset on Papillary Thyroid Carcinoma (PTC) [7] which can be accessed Data were collected over 15 years, with a follow-up duration of no less than 10 years for each patient. It comprises 383 samples and includes 17 variables, categorized as follows: 5 demographic and patient history variables, 3 clinical and physical examination variables, 3 pathological

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

variables, 4 staging variables (TNM classification), and 2 outcome variables. The outcome variables include the patient's response to therapy and the target label indicating PTC recurrence. Table 1 presents the operational definitions for each variable used in this study.

Table 1. Definition of Variables in the Papillary Thyroid Carcinoma (PTC) Dataset

Feature Name	Description
Age	Age of the patient in years.
Gender	Gender of the patient (F = Female, M = Male).
Smoking	Whether the patient is currently smoking (Yes/No).
Hx Smoking	History of smoking (Yes if the patient has smoked before, No otherwise).
Hx Radiotherapy	History of undergoing radiotherapy (Yes/No).
Thyroid Function	Thyroid function status (e.g., Euthyroid, Hypothyroid).
Physical Examination	Findings from the physical examination of the thyroid gland (e.g., Single nodular goiter, Multinodular goiter).
Adenopathy	Presence of lymph node enlargement (Yes/No).
Pathology	Type of pathology of the thyroid tissue (e.g., Micropapillary).
Focality	Tumor focality, indicating whether the tumor is Uni-Focal or Multi-Focal.
Risk	Cancer risk classification (Low/High).
T (Tumor Stage)	Tumor size and characteristics based on TNM staging (e.g., T1a, T1b).
N (Node Involvement)	Status of lymph node involvement in TNM staging (e.g., N0, N1).
M (Metastasis Status)	Whether the cancer has spread to other organs (M0 = No metastasis, M1 = Metastasis present).
Stage	Overall cancer stage based on TNM classification (e.g., I, II, III, IV).
Response	Patient's response to treatment (e.g., Excellent, Indeterminate).
Recurred	Whether the cancer has recurred after treatment (Yes/No).

2.2. Pre-processing

Pre-processing serves as the preliminary step in data preparation before its utilization in a machine learning (ML) model. The objective of pre-processing is to enhance data quality, which allow the ML model to learn more precisely and effectively [16]. In this study, the pre-processing stage utilizes the One-Hot Encoding (OHE) technique. OHE is selected due to the predominantly categorical nature of the PTC dataset. This method offers significant advantages in handling categorical data by preventing the model from incorrectly interpreting categorical values as ordinal relationships [17]. Compared to retaining categorical data in its raw form, OHE ensures that each category is distinctly represented, which strengthens the model's pattern recognition performance [18]. OHE is a data transformation technique used to reform categorical variables into a binary numerical representation. This technique is widely applied in ML to ensure that the ML model can treat categorical data without assuming an inherent sequence among the categories. In OHE, each unique category of a feature is encoded as a binary vector with a length corresponding to the total number of unique categories in that feature. Each entry in the vector contains a value of 1 (true) if the sample belongs to a specific category and 0 (false) otherwise. Mathematically, the OHE transformation can be expressed as follows [19]:

Given a feature X with k unique category variations (1):

$$X = \{x_1, x_2, \dots, x_n\} \quad (1)$$

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

Thus, the One-Hot Encoding function can be defined as follows (2):

$$OHE(x_i) = V_i \in \mathbb{R}^k \quad (2)$$

The OHE function in Equation (1) indicates that the vector V_i resides in a k dimensional space containing real numbers $\mathbb{R}$. The vector V_i represent the elements $V_i = \{v_{i1}, v_{i2}, \dots, v_{ik}\}$, where each element v_{ik} is defined as follows (3):

$$v_{ik} = \begin{cases} 1 \ (true), & \text{if } x_i = c_k \\ 0 \ (false), & \text{if } x_i \neq c_k \end{cases} \quad (3)$$

Each element is defined based on the correspondence between x_i and category c_k. If they match, then v_{ik} is assigned a value of 1 (true); otherwise, it is assigned 0 (false).

2.3. Synthetic Minority Over-sampling Technique (SMOTE)

The majority of patients with papillary thyroid carcinoma (PTC) has a favourable prognosis. Consequently, the availability of recurrence data, particularly for high-risk recurrence cases, is highly limited. To mitigate this limitation, this study employs the Virtual Sample Generation (VSG) method, specifically the SMOTE, which is widely utilized to mitigate class imbalance within datasets [20][21].

The application of SMOTE involves generating synthetic data by linearly interpolating minority class samples. The determination of new synthetic data using SMOTE can be formulated using Equation (4) [22].

$$x_{new} = x_i + \lambda \cdot (x_k - x_i) \quad (4)$$

From Equation (4), it can be inferred that the determination of the synthetic data x_{new} is performed by utilizing a minority class sample x_i and its neighbouring sample x_k. The neighbouring sample x_k is identified based on k-nearest neighbours (typically $k = 5$) of the minority sample x_i using the Euclidean distance. Meanwhile, λ is a random value ranging between 0 and 1.

2.4. Random Forest (RF)

Random Forest (RF) is an ensemble-based ML algorithm developed by Breiman [23], widely used in various machine learning applications, including classification and regression. This algorithm constructs multiple randomly generated decision trees and combines their predictions using an ensemble approach to improve accuracy and reduce the risk of overfitting.

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \quad (5)$$

Given a dataset D in equation (5) with x_n and y_n, representing the features of PTC pathology, where n is the number of data samples. The initial stage of random forest is generating the b subset utilizing bootstrap sampling as follow (6):

$$D_b \subset D \ where \ |D_b| = n \quad (6)$$

Each subset D_b is drawn with replacement, meaning a sample may appear multiple times in a single subset. Each subset then used to train one decision tree (DT) model (T_b).

The construction of the T_b begin with splitting the subset D_b based on chosen feature X_j that provides the best separation. At each node v, the best feature X_j^* is selected by reducing impurity using Gini Impurity (7) or Entropy (8) [24][25]:

$$G(X_j) = 1 - \sum p_i^2 \quad (7)$$

$$H(X_j) = -\sum p_i \, log_2 \, p_i \quad (8)$$

$$p_i = \frac{D_v^i}{D_v} \quad (9)$$

where p_i in (9) represents the probability of class i, and D_v^i represents sample in class i at node v. The best feature X_j^* for splitting node v is selected from the one that minimizes the weighted impurity using equation (10).

$$X_j^* = \arg\min_{X_j}[H(D_L) + H(D_R)] \quad (10)$$

After each decision tree T_b model generated, then the final prediction can be decided based on the majority voting of prediction result from each T_b as displayed in (11). Each T_b predicts an output $\hat{y}_b$ for a given input x.

$$\hat{y}_b = T_b(x) \quad (11)$$

Based on these prediction result $\hat{y}_b$ of each T_b, the majority voting can be expressed in (12) [26].

$$\hat{y} = \arg\min_{y} \sum_{b=1}^{B} 1(T_b(x) = y) \quad (12)$$

$$1(T_b(x) = y) = \begin{cases} 1, & \text{if } T_b(x) = y \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

Where in equation (13), 1 is an indicator function, and $\hat{y}$ represents the final prediction result of the Random Forest, which corresponds to the predicted risk of recurrence in Papillary Thyroid Carcinoma (PTC).

2.5. Extreme Gradient Boost (XGB)

Extreme Gradient Boosting (XGB) is a model developed based on the gradient boosting approach [27]. It is constructed using decision trees as the foundation for its classification process. XGB operates by iteratively building tree models, where each subsequent model focuses on rectifying the mistakes of the previous one. This process is carried out by minimizing the loss function using gradient descent while incorporating regularization to prevent overfitting [28]. XGB is widely utilized in machine learning development due to its high speed, efficiency, and scalability, enabling it to deliver optimal performance [29].

In general, XGB prediction process in t iteration can be expressed using formula (14) [28].

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + f_t(x) \quad (14)$$

where $\hat{y}^{(t)}$ represents the most recent prediction result, while $\hat{y}^{(t-1)}$ denotes the previous prediction result. The function $f_t(x)$ corresponds to the newly added decision tree model, which aims to correct the errors from the previous iteration. In each iteration, XGB enhances the decision tree model by optimizing the objective function ($\mathcal{L}$), which consists of a loss function and a regularization term.

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \quad (15)$$

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

From Equation (15), $l(y_i, \hat{y}_i)$ represents the loss function, which measure the difference between the predicted result $\hat{y}_i$ and the actual result y_i. Meanwhile, the $\Omega(f_t)$ is the regularization term used to control the complexity of the model, which is generally formulated as follow (16).

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (16)$$

where T is the leaf numbers in the decision tree, w_j denotes the weight of each leaf, γ is a penalized parameter for the number of leaves, while λ help the L2 regularization on leaf weights. This regularization controls overfitting and strengthens the model's generalization performance.

2.6. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a ML algorithm designed for handling classification and regression tasks. The essential principle of SVM is to determine the optimal hyperplane, which defines a hyperplane that maximizes the margin between two distinct data classes [30]. In a two-dimensional feature space, the hyperplane corresponds to a line; in higher-dimensional spaces, it generalizes to a flat surface that separates the classes.

$$\omega^T x + b = 0 \quad (17)$$

From Equation (17), the optimal hyperplane is defined by the dot product of the weight vector ω which defines the orientation of the decision boundary (hyperplane) and the input feature vector x followed by the addition of the bias b which shifts the decision boundary. In order to ensure the data is separated with the maximum margin, SVM maximize the margin M between the two classes with the condition in formula (18).

$$y_i(\omega^T x_i + b) \geq 1, \ \forall i \quad (18)$$

The y_i is the actual class label of the i-th data point, where $y_i \in \{-1, 1\}$ typically used in SVM to distinguish between two classes. Then, SVM minimize the objective function (19) to obtain bigger margin M.

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 \quad (19)$$

However, in many cases, data cannot be linearly separated. To address this limitation, SVM employs a kernel function that lifting the data into a higher-dimensional representation [31]. Radial Basis Function (RBF) is one of kernel commonly used in SVM particularly in handling the non-linear data [32].

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (20)$$

This transformation using $K(x_i, x_j)$ in Equation (20) allows data that are not separable in their original space to be effectively distinguished in the new feature space. This approach enhances SVM's capability to support diverse data types, including those with complex patterns, making it a powerful algorithm for various machine learning applications.

2.7. Experimental Setup and Evaluation

The experiments were conducted using Python 3.10 within the Google Collaboratory environment. The machine learning (ML) models utilized in this study were executed using several Python programming libraries: Scikit-learn (v1.1.3) for One Hot Encoding (OHE), Random Forest (RF),

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

and Support Vector Machine (SVM); XGBoost (v1.7.6) for Extreme Gradient Boosting (XGB); and Imbalanced-learn (v0.11.0) for the Synthetic Minority Over-Sampling Technique (SMOTE). All ML models were trained with default hyperparameters and a predetermined random seed of 42 to ensure reproducibility. Prior to modelling, the dataset was transformed using the One Hot Encoding technique, then divide it into training and testing datasets at an 80:20 ratio using a random split. Afterwards, the SMOTE was applied only to the training set to preserving the integrity of model evaluation.

After training machine learning models (RF, XGB, and SVM) using a synthetic dataset generated through SMOTE to predict the recurrence risk of papillary thyroid carcinoma (PTC), the next step is to evaluate the model's performance using the training data. This evaluation yields several ML performance metrics, including accuracy, precision, recall (sensitivity), and F1-score, which are generally used to examine the effectiveness of the trained ML models [33].

Accuracy measures the extent to which the model correctly classifies all target classes. Precision indicates the proportion of positive predictions that are truly positive. Recall or sensitivity quantifies the percentage of actual positive cases that are correctly identified by the model. Meanwhile, the F1-score combines precision and recall (sensitivity) to provide a balanced assessment of the model's performance.

In this study, the results of model training using the synthetic dataset generated by SMOTE are compared with those obtained from training the same model using the original dataset without any pre-processing. This comparison aims to evaluate the impact of SMOTE on improving the performance of machine learning models, particularly in addressing class imbalance within the dataset.

3. RESULT

The prediction of Papillary Thyroid Carcinoma (PTC) recurrence has been conducted using a machine learning approach. To ensure that categorical variables are not misinterpreted as ordinal data by the model, the dataset was preprocessed using One-Hot Encoding (OHE). This transformation resulted in a refined dataset comprising 383 records and 120 features. Subsequently, the dataset was partitioned into two subsets (training and testing sets) based on 80:20 ratio using random split. The training set consists of 306 samples, with 77 samples designated as the testing set.

The distribution of the training set consist of 89 samples from minority class, while the majority class has 217 samples. This represents a 41% difference between the two classes, indicating a notable class imbalance. To address the issue of class imbalance within the training dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. Prior to resampling, the distribution of the target variable was skewed, with a significantly lower number of samples in the "Yes" class (indicating recurrence of PTC) compared to the "No" class (non-recurrence). This imbalance poses a risk of biasing the classifier toward the majority class, potentially reducing its sensitivity in detecting recurrence cases.

After the application of SMOTE, the minority class was synthetically augmented to match the majority class (both consist of 217 samples), resulting in a more balanced training set. Figure 2 illustrates the class distributions before and after the application of SMOTE. As shown, the number of samples in the "Yes" class increased substantially, achieving near parity with the "No" class. Therefore, the resampled training set comprises 434 samples, which corresponds to a 41.83% increase from the original training set size. This balanced distribution is expected to improve the model's ability to detect recurrence cases by providing a more representative learning space.

In this study, three supervised machine learning algorithms, namely Random Forest (RF), Extreme Gradient Boosting (XGB), and Support Vector Machine (SVM) were employed to perform the classification task for predicting the recurrence of PTC. Each model was trained and evaluated using both the original (imbalanced) and resampled (SMOTE-balanced) versions of the training dataset in

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

order to assess the impact of class balancing on predictive performance. All ML models were trained using default hyperparameters and a fixed random seed of 42 to guarantee reproducibility.

After training, the ML models were tested using the testing set to evaluate their performance, particularly in predicting PTC recurrence. The performance metrics used to assess the models' capabilities included accuracy, recall (sensitivity), precision, and F1-score. The evaluation results based on the testing set and these metrics are presented in Table 2.

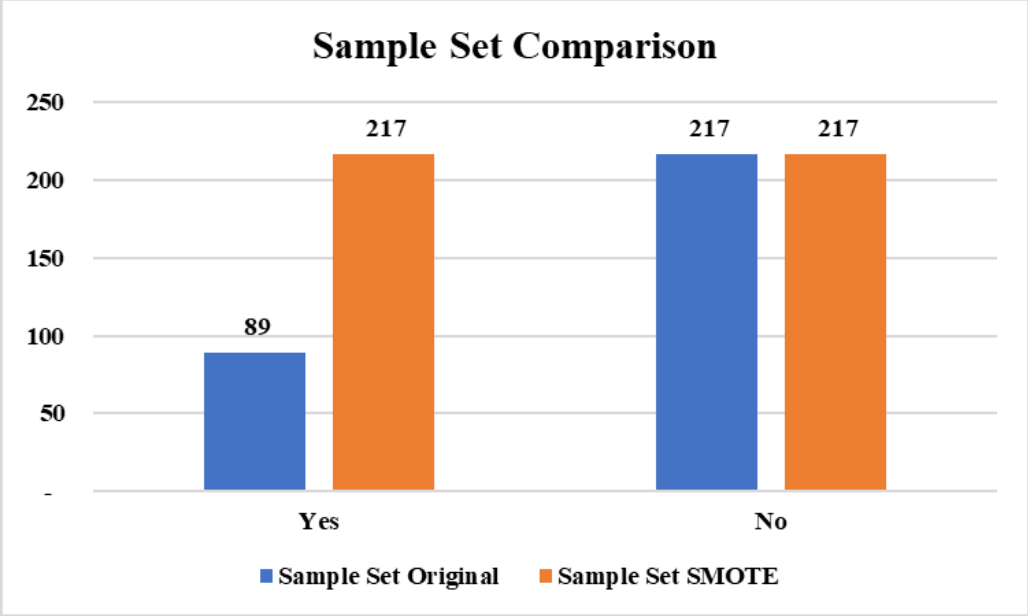

Figure 2. The Comparison of Sample Set after SMOTE application

Table 2. Evaluation Result

PTC Occurrence	Model	Accuracy		Precision		Recall		F1-score	
		Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
Yes	RF	0.97	0.98	1.00	1.00	0.89	0.95	0.94	0.97
	XGB	0.96	0.97	0.90	0.95	0.95	0.95	0.91	0.97
	SVM	0.96	0.98	1.00	1.00	0.84	0.95	0.92	0.95
No	RF	0.97	0.98	0.97	0.98	1.00	1.00	0.98	0.99
	XGB	0.96	0.97	0.98	0.98	0.97	0.98	0.97	0.98
	SVM	0.96	0.98	0.95	0.98	1.00	1.00	0.97	0.99

As presented in Table 2, all models demonstrated performance result across both datasets, with slight improvements observed after applying SMOTE. Figure 3 presents a comparison of classification accuracy achieved by three ML algorithms (RF, XGB, and SVM) when applied to a dataset related to PTC.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

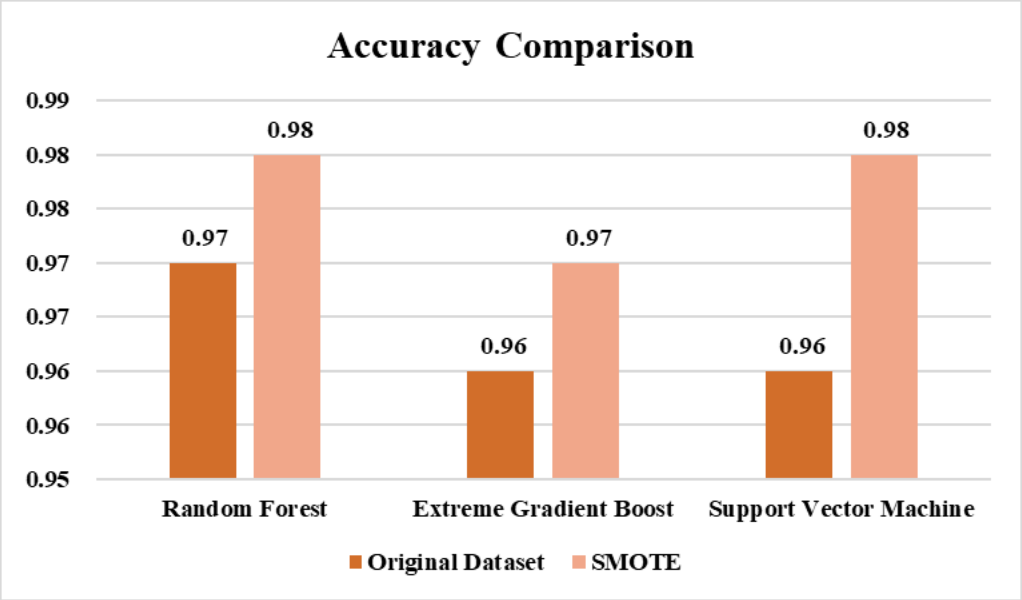

Figure 3. The Impact of SMOTE on Classification Accuracy Across Different Models

The application of SMOTE resulted in improved model performance across all algorithms. Specifically, the accuracy of RF increased from 0.97 to 0.98 (1.03%), XGB from 0.96 to 0.97 (1.04%), and SVM from 0.96 to 0.98 (2.08%). These results demonstrate that SMOTE effectively mitigates class imbalance in PTC data and enhances overall classification performance, with an average improvement of 1.38%, underscoring its utility in medical data analysis where imbalanced class distributions are common.

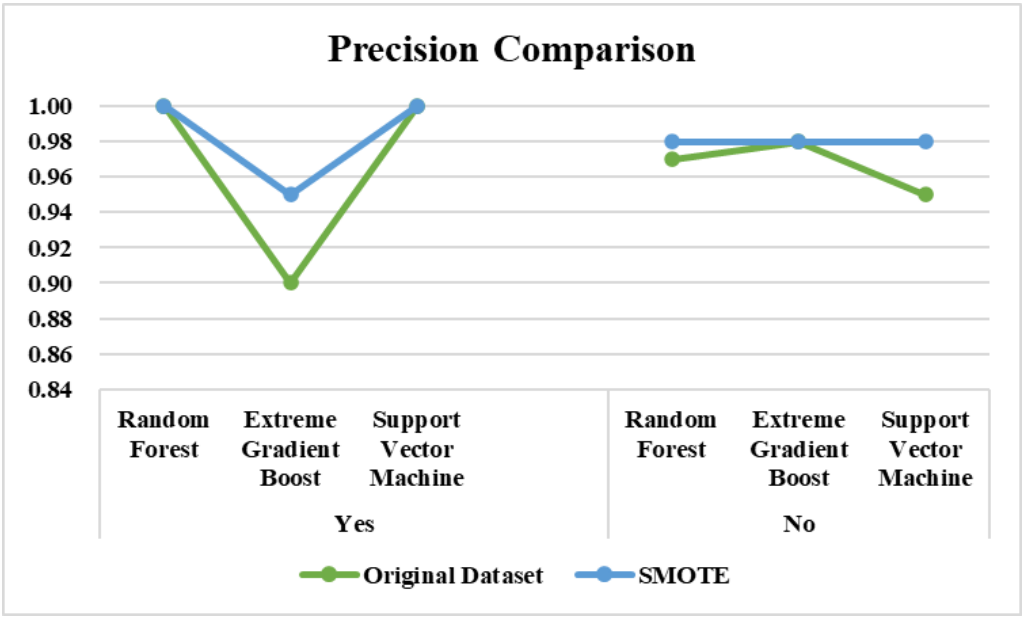

Figure 4. Comparison of Precision performance before and after SMOTE balancing on the PTC dataset.

Figure 4 further supports the findings by comparing the precision of the same models under both conditions. For the "Yes" class (recurrence class), XGB showed the highest precision improvement at 5.56%, while RF and SVM remained constant. The average improvement across models in this class was 1.85%. For the "No" class, SVM improved the most at 3.16%, followed by RF at 1.03%, while XGB showed no change. The average precision improvement for this class was 1.40%. The substantial

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

improvement in precision of trained models after SMOTE application highlights the SMOTE effectiveness in enabling the model to more accurately identify true positive cases in an imbalanced medical dataset.

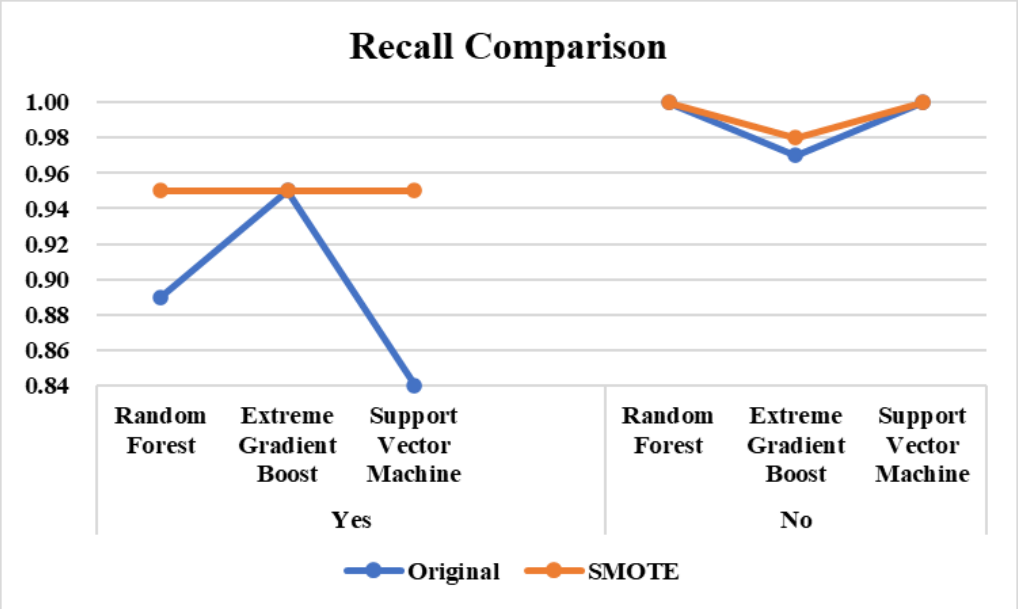

Figure 5. Comparison of recall values demonstrating the impact of SMOTE on model sensitivity across classifiers.

Moreover, Figure 5 highlights the comparison of recall (sensitivity) scores. The application of SMOTE resulted in a significant recall (sensitivity) improvement in the “Yes” class, with an average increase of 6.61% across models. In contrast, the “No” class recall (sensitivity) remained largely unchanged, with an average improvement of only 0.34%. The recall (sensitivity) results show a significant improvement in the "Yes" class (PTC recurrence), indicating that SMOTE effectively enhances data quality by balancing class distribution while preserving critical minority class information. Consequently, the model's ability to detect PTC recurrence improves notably, enabling more accurate identification of positive cases and reducing the risk of false negatives.

Figure 6 presents the comparison of F1-scores, which balances both precision and recall. The F1-scores for the "Yes" class (PTC recurrence) improved across all models after applying SMOTE, with RF increasing from 0.94 to 0.97, XGB from 0.91 to 0.97, and SVM from 0.92 to 0.95. These improvements indicate enhanced model performance in detecting recurrence cases, particularly in the minority class. For the "No" class (non-recurrence), F1-scores also increased or remained high, with RF rising from 0.98 to 0.99, XGB from 0.97 to 0.98, and SVM from 0.97 to 0.99. Overall, SMOTE not only improved recall but also contributed to a more balanced predictive performance across classes, particularly enhancing the detection of the minority class. The consistent upward trends in F1-scores further indicate that SMOTE effectively increased classifier sensitivity toward relapse cases, resulting in improved overall performance.

Collectively, these findings demonstrate that SMOTE plays a significant role in improving the predictive capabilities of ML models, especially in addressing the imbalance class commonly appear in medical datasets. The enhancement in performance matrics and overall consistency make SMOTE a valuable preprocessing step in building reliable diagnostic tools for Papillary Thyroid Carcinoma classification.

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

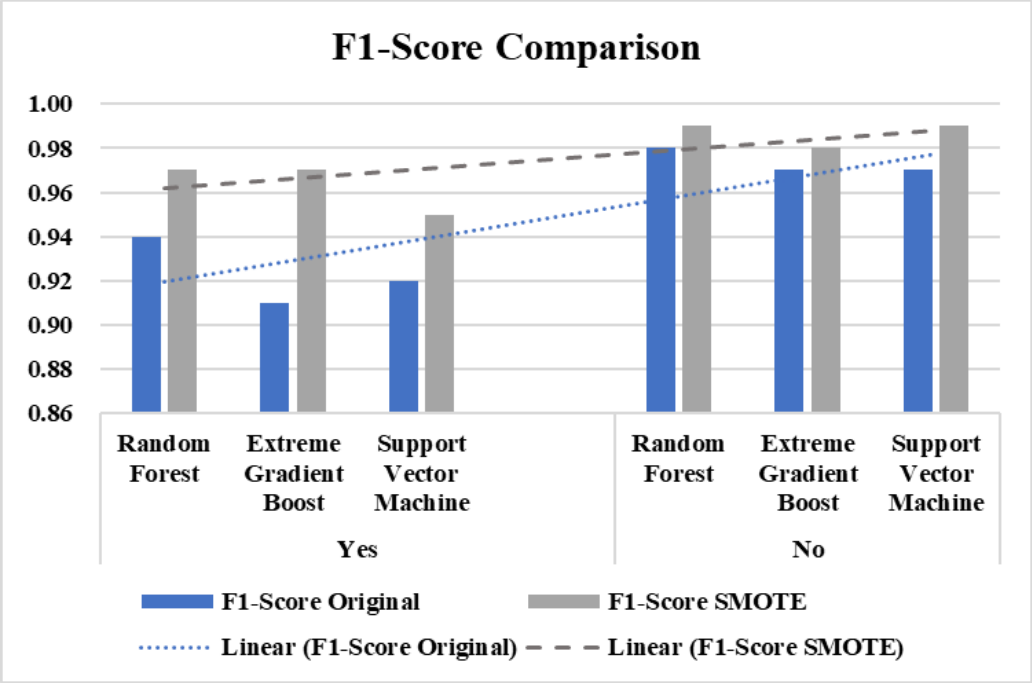

Figure 6. F1-score trends illustrating the enhancement in model balance between precision and recall after SMOTE implementation.

4. DISCUSSIONS

In clinical practice, especially when predicting recurrence in Papillary Thyroid Carcinoma (PTC), recall plays a vital role. Missing a recurrence case could lead to delays in treatment, which may compromise patient outcomes. Therefore, improving recall directly supports earlier detection and enables timely, more accurate clinical decisions [34].

At the same time, precision is equally important. A model with high precision helps ensure that patients who are not experiencing recurrence are not subjected to unnecessary diagnostic procedures or treatments [35]. This avoids undue stress, potential side effects, and additional healthcare costs. Striking the right balance between precision and recall is essential and this balance is best reflected in the F1-score, which considers both aspects.

In this study, applying the Synthetic Minority Over-sampling Technique (SMOTE) improved the models' sensitivity to the minority class without significantly compromising performance on the majority class. This is particularly important given the typical imbalance in medical datasets, which can cause models to favor the dominant class and overlook critical minority instances.

Table 3. Performance Comparison with State-of-the-art

ref.	Years	Proposed Method	Performance Accuracy
[11]	2022	Explainable Artificial Neural Network (EANN) and SMOTE Edited Nearest Neighbors (SMOTEENN)	0.94
[7]	2023	Support Vector Machine	0.96
[36]	2024	Random Forest	0.97
This Study	**2025**	**Random Forest and SMOTE**	**0.98**

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

From Table 3, which summarizes the results of several prior methods tested, RF with SMOTE demonstrated superior performance with 0.98 of accuracy compared to the previous studies. This improvement can be attributed to the robustness of RF in handling non-linear relationships and noisy data, as well as its ensemble nature, which enhances generalization. The presence of SMOTE enhanced the RF model's ability to detect recurrence cases by mitigating class imbalance, thereby facilitating the learning of more robust and clinically meaningful decision boundaries. While [11] using EANN with SMOTE with Edited Nearest Neighbors (SMOTEENN) achieved an accuracy of 0.94, and subsequent research using Support Vector Machine [7] and standard Random Forest [36] showed incremental improvements with accuracies of 0.96 and 0.97 respectively. In contrast, the proposed method of this study achieved the highest accuracy of 0.98. This improvement over SMOTEENN-based methods suggests that SMOTE provides a more effective solution for addressing class imbalance in PTC recurrence prediction. SMOTE likely outperformed SMOTEENN in this context because the under-sampling component of SMOTEENN may have removed valuable majority class samples that contained important decision boundary information [37]. Additionally, papillary thyroid carcinoma datasets typically contain subtle prognostic indicators with limited noise, conditions where pure over-sampling approaches like SMOTE excel by preserving all original data points while adding synthetic minority samples.

In conclusion, Random Forest with SMOTE is recommended as the most effective model for detecting recurrence in PTC, offering strong and balanced performance across all key evaluation metrics. Clinically, this approach holds promise for integration into screening workflows and clinical decision support systems. Its compatibility with structured data enables feasible deployment within Electronic Health Record (EHR) environments, where it could provide automated risk stratification for recurrence.

Nonetheless, several limitations of this study warrant consideration. The relatively limited dataset size ($n = 383$) may constrain the model's capacity for generalization, particularly when applied to broader or more heterogeneous patient populations. Furthermore, the absence of external validation or the use of more rigorous cross-validation strategies may affect the reliability and stability of the reported performance metrics. Although the implementation of SMOTE effectively mitigated class imbalance, the introduction of synthetic samples may inadvertently increase the risk of overfitting, especially in smaller datasets. To address these concerns, future research should incorporate larger, multi-institutional datasets and implement external validation protocols to enhance the model's robustness and clinical applicability. Additionally, the integration of explainable artificial intelligence (XAI) approaches, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) is recommended to improve the interpretability of model outputs, thereby supporting transparent and clinically meaningful decision-making for healthcare practitioners.

5. CONCLUSION

The improvement of recall (sensitivity) in papillary thyroid carcinoma (PTC) recurrence prediction has been successfully achieved using SMOTE with several machine learning models. From our comprehensive evaluation results, the application of SMOTE in RF, XGB, and SVM was capable of significantly improving overall performance parameters. This study provides the first evidence that SMOTE outperforms SMOTEENN in recall performance on PTC datasets. Among these models, Random Forest combined with SMOTE demonstrated superior performance, achieving an accuracy of 0.98. This represents a 4.26% improvement over EANN with SMOTEENN (0.94), a 2.08% increase over standard SVM (0.96), and a 1.03% improvement compared to conventional RF without SMOTE (0.97). Additionally, this combination yielded perfect precision (1.00), high recall (0.95), and a strong F1-score (0.97) in identifying recurrence cases, highlighting its effectiveness in addressing class

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

imbalance and improving predictive performance. SMOTE outperformed SMOTEENN likely because it preserved all original data points while effectively addressing class imbalance without removing potentially valuable majority class samples that contained important decision boundary information. These findings represent a significant advancement in developing reliable clinical decision support tools, potentially allowing clinicians to identify high-risk PTC patients more accurately and implement appropriate monitoring and intervention strategies earlier. Importantly, this study contributes to the field of AI in oncology by providing empirical evidence of the utility of machine learning models in improving risk stratification and supporting precision oncology efforts.

Additionally, prospective validation studies across multiple institutions with diverse patient populations are essential to confirm the generalizability of our approach. Finally, developing an interpretable clinical decision support system that not only predicts recurrence but also provides actionable insights into patient-specific risk factors would significantly enhance the practical utility of this work in thyroid cancer management.

ACKNOWLEDGEMENT

This research was supported by Universitas Dian Nuswantoro through research grant number 005/A.38-04/UDN-09/I/2025. The authors would like to express their sincere gratitude to Universitas Dian Nuswantoro for providing the financial support that made this study possible.

REFERENCES

[1] S. Yao and H. Zhang, "Papillary thyroid carcinoma with Hashimoto's thyroiditis: impact and correlation," *Front. Endocrinol. (Lausanne).*, vol. 16, Apr. 2025.

[2] Y. Ito, M. Yamamoto, M. Kihara, N. Onoda, A. Miya, and A. Miyauchi, "Establishment of novel prognostic groups for papillary thyroid carcinoma using a modified risk classification based on tumor extension in the guidelines of the Japan Association of Endocrine Surgery," *Endocr. J.*, vol. 72, no. 6, pp. EJ24-0610, 2025.

[3] A. A. Póvoa *et al.*, "Clinicopathological Features as Prognostic Predictors of Poor Outcome in Papillary Thyroid Carcinoma," *Cancers (Basel).*, vol. 12, no. 11, p. 3186, Oct. 2020.

[4] J. Zhang and S. Xu, "High aggressiveness of papillary thyroid cancer: from clinical evidence to regulatory cellular networks," *Cell Death Discov.*, vol. 10, no. 1, p. 378, Aug. 2024.

[5] H. Zhong, Q. Zeng, X. Long, Y. Lai, J. Chen, and Y. Wang, "Risk factors analysis of lateral cervical lymph node metastasis in papillary thyroid carcinoma: a retrospective study of 830 patients," *World J. Surg. Oncol.*, vol. 22, no. 1, p. 162, Jun. 2024.

[6] J. Yu *et al.*, "Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics," *Nat. Commun.*, vol. 11, no. 1, pp. 1–10, 2020.

[7] S. Borzooei, G. Briganti, M. Golparian, J. R. Lechien, and A. Tarokhian, "Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study," *Eur. Arch. Oto-Rhino-Laryngology*, vol. 281, no. 4, pp. 2095–2104, 2024.

[8] Y. M. Park and B.-J. Lee, "Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence," *Sci. Rep.*, vol. 11, no. 1, p. 4948, Mar. 2021.

[9] J. Pardede and D. P. Pamungkas, "The Impact of Balanced Data Techniques on Classification Model Performance," *Sci. J. Informatics*, vol. 11, no. 2, pp. 401–412, 2024.

[10] H. Hairani, T. Widiyaningtyas, and D. Dwi Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 3, p. 1310, Sep. 2024.

[11] S. S. Aljameel, "A Proactive Explainable Artificial Neural Network Model for the Early Diagnosis of Thyroid Cancer," *Computation*, vol. 10, no. 10, p. 183, Oct. 2022.

[12] R. Bounab, B. Guelib, and K. Zarour, "A Novel Machine Learning Approach For handling Imbalanced Data: Leveraging SMOTE-ENN and XGBoost," in *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 2024, pp. 1–7.

[13] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

ENN based on Random forest for medical imbalanced data," *J. Biomed. Inform.*, vol. 107, p. 103465, Jul. 2020.

[14] I. M. Alkhawaldeh, I. Albalkhi, and A. J. Naswhan, "Challenges and limitations of synthetic minority oversampling techniques in machine learning," *World J. Methodol.*, vol. 13, no. 5, pp. 373–378, Dec. 2023.

[15] Y. Jang, "Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE, RUS, and random forest methods: a prediction study," *Ewha Med. J.*, vol. 48, no. 2, p. e32, Apr. 2025.

[16] M. Kashina, I. D. Lenivtceva, and G. D. Kopanitsa, "Preprocessing of unstructured medical data: the impact of each preprocessing stage on classification," *Procedia Comput. Sci.*, vol. 178, pp. 284–290, 2020.

[17] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y.-I. Cho, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," *Mathematics*, vol. 12, no. 16, p. 2553, Aug. 2024.

[18] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technol. J. Ilm.*, vol. 15, no. 1, p. 93, Jan. 2024.

[19] Z. Lu, Y. Liu, and Q. Li, "A Research on the Academic System in Universities Based on the One-Hot Encoding PAC Fuzzy Comprehensive Evaluation Algorithm," in *Proceedings of Innovative Computing 2024*, 2024, pp. 224–235.

[20] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," *Appl. Nanosci.*, vol. 13, no. 3, pp. 1829–1840, Mar. 2023.

[21] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction," *Sci. Program.*, vol. 2021, pp. 1–12, Mar. 2021.

[22] H. Nizam-Ozogur and Z. Orman, "A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data," *Expert Syst.*, vol. 41, no. 8, Aug. 2024.

[23] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[24] I. A. Hidayat, "Classification of Sleep Disorders Using Random Forest on Sleep Health and Lifestyle Dataset," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 3, no. 2, pp. 71–76, Aug. 2023.

[25] S. K. Tadepalli and P. P. V. Lakshmi, "An Entropy enabled Random Forest Neural Network Algorithm to Grade the Reproductive System for Efficient Early Detection of Infertility," in *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, 2023, pp. 95–100.

[26] S. A. Domínguez-Miranda, R. Rodriguez-Aguilar, and M. Velazquez-Salazar, "Modeling the Relation Between Non-Communicable Diseases and the Health Habits of the Mexican Working Population: A Hybrid Modeling Approach," *Mathematics*, vol. 13, no. 6, p. 959, Mar. 2025.

[27] W. Zhao, J. Li, J. Zhao, D. Zhao, J. Lu, and X. Wang, "XGB model: Research on evaporation duct height prediction based on XGBoost algorithm," *Radioengineering*, vol. 29, no. 1, pp. 81–93, 2020.

[28] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sens. Networks*, vol. 18, no. 6, p. 155013292211069, Jun. 2022.

[29] E. J. Kusuma, R. Nurmandhani, L. Aryani, I. Pantiawati, and G. F. Shidik, "Optimasi Model Extreme Gradient Boosting Dalam Upaya Penentuan Tingkat Risiko Pada Ibu Hamil Berbasis Bayesian Optimization (BOXGB)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 12, no. 1, pp. 111–120, Feb. 2025.

[30] N. Amaya-Tejera, M. Gamarra, J. I. Vélez, and E. Zurek, "A distance-based kernel for classification via Support Vector Machines," *Front. Artif. Intell.*, vol. 7, Feb. 2024.

[31] H. W. Gichuhi, M. Magumba, M. Kumar, and R. W. Mayega, "A machine learning approach to explore individual risk factors for tuberculosis treatment non-adherence in Mukono district," *PLOS Glob. Public Heal.*, vol. 3, no. 7, p. e0001466, 2023.

Vol. 6, No. 4, Agustus 2025, Page. 2019-2034
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4854

[32] R. Bouchouareb and K. Ferroudji, "Classification of ECG Arrhythmia using Artificial Intelligence techniques (RBF and SVM)," in *2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 2022, pp. 1–7.
[33] E. J. Kusuma, I. Pantiawati, and S. Handayani, "Melanoma Classification based on Simulated Annealing Optimization in Neural Network," *Knowl. Eng. Data Sci.*, vol. 4, no. 2, p. 97, Mar. 2022.
[34] H. Schäfer *et al.*, "The Value of Clinical Decision Support in Healthcare: A Focus on Screening and Early Detection," *Diagnostics*, vol. 15, no. 5, p. 648, Mar. 2025.
[35] A. S. Albahri *et al.*, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Inf. Fusion*, vol. 96, pp. 156–191, Aug. 2023.
[36] H. Wang *et al.*, "Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches," *BMC Cancer*, vol. 24, no. 1, pp. 1–12, 2024.
[37] V. Kumar *et al.*, "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," *Healthcare*, vol. 10, no. 7, p. 1293, Jul. 2022.