Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2024, 2024:55
https://doi.org/10.28919/cmbn/8506
ISSN: 2052-2541

PREDICTING RECURRENCE IN DIFFERENTIATED THYROID CANCER: A COMPARATIVE ANALYSIS OF VARIOUS MACHINE LEARNING MODELS INCLUDING ENSEMBLE METHODS WITH CHI-SQUARED FEATURE SELECTION

KARLI EKA SETIAWAN*

Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Abstract: Differentiated thyroid cancer (DTC) is a kind of cancer that affects the endocrine system and is the fastest-growing cancer diagnosis globally. This research focused on the development of machine learning models for predicting the presence and absence of recurrence of DTC using 23 different machine learning models, such as SVM with various kernels (linear, polynomial, and radial basis functions), logistic regression, Naïve Bayes, Decision Tree, K-Nearest Neighbor, ensemble bagging using various base learners, random forest, ensemble stacking using various base learners, ensemble boosting using various base learners with AdaBoost, and Gradient Boosting Machine, as the first research contribution. This research continued the previous research, which shared the dataset, namely Differentiated Thyroid Cancer Recurrence. The second contribution of this research was the implementation of chi square as a feature selection method for our 23 machine learning models. The best machine learning model without using feature selection (Scenario 1) was random forest with 94% precision, recall, and f1-score in predicting the presence of recurrence of DTC and 98% precision, recall, and f1-score in predicting the absence of recurrence of DTC.

*Corresponding author
E-mail address: karli.setiawan@binus.ac.id
Received February 24, 2024

Meanwhile, after the use of feature selection (Scenario 2), we found ten machine learning models that can achieve maximum potential by achieving 100% accuracy in predicting both the presence and absence of recurrence of DTC.

Keywords: differentiated thyroid cancer; machine learning; ensemble learning; chi square feature selection; artificial intelligence.

2020 AMS Subject Classification: 68T01, 68T10, 97P80.

1. INTRODUCTION

The United Nations unveiled the Sustainable Development Goals (SDGs) on September 25, 2015, as a comprehensive plan to address global concerns over a 15-year timeframe, concluding on December 31, 2030. One of those goals is SDG 3, which specifically focuses on the goal of guaranteeing good health and well-being for people of all ages. Cancer is one of non-communicable diseases conjuring up deep fears into people around the world without warning as silent killer and a significant contributor to mortality and morbidity on a global scale, affecting about 14 million individuals annually [1]. The majority of this disease burden is concentrated in low- and middle-income nations, where 57% of cancer cases and 65% of cancer deaths take place. This study focused on the recurrence of Differentiated Thyroid Cancer (DTC), a kind of cancer that affects the endocrine system. It is the fastest-growing cancer diagnosis globally [2]. It is noted that females are more commonly affected by this malignancy [3]. DTC, or differentiated thyroid cancer, is the most often diagnosed cancer among those aged 15 to 29. In 2018, there were 567,233 reported cases of DTC globally, resulting in roughly 41,000 deaths [4] . Regarding the recurrence of DTC, recent study indicates that men have a greater rate of both recurrence and death compared to women, despite women having around three times the incidence rate of thyroid cancer than men. Derived from follicular epithelial cells, DTC constitutes more than 90% of all thyroid malignancies [5]. For information, Thyroid cancer is characterized by clinical symptoms such as dysphagia, throat constriction, painful throat, hoarseness, coughing, weight loss, weakness, and excessive perspiration [6], [7].

The right to health, which is one of the basic freedoms that society must realize, became the focal point of our research endeavors, aiming to give a computer science perspective on solving challenges in the health field [8], [9]. From a computer science perspective, our approach involves utilizing machine learning techniques to address issues related to accurate value prediction, pattern classification, data filtering, data structure, and extraction of valuable features [10]. This is particularly useful when dealing with numerous irrelevant or noisy characteristics [11]. This research presents, discusses, and analyzes a comparison study of various machine learning (ML) models around 23 different models for inspiring and helping public health stakeholders detect and monitor the recurrence of DTC as our first research contribution. The 23 various machine learning models were SVM with linear kernel, SVM with polynomial kernel, SVM with Radial Basis Function (RBF) kernel, logistic regression (LR), Naive Bayes, decision tree (DT), k-nearest neighbors (KNN), ensemble bagging of SVM with linear kernel, ensemble bagging of SVM with polynomial kernel, ensemble bagging of SVM with RBF kernel, ensemble bagging of LR, ensemble bagging of Naive Bayes, ensemble bagging of DT, ensemble bagging of KNN, random forest, ensemble stacking of various base learner (SVM with linear kernel, SVM with polynomial kernel, SVM with RBF kernel, LR, Naive Bayes, DT, KNN), ensemble boosting of SVM with linear kernel with AdaBoost, ensemble boosting of SVM with polynomial kernel with AdaBoost, ensemble boosting of SVM with RBF kernel with AdaBoost, ensemble boosting of Logistic Regression with AdaBoost, ensemble boosting of Naïve Bayes with AdaBoost, ensemble boosting of Decision Tree with AdaBoost, and gradient boosting machine (GBM). Our second contribution was the implementation of the chi square test as our feature selection to enhance our proposed machine learning models. This research conducted two experiment scenarios, where the first scenario was the use of all features, we have to train our machine learning models, and the second scenario was the use of selected and only recommended features by chi square feature selection to train our machine learning models.

2. RELATED WORKS

There was a review in machine learning to detect DTC conducted by Cao et al., where they utilized radiomics, a quantitative extraction technique, and throughput features from single or multiple medical pictures to predict the presence of DTC disease in the pictures [4] . According to them, the application of radiomics in medical image analysis is a noteworthy advancement and a crucial innovation since it allows for the automatic extraction of numerous quantitative characteristics from medical pictures in a very efficient method, which can then be used as a machine learning input to predict the presence of DTC. There was another study in understanding tumor behavior and medical results for DTC conducted by Yang et al., which utilized unsupervised machine learning approach with clustering [12]. The Ensemble Algorithm for Clustering Cancer Data (EACCD) was suggested by them as an approach to develop prognostic algorithms for well-differentiated thyroid cancer. The EACCD had three primary steps: establishing initial dissimilarities using the Gehan-Wilcoxon test statistic, getting learned dissimilarities, and conducting hierarchical clustering analysis. The researchers utilized the well-differentiated thyroid cancer data diagnosed between 2004 and 2021 from the Surveillance, Epidemiology, and End Results (SEER) database of the National Cancer Institute. Using the same source of dataset as SEER database-related demographic and clinicopathological characteristics of patients suffering thyroid cancer during 2010 to 2015, there was research in supervised machine learning approaches such as SVM, LR, XGBoost, DT, RF, and KNN conducted by Liu et al. [13]. Their research primarily aimed at creating six machine learning models to forecast lung metastasis in thyroid cancer, with the RF model emerging as the most effective.

Another reference to the usage of machine learning approaches in thyroid cancer research was experimented by Shin et al., who performed an investigation of the supervised learning capabilities of machine learning such as artificial neural networks (ANN) and support vector machines (SVM) to differentiate follicular adenomas of thyroid glands using the data from two tertiary referral hospitals (Severance Hospital and Samsung Medical Center) patients between January 2012 and December 2015 in South Korea [14]. A supervised machine learning approach with SVM was also

used in research conducted by Masuda et al. for lymph node metastasis identification in patients who have thyroid cancer [15]. The dataset they used to train the SVM model was 117 suspected thyroid cancer patient data, and the SVM model was built to distinguish between metastatic and benign lymph nodes with 0.86 in area under the curve (AUC).

Still on the topic of thyroid cancer, Zhao et al. explored the capabilities of a deep learning approach to differentiate malignant and benign thyroid nodules [16]. The deep learning models they utilized were five different CNN models, such as Xception, SE-ResNeXt50, DenseNet169, DenseNet121, and ResNet50, to train and test thyroid nodule computed tomography (CT) images. Their data was from Northern Jiangsu People's Hospital, which contained a total of 1527 patients between July 2017 and December 2019. Another study on DTC with a deep learning approach was done by Chan et al. to facilitate the diagnosis of DTC by classifying thyroid tumors as malignant or benign [17]. They used an image dataset with thyroid ultrasound images of 421 DTCs and 391 benign patients as a dataset for developing three CNN-based models such as VGG19, ResNet101, and InceptionV3. The best deep learning for diagnosing DTC was InceptionV3 with an 83.7% recall, while the results of ResNet101 and VGG19 were 72.5% and 66.2%, respectively. Their dataset was obtained from the cancer registry of Chang Gung Memorial Hospital between January 2008 and July 2020.

3. RESEARCH METHODOLOGY

This research methodology executed several steps, as depicted in Figure 1. The first step was the preprocessing step by performing exploratory data analysis and feature engineering with our dataset, namely the Differentiated Thyroid Cancer Recurrence dataset, where the detailed information about the dataset is described in the Dataset section. Due to our curiosity in treating datasets as inputs for our machine learning models, this research explored and investigated the impact of using Chi Square as feature selection by comparing the machine learning model prediction result using feature selection and the machine learning model prediction result without using feature selection. So that, this research had two scenarios, where scenario 1 was the use of

all features from the dataset for our machine learning models and scenario 2 was the use of chi square as feature selection to select the best feature to use. The third step was the search for the best hyperparameter set for our different machine learning models to perform at their best. The fourth step was machine learning training and testing. Finally, the last step of this research was the model evaluation.

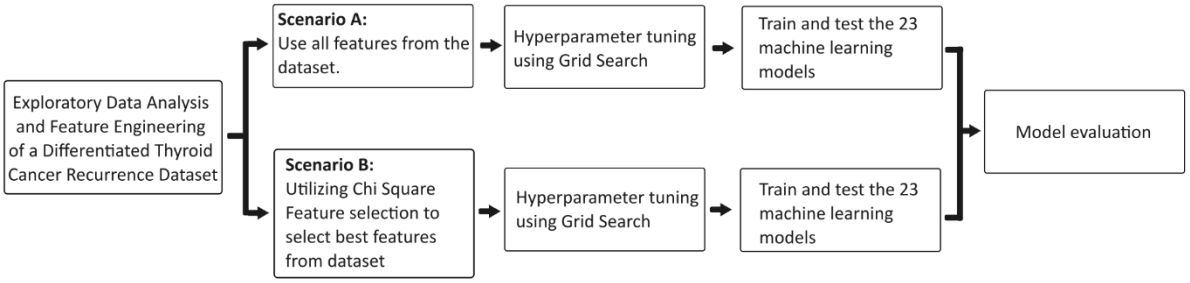

FIGURE 1. Research Methodology.

3.1. Dataset

The dataset investigated in this research paper was obtained from the research conducted by Borzooei et al. [18]. This research obtained this relatively new dataset, which has been publicly available since October 30th, 2023, through the UC Irvine Machine Learning Repository with the title Differentiated Thyroid Cancer Recurrence. The dataset contains 17 health information about 383 patients who have a track record of thyroid cancer observed for a minimum duration of 10 years and until 15 years. From the 17 information of patients, this study used 16 features such as age of patients; gender of the patients; the information if the patient is currently smoking or not; the history of smoking; the history of radiotherapy; the category thyroid function of patients; the category of physical examination of the patients; the categorical information about patients' adenopathy; the information of different pathological subtypes of thyroid cancer; the information about cancerous growth if it is unifocal (single nodule) or multifocal (multiple nodules); the category risk type; T: the size of the tumor and how far it has spread in the categorical stage, with T1 and T2 being smaller tumors and T3 and T4 being larger and more spread out [19]; N: The number of lymph nodes in the area that are affected; M: The presence or absence of distant metastasis; the stage of the thyroid cancer; and the response information; to train our machine

learning models to predict the recurrence of thyroid cancer as our label or target prediction. The illustration of the dataset used in this research is illustrated in Figure 2.

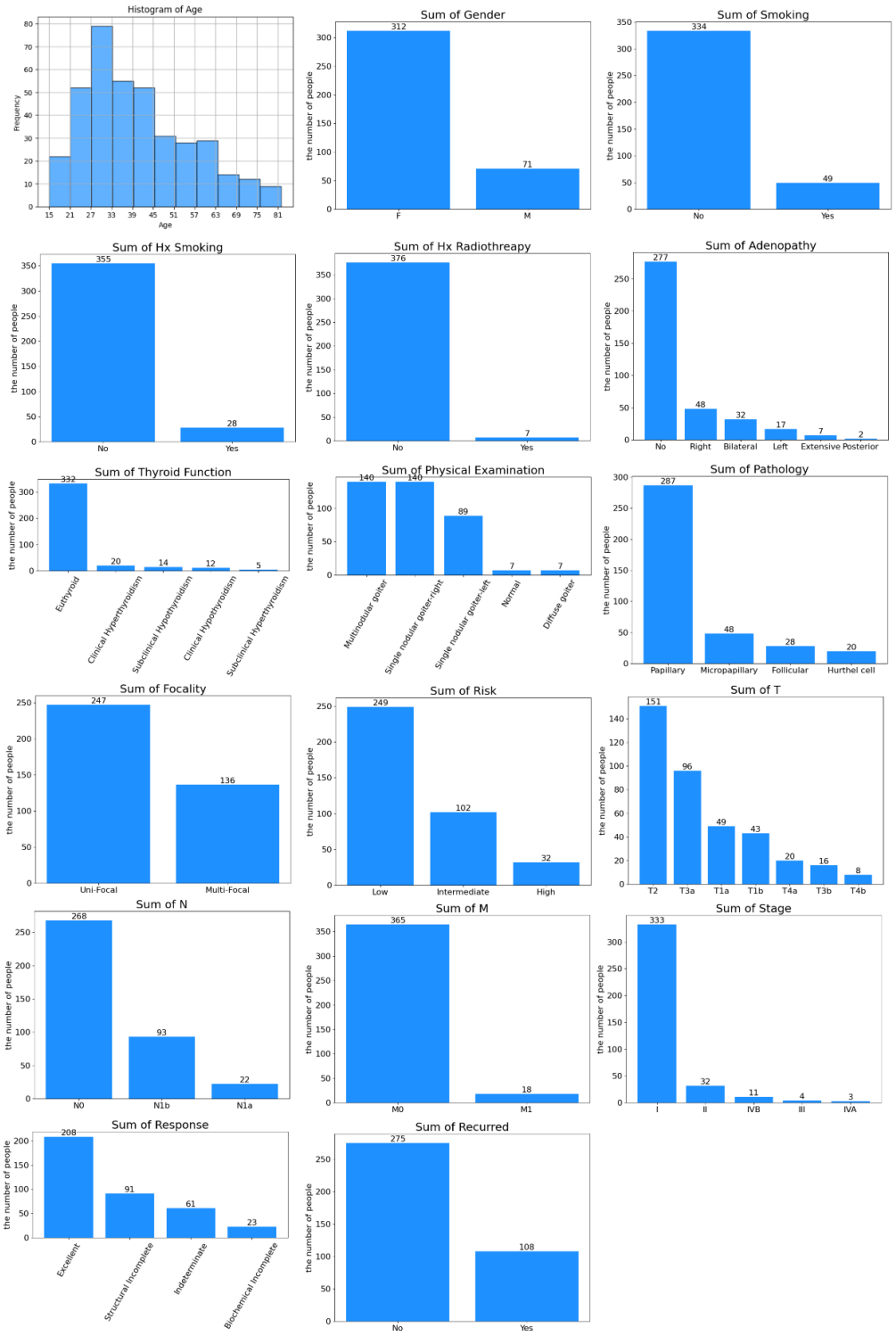

FIGURE 2. An Exploratory Data Analysis from the Dataset.

3.2. Feature Selection and Research Scenario

The chi-square test is a statistical method commonly used for feature selection, particularly with categorical data [20]. It assesses the independence between categorical variables, aiming to determine whether there is a noteworthy association between each feature and the target variable in a dataset. This the second scenario, this research using chi square test for feature selection with the result as table 1.

TABLE 1. Feature Selection using Chi Square.

Features performed Chi-Square Test	**Chi-Square Statistic**	**P-value**	**Feature Selection Decision**
Age	92.7796	0.010834025074581336	Include
Gender	39.3966	3.4588517738589313e-10	Include
Smoking	40.3440	2.129503925882454e-10	Include
Hx Smoking	5.9774	0.014489737063609615	Include
Hx Radiothreapy	8.9360	0.002795966115812055	Include
Thyroid Function	5.1486	0.27237868540391813	**Exclude**
Physical Examination	12.9743	0.011401661887227106	Include
Adenopathy	5.1486	0.27237868540391813	**Exclude**
Pathology	23.2704	3.546645063731057e-05	Include
Focality	54.6416	1.4463803836917104e-13	Include
Risk	208.8262	4.50781577510179e-46	Include
T	141.2902	5.353537306538777e-28	Include
N	153.1867	5.443985431857947e-34	Include
M	44.4444	2.616836848791985e-11	Include
Stage	97.6179	3.1610731210091906e-20	Include
Response	309.4723	8.863123780463506e-67	Include

For the second scenario, this research excluded thyroid function and adenopathy features because the P-value was below 0.05, as shown in Table 1. So, this research also tried to compare the prediction result between scenario 1, where all features were used to train the machine learning models, and scenario 2, where this research performed feature selection using chi squared to train the machine learning models.

3.3. Support Vector Machine

The Support Vector Machine (SVM) was initially introduced by Cortes and Vapnik as a method for identifying the most effective hyperplane in a space with many dimensions [21]. It has the ability to divide data into two distinct groups for binary classification, maximizing the separation between the two classes [22]. The SVM algorithm is frequently acknowledged for its superior performance compared to other classifiers [23]. However, the primary motivation for using an SVM is often the presence of a non-linearly separable issue. Therefore, an SVM with a non-linear kernel, such as the Radial Basis Function (RBF), would be appropriate in such scenario. Another pertinent rationale for employing Support Vector Machines (SVMs) is when one is operating inside a high-dimensional space. A kernel function is a mapping process applied to the training set in order to enhance its similarity to a data set that can be separated linearly. Mapping serves the objective of augmenting the dimensionality of the data collection, and this process is carried out with efficiency by employing a kernel function. In this research, we tried to solve our problem using three different kernels such as linear, polynomial, and RBF kernel.

If summarized, SVM models offer several benefits, including the utilization of kernel functions like the Gaussian basis function or polynomials, which facilitate the classification of non-linearly separable classes. It functions well as a linear classifier, demonstrates strong performance on datasets with several features, ensures appropriate segregation of data, and excels with heterogeneous distributed points. SVMs also possess certain limitations, such as the fact that they exhibit computational intensity when handling unlabeled datasets, and they also face limitations in terms of speed and size during both the training and testing stages of the algorithm. Additionally, SVM encounters speed limitations in the selection of kernel function parameters.

3.4. Logistic Regression

Logistic regression is a parametric machine learning approach that involves a mathematical model that maps input characteristics $x_1, x_2, \dots, x_p$ to an output (p). Logistic regression is a variant of linear regression that generates a logistic curve, as seen in Figure 2 [24]. Logistic regression is not a totally black box model as its purpose in machine learning is to determine the

weights $b_0, b_1, b_2, \ldots, b_p$ that accurately translate input features to the proper output target or label (p), as shown in equation 1 [25]. This research was selected to examine the efficacy of logistic regression because of its capacity to effectively model well-defined datasets.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}} \tag{1}$$

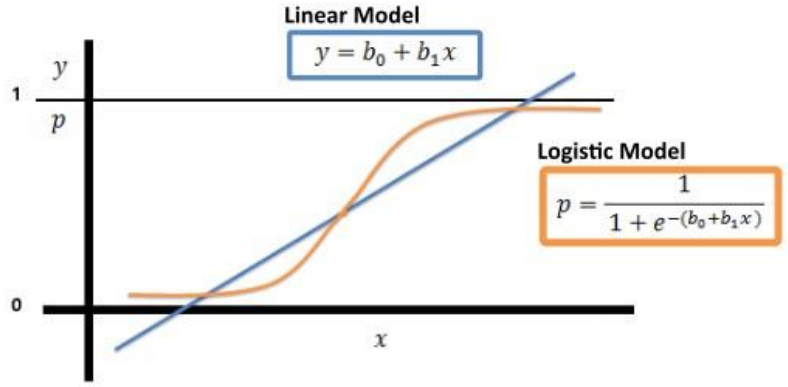

FIGURE 3. Linear Model vs Logistic Model.

3.5. Naïve Bayes

Naïve Bayes (NB) is a parametric machine learning model that utilizes conditional probability through the application of Bayes theorem, as instead of classification rules [26]. Equation 2 represents the notation of NB, which includes the posterior probability of the class (target/label) given the prediction (attribute) as $P(h|X)$, the prior probability of the class as $P(h)$, the likelihood which is the probability of the predictor given the class as $P(X|h)$, and the prior probability of the predictor as $P(X)$.

$$P(h|X) = \frac{P(X|h)P(h)}{P(X)} \tag{2}$$

Naïve Bayes (NB) is a widely recognized technique for probabilistic categorization [27]. This method is straightforward yet highly effective, and it has a broad range of practical uses. The NB classification approach is highly resilient, which is one of its key attributes. Although data may deviate from the basic assumptions of the NB model, NB can nonetheless demonstrate

extraordinary performance. The NB method has several advantages, including its ability to handle missing values by utilizing the complete set of characteristics for making predictions [28]. This implies that even if certain values are missing, the system may still utilize the existing information. Another benefit of this method is its low variance, which is attributed to the absence of a search function.

3.6. Decision Tree

A decision tree (DT) is a robust non-parametric machine learning method used to identify patterns in data for classification tasks and predictions [29]. The utilization of a recursive partition in the instance space can establish a decision tree classifier. DT employ several key terminologies, as illustrated in Figure 4. These include the root node, which marks the starting point of the DT and lacks any incoming edges. Decision nodes, on the other hand, possess outgoing edges and a solitary incoming edge. Leaf nodes, which signify the conclusion of the DT, lack outgoing edges but have precisely one incoming edge. Splitting refers to the act of dividing a node into sub-nodes, while pruning involves the removal of a sub-node.

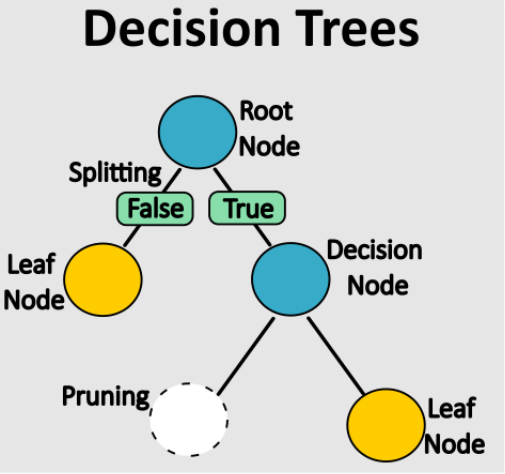

FIGURE 4. An Illustration of Decision Tree Model.

The functionality of a decision tree may be elucidated as follows: Commence by examining the characteristics or inputs in our dataset. Subsequently, divide the dataset depending on whether the rule result is true or false. Repeat this process till the class percentage is uniform and forms the leaf. Entropy and Gini may be utilized to assess the uniformity of our data inside a node.

The DT method comes to the supervised learning category and is applicable not only for classification problems but also for solving regression challenges [29]. Decision Trees (DT) offer several advantages, including their simplicity for human comprehension, their ability to be easily translated into production principles, their capability to classify both categorical (classification) and numerical outcomes (regression), and their lack of reliance on a priori hypotheses to evaluate the quality of results. However, DT also have some drawbacks, such as the potential for suboptimal decision-making and subsequent incorrect decisions, the possibility of having numerous layers in the DT, and the potential for increased complexity.

3.7. K-Nearest Neighbors

The KNN algorithm is a technique used in pattern recognition to categorize objects by comparing them to their nearest training samples in the feature space [23]. Data inside a collection of datasets is categorized based on the majority vote of its neighboring objects. Specifically, the object is assigned to the class that is most prevalent among its k-nearest neighbors, where k is a positive whole number. The KNN algorithm determines the classification of a new test feature vector based on the classes of its k-nearest neighbors. The KNN algorithm utilized Euclidean distance metrics denoted as $d(x, y)$ to locate the nearest neighbor of a data, where x and y are the observed two data explained in detail in equation 3 and Figure 5. N is the number of available features in dataset.

$$d(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \tag{3}$$

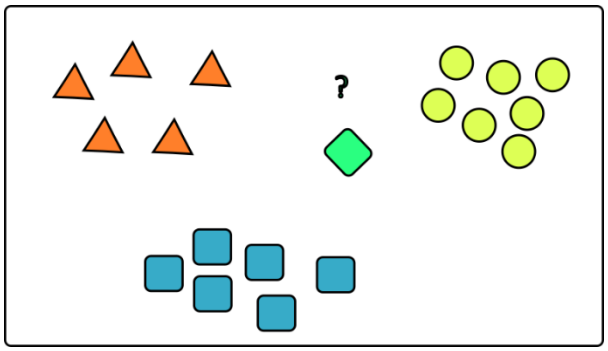

FIGURE 5. An Illustration of KNN Algorithm in Deciding Member of a Group.

KNN possesses some advantages, such as its simplicity in implementation and comprehension [23]. Additionally, it exhibits a substantial computational complexity due to the requirement of computing the Euclidean distance between the input feature and all features in the database. Nevertheless, during the training phase, it does not require any cost, but during the classification phase, it requires significant processing resources. Additionally, it demonstrates excellent performance while dealing with diverse dispersed points. Nevertheless, KNN has several drawbacks, including poor scalability when confronted with a dataset including a million labeled samples, resulting in a considerable amount of time required to identify the K nearest neighbors.

The KNN technique is not only applicable for classification, but it also can be used for regression, and search tasks. It is beneficial for resolving problems that need the identification of similar entities in order to find solutions.

3.8. Ensemble Bagging

Bagging is a widely used ensemble technique that involves training individual learners on randomly picked portions of the primary training dataset [30]. Ensemble learning using bagging might be referred to as the training of independent base learners [31]. Figure 6 illustrates that ensemble learning with bagging comprises many independent models that operate without influence from other weak learners. In this study, decision trees were employed as the weak learner. This model utilized bootstrapping, a technique that involves resampling data with replacement. In Figure 6, during image bagging with DT, each DT model was trained on a distinct subset of the dataset. Each DT model within the Bagging models is distinct from one another. The Scikit Learn package is utilized to retrieve the results of the bagging model by aggregating the soft voting outcomes of each weak learner. Bagging has the benefit of decreasing variance, hence removing the problem of overfitting [32]. Additionally, it has strong performance when applied to data with a high number of dimensions. Bagging is associated with several drawbacks, including its high computational cost, significant bias, and the loss of model interpretability.

This research developed seven different ensemble bagging models, such as bagging with linear kernel SVM, bagging with polynomial kernel SVM, bagging with RBF kernel SVM,

bagging with KNN, bagging with logistic regression, bagging with Naïve Bayes, and bagging with DT, to be investigated in solving this binary classification research.

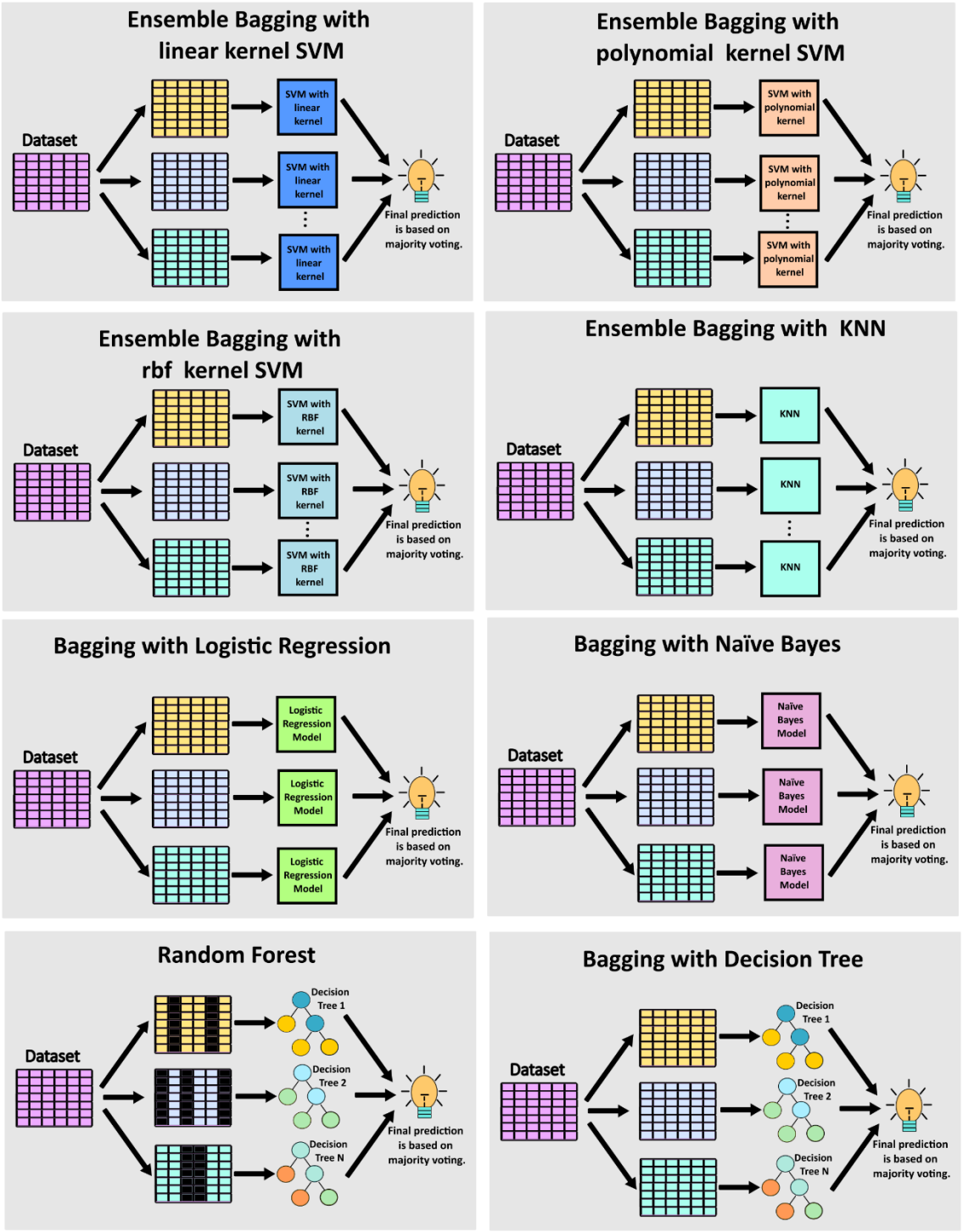

FIGURE 6. An Illustration of Ensemble Bagging Model and Random Forest Model Used in This Research.

3.9. Random Forest

Random forest is an adaptation of ensemble learning with decision trees, with the additional feature of random selection of features from the dataset [30], [31]. This distinguishes it from ensemble learning with bagging especially with ensemble bagging with DT. In order to ensure that just a certain set of characteristics is utilized by each individual poor learner inside this framework, as seen in Figure 6, the features that are not utilized are represented by the color black. The term "Random Forest" is coined based on the inclusion of many decision trees with diverse characteristics and the utilization of bootstrapping sampling techniques. The ultimate prediction remains unchanged in decision tree bagging, as it is determined by the majority voting of all base learners.

Random Forest (RF) has several benefits, including the guarantee that the generalization error consistently decreases, even when the number of trees increases [23]. Additionally, RF mitigates the risk of overfitting on a specific feature. Nevertheless, the issue of overfitting to training data persists, and it frequently leads to speedier achievement of outcomes. A drawback of RF is that bigger input datasets will result in increased classification times.

3.10. Ensemble Stacking

The ensemble stacking approach, often referred to as Stacked Generalization, is a methodology for combining information from numerous predictive models to create a new model, known as a meta-model [32]. A stacking model comprises several level-0 models, which are the basis models, and a level-1 model that combines the projections from these base models. Level 0 models, also known as basic models, are models that are trained on the training data and their predictions are aggregated. However, at the level 1 model (meta-model), the model acquires the optimal method of combining the predictions from the base models. This research implemented some machine learning models to be level 0 models or base learners, such as linear kernel SVM, polynomial kernel SVM, RBF kernel SVM, KNN, logistic regression, Naïve Bayes, and decision tree models, while the chosen linear kernel SVM as level 1 model or meta learner was a suggestion from the grid search approach in finding the best model to become a meta learner model, as

explained in the chapter of analysis and results of experimental results. The ensemble stacking model built for this research is depicted in Figure 7.

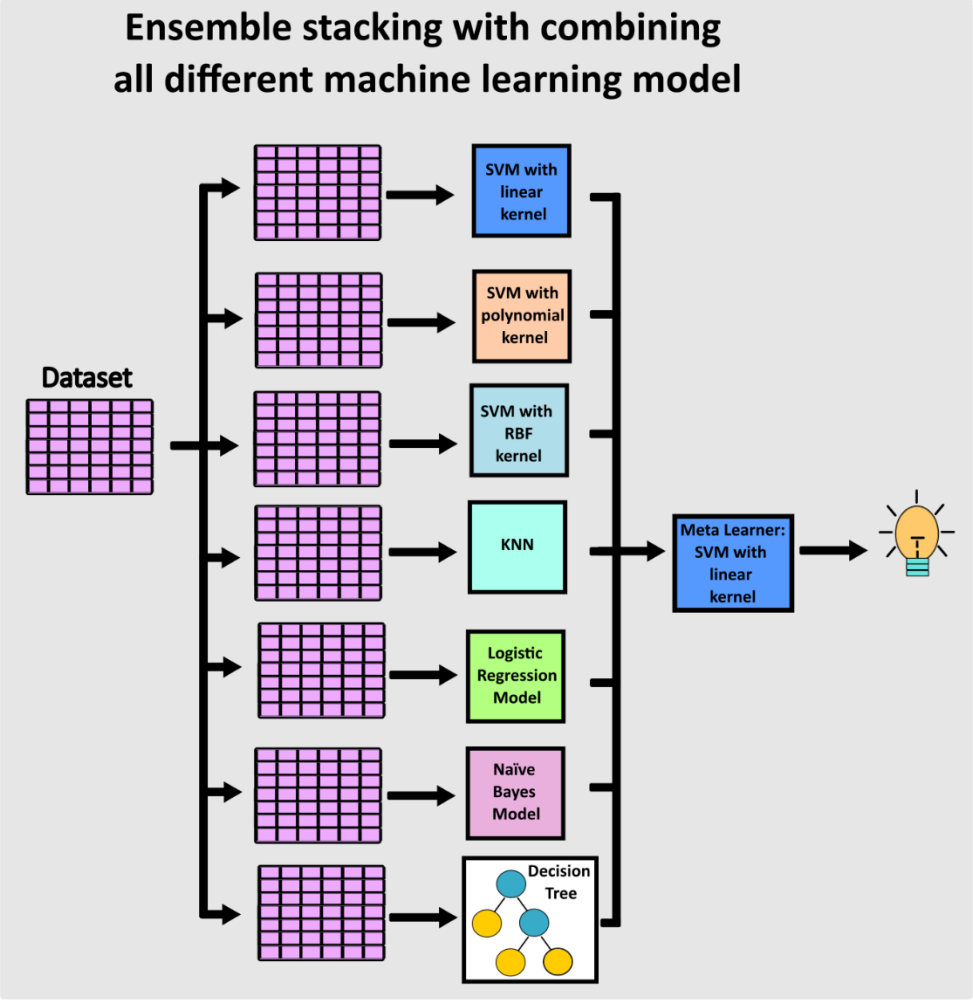

FIGURE 7. An Illustration of Ensemble Stacking Model Used in This Research.

3.11. Ensemble Boosting with Modified AdaBoost

Adaptive boosting (AdaBoost) is a member of Ensemble learning family, which architecture models were built to be adaptive minimizing the error from previous base learners [33]. The illustration about how our modified AdaBoost works is depicted in Figure 8. The AdaBoost learning method entails assigning more weights to misclassified data points and lesser weights to properly classified data points from the previous round of the base learner. The weight update procedure guarantees that future base learners prioritize the previously misclassified data points,

hence enhancing the overall performance of the ensemble model. This research proposed seven different AdaBoost models such as AdaBoost with linear kernel SVM, AdaBoost with polynomial kernel SVM, AdaBoost with RBF kernel SVM, AdaBoost with KNN, AdaBoost with logistic regression, AdaBoost with Naïve Bayes, and AdaBoost with decision tree.

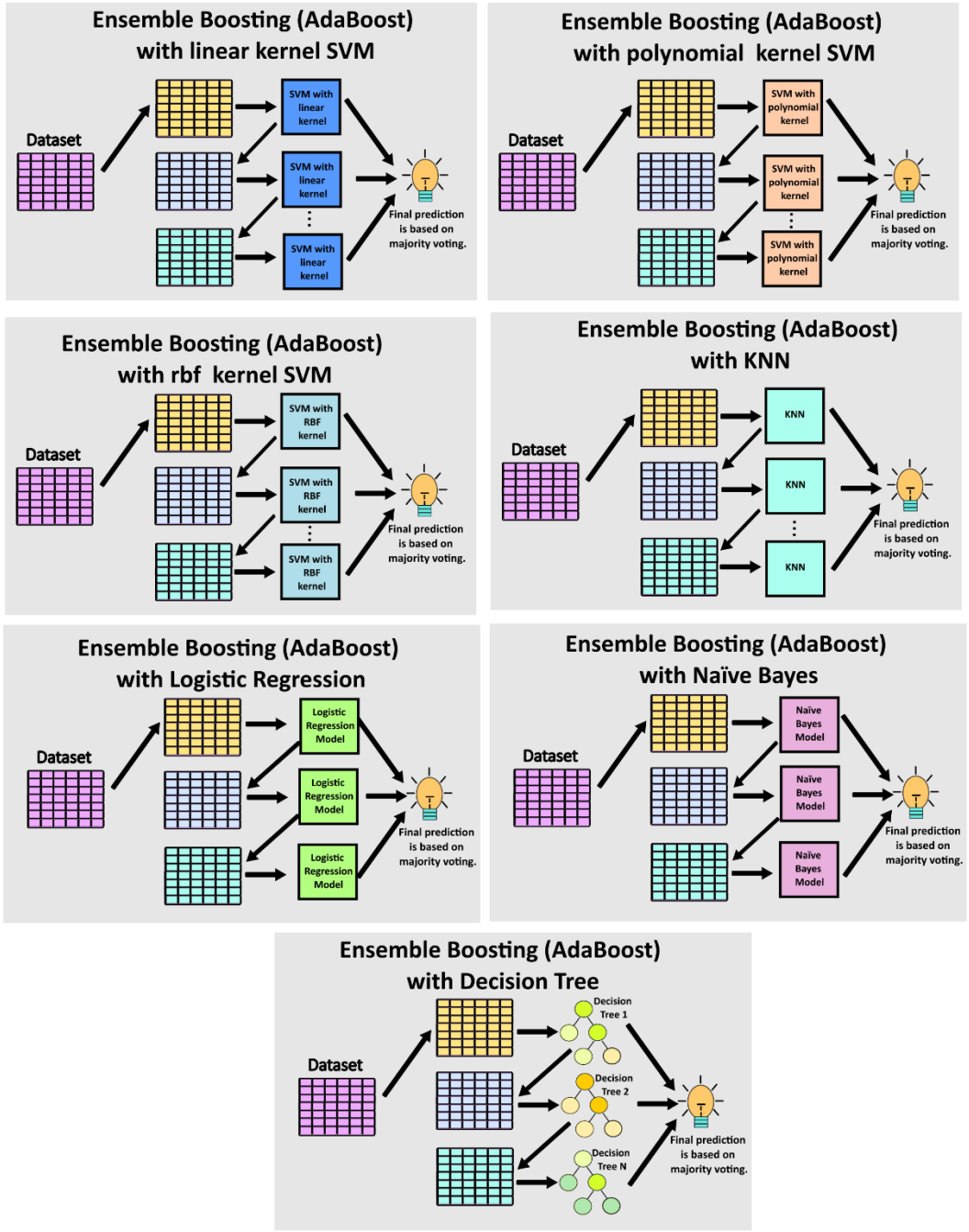

FIGURE 8. An Illustration of Ensemble Boosting with the Modified AdaBoost Model Used in This Research.

AdaBoost has several benefits, including its simplicity in comprehension and visualization, requiring just a limited number of hyper-parameters to be adjusted [33]. It also demonstrates considerable resilience to overfitting in low-noise datasets and can be applied to both regression and classification problems. However, AdaBoost does have several limitations, such as being very susceptible to noisy data and performing worse than RF and Extreme Gradient Boosting (XGBoost) when dealing with irrelevant characteristics. Additionally, it is not specifically designed for optimizing performance.

3.12. Ensemble Boosting with Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is an ensemble learning technique that uses decision trees. It creates a sequence of trees to minimize a loss function by utilizing gradient descent optimization. Unlike AdaBoost, which primarily emphasizes the allocation of weights to individual data points, GBM modifies the weights of the learners themselves to rectify mistakes made by preceding trees [33]. GBM employs an optimization process to enhance the prediction performance by effectively combining predictors both within individual trees and across multiple trees.

GBM is versatile since it can be applied to both classification and regression issues. Nevertheless, when employing GBM, the outcome of the prediction may suffer from overfitting, lack interpretability, and require substantial time and memory resources.

3.13. Evaluation Metrics

In order to assess and compare our machine learning models, this study employed various assessment measures including accuracy, precision, recall, and F1-score [34]. These metrics are represented by equations 4, 5, 6, and 7. The accuracy demonstrates the accuracy of all categories for both positive and negative classes. The recall demonstrates the sensitivity by quantifying the number of accurately identified positive cases. Precision quantifies the level of trust in a model by measuring its accuracy. The F1 score is determined by evaluating the trade-off between recall and accuracy.

A genuine positive (TP) data point in the confusion matrix occurs when the predicted positive outcome matches the actual result [35]. A false positive (FP), or Type 1 error, is a data point in the

confusion matrix when a positive outcome is expected, but the actual outcome is negative. A false negative (FN) in the confusion matrix refers to a situation when a negative result is anticipated, but the actual result turns out to be positive. This situation is classified as a Type 2 error, which is equally perilous as a Type 1 error. In the confusion matrix, a data point is classified as a true negative (TN) when the predicted outcome is negative and the actual outcome is also negative.

$$Accuracy = \frac{(True\ Positive\ + True\ Negative)}{(True\ Positive + True\ Negative + False\ Negative + False\ Positive)} \quad (4)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

$$Recall\ \ = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (6)$$

$$F1\ Score\ \ = \frac{(2\ \times Precision\ \times Recall)}{(Precision + Recall)} \quad (7)$$

4. Analysis and Results of Experimental Results

This research investigated the capability of various machine learning models—around 23 different models—to predict the recurrence of differentiated thyroid cancer. To obtain the best performance of each different machine learning model, this research performed the searching of hyperparameter tuning using grid search, where every grid search performed in this research was set with a different parameter due to the fact that all machine learning models work with different algorithms. Due to our machine learning models being 23 and this research having two scenarios, this research performed 46 grid searching processes, and the results are noted in Table 2. This research split our dataset into two parts: the first part was for training the machine learning models with 80 percent of the dataset, and the second part was for testing the machine learning models with 20 percent of the dataset for evaluation purposes.

TABLE 2. Proposed Machine Learning Models.

No	Proposed Machine Learning Model	Model Details	Optimal Hyperparameter Setting Suggested via Grid Search for scenario 1	Optimal Hyperparameter Setting Suggested via Grid Search for scenario 2
1	L-SVM	SVM with linear kernel	C=100, gamma= scale, kernel=linear	C=100, gamma= scale, kernel=linear
2	P-SVM	SVM with polynomial kernel	C= 1, degree= 2, gamma= auto, kernel= poly	C= 100, degree= 2, gamma= auto, kernel= poly
3	RBF-SVM	SVM with RBF kernel	C=100, gamma= scale, kernel= RBF	C=100, gamma= scale, kernel= RBF
4	LR	Logistic Regression	Default	Default
5	NB	Naive Bayes	Default	Default
6	DT	Decision Tree	Default	Default
7	KNN	K-Nearest Neighbors	n_neighbors = 7, weights= distance	n_neighbors = 3, weights= distance
8	L-SVM Bagging	Ensemble Bagging of SVM with linear kernel	n_estimators=75	n_estimators=10
9	P-SVM Bagging	Ensemble Bagging of SVM with polynomial kernel	n_estimators=50	n_estimators=25
10	RBF-SVM Bagging	Ensemble Bagging of SVM with RBF kernel	n_estimators=50	n_estimators=25
11	LR Bagging	Ensemble Bagging of Logistic Regression	n_estimators=10	n_estimators=25
12	NB Bagging	Ensemble Bagging of Naive Bayes	n_estimators=10	n_estimators=50
13	DT Bagging	Ensemble Bagging of Decision Tree	n_estimators=125	n_estimators=10
14	KNN Bagging	Ensemble Bagging of K-Nearest Neighbors	n_neighbors = 7, weights= distance, n_estimators= 100	n_neighbors = 7, weights= distance, n_estimators= 75

No	Proposed Machine Learning Model	Model Details	Optimal Hyperparameter Setting Suggested via Grid Search for scenario 1	Optimal Hyperparameter Setting Suggested via Grid Search for scenario 2
15	RF	Random Forest	Criterion= gini, max_depth= 40, n_estimators= 25	Criterion= gini, max_depth= 3, n_estimators= 10
16	Ensemble Stacking	Ensemble Stacking of L-SVM, P-SVM, RBF-SVM, LR, NB, DT, KNN	final_estimator= L-SVM	final_estimator= L-SVM
17	L-SVM AdaBoost	Ensemble Boosting of SVM with linear kernel with AdaBoost.	C=100, gamma= scale, kernel=linear, n_estimators=10	C=100, gamma= scale, kernel=linear, n_estimators=10
18	P-SVM AdaBoost	Ensemble Boosting of SVM with polynomial kernel with AdaBoost.	C= 1, degree= 2, gamma= auto, kernel= poly, n_estimators=10	C= 100, degree= 2, gamma= auto, kernel= poly, n_estimators=10
19	RBF-SVM AdaBoost	Ensemble Boosting of SVM with RBF kernel with AdaBoost.	C=100, gamma= scale, kernel= RBF, n_estimators=25	C=100, gamma= scale, kernel= RBF, n_estimators=25
20	LR AdaBoost	Ensemble Boosting of Logistic Regression with AdaBoost.	n_estimators=10	n_estimators=10
21	NB AdaBoost	Ensemble Boosting of Naïve Bayes with AdaBoost.	n_estimators= 25	n_estimators= 25
22	DT AdaBoost	Ensemble Boosting of Decision Tree with AdaBoost.	n_estimators= 150	n_estimators= 10
23	GBM	Gradient Boosting Machine	max_depth= 2, n_estimators= 150	max_depth= 2, n_estimators= 10

TABLE 3. Machine Learning Models Testing result based on Scenario 1.

Model	Machine Learning Models	Overall		Class 0 (The absence of recurred of DTC)				Class 1 (The presence of recurred of DTC)			
		Accuracy	Support	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
1	L-SVM	0.92	77	0.94	0.97	0.95	60	0.87	0.76	0.81	17
2	P-SVM	0.87	77	0.86	**1.00**	0.92	60	1.00	0.41	0.58	17
3	RBF-SVM	0.94	77	0.94	0.98	0.96	60	0.93	0.76	0.84	17
4	LR	0.91	77	0.93	0.95	0.44	60	0.81	0.76	0.79	17
5	NB	0.86	77	0.92	0.90	0.91	60	0.67	0.71	0.69	17
6	DT	0.91	77	0.96	0.92	0.94	60	0.75	0.88	0.81	17
7	KNN	0.90	77	0.91	0.97	0.94	60	0.85	0.65	0.73	17
8	L-SVM Bagging	0.94	77	0.95	0.97	0.96	60	0.88	0.82	0.85	17
9	P-SVM Bagging	0.87	77	0.86	**1.00**	0.92	60	1.00	0.41	0.58	17
10	RBF-SVM Bagging	0.86	77	0.85	**1.00**	0.92	60	1.00	0.35	0.52	17
11	LR Bagging	0.92	77	0.95	0.95	0.95	60	0.82	0.82	0.82	17
12	NB Bagging	0.86	77	0.92	0.90	0.91	60	0.67	0.71	0.69	17
13	DT Bagging	0.96	77	0.98	0.97	0.97	60	0.89	**0.94**	0.91	17
14	KNN Bagging	0.87	77	0.89	0.95	0.92	60	0.77	0.59	0.67	17
15	RF	0.97	77	**0.98**	**0.98**	**0.98**	60	**0.94**	**0.94**	**0.94**	17
16	Ensemble Stacking	0.91	77	0.96	0.92	0.84	60	0.75	0.88	0.81	17
17	L-SVM AdaBoost	0.91	77	0.93	0.95	0.94	60	0.81	0.76	0.79	17
18	P-SVM AdaBoost	0.78	77	0.78	**1.00**	0.88	60	0	0	0	17
19	RBF-SVM AdaBoost	0.86	77	0.87	0.97	0.91	60	0.80	0.47	0.59	17
20	LR AdaBoost	0.90	77	0.91	0.97	0.94	60	0.85	0.65	0.73	17
21	NB AdaBoost	0.78	77	0.78	**1.00**	0.88	60	0	0	0	17
22	DT AdaBoost	0.92	77	0.97	0.93	0.95	60	0.79	0.88	0.83	17
23	GBM	0.95	77	0.98	0.95	0.97	60	0.84	**0.94**	0.89	17

The testing result of machine learning models for scenario 1 is noted in Table 3. This research is in the health field, so our analysis should consider the sensitivity to detect a positive case by noticing the recall score. It can be seen from Table 3 that some machine learning models, such as SVM with polynomial kernel (model 2), ensemble bagging of SVM with polynomial kernel (model 9), and ensemble bagging with RBF kernel (model 11), were under 50 percent accurate in predicting all positive classes of the presence of recurred differentiated thyroid cancer. Whereas ensemble boosting of SVM with polynomial kernel with AdaBoost (model 18) and ensemble boosting of Naïve Bayes with AdaBoost (model 21) became the worst models due to their inability to predict all positive classes with 0 percent recall. Meanwhile, six of the twenty-three machine learning models achieved above 85 percent in recall scores, such as DT (model 6), ensemble stacking (model 16), and ensemble boosting of DT with AdaBoost (model 22), with 88 percent in recall score, and ensemble bagging with DT (model 13), random forest (model 15), and GBM (model 23), with 94 percent in recall score.

It can be seen in Table 3 that all the proposed machine learning models performed well in predicting the negative class or the absence of recurrence of differentiated thyroid cancer, with a recall score above 90 percent. To decide the best machine learning models for predicting positive class in this research based on Table 3, the random forest (RF) model (model 15) was the best model in scenario 1, with 94 percent recall, precision, and f1-score for predicting positive class and 98 percent recall, precision, and f1-score for predicting negative class.

The testing result of machine learning models for scenario 2 is noted in Table 4. The implementation of feature selection, especially using chi square, by removing two features, such as thyroid function and adenopathy, can significantly improve some machine learning models to achieve the best performance until 100 percent prediction in the presence of a recurrence of DTC in the recall score. Some machine learning models achieved the best performance in predicting both the presence and absence of recurrence of DTC in all metrics, including precision, recall, and f1-score: SVM with linear kernel (model 1), SVM with polynomial kernel (model 2), SVM with RBF kernel (model 3), DT (model 6), ensemble bagging of SVM with linear kernel (model 8),

ensemble bagging of DT (model 13), ensemble stacking (model 16), ensemble boosting of SVM with linear kernel with AdaBoost (model 17), ensemble boosting of DT with AdaBoost (model 22), and GBM (model 23).

TABLE 4. Machine Learning Models Testing result based on Scenario 2.

Model	Machine Learning Models	Overall		Class 0 (The absence of recurred of DTC)				Class 1 (The presence of recurred of DTC)			
		Accuracy	Support	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
1	L-SVM	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
2	P-SVM	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
3	RBF-SVM	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
4	LR	0.99	77	0.98	1.00	0.99	60	1.00	0.94	0.97	17
5	NB	0.94	77	0.98	0.93	0.96	60	0.80	0.94	0.86	17
6	DT	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
7	KNN	0.92	77	0.94	0.97	0.95	60	0.87	0.76	0.81	17
8	L-SVM Bagging	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
9	P-SVM Bagging	0.88	77	0.87	1.00	0.93	60	1.00	0.47	0.64	17
10	RBF-SVM Bagging	0.86	77	0.85	1.00	0.92	60	1.00	0.35	0.52	17
11	LR Bagging	0.99	77	0.98	1.00	0.99	60	1.00	0.94	0.97	17
12	NB Bagging	0.94	77	0.98	0.93	0.96	60	0.80	0.94	0.86	17
13	DT Bagging	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
14	KNN Bagging	0.88	77	0.89	0.97	0.93	60	0.83	0.59	0.69	17
15	RF	0.99	77	0.98	1.00	0.99	60	1.00	0.94	0.97	17
16	Ensemble Stacking	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
17	L-SVM AdaBoost	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
18	P-SVM AdaBoost	0.87	77	0.86	1.00	0.92	60	1.00	0.41	0.58	17
19	RBF-SVM AdaBoost	0.88	77	0.87	1.00	0.93	60	1.00	0.47	0.64	17
20	LR AdaBoost	0.96	77	0.95	1.00	0.98	60	1.00	0.82	0.90	17
21	NB AdaBoost	0.78	77	0.78	1.00	0.88	60	0	0	0	17
22	DT AdaBoost	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17
23	GBM	**1.00**	77	**1.00**	**1.00**	**1.00**	60	**1.00**	**1.00**	**1.00**	17

Based on table 4, still there are some machine learning models that achieved under 60 percent in recall score in predicting the presence of recurred DTC, such as ensemble bagging of SVM with polynomial kernel (model 9), ensemble bagging of SVM with RBF kernel (model 10), ensemble bagging of KNN (model 14), ensemble boosting of SVM with polynomial kernel with AdaBoost (model 18), ensemble boosting of SVM with RBF kernel with AdaBoost (model 19), and ensemble boosting of Naïve Bayes with AdaBoost (model 21). Comparing both tables 3 and 4, all mentioned machine learning before had slightly improved to be better after the implementation of feature selection using chi squared. The ensemble boosting of Naïve Bayes with AdaBoost (model 21) became the worst machine learning model in predicting the presence of recurred DTC in both scenarios. There is an opportunity in the future to continue this research by utilizing a deep learning approach [36].

5. CONCLUSION

This study had developed 23 machine learning models to predict the recurrence of differentiated thyroid cancer (DTC). Using all the features we have in the dataset as scenario 1, the best machine learning model was random forest (model 15), with 94 percent for predicting the presence of recurrence of TDC in precision, recall, and f1-score and 98 percent for predicting the absence of recurrence of TDC in precision, recall, and f1-score. This study also showed that the implementation of feature selection with chi squared can improve all machine learning models where it cause our ten machine learning models such as SVM with linear kernel (model 1), SVM with polynomial kernel (model 2), SVM with RBF kernel (model 3), DT (model 6), ensemble bagging of SVM with linear kernel (model 8), ensemble bagging of DT (model 13), ensemble stacking (model 16), ensemble boosting of SVM with linear kernel with AdaBoost (model 17), ensemble boosting of DT with AdaBoost (model 22), and GBM (model 23) can achieved its maximum potentiality to predict both the presence and absence of recurred of TDC with 100 percent in precision, recall, and f1-score. The author hopes this research can inspire all health researchers that feature selection may improve the utilization of machine learning, especially in

predicting the recurrence of TDC and other diseases.

For future research, this study can be continued by doing some investigation into the capability of some predictive models using deep learning approaches such as convolutional neural network (CNN)-based models, recurrent neural network (RNN)-based models, and others to improve research in the prediction of recurrence of TDC. In data science, there are other techniques that may improve the results of machine learning models, for instance, feature importance and dimension reduction. In the research of machine learning and deep learning, a lot of data can increase the prediction quality of predictive models, so it would be great if the health researchers continued this research by providing more data about the research in the recurrence of TDC.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] WHO, Cancer control: a global snapshot in 2015: Summary of results from the 2015 WHO NCD country capacity survey, World Health Organization, 2016.

[2] A. Matrone, M.C. Campopiano, A. Nervo, et al. Differentiated thyroid cancer, from active surveillance to advanced therapy: toward a personalized medicine, Front. Endocrinol. 10 (2020), 884. https://doi.org/10.3389/fendo.2019.00884.

[3] S. Vaccarella, J. Lortet-Tieulent, M. Colombet, et al. Global patterns and trends in incidence and mortality of thyroid cancer in children and adolescents: a population-based study, Lancet Diab. Endocrinol. 9 (2021), 144-152. https://doi.org/10.1016/s2213-8587(20)30401-0.

[4] Y. Cao, X. Zhong, W. Diao, et al. Radiomics in differentiated thyroid cancer and nodules: explorations, application, and limitations, Cancers 13 (2021), 2436. https://doi.org/10.3390/cancers13102436.

[5] L.A. Stewart, J.H. Kuo, Advancements in the treatment of differentiated thyroid cancer, Therap. Adv. Endocrinol. 12 (2021), 204201882110002. https://doi.org/10.1177/20420188211000251.

[6] Y.H. Sharifovna, Thyroid cancer diagnostics, classification, staging, Ijtimoiy Fanlarda Innovasiya Onlayn Ilmiy Jurnali, 1 (2021), 63-69.

[7] M. Nguyen, G. He, A.K.Y. Lam, Clinicopathological and molecular features of secondary cancer (metastasis) to the thyroid and advances in management, Int. J. Mol. Sci. 23 (2022), 3242. https://doi.org/10.3390/ijms23063242.

[8] P. Aryawibowo, A.F. Hidayanto, Y.M. Toemali, et al. Intelligent monitoring and diagnosing capability in healthcare: systematic literature review, in: 2023 International Conference on Information Management and Technology (ICIMTech), IEEE, Malang, Indonesia, 2023: pp. 627–632. https://doi.org/10.1109/ICIMTech59029.2023.10277846.

[9] K.E. Setiawan, A. Kurniawan, A. Chowanda, et al. Clustering models for hospitals in Jakarta using fuzzy c-means and k-means, Procedia Computer Sci. 216 (2023), 356-363. https://doi.org/10.1016/j.procs.2022.12.146.

[10] N. Ismail, U.K. Yusof, Recent trends of machine learning predictions using open data: a systematic review, J. Inform. Commun. Technol. 21 (2022), 337-381. https://doi.org/10.32890/jict2022.21.3.3.

[11] R.G. Jimoh, O.A. Abisoye, M.M.B. Uthman, Ensemble feed-forward neural network and support vector machine for prediction of multiclass malaria infection, J. Inform. Commun. Technol. 21 (2021), 117-148. https://doi.org/10.32890/jict2022.21.1.6.

[12] C.Q. Yang, L. Gardiner, H. Wang, et al. Creating prognostic systems for well-differentiated thyroid cancer using machine learning, Front. Endocrinol. 10 (2019), 288. https://doi.org/10.3389/fendo.2019.00288.

[13] W. Liu, S. Wang, Z. Ye, et al. Prediction of lung metastases in thyroid cancer using machine learning based on SEER database, Cancer Med. 11 (2022), 2503-2515. https://doi.org/10.1002/cam4.4617.

[14] I. Shin, Y.J. Kim, K. Han, et al. Application of machine learning to ultrasound images to differentiate follicular neoplasms of the thyroid gland, Ultrasonography 39 (2020), 257-265. https://doi.org/10.14366/usg.19069.

[15] T. Masuda, T. Nakaura, Y. Funama, et al. Machine learning to identify lymph node metastasis from thyroid cancer in patients undergoing contrast-enhanced CT studies, Radiography 27 (2021), 920–926. https://doi.org/10.1016/j.radi.2021.03.001.

[16] H. Zhao, C. Liu, J. Ye, et al. A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on CT images, Endokrynol. Polska 72 (2021), 217-225. https://doi.org/10.5603/EP.a2021.0015.

[17] W.K. Chan, J.H. Sun, M.J. Liou, et al. Using deep convolutional neural networks for enhanced ultrasonographic image diagnosis of differentiated thyroid cancer, Biomedicines 9 (2021), 1771. https://doi.org/10.3390/biomedicines9121771.

[18] S. Borzooei, G. Briganti, M. Golparian, et al. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study, Eur. Arch. Otorhinolaryngol. 281 (2024), 2095-2104. https://doi.org/10.1007/s00405-023-08299-w.

[19] A.R. Shaha, TNM classification of thyroid carcinoma, World J. Surg. 31 (2007), 879-887. https://doi.org/10.1007/s00268-006-0864-0.

[20] K. Dissanayake, M.G. Md Johar, Comparative study on heart disease prediction using feature selection techniques on classification algorithms, Appl. Comput. Intell. Soft Comput. 2021 (2021), 1-17. https://doi.org/10.1155/2021/5581806.

[21] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995), 273-297. https://doi.org/10.1007/BF00994018.

[22] B. Gaye, D. Zhang, A. Wulamu, Improvement of support vector machine algorithm in big data background, Math. Probl. Eng. 2021 (2021), 5594899. https://doi.org/10.1155/2021/5594899.

[23] E.Y. Boateng, J. Otoo, D.A. Abaye, Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: a review, J. Data Anal. Inform. Process. 08 (2020), 341-357. https://doi.org/10.4236/jdaip.2020.84020.

[24] P. Schober, T.R. Vetter, Logistic regression in medical research, Stat. Minute 132 (2021), 365-366.

[25] H. Jain, A. Khunteta, S. Srivastava, Churn prediction in telecommunication using logistic regression and logit boost, Procedia Computer Sci. 167 (2020), 101-112. https://doi.org/10.1016/j.procs.2020.03.187.

[26] R.V.B. Vangara, K. Thirupathur, S.P. Vangara, Opinion mining classification using Naive Bayes algorithm, Int. J. Innov. Technol. Explor. Eng. 9 (2020), 495-498. https://doi.org/10.35940/ijitee.e2402.039520.

[27] I. Wickramasinghe, H. Kalutarage, Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation, Soft Comput. 25 (2020), 2277-2293. https://doi.org/10.1007/s00500-020-05297-6.

[28] R. Irmanita, S.S. Prasetiyowati, Y. Sibaroni, Classification of malaria complication using CART (Classification

and Regression Tree) and Naïve Bayes, J. RESTI (Rek. Sist. Teknol. Inf.) 5 (2021), 10-16. https://doi.org/10.29207/resti.v5i1.2770.

[29] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, J. Appl. Sci. Technol. Trends 2 (2021), 20-28. https://doi.org/10.38094/jastt20165.

[30] G. Ngo, R. Beard, R. Chandra, Evolutionary bagging for ensemble learning, Neurocomputing 510 (2022), 1-14. https://doi.org/10.1016/j.neucom.2022.08.055.

[31] S. González, S. García, J. Del Ser, et al. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities, Inform. Fusion 64 (2020), 205-237. https://doi.org/10.1016/j.inffus.2020.07.007.

[32] A. Mohammed, R. Kora, A comprehensive review on ensemble deep learning: Opportunities and challenges, J. King Saud Univ. – Computer Inform. Sci. 35 (2023), 757-774. https://doi.org/10.1016/j.jksuci.2023.01.014.

[33] H. Jafarzadeh, M. Mahdianpari, E. Gill, et al. Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: a comparative evaluation, Remote Sensing 13 (2021), 4405. https://doi.org/10.3390/rs13214405.

[34] S.Y. Prasetyo, A. Kurniawan, E.F.A. Sihotang, et al. Heart disease risk prediction using K-nearest neighbor: A study of Euclidean and cosine distance metrics, in: 2023 3rd International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS), IEEE, Bali, Indonesia, 2023: pp. 236–240. https://doi.org/10.1109/ICON-SONICS59898.2023.10435299.

[35] Ž.Đ. Vujovic, Classification model evaluation metrics, Int. J. Adv. Computer Sci. Appl. 12 (2021), 599-606. https://doi.org/10.14569/IJACSA.2021.0120670.

[36] K.E. Setiawan, G.N. Elwirehardja, B. Pardamean, Comparison of deep learning sequence-to-sequence models in predicting indoor temperature and humidity in solar dryer dome, Commun. Math. Biol. Neurosci. 2022 (2022), 98. https://doi.org/10.28919/cmbn/7655.