



Enhancing Predictive Accuracy for Differentiated Thyroid Cancer (DTC) Recurrence Through Advanced Data Mining Techniques

Imelda Juliana BR. Sibarani^{*}, Katherina Meylda Loy S, Suharjito

Industrial Engineering Department, BINUS Graduate Program, Master of Industrial Engineering, Bina Nusantara University, Jakarta 11480, Indonesia

Email: ^{1,*}imelda.sibarani@binus.ac.id, ²katherina.loy@binus.ac.id, ³suharjito@binus.ac.id

Correspondence Author Email: imelda.sibarani@binus.ac.id

Abstract—Thyroid cancer is becoming more common, and its 20% recurrence rate of which almost half are discovered more than five years after surgery, highlights how difficult it is to distinguish between a true disease relapse and chronic disease brought on by insufficient initial treatment. This ambiguity highlights the complicated dynamics that drive the mortality rates in patients with thyroid cancer. The purpose of this study is to be refining these predictions to control Differentiated Thyroid Cancer recurrence and minimize the risk of recurrence. The dataset was obtained by monitoring a total of 383 patients with 17 attributes. This study adopted a data mining modelling strategy to evaluate the performance, classification accuracy, and cluster distribution, utilizing the Orange data mining software. The Exploratory Data Analysis was conducted to pinpoint the most significant contributors. Subsequently, a variety of supervised techniques were applied to assess the precision of both single and ensemble models in classification. For cluster determination, we implemented several unsupervised learning techniques, including k-means, hierarchical, and Louvain Clustering. The result shows that ensemble stacking algorithm demonstrated superior performance and classification accuracy, achieving impressive scores of 0.971. The analysis of clustering methods, notably k-means and hierarchical clustering, suggested that the dataset could be segmented into two distinct clusters. The most dominant factors in influencing the recurrence of thyroid cancer with strong correlation revealed 'Response', 'Risk', 'Adenopathy', and 'N'. The refinement of the diagnostic model, through the identification of accurate models and key factors, enhances the prediction of Differentiated Thyroid Cancer recurrence.

Keywords: Thyroid Cancer; Cancer Recurrence; Prediction; Orange Software; Ensemble Model; Unsupervised Learning

1. INTRODUCTION

Cancer Recurrence after treatment is a main reason for patient deaths, showing that the cancer cells have become more aggressive and harder to cure. To help patients live longer, it's very important to correctly predict if the cancer will come back and to provide the right treatment (Fengyun Zhang, Jie Geng, De-Gan Zhang, Jinglong Gui, 2023). For each person affected by Differentiated Thyroid Cancer (DTC), the main goals of the first line of treatment include enhancing the duration and quality of life while preventing the recurrence of the condition. Significant advancements have been made in the treatment of DTC (Differentiated Thyroid Cancer) over the last forty years. Patients with low-risk DTC face a less than 1% risk of mortality from thyroid cancer, and most recurrences in these cases are treatable. However, patients with high-risk DTC, who constitute 5–10% of all cases, are more likely to experience frequent recurrences and most thyroid cancer-related deaths occur within this subgroup. Consequently, the approach to managing patients with intermediate-risk DTC is critical and often determined on an individual basis (Schlumberger & Leboulleux, 2021).

The most common type of DTC is Papillary Thyroid Cancer (PTC), making up more than 85% of cases. Next, Follicular Thyroid Cancer (FTC) comes in three forms, each more aggressive than the last: one with invasion just of the capsule, another that invades blood vessels but is still encapsulated, and a third that is widely invasive. The third and fourth categories of DTC include Hürthle Cell Cancer and Poorly Differentiated Thyroid Cancer (Schlumberger & Leboulleux, 2021). Therefore, enhancing prognostic factors is also essential. In classifying the prognosis of DTC, two main risk factors are considered: the risk of dying from the cancer, which is determined using the TNM classification system, and the risk of the cancer coming back in a physical form, which is assessed using the risk stratification guidelines provided by the American Thyroid Association (ATA) (Schlumberger & Leboulleux, 2021).

The incidence of thyroid cancer has been steadily rising over the last several decades. Between 2005 and 2015, a publication found that while thyroid cancer incidence climbed significantly, thyroid cancer mortality increased slightly in China (J. Wang et al., 2020). Similar to in the US, the number of cases of thyroid cancer has increased over time. The study examined secular changes in the incidence of thyroid cancer between 1975 and 2009. Based on their research, the primary cause of the increase in thyroid cancer incidence is the identification of tiny papillary lesions. (Davies & Welch, 2014). In 2018, the most recent year for which data is available, nearly 900,000 people in the United States were diagnosed with thyroid cancer. Despite this, over 2,000 patients die every year from thyroid cancer (Agosto Salgado et al., 2023). By 2020, the incidence of well-differentiated thyroid carcinoma is expected to rise by 30%, surpassing the rate of increase in all other malignancies (Chan et al., 2020). There are no known risk factors for thyroid cancer, and the incidence rate of the disease is expected to rise dramatically. The observed rise in incidence rates is frequently attributed to overdiagnosis of thyroid cancer (Smittenaar et al., 2016). Practitioners in the medical field endeavor to diagnose diseases in their early stages, allowing for timely and cost-effective treatment (Javaid et al., 2022).

However, numerous studies have shown that 20% of patients with this illness experience recurrences, and of those, nearly half were discovered more than five years following the initial surgery. It is frequently unclear whether these occurrences constitute a true relapse after a time of disease-free status or chronic disease following incomplete therapy (Chatchomchuan et al., 2021)(Sapuppo et al., 2018). Even complete initial treatment, with no leftover malignant cells, cannot allow for recurrence (Sapuppo et al., 2018). Several observational studies suggest that low-volume



recurrent nodal disease might be slow-moving, necessitating close monitoring in order to treat it (Haugen et al., 2016). Davies and Welch's Journal (Davies & Welch, 2014) lists a number of risk factors for recurring thyroid cancer, such as age, sex, tumour features, histological subtype, inadequate first treatment, involvement of lymph nodes(N), genetic mutations, radiation exposure, incomplete treatment and etc. Several articles cannot validate the precise risk factors that may affect the recurrence of thyroid cancer (J. Wang et al., 2020),(Sapuppo et al., 2018).

Numerous factors can impact the rate of metastasis and recurrence, which are the primary causes of death in thyroid cancer. Integrating Big Data into medical studies can improve human health on a deeper and bigger scale (Hong et al., 2018). Health care data is one of the driving drivers behind big data. With enhanced data generation technology, there is an exponential increase in the volume of data. Clinical actions create multiple records, including patient identity, diagnosis, medication schedule, physician notes, and sensor data(L. Wang & Ann Alexander, 2013). Machine learning has been applied to identify many diseases as prognostic. Researchers utilize a variety of classification methods ANN, including Bayesian network (BN), SVM, neural network, decision tree (DT), Naive Bayes, Random Forest and K-nearest neighbor (KNN) (Abbad Ur Rehman, Lin, Mushtaq, et al., 2021),(Chaganti et al., 2022),(Chaubey et al., 2021). Recent developments in data processing and computation have led to the application of machine learning approaches for predicting Thyroid cancer and the recurrence. We studied the papers and selected 10 that are highly relevant to our job. Related study with single model proposed Naive Bayes and KNN Alogrithm as the highest Accuracy shows in (Abbad Ur Rehman, Lin, & Mushtaq, 2021; Abbad Ur Rehman, Lin, Mushtaq, et al., 2021; Chaubey et al., 2021; Jha et al., 2022) and with ensemble model spesificially bagging proposed Random Forest Alogrithm as the highest Accuracy shows in (Alyas et al., 2022; Chaganti et al., 2022; Idarraga et al., 2021).

Rehman et al. (2021) suggests using machine learning algorithms to detect and diagnose thyroid disease, with emphasis on accuracy and performance evaluation measures. It was found that Naïve Bayes achieved 100% accuracy in all three parts of the experiment(Abbad Ur Rehman, Lin, Mushtaq, et al., 2021). Chaubey et al. (2021) emphasizes the importance of machine learning techniques in healthcare for thyroid disease diagnosis and treatment. Results indicate that the kNN classifier is more effective for this data set. The accuracy calculated here is 96.875% (Chaubey et al., 2021). Rehman et al. (2021) analyzed a thyroid disease dataset from the District Headquarters Teaching Hospital in Dera Ghazi Khan, Pakistan (Teaching Hospital 2020), aiming to determine the most effective KNN distance function used in the research. The study highlighted the benefit of using KNN for its variety of distance functions and adjustable k values, which can be tailored to improve model performance. (Abbad Ur Rehman, Lin, & Mushtaq, 2021). Jha et al. (2022) propose a modelling strategy for predicting thyroid disease, enabling society to benefit from advancements in computer research. The thyroid disease dataset includes 3152 instances, 23 features, and a class that predicts illness status. The K Nearest Neighbor classifier was tested with various K values (3, 5, 7, and 9), and the best results were presented. Finally, the proposed strategies, the first with feature reduction shows an accuracy of 98.7%, and the second with data augmentation technique delivers an accuracy of 99.95%, exceed all the others (Jha et al., 2022).

Alyas et al. (2022) suggests a classification of Thyroid disorders challenging that can be tackled using data mining techniques. The research applies various machine learning algorithms, including decision trees, random forests, KNN, and artificial neural networks, for the classification of thyroid disorders. The random forest method achieved the highest accuracy (94.8%) and specificity (91%) (Alyas et al., 2022). Chaganti et al. (2022) suggests a technique for predicting thyroid disease that incorporates feature selection, machine learning, and deep learning models. Approach by considering feature selection and assessing how well different models perform. Extra tree classifier-based features outperform the RF model with an accuracy of 0.99. (Chaganti et al., 2022). Idarraga et al. (2021) reviewed medical records retrospectively and used scikit-learn to create linear, non-linear, and ensemble models. Machine learning methods, such as Random Forest, may increase the detection of cancer in thyroid nodules and aid in prompt diagnosis and treatment (Idarraga et al., 2021). Wang, et. Al (2024) analyzed a large dataset of PTC patients in West China Hospital. Random Forest shows the best performance to predict recurrence with generally good discrimination, calibration and interpretability in this study (H. Wang et al., 2024). Raj, M.G (2024) demonstrates how these hybrid ML algorithms can efficiently process and interpret metabolomic data, leading to enhanced diagnostic accuracy. Random Forest model achieve accuracy 99.45% (Raj, 2024).

Chen et al. (2020) have primary goal is to identify US characteristics that are strongly linked with malignancy and to build an effective grading system to assist doctors in appropriately recognizing thyroid cancer. The study employed a logistic regression (LR) model with the LASSO (least absolute shrinkage and selection operator) technique to identify and evaluate ultrasound (US) characteristics associated with cancer. However, the high occurrence rate of cancer might influence the precision of predicting non-cancerous nodules, leading to low negative predictive values (NPV) for several classifiers: RF (61.2%), LR (60.0%), SVM (58.2%), NET (60.8%), ELM (52.4%), KNN (51.7%), NB (56.9%), ADAB (62.9%), LOG (56.5%), and LDA (55.6%) (Chen et al., 2020). Ruiz et al. (2020) create a gene panel using a machine learning model to predict lymph node metastasis and recurrence in PTC. Most thyroid cancers are well-differentiated papillary thyroid carcinomas (PTC), which are treated with surgical resections and radioactive iodine therapy. The gene panel demonstrated a sensitivity of 86%, specificity of 62%, positive predictive value of 93%, and negative predictive value of 42% in predicting lymph-node metastases (Ruiz et al., 2020). Ding et al. (2019) aims to address the challenge in accurately predicting event-free survival (EFS) for individuals with early-stage Papillary Thyroid Cancer (PTC) across TNM stages I, II, and III. The effectiveness of the nomogram was evaluated using the concordance index (C-index), which measures the model's ability to distinguish outcomes. When compared to the American Joint Committee on Cancer (AJCC) staging system, the nomogram demonstrated superior performance, with



a C-index of 0.70 (95% CI, 0.64-0.76), markedly outperforming the AJCC's C-index of 0.52, indicating the nomogram's enhanced predictive accuracy for this patient population (Ding et al., 2019).

Table 1. Demonstrates briefly recent studies on thyroid prediction using various datasets with the various accuracy and algorithm proposed

Ref.	Dataset Source	Accuracy	Algorithm
(Abbad Ur Rehman, Lin, Mushtaq, et al., 2021)	The First Affiliated Hospital of Kunming Medical University in China	100%	Naïve Bayes
(Chaganti et al., 2022)	UCI	99%	Random Forest
(Chaubey et al., 2021)	UCI thyroid repository	96.875 %	kNN classifier
(Chen et al., 2020)	The First Affiliated Hospital of Kunming Medical University in China	61.2%	LR combine Random Forest
(Idarraga et al., 2021)	Data FNA cytology	95%	Random Forest
(Abbad Ur Rehman, Lin, & Mushtaq, 2021)	District Headquarters (DHQ) Teaching Hospital	100%	kNN
(Jha et al., 2022)	UCI	99,5%	kNN
(Ruiz et al., 2020)	Carcinoma Genome Atlas database	93%	Kruskal-Wallis tests
(Ding et al., 2019)	Hospital of Zhejiang University	95%	Univariate & multivariate Cox regression analysis
(Alyas et al., 2022)	UCI	94.8%	Random Forest

The gaps in this study compared to previous research such as 1) It is the first to apply the orange data mining software for ensemble stacking and unsupervised clustering with DTC Recurrence datasets; 2) Few studies have focused on optimizing model analysis and classification accuracy for all types of classifications; 3) There is limited research on validating the number of clusters using various clustering techniques on the same dataset.

This study concentrated to explore how accurately it can classify, how many clusters it can identify by using the orange data mining software, and its overall effectiveness with DTC Recurrence datasets in unsupervised machine learning. Given its significance, previous research has explored various methods to diagnose thyroid cancer but has not fully addressed the potential for differentiated thyroid cancer (DTC) recurrence. The goal of this study is to forecast the likelihood of DTC recurrence using patient medical data, while ensuring the protection of personal information, and to identify which factors are most influential in predicting DTC recurrence. To enhance the precision of DTC recurrence forecasts, we introduce a method involving hyperparameter optimization within an ensemble prediction model. Section 3 describes the data exploration and research methodology; Section 4 discusses the findings as result and discussion; and Section 5 concludes the study by summarizing the results and suggesting directions for future research.

2. RESEARCH METHODOLOGY

In this study, a hyperparameter model utilized from various supervised and unsupervised algorithm shown in **Figure 1.** is proposed to predict DTC Recurrence. To evaluate the predictive performance of this model by comparing prediction accuracy five single model and five ensemble model in supervised learning then comparing the number clustering of unsupervised learning. Model framework in this study shown in **Figure 1.**

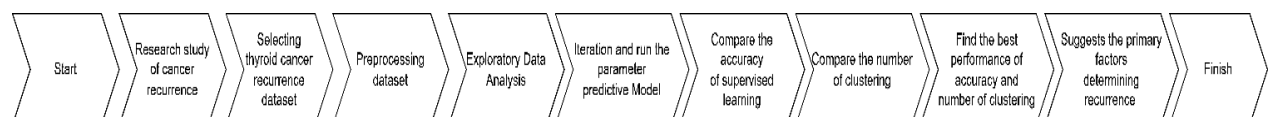


Figure 1. Model framework

2.1 Dataset and Preprocessing

This dataset derived from the UCI Machine Learning Repository, serves as a rich resource for investigating the recurrence of DTC Recurrence. It encompasses 13 clinicopathological attributes, designed to forecast the recurrence of well-differentiated thyroid cancer. The collection of this dataset spanned over 15 years, with a follow-up period of at least 10 years for each patient involved. It encompasses a range of attributes that are pivotal in understanding and modelling the indication of thyroid cancer. The attributes, presented as integer-encoded values for computational



efficiency, cover various clinical and pathological aspects, including patient demographics, medical history, tumour characteristics, and treatment outcomes. It also contains 383 patients with 17 attributes. Preprocessing operation is a crucial step in machine learning, which significantly improves the model performance. In this study, we employ categorical data encoding as part of our preprocessing which is describe in **Table 2**.

Table 2. Attribute Dataset dan Preprocessing.

Attribute	Type Data	Description Attribute	Pre-processing
Age	Integer	The age of the patient	Min = 15, max = 82
Gender	Integer	The gender of the patient	1 = Male, 0 = Female
Smoking	Integer	Indicates whether the patient is a current smoker	1 = Yes, 0 = No
Hx Smoking	Integer	Indicates a history of smoking	1 = Yes, 0 = No
Hx Radiothreapy	Integer	Indicates whether the patient has a history of radiotherapy	1 = Yes, 0 = No
Thyroid Function	Integer	Indicate the condition of the thyroid function	1= Euthyroid, 2= Subclinical Hypothyroidism, 3= Subclinical Hyperthyroidism, 4= Clinical Hypothyroidism, 5= Clinical Hyperthyroidism
Physical Examination	Integer	The findings of the physical examination	1= Normal, 2= Single nodular goiter-left, 3= Single nodular goiter-right, 4= Diffuse goiter, 5= Multinodular goiter
Adenopathy	Integer	Indicates the location and distribution of lymph node	1= No, 2= Left, 3= Right, 4= Bilateral, 5= Posterior, 6= Extensive
Pathology	Integer	Indicates the classifications of thyroid cancer	1= Papillary, 2= Micropapillary, 3= Follicular, 4= Hurthel cell
Focality	Integer	Describes whether the cancer is Uni-Focal or Multi-Focal	1= Uni-Focal, 2= Multi-Focal
Risk	Integer	Indicates the risk level associated with the cancer	1= Low, 2= Intermediate, 3= High
T	Integer	Indicates the size and location of tumor (T)	1= T1a, 2= T1b, 3= T2, 4= T3a, 5= T3b, 6= T4a, 7= T4b
N	Integer	Describe whether the cancer has spread to nearby lymph nodes (N)	1= N0, 2= N1a, 3= N1b
M	Integer	Describe the extent of cancer spread	1= M0, 2= M1
Stage	Integer	Indicates the stage of thyroid cancer	1= I, 2= II, 3= III, 4= IVA, 5= IVB
Response	Integer	Indicates the degree of effectiveness of the treatment on differentiated thyroid cancer	1= Excellent, 2= Indeterminate, 3= Biochemical Incomplete, 4= Structural Incomplete
Recurred	Integer	Indicates whether the cancer has recurred or returned after a period of treatment or remission	1 = Yes, 0 = No

2.2 Exploration Data Analysis

The computing of descriptive statistical measures to examine the attributes of all variables within the dataset. This analysis encompasses the total number of observations (count), the average value (mean), the measure of variability (standard deviation, std), the range of values (minimum and maximum, min/max), and the quartile values (25th/50th/75th percentiles). A segment of the comprehensive dataset is presented in Table 3. Dataset descriptive statistics.

Table 3. Dataset descriptive statistics.

Attribute	Count	Mean	Std	Min	Max	25%	50%	75%
Age	364.0	41.250.000	15.314.360	15.0	82.0	30.0	38.0	52.00
Gender	364.0	0.195055	0.396788	0.0	1.0	0.0	0.0	0.00
Smoking	364.0	0.134615	0.341782	0.0	1.0	0.0	0.0	0.00
Hx Smoking	364.0	0.076923	0.266836	0.0	1.0	0.0	0.0	0.00
Hx Radiotherapy	364.0	0.019231	0.137524	0.0	1.0	0.0	0.0	0.00
Thyroid Function	364.0	1.384.615	1.060.373	1.0	5.0	1.0	1.0	1.00
Physical Examination	364.0	3.480.769	1.262.803	1.0	5.0	2.0	3.0	5.00
Adenopathy	364.0	1.692.308	1.207.801	1.0	6.0	1.0	1.0	2.00
Pathology	364.0	1.442.308	0.855687	1.0	4.0	1.0	1.0	2.00
Focality	364.0	1.373.626	0.484432	1.0	2.0	1.0	1.0	2.00

Attribute	Count	Mean	Std	Min	Max	25%	50%	75%
Risk	364.0	1.456.044	0.651966	1.0	3.0	1.0	1.0	2.00
T	364.0	3.241.758	1.359.312	1.0	7.0	3.0	3.0	4.00
N	364.0	1.571.429	0.870614	1.0	3.0	1.0	1.0	3.00
M	364.0	1.049.451	0.217105	1.0	2.0	1.0	1.0	1.00
Stage	364.0	1.255.495	0.791203	1.0	5.0	1.0	1.0	1.00
Response	364.0	2.043.956	1.258.904	1.0	4.0	1.0	1.0	3.25
Recurred	364.0	0.296703	0.457433	0.0	1.0	0.0	0.0	1.00

Figure 2 provide a correlation matrix heatmap, which visually depicts the relationships between all variables within the dataset. The colour gradient, ranging from grey to intense blue, indicates the strength and direction of the correlation. Black indicates no correlation, orange indicates a positive or direct correlation (where an increase in one variable corresponds with an increase in another), and soft blue indicates a negative or indirect correlation (where an increase in one variable corresponds with a decrease in another).

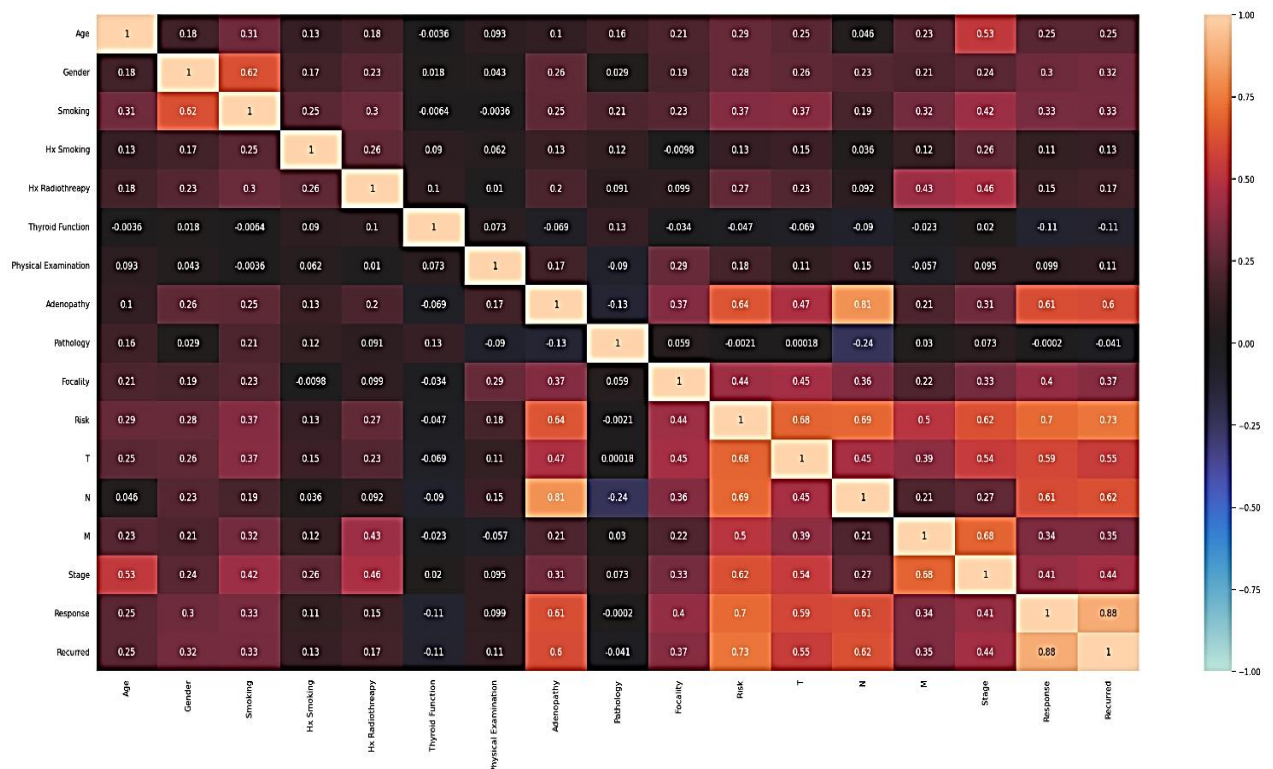


Figure 2. Correlation heatmap.

Based on the heatmap's colour intensity Figure 2, we can discern that variable with high positive correlations (ranging from 0.8 to 1) or high negative correlations (ranging from -1 to -0.2) are significantly related to each other.

1. “Response” and “Recurred” have a strong correlation (0.88): indicating a potentially significant relationship between the response to treatment and the recurrence of the condition.
2. “Adenopathy” exhibits a relatively high correlation with “N” (0.81), showing a strong relationship between the presence of lymph node abnormalities and the 'N' parameter, which typically represents lymph node involvement.
3. “Risk” and “Recurred” with (0.73), indicating a significant relationship between the risk factor and the recurrence of Differentiated thyroid cancer. The high correlation value suggests that as the risk factors increase, there may be a correspondingly higher likelihood of the cancer returning.
4. “Response” and “Risk” has correlation (0.7), this implies that as the risk associated with the cancer increases, the effectiveness of the treatment on differentiated thyroid cancer as indicated by the 'Response' tends to increase as well. Essentially, it indicates that higher risk levels are likely associated with more aggressive or effective treatment responses.

The attributes "Response", "Risk", "Adenopathy", and "N" are the most dominant factors in influencing the recurrence of thyroid cancer. This is further elucidated in Figure 3.

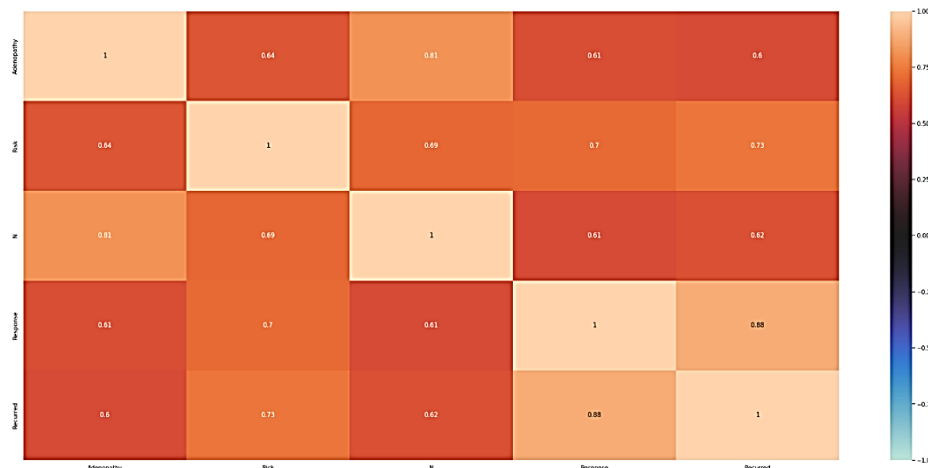


Figure 3. The attributes "Response", "Risk," "Adenopathy", "N" and Recurred.

The following set of visualizations provides an insightful overview of various aspects related to thyroid cancer recurred within a patient cohort. Each subplot of the dominant attribute is tailored to display the distribution of a specific variable, offering a clear depiction of its frequency within the dataset.

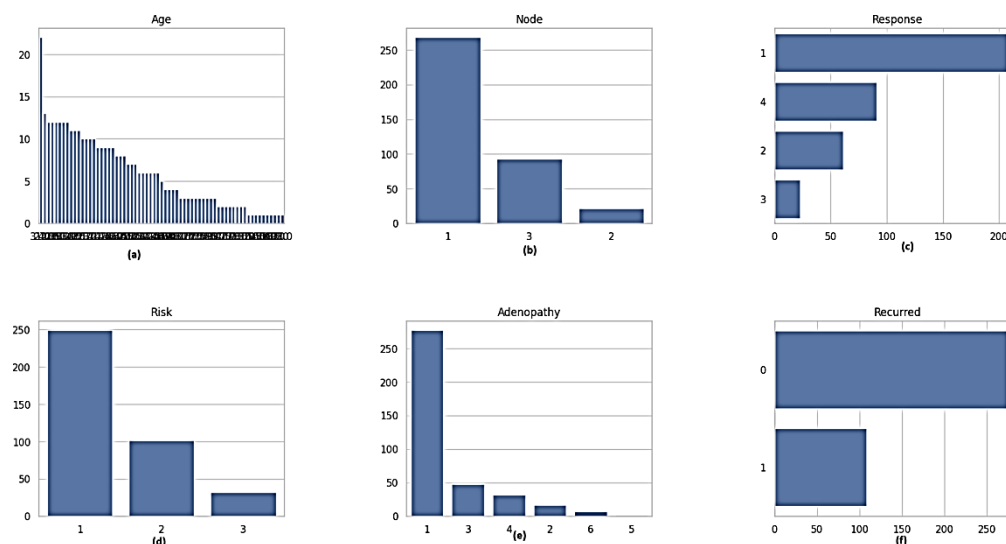


Figure 4. (a) Age, (b) N, (c) Response, (d) Risk, (e) Adenopathy, and (f) Recurred.

- Figure 4a. Age: A bar chart of age presents the age distribution of the patient with its height reflecting 383 patients. The distribution appears to be skewed toward younger patients, with the largest count in the lower age groups.
- Figure 4b. N (Lymph Node Involvement): The chart shows the number of patients with different levels of lymph node involvement. The 'N0' category has the most patients, indicating that most patients fall into this level of nodal involvement.
- Figure 4c. Response: This chart displays the response of patients to treatment, with the number of patients shown for each response category. Categories 'Excellent' and 'Structural Incomplete' appear to be the most common.
- Figure 4d. Risk: This bar chart illustrates the distribution of patients according to their thyroid cancer recurred risk level. Most patients indicate a low risk.
- Figure 4e. Adenopathy: The chart depicts the number of patients with varying conditions of adenopathy. Category 'No' has the highest count, which indicate no distribution of lymph node in most patients.
- Figure 4f. Recurred: The chart illustrates the number of patients who have experienced a recurrence of thyroid cancer after a period of treatment. Most patients fall into category 'No' which indicate they did not have thyroid cancer recurrence.

These graphs collectively serve as a comprehensive dashboard, facilitating a quick and informative evaluation of key data points that are vital for medical prognostic assessment.

2.3 Predictive Model Building

In this investigation, we employed a suite of machine learning algorithms within the Orange Data Mining to construct our predictive models. The methodology involved the utilization of a diverse array of classifiers, single model: LR,

SVM, kNN, ANN, Naïve Bayes, AdaBoost, and ensemble model RF, and SGD, Ensemble1, Ensemble2 and Ensemble3, to evaluate and ascertain the most efficacious model based on performance analysis and classification accuracy metrics. The overarching objective was to conduct a comparative analysis to determine the optimal classifier for addressing the Thyroid Cancer Recurrence. The classifier exhibiting superior performance in terms of classification results was subsequently selected for the prediction phase. Parameters for each predictive model are detailed in Table 4.

Table 4. Parameters of each predictive model for supervised algorithm.

Supervised Algorithm	Abbreviation	Type Model	Model Parameter
Logistic Regression	LR	Single	Regularization type: Lasso (L1) C: 1 Kernel: Polynomial C: 1
Support Vector Machine	SVM	Single	g: Auto d: 3 Iteration Limit: 100 Matric: Euclidean
k-Nearest Neighbor	kNN	Single	Wight: Uniform Number of neighbors: 1 Hidden layers: 10, 20, 30, 40, 50
Artificial Neural Network	ANN	Single	Activation: ReLu Solver: SGD Max Iteration: 300 α : 0.007 Estimators: 50
AdaBoost	AdaBoost	Single	Learning rate: 1 Classification algorithm: SAMME.R Regression loss function: Exponential Growth control split: 5
Random Forest	RF	Bagging	Total each split: 7 Trees: 17 Classification: Hinge Regression: Square Loss Regularization: Lasso (L1) α : 0.00001
Stochastic Gradient Descent	SGD	Boosting	Optimization learning rate: Constant η_0 : 0.01 Number Iteration: 1000 Tolerance: 0.001
Ensemble 1	RF - LR	Stacking	Random Forest – Logistic Regression
Ensemble 2	SGD – LR	Stacking	Stochastic Gradient Descent – Logistic Regression
Ensemble 3	SVM - LR	Stacking	Support Vector Machine – Logistic Regression

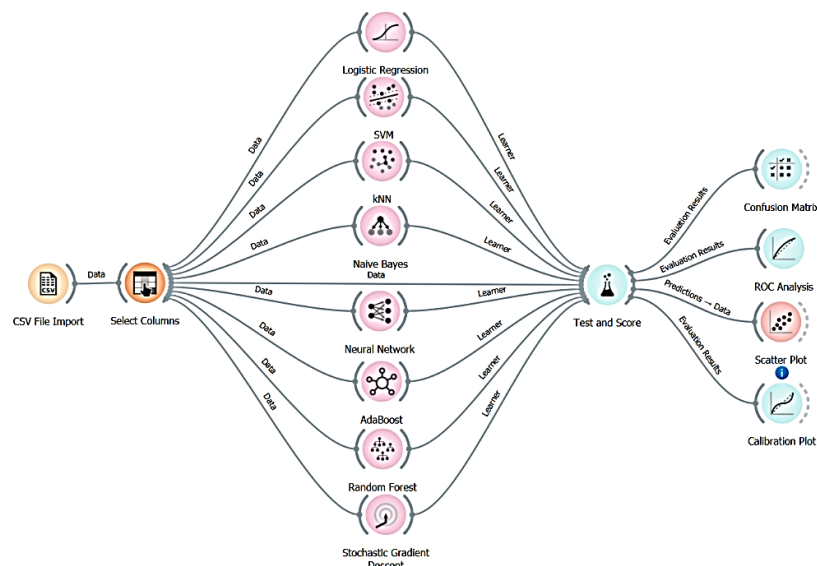


Figure 5. Predictive supervised algorithm

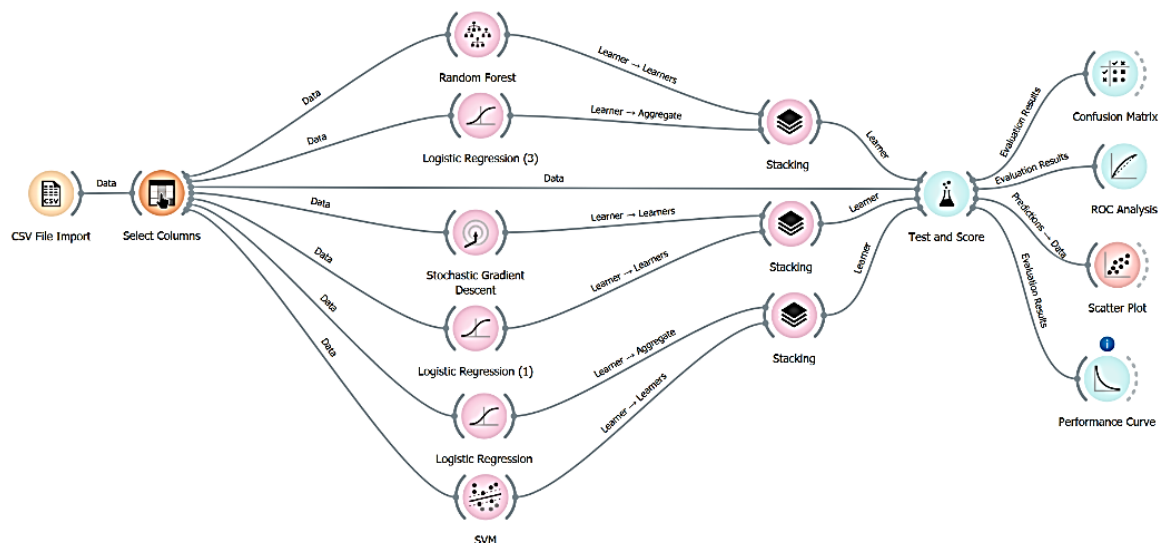


Figure 6. Predictive stacking ensemble learning supervised algorithm

Following the initial analysis, we implemented several unsupervised learning clustering techniques, including k-means, hierarchical, and Louvain clustering, to establish the optimal cluster count within the cohort of Thyroid Cancer Recurrence patients. All analytical processes were conducted using the Orange Data Mining software. The parameter model is shown in Table 5.

Table 5. Parameters of each predictive model for unsupervised algorithm.

Unsupervised Algorithm	Model Parameter
k-Means	Number of Cluster: 2
	Preprocessing: Normalize Columns
	Re-runs: 10
	Max Iteration 300
Hierarchical Clustering	Distance Compare: Rows
	Distance Matric: Euclidean (normalized)
	Bar width: 3
Louvain Clustering	7 Cluster found.
	Preprocessing Normalize data
	PCA Components: 16
	Distance metric: Euclidean
	k-neighbors: 30
	Resolution: 1

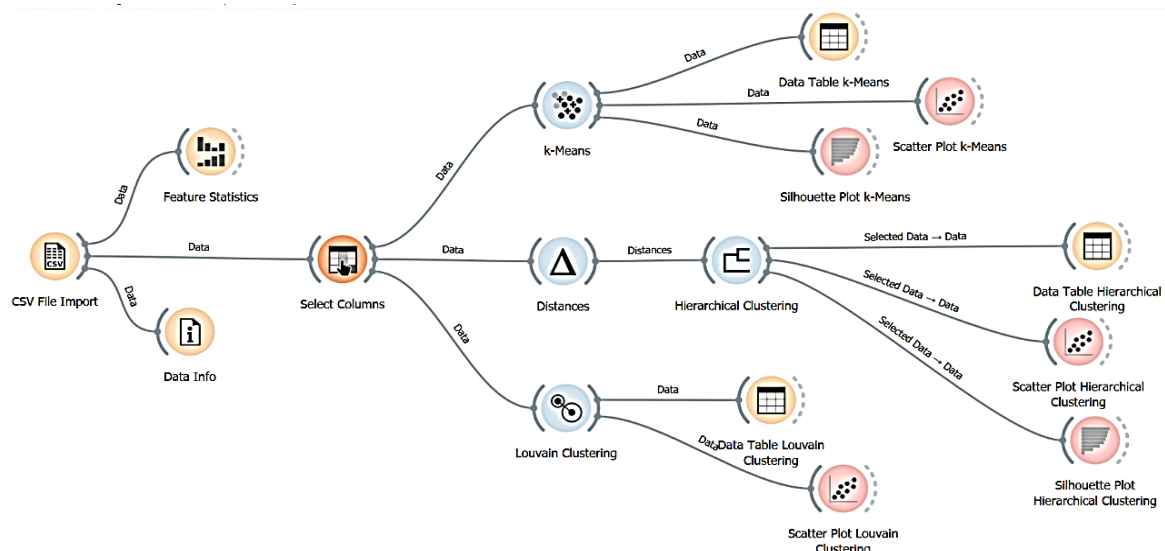


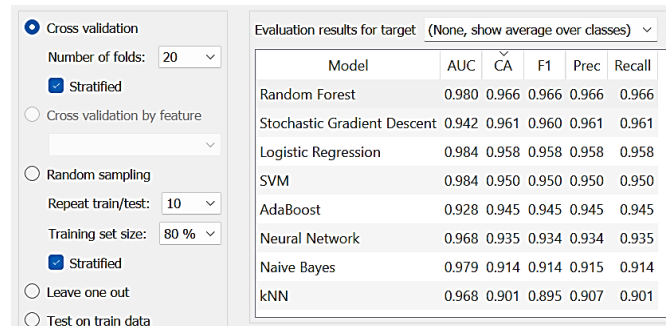
Figure 7. Predictive unsupervised algorithm

3. RESULTS AND DISCUSSION

3.1 Analysis and classification accuracy supervised learning

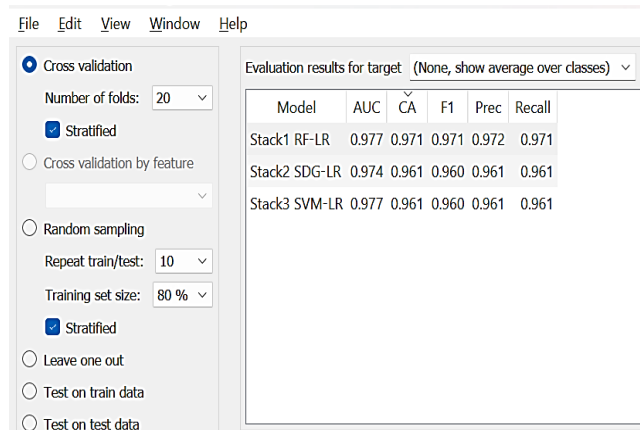
Figures 5 and 6 shows the workflow demonstrates process in machine learning for importing data, selecting features, training various models, and then evaluating and visualizing the performance of those models. Each parameter if of those algorithm model shown in Table 4.

The findings depicted in Figures 8 and 9 demonstrate that Stack-1 and the Random Forest Algorithm emerge as the top-performing techniques for classifying DTC Recurrence predictive data.



Model	AUC	CA	F1	Prec	Recall
Random Forest	0.980	0.966	0.966	0.966	0.966
Stochastic Gradient Descent	0.942	0.961	0.960	0.961	0.961
Logistic Regression	0.984	0.958	0.958	0.958	0.958
SVM	0.984	0.950	0.950	0.950	0.950
AdaBoost	0.928	0.945	0.945	0.945	0.945
Neural Network	0.968	0.935	0.934	0.934	0.935
Naive Bayes	0.979	0.914	0.914	0.915	0.914
kNN	0.968	0.901	0.895	0.907	0.901

Figure 8. Comparative result of predictive supervised algorithm



Model	AUC	CA	F1	Prec	Recall
Stack1 RF-LR	0.977	0.971	0.971	0.972	0.971
Stack2 SDG-LR	0.974	0.961	0.960	0.961	0.961
Stack3 SVM-LR	0.977	0.961	0.960	0.961	0.961

Figure 9. Comparative result of ensemble-stacking algorithm

In the study, ensemble models that combined RF, SGD, SVM, and Logistic Regression showed highest accuracy in predicting the recurrence of DTC compared to single models. Figure 10. illustrates the accuracy scores of each algorithm, with distinct colors for each bar to clearly differentiate them. The ensemble models, particularly Stack1 RF-LR, and Random Forest, performed better than single models, where the accuracy is 97,7% and 96,6%, demonstrating the effectiveness of using multiple algorithms in unison for predictive accuracy in DTC recurrence. As a result of the analysis supervised learning the ensemble model using Random Forest and Logistic Regression model that creates an optimal algorithm showed the best performance of all indicators such as accuracy (0.977) F1-score (0.971) precision (0.972), and recall (0.971). Compared to previous studies using different datasets (Chaubey et al., 2021; Idarraga et al., 2021) this study revealed the highest accuracy using a stacking ensemble algorithm. Then followed by Random Forest, Stack2 SDG-LR, Stack3 SVM-LR also showed high performance Figure 10.

3.2 Analysis total cluster of unsupervised learning

The evaluation of the performance and validity of clustering results using various internal and external validity indices, such as Silhouette Score, or the Adjusted Rand Index. This analysis can help determine the most appropriate number of clusters and the overall quality of the clustering solution (Brun et al., 2007; Kayaalp & Erdogmus, 2020). The workflow illustrates the steps in machine learning for unsupervised learning demonstrated in Figure 7. Table 6 Shows that k-means and Hierarchical clustering result has 2 cluster where the score range almost similar. k-Means algorithm has identified 2 clusters in the DTC Recurrence dataset. The Silhouette score for Cluster 1 (C1) is 0.243, indicating a fair level of separation as Silhouette scores range from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Hierarchical Clustering method also suggests 2 clusters, with Silhouette scores of 0.255 for C1 and -0.105 for C2. Similar to the k-Means results, one cluster seems fairly well defined while the other does not, as indicated by the negative Silhouette score. Louvain Clustering algorithm has the Silhouette scores range from -0.463 to 0.381, indicating a varied degree of fit across different clusters. Conversely,

Clusters C1, C6, and C7 have negative scores, where C1 being particularly low at -0.463, which might indicate that the data points in these clusters are quite dispersed or incorrectly assigned, leading to poor cluster cohesion and separation.

Based on the Silhouette scores provided, both k-Means and Hierarchical Clustering identify only two clusters, where both silhouette scores are also positive for one of the clusters, indicating a better fit for the data points within this cluster. Both k-Means and Hierarchical Clustering resulted in the same number of clusters, which could indicate that there is a natural division of the dataset into two distinct groups for prediction of DTC Recurrence dataset.

Table 6. Comparative result of unsupervised algorithm.

Unsupervised Algorithm	Total Clusters	Distance Matric	Silhouette Score						
			C1	C2	C3	C4	C5	C6	C7
k-Means	2	Euclidean	0.243	-0.102					
Hierarchical Clustering	2	Euclidean	0.255	-0.105					
Louvain Clustering	7	Euclidean	-0.463	0.314	0.327	0.381	0.298	-0.415	-0.237

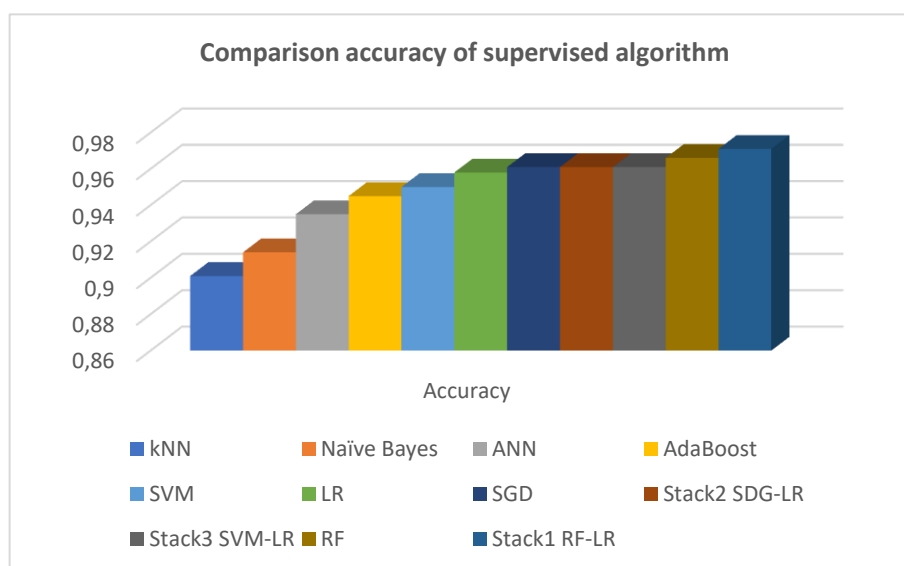


Figure 10. Result comparison accuracy of supervised algorithm

4. CONCLUSION

Prediction of thyroid cancer recurrence is important information for planning post-surgical follow-ups. UCI Machine Learning Repository DTC Recurrence dataset was used to construct the recurrence of Differentiated Thyroid Cancer Recurrence. The dataset contains 383 patients with 17 attributes. A hyperparameter model optimized from various supervised and unsupervised algorithm to predict Differentiated Thyroid Cancer Recurrence. To evaluate the predictive performance of this model by comparing prediction accuracy five single model and five ensemble model in supervised learning then comparing the number clustering of unsupervised learning. LR, SVM, kNN, ANN, Naïve Bayes, AdaBoost as single model and ensemble model RF, and SGD, Stack1 RF-LR, Stack2 SGD-LR and Stack3 SVM-LR were constructed. Subsequently, for number of cluster determination we optimized various unsupervised learning clustering methods, including k-means, hierarchical, and Louvain Clustering. The contribution of each algorithm to DTC Recurrence prediction is the stacking ensemble prediction model showed as the highest performance. The ensemble model using Random Forest and Logistic Regression performed highest accuracy 97,1%. The analysis of clustering methods, notably k-means and hierarchical clustering, suggested that the DTC Recurrence datasets could be segmented into two distinct clusters. Based on the Exploratory Data Analysis (EDA) shows that the most significant attributes or contributors such as "Response", "Risk", "Adenopathy", and "N" as the most dominant factors in influencing the recurrence of thyroid cancer with strong correlation. As the first study to apply a stacking ensemble model to predict DTC Recurrence and optimized various unsupervised learning clustering methods to determine the number of clustering significantly suggest the combination of hyperparameter optimization that shows state-of-the-art performance. Through the identification of accurate models and key factors, enhances the prediction of patients at risk of DTC Recurrence, thereby contributing to more precise and effective patient care strategies. The propose of ideas for further exploration of an unsupervised Machine Learning dataset on DTC Recurrence. Future Research can explore the dimensionality reduction method to investigate the impact of advanced feature engineering techniques, including interaction terms and polynomial features, on clustering performance. Use dimensionality reduction techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), or Uniform Manifold Approximation and Projection (UMAP) to visualize high-dimensional data and potentially improve clustering outcomes.



REFERENCES

- Abbad Ur Rehman, H., Lin, C. Y., & Mushtaq, Z. (2021). Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease. *Journal of the Chinese Institute of Engineers, Transactions of the Chinese Institute of Engineers, Series A*, 44(1), 77–87. <https://doi.org/10.1080/02533839.2020.1831967>
- Abbad Ur Rehman, H., Lin, C. Y., Mushtaq, Z., & Su, S. F. (2021). Performance Analysis of Machine Learning Algorithms for Thyroid Disease. *Arabian Journal for Science and Engineering*, 46(10), 9437–9449. <https://doi.org/10.1007/s13369-020-05206-x>
- Agosto Salgado, S., Kaye, E. R., Sargi, Z., Chung, C. H., & Papaleontiou, M. (2023). Management of Advanced Thyroid Cancer: Overview, Advances, and Opportunities. *American Society of Clinical Oncology Educational Book*, 43, 1–10. https://doi.org/10.1200/edbk_389708
- Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N., & Ahmad, A. (2022). Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach. *BioMed Research International*, 2022. <https://doi.org/10.1155/2022/9809932>
- Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., & Dougherty, E. R. (2007). Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40(3), 807–824. <https://doi.org/10.1016/j.patcog.2006.06.026>
- Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J. L. V., Rodríguez, C. L., & Ashraf, I. (2022). Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. *Cancers*, 14(16), 1–23. <https://doi.org/10.3390/cancers14163914>
- Chan, S., Karamali, K., Kolodziejczyk, A., Oikonomou, G., Watkinson, J., Paleri, V., Nixon, I., & Kim, D. (2020). Systematic Review of Recurrence Rate after Hemithyroidectomy for Low-Risk Well-Differentiated Thyroid Cancer. *European Thyroid Journal*, 9(2), 73–84. <https://doi.org/10.1159/000504961>
- Chatchomchuan, W., Thewjitcharoen, Y., Karndumri, K., Porramatikul, S., Krittiyawong, S., Wanothayaroj, E., Vongterapak, S., Butadej, S., Veerasomboonsin, V., Kanchanapitak, A., Rajatanavin, R., & Himathongkam, T. (2021). Recurrence Factors and Characteristic Trends of Papillary Thyroid Cancer over Three Decades. *International Journal of Endocrinology*, 2021. <https://doi.org/10.1155/2021/9989757>
- Chaubey, G., Bisen, D., Arjaria, S., & Yadav, V. (2021). Thyroid Disease Prediction Using Machine Learning Approaches. *National Academy Science Letters*, 44(3), 233–238. <https://doi.org/10.1007/s40009-020-00979-z>
- Chen, D., Hu, J., Zhu, M., Tang, N., Yang, Y., & Feng, Y. (2020). Diagnosis of thyroid nodules for ultrasonographic characteristics indicative of malignancy using random forest. *BioData Mining*, 13(1), 1–21. <https://doi.org/10.1186/s13040-020-00223-w>
- Davies, L., & Welch, H. G. (2014). Current thyroid cancer trends in the United States. *JAMA Otolaryngology - Head and Neck Surgery*, 140(4), 317–322. <https://doi.org/10.1001/jamaoto.2014.1>
- Ding, Y., Mao, Z., Ruan, J., Su, X., Li, L., Fahey, T. J., Wang, W., & Teng, L. (2019). Nomogram-Based New Recurrence Predicting System in Early-Stage Papillary Thyroid Cancer. *International Journal of Endocrinology*, 2019. <https://doi.org/10.1155/2019/1029092>
- Fengyun Zhang, Jie Geng, De-Gan Zhang, Jinglong Gui, R. S. (2023). Prediction of cancer recurrence based on compact graphs of whole slide images. *Computers in Biology and Medicine*, 167. <https://doi.org/10.1016/j.compbio.2023.107663>
- Haugen, B. R., Alexander, E. K., Bible, K. C., Doherty, G. M., Mandel, S. J., Nikiforov, Y. E., Pacini, F., Randolph, G. W., Sawka, A. M., Schlumberger, M., Schuff, K. G., Sherman, S. I., Sosa, J. A., Steward, D. L., Tuttle, R. M., & Wartofsky, L. (2016). 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid*, 26(1), 1–133. <https://doi.org/10.1089/thy.2015.0020>
- Hong, L., Luo, M., Wang, R., Lu, P., Lu, W., & Lu, L. (2018). Big Data in Health Care: Applications and Challenges. *Data and Information Management*, 2(3), 175–197. <https://doi.org/10.2478/dim-2018-0014>
- Idarraga, A. J., Luong, G., Hsiao, V., & Schneider, D. F. (2021). False Negative Rates in Benign Thyroid Nodule Diagnosis: Machine Learning for Detecting Malignancy. *Journal of Surgical Research*, 268(268), 562–569. <https://doi.org/10.1016/j.jss.2021.06.076>
- Javaid, M., Haleem, A., Pratap Singh, R., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3(February), 58–73. <https://doi.org/10.1016/j.ijin.2022.05.002>
- Jha, R., Bhattacharjee, V., & Mustafi, A. (2022). Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society. *Wireless Personal Communications*, 122(2), 1921–1938. <https://doi.org/10.1007/s11277-021-08974-3>
- Kayaalp, F., & Erdogmus, P. (2020). Benchmarking the Clustering Performances of Evolutionary Algorithms: A Case Study on Varying Data Size. *Irbm*, 41(5), 267–275. <https://doi.org/10.1016/j.irbm.2020.06.002>
- Raj, M. G. (2024). Enhancing Thyroid Cancer Diagnostics Through Hybrid Machine Learning and Metabolomics Approaches. *International Journal of Advanced Computer Science & Applications*, Vol 15, Issue 2, p282. <https://doi.org/10.14569/ijacsa.2024.0150230>
- Ruiz, E. M. L., Niu, T., Zerfaoui, M., Kunnimalaiyaan, M., Friedlander, P. L., Abdel-Mageed, A. B., & Kandil, E.



- (2020). A novel gene panel for prediction of lymph-node metastasis and recurrence in patients with thyroid cancer. *Surgery (United States)*, 167(1), 73–79. <https://doi.org/10.1016/j.surg.2019.06.058>
- Sapuppo, G., Tavaralli, M., Belfiore, A., Vigneri, R., & Pellegriti, G. (2018). Time to Separate Persistent from Recurrent Differentiated Thyroid Cancer: Different Conditions with Different Outcomes. *Journal of Clinical Endocrinology and Metabolism*, 104(2), 258–265. <https://doi.org/10.1210/jc.2018-01383>
- Schlumberger, M., & Leboulleux, S. (2021). Current practice in patients with differentiated thyroid cancer. *Nature Reviews Endocrinology*, 17(3), 176–188. <https://doi.org/10.1038/s41574-020-00448-z>
- Smittenaar, C. R., Petersen, K. A., Stewart, K., & Moitt, N. (2016). Cancer incidence and mortality projections in the UK until 2035. *British Journal of Cancer*, 115(9), 1147–1155. <https://doi.org/10.1038/bjc.2016.304>
- Wang, H., Zhang, C., Li, Q., Tian, T., Huang, R., Qiu, J., & Tian, R. (2024). Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches. *BMC Cancer*, 24(1), 1–12. <https://doi.org/10.1186/s12885-024-12146-4>
- Wang, J., Yu, F., Shang, Y., Ping, Z., & Liu, L. (2020). Thyroid cancer: incidence and mortality trends in China, 2005–2015. *Endocrine*, 68(1), 163–173. <https://doi.org/10.1007/s12020-020-02207-6>
- Wang, L., & Ann Alexander, C. (2013). Applications of Automated Identification Technology in EHR/EMR. *International Journal of Public Health Science (IJPHS)*, 2(3), 109–122. <https://doi.org/10.11591/ijphs.v2i3.3300>