Jurnal Aplikasi Sains Data

Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

JURNAL APLIKASI SAINS DATA

Feature Importance-Guided Ensemble Classification for Predicting Recurrence in Differentiated Thyroid Cancer

Muhammad Ghinan Navsih[1], Hikmata Tartila[2], Wahyu Putra Pratama[3],Dwi Arman Prasetya[4],Tresna Maulana Fahrudin[5]
[1,2,3,4]Data Science Program Study of Universitas Pembangunan Negara "Veteran" Jawa Timur,
[5] Department of Information and Communication Systems, Okayama University.
22083010057@student.upnjatim.ac.id[1], 22083010082@student.upnjatim.ac.id[2], 22083010092@student.upnjatim.ac.id[3],
arman.prasetya.sada@upnjatim.ac.id[4*], tresnamf@s.okayama-u.ac.jp[5]

ABSTRACT

Accurate prediction of cancer recurrence is critical for improving patient monitoring and personalized treatment planning. In this study, we propose a machine learning framework to predict recurrence in patients with differentiated thyroid cancer using statistically selected clinical features. Feature relevance was assessed using ANOVA for ordinal/numerical variables and the Chi-square test for one-hot encoded categorical variables, allowing us to identify the most informative predictors. We then trained three distinct classifiers—Random Forest, Logistic Regression, and XGBoost—and combined them using a hard voting ensemble strategy. The proposed ensemble achieved an accuracy of 98.7% on the test set, with particularly strong precision and recall scores for the recurrent class, indicating its potential clinical utility. Interestingly, all three base classifiers produced identical predictions on the test data, suggesting the dataset's strong internal structure and the effectiveness of our feature selection process. This work highlights the value of integrating statistical feature selection with ensemble modeling for robust and interpretable prediction in clinical oncology applications.

Keywords: Thyroid cancer, recurrence prediction, feature selection, ensemble learning, machine learning

I. INTRODUCTION

Thyroid cancer is one of the most commonly diagnosed cancers affecting the neck region. Among its various forms, differentiated thyroid cancer (DTC) accounts for the majority of cases and is typically associated with favorable long-term outcomes [1]. However, a portion of patients experience cancer recurrence, which requires additional treatment and prolonged monitoring. Early prediction of recurrence is therefore an important step in improving patient management and optimizing healthcare resources.

In recent years, machine learning (ML) has emerged as a powerful tool in medical data analysis, enabling the discovery of patterns that are often difficult to detect using conventional statistical methods [2]. Among various ML techniques, ensemble learning—which combines the predictions of multiple base models—has been shown to improve both accuracy and stability across various domains, including healthcare [3][4].

To ensure that predictive models are both accurate and interpretable, feature selection is a critical preprocessing step. Statistical techniques such as Analysis of Variance (ANOVA) for numerical features and the Chi-square test for categorical features allow us to identify which variables are most strongly

* *Corresponding author.*
E-mail address: arman.prasetya.sada@upnjatim.ac.id

http://dx.doi.org/

associated with the target outcome—in this case, recurrence [5], [6], [7]. Selecting only the most relevant features helps reduce dimensionality and enhances model performance.

In this study, we propose a machine learning framework for predicting recurrence in patients with differentiated thyroid cancer [8]. We apply statistical feature selection to identify significant clinical variables, and then train three classification models—Random Forest, Logistic Regression, and XGBoost. These models are integrated into a hard voting ensemble, where the final prediction is based on majority agreement [9]. Our experiments demonstrate that the ensemble model achieves high predictive accuracy, with all base classifiers showing strong alignment in their predictions [10]. These findings suggest that the selected features carry strong predictive signals and that ensemble learning can provide reliable decision support in clinical prediction tasks.

II. PROCEDURE FOR PAPER SUBMISSION

A. *Data Description*

The dataset used in this study consists of clinical records from 383 patients diagnosed with differentiated thyroid cancer. Each record includes demographic, clinical, and diagnostic variables relevant to recurrence prediction. The dataset was collected from a single source and includes both categorical and numerical features, along with a binary target variable labeled Recurred, indicating whether the patient experienced cancer recurrence after treatment.

Table I presents a complete list of the original features in the dataset along with their descriptions and data types. These attributes served as the basis for further preprocessing and feature selection prior to model training.

Table I. DESCRIPTION OF ORIGINAL DATASET ATTRIBUTES

ID	Attribute	Description	Data Type
1	Age	Age of the patient	Numeric
2	Gender	Biological sex of the patient	Categorical (Binary)
3	Smoking	Whether the patient has a history of smoking	Categorical (Binary)
4	Hx Smoking	Patient’s prior history of smoking	Categorical (Binary)
5	Hx Radiotherapy	History of radiotherapy treatment	Categorical (Binary)
6	Thyroid Function	Thyroid function status	Categorical (Nominal)
7	Physical Examination	Physical examination result	Categorical (Nominal)
8	Adenopathy	Lymph node involvement	Categorical (Nominal)
9	Pathology	Cancer cell type	Categorical (Nominal)
10	Focality	Tumor focality (single/multifocal)	Categorical (Binary)
11	Risk	Risk category (low/intermediate/high)	Ordinal
12	T	Tumor size and extent	Ordinal
13	N	Lymph node spread	Ordinal
14	M	Metastasis presence	Ordinal
15	Stage	Overall cancer stage	Ordinal
16	Response	Post-treatment response category	Ordinal
17	Recurred	Whether the cancer recurred (target variable)	Categorical (Binary)

B. Data Preprocessing

Before applying machine learning models, several preprocessing steps were performed to ensure the dataset was in a suitable format for analysis. First, any missing or inconsistent values were identified and handled appropriately. In this case, no missing values were detected in the dataset, allowing all records to be retained.

Next, all features were transformed into numerical format to meet the input requirements of most machine learning algorithms. This process involved two main strategies:

- Label encoding was applied to ordinal and binary categorical features such as Risk, Stage, T (Tumor size/extent), N (Lymph node involvement), M (Metastasis status), Focality, Gender, Smoking, and Recurred. This preserved the inherent ordering or binary nature of these variables.
- One-hot encoding was used for nominal categorical features without a meaningful order, such as Pathology, Thyroid Function, Physical Examination, and Adenopathy. This created new binary columns representing each unique category.

Finally, all boolean values resulting from one-hot encoding were converted to integers (0 and 1) to ensure compatibility with downstream model training. At the end of preprocessing, the dataset contained only numerical features and was ready for statistical analysis and model training.

C. Feature Selection

To improve model efficiency and interpretability, statistical feature selection was conducted prior to training. Two different methods were used based on the type of feature.For numerical and ordinal features, a one-way Analysis of Variance (ANOVA) test was performed. This statistical method evaluates whether there are significant differences in the mean values of a numerical feature across the two classes of the target variable (Recurred). Features with a p-value less than 0.05 were considered statistically significant and retained for modeling.

Table II. ANOVA TEST RESULTS FOR NUMERICAL AND ORDINAL FEATURES

No.	Feature	p-value	Significant ($p < 0.05$)
1	Response	1.0086×10^{-124}	Yes
2	Risk	7.7234×10^{-66}	Yes
3	N	3.7103×10^{-44}	Yes
4	T	1.7417×10^{-32}	Yes
5	Stage	2.0633×10^{-20}	Yes
6	Focality	6.9033×10^{-16}	Yes
7	M	8.9687×10^{-13}	Yes
8	Smoking	2.1921×10^{-11}	Yes
9	Gender	4.5468×10^{-11}	Yes
10	Age	2.7765×10^{-7}	Yes
11	Hx Radiotherapy	6.0724×10^{-4}	Yes
12	Hx Smoking	7.6596×10^{-3}	Yes

All 12 features tested in the ANOVA procedure—including Response, Risk, T, N, M, and Stage—showed p-values well below the 0.05 threshold. This indicates that these features are statistically associated with cancer recurrence and were retained for modeling.

For categorical features encoded using one-hot encoding, a Chi-square test of independence was used to determine whether the presence or absence of a particular category was significantly associated with recurrence. As with ANOVA, a p-value threshold of 0.05 was applied to identify significant associations.

Table III. CHI-SQUARE TEST RESULTS FOR ONE-HOT ENCODED CATEGORICAL FEATURES

No.	Feature	p-value	Significant ($p < 0.05$)
1	Adenopathy_Bilateral	1.6368×10^{-12}	Yes
2	Adenopathy_No	3.2634×10^{-10}	Yes
3	Adenopathy_Right	1.2815×10^{-7}	Yes
4	Path_Micropapillary	1.4134×10^{-5}	Yes
5	Adenopathy_Extensive	2.4229×10^{-5}	Yes
6	Adenopathy_Left	1.0265×10^{-4}	Yes
7	Exam_Multinodular goiter	1.8673×10^{-2}	Yes
8	Adenopathy_Posterior	2.4028×10^{-2}	Yes
9	Exam_Single nodular goiter-right	3.1097×10^{-2}	Yes
10	Path_Follicular	8.4738×10^{-2}	No
11	Exam_Diffuse goiter	9.7308×10^{-2}	No
12	Thyroid_Subclinical Hypothyroidism	1.6112×10^{-1}	No
13	Thyroid_Clinical Hyperthyroidism	1.8596×10^{-1}	No
14	Path_Papillary	2.3409×10^{-1}	No
15	Thyroid_Clinical Hypothyroidism	3.7460×10^{-1}	No
16	Thyroid_Subclinical Hypothyroidism	5.3199×10^{-1}	No
17	Thyroid_Euthyroid	5.9308×10^{-1}	No
18	Exam_Single nodular goiter-left	8.3147×10^{-1}	No
19	Path_Hürthle cell	8.5789×10^{-1}	No
20	Exam_Normal	9.8250×10^{-1}	No

Out of the 20 one-hot encoded features evaluated using the Chi-square test, 9 features—such as Adenopathy_Bilateral, Adenopathy_Right, and Path_Micropapillary—were identified as statistically significant ($p < 0.05$). These features were retained for use in the final model.

By applying these tests, the dataset was reduced to include only those features that demonstrated a statistically significant relationship with the target variable. This process not only reduced dimensionality but also eliminated noisy or redundant information that could affect model performance. The final set of features used for modeling consisted of both statistically relevant ordinal/numeric variables and one-hot encoded categorical variables with proven significance.

D. Model Design and Ensemble

This study uses three machine learning classification models: Random Forest, Logistic Regression, and XGBoost, which are later combined using a hard voting ensemble method:

- Random Forest: An ensemble of decision trees that operates by constructing multiple trees during training and outputs the mode (majority vote) of their predictions.
- Logistic Regression: A linear model that estimates the probability of a binary outcome based on a weighted combination of input features, using the sigmoid function.
- XGBoost: A gradient boosting algorithm that builds decision trees sequentially, with each new tree aiming to correct the errors made by the previous ones, optimized for speed and performance.

These models were selected based on their strengths in handling classification problems, especially with structured clinical data:

- Random Forest: Known for high accuracy, robustness to overfitting, and effectiveness with both categorical and numerical features.

- Logistic Regression: Interpretable and effective for binary classification tasks, providing a strong baseline with linear assumptions.
- XGBoost: Powerful in capturing complex feature interactions and often outperforms other models in structured datasets due to its boosting mechanism.

To further enhance predictive performance, the three models were combined using a hard voting ensemble strategy. In this method, each base classifier votes for a class label, and the final prediction is made based on the majority vote. Hard voting was chosen for its simplicity and effectiveness, especially when base models are strong and complementary. In this study, all three classifiers achieved similarly high individual performance, and their combined output via hard voting maintained high overall accuracy while reinforcing prediction reliability.

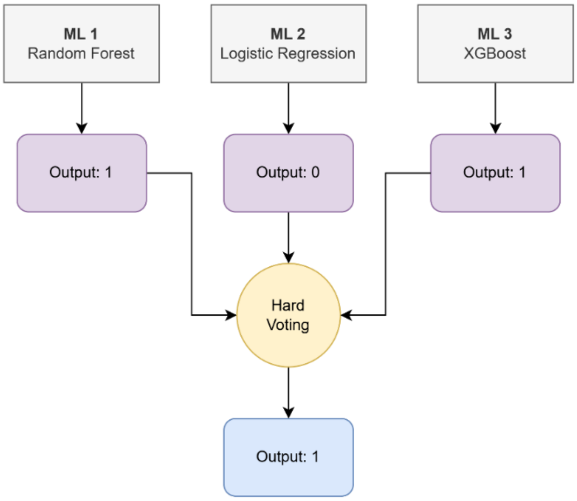


Figure 1. Hard Voting Ensemble Workflow

E. Training and Evaluation

The dataset was split into training and testing sets using an 80:20 ratio to ensure generalizability and prevent data leakage. All three models—Random Forest, Logistic Regression, and XGBoost—were trained using identical train-test splits for consistent evaluation. Only features that were found to be statistically significant (p-value < 0.05) from the ANOVA and Chi-square tests were used as input variables.
Each model was trained using its default hyperparameters, with random_state=42 applied to ensure reproducibility and comparability. After individual training, the models were combined using a hard voting ensemble, where each classifier casts a single vote and the majority vote determines the final prediction.

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score, which are appropriate for imbalanced binary classification problems. In addition to scalar metrics, a confusion matrix was prepared to visualize true positives, true negatives, false positives, and false negatives. These evaluation results are presented and discussed in Section III.

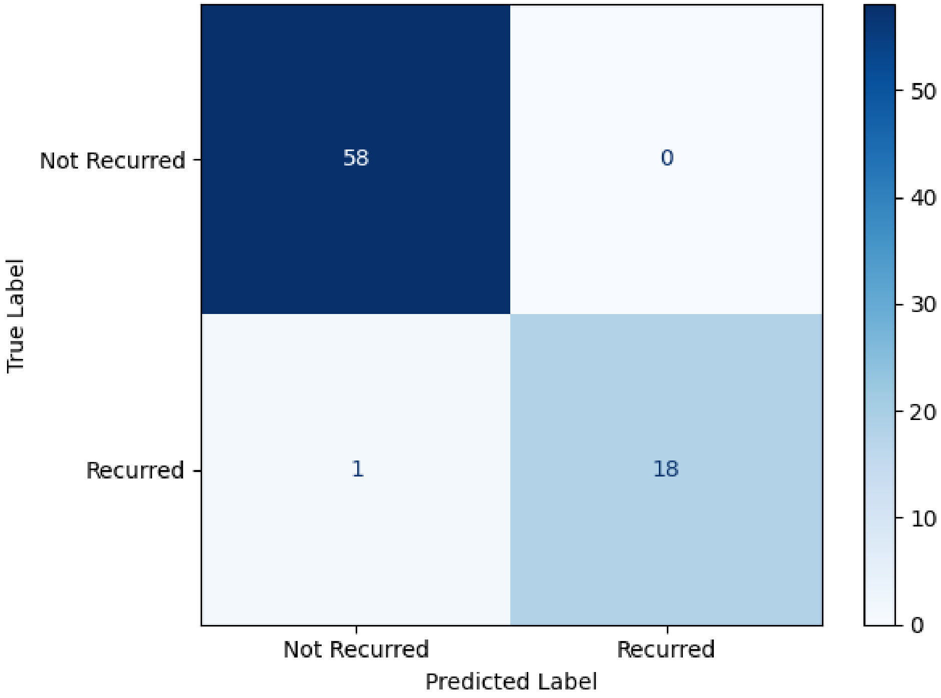
III. RESULT AND DISCUSSION

After training the individual models and combining them through hard voting, the ensemble classifier was evaluated on the test dataset. The performance of the model was assessed using standard classification metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of how well the model performs across both classes—*Recurred* and *Not Recurred*—in a clinical prediction context.

The evaluation results are summarized in **Table IV**. As shown, the ensemble model achieved an overall accuracy of **98.7%**. The precision for predicting recurrence (Class 1) was **1.00**, indicating that all cases predicted as recurrence were indeed correct. The recall for the same class was **94.7%**, showing that the model successfully identified most recurrence cases. For the non-recurrence class (Class 0), the model achieved perfect recall (**1.00**) and high precision (**98.3%**), suggesting excellent performance in minimizing false negatives and false positives.

Table IV. CLASSIFICATION REPORT OF VOTING ENSEMBLE MODEL

Class	Precision	Recall	F1-Score	Support
0 (Not Recurred)	0.983	1.0	0.991	58.0
1 (Reccured)	1.0	0.947	0.973	19.0
accuracy		0.987		77.0
macro avg	0.992	0.974	0.982	77.0
weighted avg	0.987	0.987	0.987	77.0



To further understand how predictions were distributed, a confusion matrix was generated, as shown in Figure 2. Out of 77 total test samples, the model misclassified only one instance—specifically, one recurrence case was predicted as non-recurrence. All other predictions were correct, including 58 non-recurrence cases and 18 recurrence cases. This result demonstrates that the model exhibits both high sensitivity and specificity, which is crucial in medical decision-support systems where both false negatives and false positives carry significant clinical consequences.

CONCLUSION

This study presented a feature importance–guided ensemble learning approach to predict recurrence in patients with differentiated thyroid cancer. Statistically significant features were selected using ANOVA and Chi-square tests, and three classifiers—Random Forest, Logistic Regression, and XGBoost—were combined using a hard voting strategy. The ensemble model achieved an accuracy of **98.7%,** with a precision of **1.00** and recall of **94.7%** for the recurrence class, correctly identifying **18 out of 19** recurrence cases. These results demonstrate strong overall performance and high sensitivity to the minority class, making the model well-suited for supporting clinical decisions. However, the study is limited by its reliance on a single dataset, and the absence of external validation may affect generalizability to other clinical settings or populations.

REFERENCES

[1] American Cancer Society, “What Is Thyroid Cancer?” [Online]. Available: https://www.cancer.org/cancer/thyroid-cancer/about/what-is-thyroid-cancer.html

[2] J. R. Sijmons, L. W. M. P. Timmers, et al., "Management of thyroid cancer: a practical guide for clinicians," *Cancer Treatment Reviews*, vol. 41, no. 6, pp. 501–510, 2015.

[3] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.

[5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 2016, pp. 785–794.

[6] R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936. (Cited for ANOVA)

[7] K. Pearson, “On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that It Can be Reasonably Supposed to Have Arisen from Random Sampling,” *Philosophical Magazine*, vol. 50, no. 302, pp. 157–175, 1900. (Cited for Chi-square test)

[8] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, CRC Press, 2012.

[9] A. Muhaimin, D. D. Prastyo and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 802-809, doi: 10.1109/Confluence51648.2021.9376880.

[10] A. Muhaimin, W. Wibowo, and P. A. Riyantoko, “Multi-label Classification Using Vector Generalized Additive Model via Cross-Validation”, JICT, vol. 22, no. 4, pp. 657–673, Oct. 2023.