The Prediction of Thyroid Cancer Recurrence with the XGBoost Method: The Clinicopathological Feature-Based Approach

Tuti Alawiyah1,*, Taufik Wibisono2, Yani Sri Mulyani3
1,2,3Universitas Bina Sarana Informatika, Tasikmalaya, Indonesia
1tuti.tah@bsi.ac.id, 2taufik.tik@bsi.ac.id, 3yani.ymn@bsi.ac.id

ABSTRACT

This research aims to develop a thyroid cancer recurrence prediction model using the XGBoost method with a clinicopathological feature-based approach. Thyroid cancer is one of the cancers that have a significant recurrence rate after initial treatment. Therefore, thyroid cancer recurrence prediction is important in determining treatment plans and patient management. In this study, we used a dataset containing 383 records of clinicopathological information on thyroid cancer patients who had undergone treatment. The features include various clinical and pathological parameters that are considered important in recurrence prediction. We used the XGBoost algorithm, which has proven effective in various classification tasks, to build a prediction model. The model evaluation results show good consistency in predicting the thyroid cancer recurrence with an average accuracy value of around 97.74% and an average F1-score value of around 95.94%. The results show that the XGBoost model can provide thyroid cancer recurrence prediction with good accuracy, with the ability to effectively detect both classes (recurrence and non-recurrence). The model is expected to be a valuable tool in supporting clinical decision-making related to the management of thyroid cancer patients.

Keywords: Thyroid Cancer; Xgboost; Clinicopathological Feature

1. INTRODUCTION

Thyroid cancer is a type of neoplasm that affects the thyroid gland and occurs due to the uncontrolled proliferation of thyroid cells (Andinata, n.d.). This condition is more common in women (Siswandi et al., 2020) (Nur et al., 2023), who are affected by the menstrual cycle, pregnancy, childbirth, and breastfeeding, which impact the hormone production of the thyroid gland. This gland is responsible for the secretion of thyroid hormones, which have essential functions in the regulation of heart rate, body temperature, and blood pressure. These hormones also influence metabolism and various other biological functions, so disruptions to the thyroid gland can have serious health implications. Thyroid cancer is one type of endocrine cancer whose prevalence has increased significantly in recent decades. Several risk factors that are thought to influence the prevalence of thyroid cancer include radiation exposure, genetic mutations, heredity, iodine deficiency, elevated TSH levels, autoimmune conditions, the presence of thyroid nodules, environmental factors, lifestyle, and estrogen levels (Dwi Arista et al., 2023). Other studies mention the female gender, living in mountainous areas and aged adults to the elderly are the most sufferers of thyroid cancer (Parura et al., 2016). Vitamin D deficiency is one of the risk factors for the malignancy of this cancer (Bogar et al., 2024). Sri Hartanti in her research to classify the level of malignancy of thyroid cancer using the classification and regression tree method states that the results of clinical examination are the most influential factor (Handayani et al., 2014).

Although thyroid cancer cure rates are generally high, the risk of relapse remains a serious clinical challenge. Thyroid cancer relapse can occur several years after initial treatment, requiring close monitoring and accurate prediction strategies to identify patients at high risk of relapse. In the context of thyroid cancer relapse prediction, the patient's medical history is an invaluable source of data. This data includes information on clinical factors, laboratory test results, diagnostic procedures and response to therapy. Using this data effectively to predict the likelihood of recurrence requires a sophisticated and precise analytical approach. The implementation of XGBoost in thyroid cancer relapse prediction based on patient medical history offers great potential to improve the quality of clinical management and decision-making. Through proper modeling, accurate predictions can be generated, assisting clinicians in designing better and personalized treatment strategies for patients at high risk of relapse.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).
This study aims to develop and evaluate a thyroid cancer relapse prediction model using the XGBoost algorithm, as well as to identify key factors that contribute to the risk of relapse. This research focuses not only on developing a predictive model, but also on an in-depth analysis of relevant clinical variables, so as to provide better insights for the medical community in their efforts to address the challenges posed by thyroid cancer recurrence. With this approach, it is hoped to create a reliable and applicable predictive tool in clinical practice, so as to reduce the burden of disease and improve the prognosis of patients with thyroid cancer.

The prediction of thyroid cancer recurrence has a significant impact on clinical management. Through more personalized care, better management of resources, improvement of the patient quality of life, development of new treatment strategies, and the reduction of patient stress, these predictions are becoming an important tool in efforts to provide better and more effective care for thyroid cancer patients. With accurate predictions, doctors can customize a specific treatment plan for each patient. Patients at high risk of recurrence may be given more intensive treatment and closer monitoring, while patients at low risk may require a lighter approach. This allows for more effective and efficient treatment, avoiding over-treatment or under-treatment. By continuing to develop and refine prediction models, we can hope for better outcomes in thyroid cancer management in the future.

Approaches to predicting thyroid cancer recurrence have evolved from traditional clinical and epidemiological analysis to the use of some machine learning technologies and some multi-omics data integration. This modern approach utilizing the XGBoost method allows for more accurate predictions and more personalized treatment, which in turn can improve clinical outcomes and the patient quality of life. Integration of these models in medical practice can provide more accurate predictions and better personalization of treatment, which is expected to continue to develop and make significant contributions at the management of thyroid cancer in the future.

2. LITERATURE REVIEW

Thyroid cancer is a type of cancer that affects the thyroid gland, which produces thyroid hormones that regulate body temperature, blood pressure and heart rate. Although thyroid cancer is less common than cancers of other organs in its early stages, it usually does not cause any distinctive symptoms. However, if the thyroid gland becomes enlarged, it may cause symptoms such as a lump at the front of the neck and difficulty swallowing. The exact cause of thyroid cancer is not yet known, but thyroid cancer patients are generally found with clinicopathological features including being between 55 and 64 years old, female, without a family history of cancer, having a normal BMI range, having a single nodule with a size of more than 4 cm, solid, papillary, and most undergo surgery in the form of total thyroidectomy (Fachri, 2024). Based on its severity and stage of development, thyroid cancer is classified into four stages based on the TNM classification (tumor, nodule, and metastasis).

XGBoost (Extreme Gradient Boosting) has proven to be some of the most effective machine learning algorithms in various medical prediction applications. It is known for its ability to manage high-dimensional data and handle class imbalance often found in medical data. By utilizing boosting techniques, XGBoost is able to improve prediction accuracy by combining the strengths of many simple predictive models to form a robust and reliable model. Many researchers have used the XGBoost algorithm in classification and prediction in the medical field. The accuracy result in predicting diabetes using XGBoost is 76% (Salsabil et al., 2024) and the classification of Parkinson's disease using XGBoost optimized with PSO gets an AUC value of 0.9483 (Kurnia et al., 2023). Another study used XGBoost for air quality prediction and obtained an AUC-ROC value of 99.48% (Nababan et al., 2023). In another study, XGBoost produced an accuracy of 80.039%, a precision of 81.338%, and a recall value of 96.854% (Yulianti et al., 2022). From several studies, it appears that the XGBoost algorithm obtained the highest average value in making predictions (Nugraha, 2021) (Andriansyah & Fridayanthie, 2023; Ravly Andryan et al., 2022).

3. METHOD

This study aims to develop a well-differentiated thyroid cancer recurrence prediction model using the XGBoost algorithm. The dataset used was obtained from the UCI Machine Learning Repository containing 13 clinicopathological features collected over a 15-year period, with each patient followed for a minimum of 10 years. The stages carried out in this study include understanding the problem, collecting data, data cleaning, exploratory data analysis (EDA), data preprocessing, data modeling, model evaluation.

At the stage of understanding the problem, researchers identify the problem to be solved and determine the purpose of this research. The researcher found a problem in predicting thyroid cancer recurrence after initial treatment. Although medical treatments have made significant progress, thyroid cancer recurrence remains a serious problem.

* Corresponding author
This recurrence can lead to a decrease in the patient's quality of life, increased healthcare costs, and lower survival rates. Without accurate predictions, it is difficult for doctors to design effective and personalized treatment plans for each patient, which could lead to the need for more intensive monitoring or unnecessary medical interventions. This study aimed to develop a thyroid cancer recurrence prediction model using the XGBoost algorithm. Researchers analyzed clinical and pathology data to identify key risk factors that contribute to thyroid cancer recurrence.

Data was collected from the UCI Machine Learning Repository, which includes clinicopathological information from patients with well-differentiated thyroid cancer. Features collected include (1) patients age at the time of diagnosis or treatment, (2) patient gender, (3) patient smoking status, (4) patient smoking history, (5) radiotherapy treatment history, (6) thyroid function status, (7) patient physical examination findings, including palpation of the thyroid gland and surrounding structures, (8) presence of enlarged lymph nodes in the neck region, (9) types of thyroid cancer determined by pathology examination of biopsy samples, (10) focialities, (11) cancer risk category based on various factors, such as tumor size, the extent of spread, and histological type, (12) tumor classification, (13) nodule classification, (14) metastasis classification, (15) cancer stage, (16) response to treatment, (17) cancer recurrence after initial treatment.

Data cleaning takes place during this stage. Data cleaning is the process of finding and repairing flaws to make the data clean and of high quality, such as dealing with empty, inconsistent, or duplicate data. Mean, median, and standard deviation are examples of descriptive statistics used in exploratory data analysis to understand data distribution. Furthermore, data exploration has been conducted by examining the distribution, trends, and anomalies in the data.

Data transformations such as normalization or standardization are conducted during the data preprocessing stage to guarantee that the data is on a consistent scale, and encoding is used to turn categorical data into a numerical format that modeling algorithms can use. In addition, the data set has been divided into three parts: training, validation, and testing. This was done to ensure that the model being created could be appropriately evaluated while avoiding overfitting. Thus, data pretreatment guarantees that the data used in the following stage is clean, structured, and ready for detailed analysis and predictive modeling.

Data modeling is done using the XGBoost algorithm. Before conducting training and testing, the dataset is divided into train data and test data with a proportion of 72:28. This proportion is done according to the proportion of the label. The model's performance will be evaluated using metrics such as accuracy, precision, and recall to measure the model's ability to predict thyroid cancer recurrence.

### 4. RESULT

**Data Cleaning**

The data used comes from the UCI machine learning repositories at link [https://www.kaggle.com/datasets/jainaru/thyroid-disease-data](https://www.kaggle.com/datasets/jainaru/thyroid-disease-data) This dataset included 13 clinicopathological features used to predict the recurrence of well-differentiated thyroid cancer as seen in table 1. The data was collected over 15 years, with each patient monitored for a minimum of 10 years for a total of 383 records.

<table>
<thead>
<tr>
<th>Table 1. Thyroid Disease Data</th>
</tr>
</thead>
</table>

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).
Checking the data content of each feature at this stage. Checking whether any fields are not filled in (missing value) at this stage. No missing values are seen in all fields in each record as shown in Figure 1. Checking and removing duplicated data. The resulting amount of data becomes 348 records after deleting duplicate data, with the number of unique data points from each field shown in Figure 2.

Fig. 1 Missing Value Check Results
EXPLORATORY DATA ANALYSIS (EDA)

This stage is done by visualizing the distribution, patterns, and anomalies in the data for each feature to see whether the feature has an effect on thyroid cancer recurrence or not. Figure 3 shows the results of visualizing the smoking feature on thyroid cancer recurrence. As seen in Figure 3, the smoking feature is very influential on recurred. From the visualization results, the values of adenopathy can be simplified into yes and no, because the content of the no data indicates a high probability of not recurring and the other indicates a high probability of recurrence.

Fig. 2 Checking and removing duplicated data

Fig. 3 Visualization of Smoking Features against Recurred
Figure 4 shows the distribution of thyroid cancer patients based on adenopathy features and recurrence status. From this bar chart, it can be seen that patients without adenopathy (category "No") have a lower probability of recurrence. Patients with right-sided and bilateral adenopathy have a higher risk of recurrence. Other categories such as "Left", "Extensive", and "Posterior" also showcases recurrence, but in smaller numbers. These data indicate that the presence of adenopathy, especially right-sided and bilateral, may be associated with an increased risk of thyroid cancer recurrence.

**PREPROCESSING DATA**

At this stage, categorical data is converted into numerical form using encoding techniques. Some features with categorical data are changed using the label encoding technique and some use the ordinal encoding technique. The label encoding technique is used on features that do not consider the order of some data such as Gender, Smoking, Hx Smoking, Hx radiotherapy, and Recurred features. While features with other category data are converted into numeric form using the ordinal encoding technique because this change in data form considers the order of the category. These features consist of Thyroid Function, Physical Examination, Adenopathy, Pathology, Focality, Risk, T, N, M, Stage, and Response features. The results of this data transformation can be seen in Table 2.

**Table 2**

Data Transformation Results

<table>
<thead>
<tr>
<th>Index</th>
<th>Age</th>
<th>Gender</th>
<th>Smoking</th>
<th>Hx Smoking</th>
<th>Hx Radiotherapy</th>
<th>Thyroid Function</th>
<th>Physical Examination</th>
<th>Adenopathy</th>
<th>Pathology</th>
<th>Focality</th>
<th>Risk</th>
<th>T</th>
<th>N</th>
<th>M</th>
<th>Stage</th>
<th>Response</th>
<th>Recurred</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>2.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.0</td>
<td>3.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>62</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>3.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>378</td>
<td>72</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0</td>
<td>1.0</td>
<td>3.0</td>
<td>1.0</td>
<td>0.0</td>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.0</td>
<td>2.0</td>
</tr>
<tr>
<td>379</td>
<td>61</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0</td>
<td>3.0</td>
<td>5.0</td>
<td>1.0</td>
<td>0.0</td>
<td>2.0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).
At this stage, the Relationship Analysis between Variables is also carried out, calculating the correlation between numerical variables to see the linear relationship as shown in Figure 5. Based on the results of the correlation analysis as shown in figure 5, the Response and Risk variables have a strong correlation with Recurrence.

![Fig. 5 Result of Analysis of Data Correlation Value against Recurred](image-url)

### Table: Correlation Analysis Results

<table>
<thead>
<tr>
<th>Index</th>
<th>Ag</th>
<th>Gender</th>
<th>Smoking</th>
<th>Hx Smoking</th>
<th>Hx Radiotherapy</th>
<th>Thyroid Function</th>
<th>Physical Examination</th>
<th>Adenopathy</th>
<th>Pathology</th>
<th>Focality</th>
<th>Risk</th>
<th>T</th>
<th>N</th>
<th>M</th>
<th>Stage</th>
<th>Response</th>
<th>Recurred</th>
</tr>
</thead>
<tbody>
<tr>
<td>380</td>
<td>72</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.0</td>
<td>3.0</td>
<td>4.0</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>4.0</td>
<td>3.0</td>
</tr>
<tr>
<td>381</td>
<td>61</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4.0</td>
<td>3.0</td>
<td>5.0</td>
<td>3.0</td>
<td>1.0</td>
<td>2.0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>382</td>
<td>67</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>3.0</td>
<td>4.0</td>
<td>1.0</td>
<td>2.0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).
Figure 6 displays that feature Risk, N (Nodal involvement), adenopathy, and response are likely to be statistically significant to Recurrence. Based on the interpretation of the graphs in Figure 5 and Figure 6, the variable that has the weakest and least significant correlation consists of Physical Examination: Very weak correlation (0.0981469) and largest p-value; Thyroid Function: Very weak and negative correlation (-0.1176354) and large p-value; Pathology: Very weak correlation (0.1517054) and large p-value; and Hx Radiotherapy: Very weak correlation (0.1693404) and large p-value. These variables may not provide much valuable information about thyroid disease recurrence and could be considered to be ignored or analyzed further with a different approach. This study will not use these four variables to model the data.

Next, perform normalization or standardization to ensure that the data is on a consistent scale. The last step in data preprocessing is the separation of the dataset into a training set and a test set with a proportion of 70% of the data will be used for training and 30% of the data will be used for testing. This is done to ensure that the model to be built can be evaluated accurately and avoid overfitting. This is done to ensure that the data used in the next stage is clean, structured, and ready for in-depth analysis and predictive modeling.

DATA MODELING

In the modeling stage of this research, we used the XGBoost algorithm to develop the classification model. We started by importing the XGBClassifier library from the xgboost package. This library provides an easy-to-use implementation of the XGBoost model in Python. We then created an instance of XGBClassifier. This model uses the default parameters provided by the XGBoost library, which has been optimized for many common cases. However, these parameters can be further customized based on the specific needs and characteristics of the dataset. With the model initialized, the next step is to train the model using the training data. The fit method in XGBClassifier is used to fit the model to the training data (X_train, y_train). This training process involves the use of a boosting algorithm to build a robust ensemble of decision trees, capable of making more accurate predictions by iteratively reducing the error.
MODEL EVALUATION
After the model is trained, we conduct a preliminary evaluation to understand the performance of the model. The test data is used to predict outcomes, and evaluation metrics such as confusion matrix and classification report are used to measure model performance. These results provide insight into the model's accuracy, precision, recall, and F1 score, which are important for understanding the model's effectiveness in a given classification task. The evaluation results of the model built using the Cross-Validation (CV) approach are shown in Table 3 while Figure 7 shows the model evaluation results using the confusion matrix and classification report. Overall, based on the data shown in Table 3, through cross-validation evaluation, the mean value and standard deviation for each evaluation metric such as precision, recall, and f1-score were obtained. The model shows good consistency in predicting a thyroid cancer recurrence with an average accuracy value of about 97.74% and an average F1-score value of about 95.94%. Figure 7 shows the evaluation using a confusion matrix and the classification report on each fold shows a consistent pattern in the model performance. This shows that the model has a stable ability to classify thyroid cancer recurrence.

The model performed well in detecting both positive (thyroid cancer recurrence) and negative (no thyroid cancer recurrence) classes. it shows the high precision, recall, and f1-score values for both classes in both evaluation approaches. The confusion matrix results show that the model tends to create fewer false positives and false negatives, indicating that the model has a good ability to avoid adverse prediction errors.

Table 3
Evaluation Results Using the Cross-Validation Approach

<table>
<thead>
<tr>
<th>Iterasi</th>
<th>tn</th>
<th>Fp</th>
<th>Fn</th>
<th>tp</th>
<th>prec 1</th>
<th>prec 0</th>
<th>recall 1</th>
<th>recall 0</th>
<th>Accuracy</th>
<th>f1_score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>59</td>
<td>1</td>
<td>1</td>
<td>26</td>
<td>0.962963</td>
<td>0.983333</td>
<td>0.962963</td>
<td>0.983333</td>
<td>0.977011</td>
<td>0.962963</td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>0</td>
<td>1</td>
<td>26</td>
<td>1.000000</td>
<td>0.983607</td>
<td>0.962963</td>
<td>1.000000</td>
<td>0.988506</td>
<td>0.981132</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>4</td>
<td>1</td>
<td>26</td>
<td>0.866667</td>
<td>0.982456</td>
<td>0.962963</td>
<td>0.933333</td>
<td>0.942529</td>
<td>0.912281</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>0</td>
<td>1</td>
<td>26</td>
<td>1.000000</td>
<td>0.983607</td>
<td>0.962963</td>
<td>1.000000</td>
<td>0.988506</td>
<td>0.981132</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.957407</td>
<td>0.983251</td>
<td>0.962963</td>
<td>0.979167</td>
<td>0.974138</td>
<td>0.959377</td>
</tr>
<tr>
<td>Std.dev</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.054528</td>
<td>0.000472</td>
<td>0.000000</td>
<td>0.027323</td>
<td>0.018843</td>
<td>0.028185</td>
</tr>
</tbody>
</table>

Fig. 7 Model Evaluation Results Using Confusion Matrix and Classification Report

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).
5. CONCLUSION

Based on the evaluation results, the XGBoost model built with a clinicopathological feature-based approach can provide excellent thyroid cancer recurrence prediction. The model can be a useful tool in assisting clinicians and healthcare professionals in making more accurate clinical decisions related to the management of thyroid cancer recurrence. However, it should be noted that further evaluation and external validation may be required to ensure the reliability and generalisability of the model in different clinical scenarios.

These findings provide a strong basis for continued research and development of more sophisticated prediction models integrated with multi-omics data and other AI technologies. To increase the validity and reliability of the developed prediction model, future research can carry out external validation with datasets from various health centers and different populations to ensure the generalization of the model in various clinical conditions, adding other features such as genetic data, imaging data, and lifestyle information patients to increase the accuracy and richness of the information used in the model as well as integrate the prediction model with the hospital and clinic information systems to facilitate its use in daily practice by doctors and medical teams.

6. REFERENCES


Medika Malahayati, 4(3), 244–248.