Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.14334 p-ISSN : 2541-044X

# Thyroid Disease Prediction Using Random Forest with KNNImputer for Missing Values

**Raffy Nicandra Putra Pratama[1)*], Sri Winarno[2)], Tan Nicholas Octavian Wijaya[3)]**
[1)2)3)]Information System, Dian Nuswantoro University, Semarang, Indonesia
[1)]112202106602@mhs.dinus.ac.id, [2)]sri.winarno@dsn.dinus.ac.id, [3)] 112202106556@mhs.dinus.ac.id

**Submitted** :Dec 13, 2024 | **Accepted** : Jan 2, 2024 | **Published** : Jan 8, 2025

**Abstract:**Thyroid disease is a health dysfunction that requires immediate and accurate diagnosis. This research seeks to design a classification model based on the Random Forest algorithm to detect the type of thyroid disease utilizing data from the UCI Repository. In the data processing stage, KNNImputer is used to handle missing data by calculating the average value of the nearest neighbors based on Euclidean distance, thus ensuring better data quality for model training. The developed model was evaluated utilizing the confusion matrix, which showed an accuracy of 98%, with precision, recall, and F1 score values reached 98% based on weighted avg.These results corroborate that the proposed model is highly reliable in detecting various types of thyroid diseases, such as Negative, Hypothyroid, and Hyperthyroid. This research makes an important contribution to the application of data mining technology for medical diagnosis, while proving that optimal data processing through methods such as KNN Imputer can significantly improve model performance.

**Keywords:** Thyroid, Classification, Random Forest, KNNImputer

## INTRODUCTION

An illness is an abnormal state of the body or mind that leads to distress and impaired function. Busyness in daily activities often makes us neglect our health, which can trigger various diseases, including thyroid disease(Khosravİ et al., 2015).

Thyroid disease is a dysfunction of the thyroid gland, an endocrine gland in the neck that produces important hormones such as thyroxine (T4) and triiodothyronine (T3). These hormones control metabolic processes, growth, development, and the immune system. Thyroid gland malfunction can lead to a variety of conditions, including goiter, hypothyroidism (low hormone production), and hyperthyroidism (excess hormone production). Thyroid disease is often associated with iodine deficiency and can affect fetal development and pregnancy The thyroid gland is controlled by Thyroid Stimulating Hormone (TSH) from the pituitary gland, which stimulates the release of T4 and T3 into the blood. TSH levels in the blood are used as a key indicator to assess thyroid function and help diagnose disorders such as hypothyroidism and hyperthyroidism. Hyperthyroidism is characterized by the overproduction of thyroid hormones, leading to increased levels of T3 and decreased TSH. This condition is diagnosed through laboratory tests that show serum TSH levels below normal and fT4 more than 24.5 pmol/l or fT3 more than 6.3 pmol/l. In contrast, Hypothyroidism is a disorder where the thyroid gland produces insufficient amounts of hormones, lowering blood levels of fT4 and fT3(Yurizali & Adhyka, 2024).

Early detection and accurate diagnosis are essential to prevent complications. Data mining is a set of systematic processes that aim to find hidden patterns or relationships in data that are difficult to identify manually. This process enables in-depth data analysis to uncover previously unknown information, resulting in new insights and useful knowledge. Through the application of data mining, various patterns related to health and disease diagnosis can be identified, which ultimately helps in making more effective and targeted medical decisions(Handayani et al., 2019).

This study employs data mining techniques in the form of classification, aiming to analyze datasets to generate rules for classifying or identifying new, previously unexamined data, as well as grouping data into specific categories. The learning process requires input data in the form of training data with class attributes and produces output in the form of a classification model (Ginantra et al., 2021).

One of the effective methods in medical data classification is the Random Forest algorithm. This algorithm is well known for its expertise in producing reliable, accurate, consistent predictions and its ability to avoid overfitting. By using an ensemble-based approach that combines many decision trees, Random Forest effectively

*name of corresponding author

suppresses the risk of overfitting and optimizes the stability and accuracy of prediction results. This makes Random Forest a good choice for processing thyroid disease data that has many features or variables (Hengl et al., 2018).

This study utilizes the thyroid disease dataset from the UCI Machine Learning Repository, which offers comprehensive data for analyzing various thyroid conditions, including hypothyroidism and hyperthyroidism. One of the challenges in medical datasets is the presence of missing values that can affect the accuracy of the model. Therefore, the K-Nearest Neighbors Imputation (KNNImputation) method is used to handle missing data, as it is able to maintain data distribution by considering the proximity between data(Supardianto et al., 2022).

There have been many studies utilizing the Random Forest algorithm, including studies focusing on predicting the likelihood of diabetes at an early stage. For example, a study that applied this algorithm managed to show excellent performance, with an accuracy rate of 97.88%. This result confirms that the Random Forest algorithm can be an effective method for predicting health conditions with high accuracy (Apriliah et al., 2021).

Furthermore, the research of Umri, Ahmadi, & Kamil, conducted a comparison of the K-Nearest Neighbor and Random Forest Methods in Predicting the Accuracy of Wart Disease Treatment Classification. Test results using the K-Nearest Neighbor method obtained an accuracy rate of 76.78%, and then the Random Forest method obtained an accuracy rate of 86.56%. These results show that Random forest is better than KNN (Erdiansyah et al., 2022). In Listiana & Zailani's research, in Determining the Feasibility of Lending at Mitra Sejahtera Cooperative using Random forest results in the random forest algorithm was more accurate than the analysis conducted by credit analysts with the results of evaluating the accuracy value of 87.88% (Zailani & Hanun, 2020). In research that uses KNNimputation to overcome missing values, KNNimputation is applied to improve SVM performance. The results showed that the use of KNNImputer successfully handled data that had missing values, thus improving SVM classification performance with evaluation results reaching 93% (Supardianto et al., 2022). In Widianti & Pratama's research, KNN-Imputer showed the best performance among all imputation methods (Widianti & Pratama, 2024).

By combining the Random Forest algorithm for classification and KNNImputation to address missing values, this research is expected to produce an accurate and effective model. This model aims to help medical professionals diagnose thyroid diseases faster and reduce the risk of misdiagnosis.

## LITERATURE REVIEW

### Thyroid Disease

Thyroid disease is a dysfunction of the thyroid gland, an endocrine gland in the neck that produces important hormones such as thyroxine (T4) and triiodothyronine (T3). These hormones control metabolic processes, growth, development, and the immune system. Thyroid gland malfunction can lead to a variety of conditions, including goiter, hypothyroidism (low hormone production), and hyperthyroidism (excess hormone production). Thyroid disease is often associated with iodine deficiency and can affect fetal development and pregnancy(Khonbuvi & Usmonovna, 2024). Thyroid disorders such as hypothyroidism and hyperthyroidism. Hyperthyroidism is characterized by the overproduction of thyroid hormones, leading to increased levels of T3 and decreased TSH. This condition is diagnosed through laboratory tests that show serum TSH levels below normal and fT4 more than 24.5 pmol/l or fT3 more than 6.3 pmol/l. In contrast, Hypothyroidism is a disorder where the thyroid gland produces insufficient amounts of hormones, lowering blood levels of fT4 and fT3(Yurizali & Adhyka, 2024).

### Random Forest

Random Forest Algorithm is well-known for its ability to create reliable, accurate, consistent predictions, and its ability to avoid overfitting. By using an ensemble-based approach that integrates multiple decision trees, Random Forest can reduce the likelihood of overfitting and optimize the stability and accuracy of prediction results. This makes Random Forest a good choice for processing thyroid disease data that has many features or variables(Hengl et al., 2018).

Random Forest is a machine learning method that utilizes a collection of decision trees as its primary components to build predictive models. This technique can be applied to both classification tasks, where the goal is to group data into specific categories, and regression tasks, which focus on predicting continuous values. By combining multiple decision trees, Random Forest enhances prediction accuracy and reduces the risk of overfitting (Primajaya & Sari, 2018).

### KNNImputer

KNNImputer or KNNImputation is a technique to handle missing values in a dataset. This method uses an Euclidean distance matrix to find the nearest neighbors. The missing values are then imputed based on the average of these nearest neighbor values. KNN Imputer has become popular due to its reliability in filling missing values over traditional methods. This approach is considered better than simply deleting the missing data, as it is able to maintain the integrity of the dataset and enhance the performance of the predictive model(Juna et al., 2022).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.14334 p-ISSN : 2541-044X

## METHOD

This research aims to classify thyroid disease datasets by applying the Random Forest algorithm, which is complemented by a data preprocessing stage using KNNImputer. This research method includes several important stages which are explained as follows :

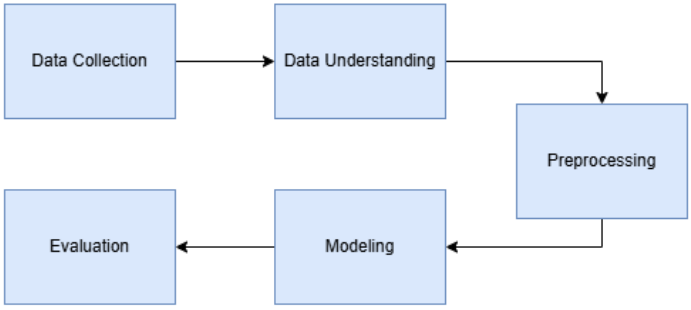


Figure. 1 Method

### Data Collection

The first stage in this research method is the data collection process. The data source in this research comes from the UCI Machine Learning Repository, which has specific datasets related to thyroid disease.

### Data Understanding

The second step in the process was to understand the data to be used. This dataset related to thyroid disease consists of 9,172 rows of data, which represents the number of samples available, and includes 31 attributes or variables. These attributes contain various relevant information such as patient characteristics, laboratory test results, and other indicators that can help in further analysis. Understanding the structure and content of this dataset is essential to ensure that the data can be processed appropriately in subsequent stages.

### Preprocessing

The data preprocessing stage aims to clean and prepare the dataset to make it more accurate, and suitable for analysis. This first stage begins with data exploration to identify information that contains missing values. Next, data cleaning is performed by handling missing values using the KNNImputer method, which preserves the distribution of data by calculating values based on nearest neighbors. In addition, the data was further processed by converting the attributes into a numeric format that was compatible with the model algorithms used. These steps ensured the data was ready to be used to build the optimal model.

### Modelling

The fourth stage in this research is the modeling stage. At this stage, a prediction model using the Random Forest algorithm is created. This algorithm was chosen because of its reliable ability to process complex data and produce accurate prediction results. The modeling process involves building and evaluating the model to ensure that the algorithm can process data effectively and provide outputs that match the research objectives.

### Evaluation

The fifth stage is evaluation, which is a crucial stage in the data analysis procedure. At this step, the model that has been built will be tested to measure its performance using various evaluation metrics, such as accuracy, precision, recall, F1-score, and confusion matrix. These metrics help provide an overall picture of the extent to which the model is able to make accurate and optimal predictions, identify errors, and assess the balance between positive and negative predicted outcomes. The outcomes of this assessment offer a basis for understanding the advantages and limitations of the model and determining the necessary corrective measures.

## RESULT

This study uses a thyroid disease dataset from the UCI Machine Learning Repository, which is made up of 9,172 rows of data with 31 attributes. The dataset includes variables such as age, gender, laboratory test results (such as TSH, T3, TT4, and FTI), as well as information related to the patient's medical history, such as goiter, tumor, and pregnancy conditions.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14334
e-ISSN : 2541-2019
p-ISSN : 2541-044X

**Table 1.** Attribute Names in Dataset

| Attribute Names |
|---|
| Age, Sex, On Thyroxine, Query on thyroxine, On antithyroid medication, Sick, Pregnant, Thyroid surgery, I131 Treatment, Query Hypothyroid, Query Hyperthyroid,Lithium, Goitre, Tumor, Hypopituitary, Psych, TSH Measured, T3 Measured, TT4 Measured, T4U Measured, FTI Measured, TBG Measured, TSH, T3, TT4,T4U, FTI, TBG,Referral source, Target, Patient id |

These variables are highly relevant for the diagnosis of thyroid disease, which includes Negative, Hypothyroid, and Hyperthyroid categories. In the data-checking process, seven attributes were found with missing values and some illogical values in the age attribute, such as age above 100 years. To handle missing values in the "Age" attribute, we use the mean rather than the median because the mean is sensitive to outliers or extreme values in the data. In contrast, the median or the middle value in the data is less sensitive to outliers or extreme values than the mean. In proving which is better between the mean and median in handling the 'Age' attribute we calculate and evaluate the skewness of the data. Skewness measures the asymmetry of the data distribution, if skewness is close to zero, the data distribution is relatively symmetrical and the mean is a good measure of central tendency. Central Tendency is a single value used to describe the center of a dataset.

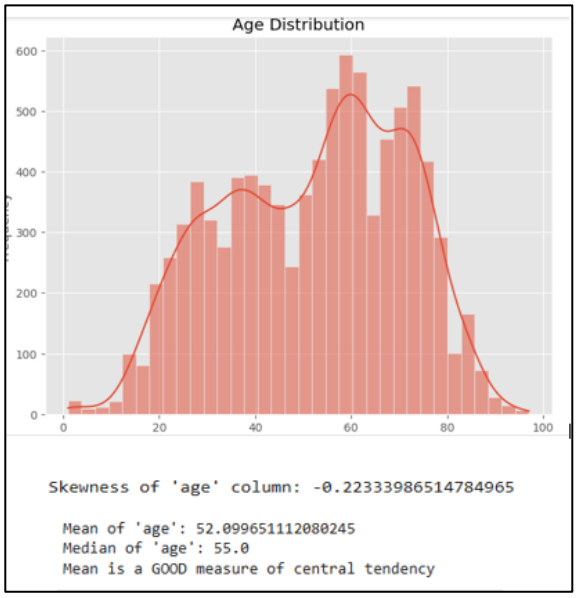


Figure. 2 Resolving missing values in the Age attribute

In addition to the age attribute, missing values were also found in sex, TT4, T3, T4U, FTI, and TSH. To overcome this problem, missing values were handled using the KNNImputer method, which calculates the average value based on the nearest neighbor using Euclidean distance.

```
knn = KNNImputer(n_neighbors=13)
knn_imputed_df = knn.fit_transform(b_fill_df)

knn_imputed_df = pd.DataFrame(knn_imputed_df, index=b_fill_df.index)
knn_imputed_df = knn_imputed_df.rename(columns=dict(zip(knn_imputed_df.columns,columns)))

df.update(knn_imputed_df)
```

Figure. 3 KNNImputer

This approach succeeded in maintaining the data distribution, resulting in better-quality data used for model training. In addition, attributes that were irrelevant or had high levels of missing values, like TSH_measured,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14334
e-ISSN : 2541-2019
p-ISSN : 2541-044X

T3_measured, TT4_measured, T4U_measured, FTI_measured, TBG_measured, referral source, patient ID, and TBG, were removed from the dataset to simplify the analysis.

Next, analysis and mapping of target attributes containing unique codes for each disease category were conducted, so that the data became more structured and easy to process. Some of the main attributes used in the analysis were sex, age, TT4, T3, TSH, FTI, T4U, goitre, and tumor. Non-numeric attributes such as sex were converted into numeric values (0 for female and 1 for male) in order to be processed by the algorithm. This preprocessing process ensures all attributes in the dataset are suitable for modeling.

| | age | tumor | sex | TT4 | T3 | T4U | FTI | TSH | goitre | target |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 28 | 0 | 0 | 206 | 361 | 455 | 259 | 43 | 0 | 2 |
| **1** | 28 | 0 | 0 | 293 | 187 | 410 | 427 | 215 | 0 | 2 |
| **2** | 40 | 0 | 0 | 154 | 369 | 95 | 500 | 420 | 0 | 2 |
| **3** | 35 | 0 | 0 | 229 | 352 | 460 | 335 | 397 | 0 | 2 |
| **5** | 59 | 0 | 0 | 139 | 235 | 225 | 269 | 393 | 0 | 2 |
| **6** | 76 | 0 | 0 | 248 | 97 | 325 | 483 | 316 | 0 | 2 |
| **7** | 27 | 0 | 0 | 239 | 406 | 420 | 324 | 84 | 0 | 2 |
| **8** | 27 | 0 | 0 | 83 | 146 | 90 | 151 | 152 | 0 | 2 |
| **9** | 27 | 0 | 0 | 90 | 110 | 132 | 168 | 260 | 0 | 2 |
| **10** | 53 | 0 | 0 | 304 | 329 | 362 | 471 | 260 | 0 | 2 |

Figure. 4 Data is Converted to Numeric

Modeling was performed using the Random Forest algorithm, which is renowned for its reliability in handling complex and diverse datasets. The dataset was divided into two parts, 80% for training data and 20% for testing data. This division aims to give the model enough data to learn the patterns and characteristics of the dataset while ensuring that the evaluation is done on data that has not been seen before. The model training and testing process involves attributes that have been selected based on their relevance to thyroid disease diagnosis.

The evaluation results show that the built model has excellent performance. Analysis using a confusion matrix shows that the Random Forest algorithm achieves 98% accuracy, precision, recall, and the F1 score reaches 98% based on the weighted average. In addition, the model also produces a precision value of High performance reflecting the model's ability to classify Negative, Hypothyroid, and Hyperthyroid disease categories with very high accuracy.

The success of the model is due not only to the advantages of the Random Forest algorithm but also to the effective data preprocessing approach, including the handling of missing values with KNNImputer. Analysis using key attributes such as TT4, TSH, and FTI helped the model to identify significant patterns in the dataset, which ultimately improved the accuracy of the predictions.

With these results, the developed model shows great potential to support the diagnosis process of thyroid disease quickly, accurately, and efficiently. This approach can be a reliable solution in the application of data mining technology in the medical field, especially in helping health workers make more informed decisions based on data.

## DISCUSSIONS

This section shows that the Random Forest algorithm with preprocessing using KNNImputer can provide impressive performance in overcoming the challenge of missing values with the original data distribution by filling in the missing values based on the similarity between the data. The missing information does not cause distortion or bias in the model training process and achieves high classification accuracy. With an accuracy rate of 97%, as well as precision, recall, and F1 score values reaching 97% based on the weighted average, the model confirms the reliability in classifying the Negative, Hypothyroid, and Hyperthyroid categories.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.14334 p-ISSN : 2541-044X

No significant overfitting detected.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Negative | 0.94 | 0.87 | 0.90 | 38 |
| Hypothyroid | 0.85 | 0.93 | 0.89 | 133 |
| Hyperthyroid | 0.99 | 0.98 | 0.99 | 1365 |
| accuracy | | | 0.98 | 1536 |
| macro avg | 0.93 | 0.93 | 0.93 | 1536 |
| weighted avg | 0.98 | 0.98 | 0.98 | 1536 |

Figure. 5 Clasification Report

In this study, the strong performance of the model can be attributed to its ability to handle complex interactions between features effectively. The main attributes used in prediction such as age, gender, hormone levels (TT4, T3, TSH), and medical indicators such as goiter and tumor are carefully selected and processed to improve predictive accuracy. The TT4, T3, and TSH attributes were chosen because they are hormone test results that are important indicators in diagnosing diseases. The goiter and tumor attributes were selected because they can both be related to the thyroid gland and cause enlargement in the neck area. The model shows that by optimally utilizing these attributes, thyroid disease classification can be performed with high precision. The transformation of the attributes into a numerical format and removing less relevant attributes ensure that the model focuses only on significant features, thereby reducing noise and improving computational efficiency.

This study strengthens the findings of the research (Supardianto et al. 2022), which shows that using KNNImputer in handling missing values can increase the accuracy of SVM models up to 93%. In this study, applying KNNImputer was even more effective when combined with Random Forest, resulting in higher accuracy. In addition, a comparison with a study (Erdiansyah et al. 2022) comparing the K-Nearest Neighbor (KNN) and Random Forest methods shows that Random Forest has a significant advantage, with a higher accuracy of 86.56% compared to KNN which only reaches 76.78%. This proves that Random Forest is more capable of handling complex and diverse data such as thyroid disease datasets.

The advantage of the Random Forest algorithm lies in its ensemble-based approach, which reduces overfitting by combining predictions from multiple decision trees. This is proven in this model with an 80-20 split, there is no significant overfitting, ensuring stability and consistency, even with complex and high-dimensional datasets such as thyroid disease data.

This model can be used as a decision support system that assists health workers in identifying patients at risk of thyroid disease. With the implementation of this model, it is still possible to provide an accurate diagnosis based on available patient data. For example, in areas with limited access to laboratories, this model allows health workers to utilize basic data such as simple hormone test results (TSH, T3, TT4) and medical history to obtain predictions that are close to expert diagnoses. This not only improves efficiency but also accelerates early detection and treatment, ultimately contributing to reduced complication rates and improved patient quality of life.

The combination of Random Forest and KNN Imputer proved to be an effective solution for thyroid disease prediction. This research highlights the importance of robust preprocessing techniques and ensemble learning in handling medical datasets. With further refinement and wider testing, this model can significantly contribute to modern healthcare, especially in facilitating accurate and efficient diagnosis.

## CONCLUSION

This study successfully implemented a classification method based on the Random Forest algorithm to detect thyroid disease using the UCI Repository dataset. One of the important processes in this research is the use of KNNImputer, which plays a role in handling missing data. This approach ensures better data quality so that the model can be trained optimally. The evaluation results showed excellent model performance, with accuracy reaching 98%. In addition, the precision, recall, and F1 score reached 98% based on the weighted average. These results show the reliability of the model in consistently classifying negative, hypothyroid, and hyperthyroid categories. This model shows great potential in supporting the diagnosis of thyroid disease quickly and accurately. Thus, this study not only contributes to the application of data mining methods in the medical field but also emphasizes the importance of proper data processing through methods such as KNNImputer to improve model performance

*name of corresponding author

## REFERENCES

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi*, *10*(1), 163. https://doi.org/10.32520/stmsi.v10i1.1129

Erdiansyah, U., Irmansyah Lubis, A., & Erwansyah, K. (2022). Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil. *Jurnal Media Informatika Budidarma*, *6*(1), 208. https://doi.org/10.30865/mib.v6i1.3373

Ginantra, N. L. W. S. R., Arifah, F. N., Wijaya, A. H., Septarini, R. S., Ahmad, N., Ardiana, D. P. Y., Effendy, F., Iskandar, A., Hazriani, H., Sari, I. Y., Gustiana, Z., Prianto, C., Gustian, D., & Negara, E. S. (2021). *Data Mining dan Penerapan Algoritma*.

Handayani, P., Nurlelah, E., Raharjo, M., & Ramdani, P. M. (2019). Liver Disease Prediction Using Decision Tree and Neural Network Methods. *Computer Engineering, Science and System Journal*, *4*(1), 55.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *2018*(8). https://doi.org/10.7717/peerj.5518

Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. *Water (Switzerland)*, *14*(17), 1–19. https://doi.org/10.3390/w14172592

Khonbuvi, H., & Usmonovna, S. G. (2024). *Thyroid diseases*. *26*, 40–43.

Khosravİ, M., Yazdanshenas, M., & Nematİ, M. H. (2015). *Design of an expert system for diagnosis of thyroid cancer*. *36*.

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, *1*(1), 27. https://doi.org/10.24014/ijaidm.v1i1.4903

Supardianto, Lalu Mutawalli, & Wafiah Murniati. (2022). Penerapan Knnimputer Dalam Mengolah Data Missing Value Untuk Membantu Meningkatkan Akurasi Support Vector Machine Klasifikasi Penyakit Tiroid. *Jurnal Informatika Teknologi Dan Sains*, *4*(4), 386–390. https://doi.org/10.51401/jinteks.v4i4.2077

Widianti, A., & Pratama, I. (2024). Penanganan Missing Values Dan Prediksi Data Timbunan Sampah Berbasis Machine Learning. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *9*(2), 242–251. https://doi.org/10.36341/rabit.v9i2.4789

Yurizali, B., & Adhyka, N. (2024). *Profil Tingkat Hormon Stimulasi Tiroid dan Kondisi Kesehatan dalam Studi Populasi Dewasa*. 124–137.

Zailani, A. U., & Hanun, N. L. (2020). Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera. *Infotech: Journal of Technology Information*, *6*(1), 7–14. https://doi.org/10.37365/jti.v6i1.61

*name of corresponding author