Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 1, March 2025, pp. 281 – 287
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Random Forest Optimization Using Recursive Feature Elimination for Stunting Classification

[1]Sophya Hadini Marpaung, [2]Frans Mikael Sinaga, [3]Khairul Hawani Rambe, [4]Fandi Presly Simamora, [5]Kelvin
[1,2,3,4,5]Informatics Faculty, Universitas Mikroskil
Email: [1]sophya.marpaung@mikroskil.ac.id, [2]frans.sinaga@mikroskil.ac.id, [3]khairul.rambe@mikroskil.ac.id, [4]fandi.simamora@mikroskil.ac.id, [5]kelvin.chen@mikroskil.ac.id

Article Info
Article history:
Received Dec 8th, 2024
Revised Feb 13th, 2025
Accepted Mar 5th, 2025

Keyword:
Classification
Mitra Medika
Optimization
RFE
Stunting

ABSTRACT
Stunting is still a major health problem in Indonesia, with a prevalence of 27% in toddlers in 2023, far from the WHO target of below 20%. RSU Mitra Medika Tanjung Mulia in Medan serves patients with various socio-economic backgrounds, which affects the quality of services, including stunting detection. Conventional methods are prone to bias and error. This study used the Random Forest algorithm and the Recursive Feature Elimination (RFE) feature selection method to improve the accuracy of stunting classification. After data preprocessing and feature selection, two main variables were identified, namely age and height. The initial Random Forest model achieved an accuracy of 94.38%, which increased to 94.42% after hyperparameter tuning. The results showed that this approach produced an accurate, efficient model that can be integrated into clinical systems, helping medical personnel identify children at risk of stunting quickly and accurately, increasing the effectiveness of interventions, and supporting government efforts to reduce the prevalence of stunting.

Corresponding Author:
Frans Mikael Sinaga,
Informatics Engineering, Mikroskil University,
Jl. M.H Thamrin No.140, Pusat Ps., Kota Medan, Sumatera Utara, Indonesia 20212
Email: frans.sinaga@mikroskil.ac.id
DOI: http://dx.doi.org/10.24014/ijaidm.v8i1.35295

1. INTRODUCTION

Stunting remains a major public health challenge in Indonesia and globally, particularly in developing countries where it affects a significant proportion of children under five years of age. Based on data from the Ministry of Health, around 27% of toddlers in Indonesia experienced stunting in 2023, a figure still far from the World Health Organization (WHO) target of below 20% [1]. The Indonesian government has set a goal to reduce the prevalence of stunting to 14% as part of the 2024 State Budget priorities [2], [3]

General Hospital or Rumah Sakit Umum (RSU) Mitra Medika Tanjung Mulia, located on Jl. KL. Medan Deli District, Medan City, North Sumatra 20241, provides various health services, including emergency installations and intensive care. This hospital serves patients from diverse socio-economic backgrounds, affecting access and quality of health services, including stunting detection and classification [4]. Conventional methods such as manual assessments and data recapitulation in simple Excel formats are prone to subjective bias and variability, which can affect diagnostic accuracy [5], [6]

Stunting is a significant public health issue that affects millions of children worldwide, particularly in developing countries, and is primarily caused by chronic malnutrition during critical growth periods. Recent studies have demonstrated the effectiveness of machine learning algorithms, such as Random Forest, in accurately classifying stunting based on various anthropometric and nutritional data [7], [8], [9]. Additionally, the integration of Recursive Feature Elimination (RFE) has shown promise in enhancing model performance by systematically selecting the most relevant features, thereby improving classification accuracy [10], [11], [12]. This research highlights the importance of implementing a more accurate and efficient stunting

classification method at RSU Mitra Medika using Random Forest technology combined with RFE. The combination of the two methods allows for the elimination of less significant features while assessing the importance of remaining features, resulting in a more optimal and accurate classification model [13], [14], [15]

Stunting is a nutritional issue in toddlers characterized by linear growth retardation, with below-average height due to chronic malnutrition and suboptimal health conditions. This condition negatively impacts children's physical and cognitive development. Prior studies have investigated various classification methods for diagnosing stunting, though many still face limitations in terms of accuracy and computational efficiency [7], [16]. For example, one study comparing several machine learning models for stunting classification in children under five years in Zambia found that Random Forest achieved the highest accuracy of 76% [17]. Another study on the application of Random Forest for stunting classification achieved an accuracy of 90.7% using features such as gender, birth weight, birth height, age, weight at measurement, height at measurement, and stunting status [18], [19], [20]. Both studies emphasize the importance of feature selection, as not all features significantly impact stunting classification [21], [22].

This research aims to optimize the Random Forest algorithm using RFE to provide a robust and accurate classification framework for stunting, ultimately contributing to better-targeted interventions and policy decisions in child nutrition. By leveraging machine learning advancements, particularly in feature selection and classification accuracy, this study seeks to enhance clinical decision-making processes and reduce misdiagnosis in stunting detection.

Furthermore, this study aims to develop a classification model that can be integrated with the clinical system at RSU Mitra Medika, to help doctors and health workers identify children at high risk of stunting more accurately. Thus, it is hoped that there will be an increase in the quality of health services and the effectiveness of the interventions carried out, which can ultimately support a significant reduction in stunting rates in the community. Based on this background, the formulation of the problem in this study is:

1. How to accurately classify stunting using the Random Forest algorithm with Recursive Feature Elimination (RFE) feature selection?
2. How can the process of developing and implementing this classification model identify significant features for stunting diagnosis?
3. How can the developed classification model be integrated into the clinical service system at RSU Mitra Medika to improve the accuracy and efficiency of diagnosis?

2. RESEARCH METHOD

This research method includes the development and training of a stunting classification model using the Random Forest algorithm with the Recursive Feature Elimination (RFE) feature selection approach. The use of RFE aims to improve classification accuracy by identifying the most relevant features from the dataset originating from RSU Mitra Medika. The research data used came from RSU Mitra Medika Medan from 2020-2024. This dataset consists of various relevant variables, including gender, age, weight, height, nutritional status, health history, and other risk factors that contribute to stunting. This data is also supplemented with information from relevant health surveys to enrich the data analysis.

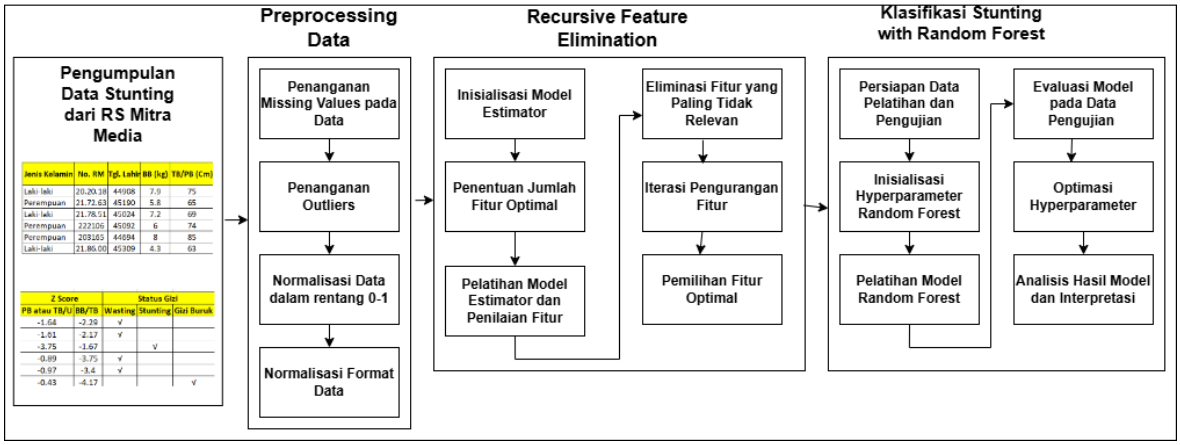

Figure 1. Research Stage

The dataset used comes from medical records at RSU Mitra Medika. This data includes anthropometric information, nutritional data, and risk factors related to stunting. Data collection was carried out systematically to ensure the accuracy and completeness of the information. After this stage, the author

moving to preprocessing data stage. This step involves several important processes, such as handling missing values and outliers, data normalization, and data format standardization. This process aims to ensure that the model can process data effectively, as well as improve the quality of input used in the classification model. Next, we have recursive feature elimination**,** the use of RFE as a feature selection method aims to identify the variables that have the most influence on stunting classification. RFE works by iteratively removing less important features and retaining features that have a significant contribution, thereby improving the performance of the Random Forest model. After all of this, the author also did Stunting Classification with Random Forest**.** The Random Forest model is designed to capture complex relationships between variables. Random Forest has the advantage of handling large and complex data, and minimizing the risk of overfitting through ensemble techniques. This model is trained with preprocessed data and selected features.

1. Model Training
   The trained model is then validated using separate validation data. The validation process aims to evaluate the model's performance in predicting new data that is not involved in training. Early stopping is used to automatically stop training if there is no increase in performance, thereby helping to prevent overfitting.

2. Model Validation
   The trained model is then validated using separate validation data. The validation process aims to evaluate the model's performance in predicting new data that is not involved in training. Early stopping is used to automatically stop training if there is no increase in performance, thereby helping to prevent overfitting.

3. Model Evaluation
   The evaluation is carried out using the accuracy, precision, recall, and Area Under Curve (AUC) metrics from the Receiver Operating Characteristic (ROC) curve. In addition, the cross-validation method is applied to ensure the stability and consistency of model predictions, as well as to avoid overfitting problems.

3. RESULTS AND ANALYSIS

The results and analysis in this research did by collectiong the data set. The data used in this study came from medical records at RSU Mitra Medika. The dataset consists of several important features, namely:

1. Gender: Identifying a child as male or female.
2. Age: The child's age in months.
3. Height (TB): Measurement of the child's height in centimeters (cm).
4. Weight (BB): Measurement of the child's weight in kilograms (kg).
5. Nutritional Status: A binary label that identifies whether the child is at risk of stunting (1) or normal (0).
6. Wasting: Category of weight status against height, which is grouped into several categories such as Underweight, Risk of Overweight, or Normal weight.
7. Z-Score: Standard deviation value that indicates the extent to which a child's growth differs from the reference population standard.

An example of the first 10 data samples in the dataset used can be seen in Table 1.

Table 1. Sample Dataset Penelitian

Gender	Age	Tall	Weight	Nutritional Status	Wasting	Z-Score
Male	20	77.7	8.5	1	Underweight	-0.73
Male	10	79	10.3	0	Risk of Overweight	-0.6
Female	2	50.3	8.3	1	Risk of Overweight	-3.47
Female	5	56.4	10.9	1	Risk of Overweight	-2.86
Male	11	76.3	5.9	0	Severely Underweight	-0.87
Male	16	80.7	9.9	0	Normal weight	-0.43
Female	15	72.6	6.5	0	Severely Underweight	-1.24
Female	18	78.4	15.6	0	Risk of Overweight	-0.66
Male	2	63.4	7	0	Risk of Overweight	-2.16
Male	6	60.4	11.5	1	Risk of Overweight	-2.46

The dataset used was then preprocessed using four different stages, namely: handling missing values, removing unused features, one-hot encoding and data normalization. The four stages carried out were:

1. Handling missing values is done by deleting if there are data rows that have a null value.

2. After the dataset is clean from missing values, it is continued by removing the wasting feature because this feature does not have a clear correlation with the nutritional status feature which is the target of the stunting classification. This can be seen in the sample dataset in Table 1 where the nutritional status feature with a value of 1 has various wasting feature values.
3. Preprocessing is then continued by changing the dataset features such as gender in the form of text categories to numeric so that they can be used in the modeling stage.
4. The preprocessing stage ends by normalizing the data on all features in the range 0-1 using the Min-Max Scaling method. This aims to equalize the scale between features so that no feature dominates the model training process.

An example of the first 10 data samples in the dataset after going through data preprocessing can be seen in Table 2.

Table 2. Research Dataset Sample Preprocessing Results

Gender	Age	Tall	Weight	Nutritional Status	Z-Score
1	20	0.650000	0.462963	1	0.650000
1	10	0.674074	0.574074	0	0.674074
0	2	0.142593	0.450617	1	0.142593
0	5	0.255556	0.611111	1	0.255556
1	11	0.624074	0.302469	0	0.624074
1	16	0.705556	0.549383	0	0.705556
0	15	0.555556	0.339506	0	0.555556
0	18	0.662963	0.901235	0	0.662963
1	2	0.385185	0.370370	0	0.385185
1	6	0.329630	0.648148	1	0.329630

The dataset that has gone through data preprocessing is then divided into two parts, namely training data and testing data. The nutritional status feature is used as a target for classification. The division of the dataset is done with a ratio of 80% for training data and 20% for testing data. This means that 80% of the data is used to train the model, while the remaining 20% is used to test the performance of the trained model.

The feature selection process is carried out using the Recursive Feature Elimination (RFE) method, which aims to select the best features based on relevance to the target (Nutritional Status). At this stage, the initial estimation model uses Random Forest as the basis for evaluating feature importance. The RFE method iteratively eliminates features that are considered less relevant until the optimal number of features for classification is obtained. Based on the results of this process, the features that have the most significant influence on stunting classification are Age and Height. Thus, the next Random Forest model is built using these two features as the main input. This feature selection is expected to improve model performance by eliminating noise from the data, so that the model focuses on the variables that are most relevant to stunting classification.

After all of this steps, the author did stuntingclassification with random forest. Two features obtained through the Recursive Feature Elimination (RFE) process, namely Age and Height, were used to build the Random Forest model. In the initial stage, a Random Forest model without optimization (base model) was built, and the evaluation results showed an accuracy of 94.38%. The complete evaluation results for the RFE-Random Forest base model are shown in Table 3.

Table 3. Evaluation Report RFE-Random Forest

	Precision	Recall	F1-Score	Support
0	0.97	0.96	0.96	14427
1	0.87	0.89	0.88	4432
accuracy			0.94	18859
macro avg	0.92	0.92	0.92	18859
weighted avg	0.94	0.94	0.94	18859

Then to optimize the previous RFE-Random Forest base model, in this study hyperparameter tuning was carried out using grid search. The following is the Random Forest hyperparameter search space using grid search in Table 4.

Table 4. Random Forest Hyperparameter Search Space

Hyperparameter	Search Space
n_estimators	50, 100, 200
max_depth	None, 10, 20
min_samples_split	2, 5, 10

Through grid search, the best hyperparameter configuration of Random Forest in this study was n_estimators 50, max_depth 20, and min_samples_split 10. Thus, the RFE-Random Forest model will be rebuilt using this configuration and an accuracy result of 94.42% was obtained. The evaluation results carried out on this new model can be seen in Table 5.

Table 5. Evaluation RFE-Random Forest-Grid Search

	Precision	Recall	F1-Score	Support
0	0.97	0.96	0.96	14427
1	0.88	0.89	0.88	4432
accuracy			0.94	18859
macro avg	0.92	0.93	0.92	18859
weighted avg	0.94	0.94	0.94	18859

Based on Table 5 above, it can be seen that the increase in accuracy obtained through hyperparameter tuning grid search is not too significant. Which means that the base model RFE-Random Forest is actually sufficient to provide results with good accuracy.

4. CONCLUSION

Stunting is still a significant public health challenge in Indonesia, with a prevalence of 27% in toddlers in 2023. This study aims to improve the accuracy and efficiency of stunting classification at RSU Mitra Medika by developing a model based on the Random Forest algorithm combined with the Recursive Feature Elimination (RFE) feature selection method. The dataset used consists of children's medical records, including variables such as age, height, weight, and nutritional status. After going through the data preprocessing process and feature selection using RFE, the two main variables that are most relevant to stunting classification were identified, namely age and height. The initial Random Forest model achieved an accuracy of 94.38%, which increased to 94.42% after hyperparameter tuning using grid search. The results of the study show that the RFE and Random Forest-based approaches can produce accurate, efficient classification models that can be integrated into the clinical service system. With this model, health workers can more quickly and accurately identify toddlers at risk of stunting, thereby increasing the effectiveness of health interventions and supporting the government's target of reducing the prevalence of stunting.

REFERENCES

[1] Kementerian Kesehatan RI, "Panduan Hari Gizi Nasional ke 64 Tahun 2024," Kemkes.go.id., Jakarta, 2024.

[2] U. M. Malang, "MIND (Multimedia Artificial Intelligent Networking Database Klasifikasi Penyakit Stunting Menggunakan Algoritma Multi-Layer Perceptron Putri Intan Ashuri, Indah Ardhia Cahyani, Christian Sri Kusuma Aditya," *Journal MIND Journal | ISSN*, vol. 9, no. 1, pp. 52–63, 2024, doi: 10.26760/mindjournal.v9i1.52-63.

[3] S. Lonang and D. Normawati, "Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination," *Jurnal Media Informatika Budidarma*, vol. 6, no. 1, p. 49, Jan. 2022, doi: 10.30865/mib.v6i1.3312.

[4] Mitra Medika, "RS Mitra Medika - Tanjung Mulia." Accessed: Sep. 11, 2024. [Online]. Available: Available: https://tanjungmulia.mitramedika.com

[5] L. Asra Laily, S. Indarjo, J. Ilmu Kesehatan Masyarakat, F. Ilmu Keolahragaan, and U. Negeri Semarang, "354 HIGEIA 7 (3) (2023) Higeia Journal Of Public Health Research And Development Literature Review: Dampak Stunting terhadap Pertumbuhan dan Perkembangan Anak," 2023, doi: 10.15294/higeia/v7i3/63544.

[6] R. Yusran, A. Nanda, A. Amalda, R. Luthvia, and R. Fadlan, "Upaya Pemenuhan Kesadaran Masyarakat dan Pemenuhan Gizi Seimbang untuk Mencegah Peningkatan Angka Stunting di Nagari Pariangan 2023," *Inovasi Jurnal Pengabdian Masyarakat*, vol. 1, no. 2, pp. 131–140, Aug. 2023, doi: 10.54082/ijpm.138.

[7] A. A. R. Reza and Muhammad Syaifur Rohman, "Prediction Stunting Analysis Using Random Forest Algorithm and Random Search Optimization," *Journal of Informatics and Telecommunication Engineering*, vol. 7, no. 2, pp. 534–544, Jan. 2024, doi: 10.31289/jite.v7i2.10628.

[8] J. R. Khan, J. H. Tomal, and E. Raheem, "Model and variable selection using machine learning methods with applications to childhood stunting in Bangladesh," *Inform Health Soc Care*, vol. 46, no. 4, pp. 425–442, 2021, doi: 10.1080/17538157.2021.1904938.

[9] M. Yunus, M. K. Biddinika, and A. Fadlil, "Classification of Stunting in Children Using the C4.5 Algorithm," *Jurnal Online Informatika*, vol. 8, no. 1, pp. 99–106, Jun. 2023, doi: 10.15575/join.v8i1.1062.

[10] W. S. Lestari, Y. M. Saragih, and Caroline, "Comparison of Deep Neural Networks and Random Forest Algorithms for Multiclass Stunting Prediction in Toddlers," *Teknika*, vol. 13, no. 3, pp. 412–417, Oct. 2024, doi: 10.34148/teknika.v13i3.1063.

[11] M. N. A. Khan and R. M. Yunus, "A hybrid ensemble approach to accelerate the classification accuracy for predicting malnutrition among under-five children in sub-Saharan African countries," *Nutrition*, vol. 108, Apr. 2023, doi: 10.1016/j.nut.2022.111947.

[12] H. Pohan *et al.*, "Penerapan Algoritma K-Medoids dalam Pengelompokan Balita Stunting di Indonesia".

[13] P. Prihandoko, D. Jollyta, G. Gusrianty, M. Siddik, and J. Johan, "Cluster Validity for Optimizing Classification Model: Davies Bouldin Index – Random Forest Algorithm," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 24, no. 1, pp. 61–72, Nov. 2024, doi: 10.30812/matrik.v24i1.4043.
[14] S. Marsya Finda and D. Wahyu Utomo, "Klasifikasi Stunting Balita menggunakan Metode Ensemble Learning dan Random Forest," *Jl. Imam Bonjol No*, vol. 15, no. 02, 2024, doi: 10.35970/infotekmesin.v15i2.2326.
[15] A. M. Priyatno and T. Widiyaningtyas, "A Systematic Literature Review: Recursive Feature Elimination Algorithms," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 9, no. 2, pp. 196–207, Feb. 2024, doi: 10.33480/jitk.v9i2.5015.
[16] F. Kesehatan Masyarakat, M. Renssca Inas, L. Widajanti, and S. Achadi Nugraheni, "Media Kesehatan Masyarakat Indonesia Hubungan Asupan Energi, Zinc, Protein pada Ibu Hamil dengan Kejadian Stunting pada Balita 7-24 Bulan di Indonesia: Literature Review", doi: 10.14710/mkmi.21.5.354-357.
[17] O. N. Chilyabanyama *et al.*, "Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia," *Children*, vol. 9, no. 7, Jul. 2022, doi: 10.3390/children9071082.
[18] N. Faoziatun Khusna *et al.*, "Implementasi Random Forest dalam Klasifikasi Kasus Stunting pada Balita dengan Hyperparameter Tuning Grid Search," *Seminar Nasional Sains Data*, vol. 2024.
[19] B. Satria, T. A. Y. Siswa, and W. J. Pranoto, "Optimasi Random Forest dengan Genetic Algorithm dan Recursive Feature Elimination pada High Dimensional Data Stunting Samarinda," *Jurnal Media Informatika Budidarma*, vol. 8, no. 3, p. 1778, Jul. 2024, doi: 10.30865/mib.v8i3.7883.
[20] H. Nalatissifa, W. Gata, S. Diantika, and K. Nisa, "Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja," *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, p. 578, Dec. 2021, doi: 10.32493/informatika.v5i4.7575.
[21] R. Rizqi Robbi Arisandi, B. Warsito, and A. Rachman Hakim, "Aplikasi Naïve Bayes Classifier (Nbc) Pada Klasifikasi Status Gizi Balita Stunting Dengan Pengujian K-Fold Cross Validation," vol. 11, no. 1, pp. 130–139, 2022, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/
[22] A. Husaini, I. Hoeronis, H. H. Lumana, and L. D. Puspareni, "Early Detection of Stunting in Toddlers Based on Ensemble Machine Learning in Purbaratu Tasikmalaya," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 11, no. 3, p. 487, Jul. 2023, doi: 10.26418/justin.v11i3.66465.

BIBLIOGRAPHY OF AUTHORS

Sophya Hadini Marpaung, The author is a lecturer at Universitas Mikroskil. The author completed Master's degree in the Master Management of Information Systems program. The author focuses on writing, which is an inseparable part of the life of a lecturer as a part of Tridharma of Higher Education (Education and Teaching, Research, and Community Service).

Frans Mikael Sinaga, S.Kom., M.Kom., Lecturer at the Department of Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. Born in Penggalangan village on October 24, 1993. The author is the third child out of 4 siblings of Mr. Waristo and Mrs. Linda. The author completed a Bachelor's degree (S1) in Informatics Engineering and a Master's degree (S2) in Information Technology at STMIK Mikroskil Medan. The author has written several book titles such as Introduction to Computer Networks and Data Mining. In addition to writing books, the author has also conducted several research projects in the fields of Data Science and Computer Vision.

Khairul Hawani Rambe, The author is a lecturer in the Undergraduate Informatics Study Program, Faculty of Informatics, Universitas Mikroskil. The author completed her Bachelor's degree in the Information Engineering program and continued her Master's studies in the International Master Program in Information Technology and Applications. The author focuses on writing, which is an essential part of a lecturer's role in fulfilling the Tri Dharma of Higher Education (Education and Teaching, Research, and community Service).

Fandi Presly Simamora, The author is a lecturer in the Undergraduate Informatics Study Program, Faculty of Informatics, Universitas Mikroskil. The author completed her Bachelor's degree in the Information Engineering program and continued her Master's studies in the International Master Program in Information Technology and Applications. The author focuses on writing, which is an essential part of a lecturer's role in fulfilling the Tri Dharma of Higher Education (Education and Teaching, Research, and community Service).

Kelvin, S.Kom., M.Kom., The author is a Software Engineer and Lecturer in the Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. He completed his bachelor's degree in Informatics Engineering at STMIK Mikroskil in 2018. Then, in 2020, the author pursued postgraduate studies in Information Technology at Mikroskil University and successfully completed them in 2021. The courses he has taught include Introduction to Algorithms, Web Design, C Programming, Object-Oriented Programming, Back-End Web Development, Artificial Intelligence, and Natural Language Processing. In addition to his academic involvement, the author has over 5 years of experience as a software engineer, working for both domestic and international companies. For more information, visit the author's LinkedIn page at https://www.linkedin.com/in/kelvinchen96