VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i3.5972

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

COMPARISON OF ENSEMBLE METHODS FOR DECISION TREE MODELS IN CLASSIFYING E. COLI BACTERIA

Alvin Rahman Al Musyaffa[1]; Yoga Pristyanto[2*]; Nia Mauliza[3]

Department of Information System[1, 2, 3]
Amikom Yogyakarta University, Yogyakarta, Indonesia[1, 2, 3]
https://home.amikom.ac.id[1, 2, 3]
alvinram@students.amikom.ac.id[1], yoga.pristyanto@amikom.ac.id[2*],
niamauliza08@students.amikom.ac.id[3]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract—*Certain strains of Escherichia coli (E. coli) can cause serious illness, so identifying dangerous strains with high accuracy is a priority in supporting public health and food safety. However, traditional machine learning methods, such as Decision Trees, are often not robust enough to handle the complexity of biological data. This research presents a solution by systematically evaluating seven ensemble methods, namely Adaboost, Gradient Boosting, XGBoost, LightGBM, Random Forest, Bagging, and Stacking, using a dataset that includes 336 E. coli samples with eight biological features. These models are evaluated based on accuracy, precision, recall, and F1 score, with parameter optimization to obtain the best results. The results show that XGBoost is superior with accuracy, recall, and F1 score of 88% and precision of 87%, outperforming other methods. This research has the advantage of a comprehensive approach in comparing various ensemble methods simultaneously, accompanied by the application of confusion matrix-based evaluation to ensure the accuracy of the results. Additionally, the ensemble approach proved to be more effective in handling complex data patterns and reducing bias in bacterial strain classification. These findings provide a significant contribution, namely a practical framework for improving laboratory diagnostics and public health surveillance, with machine learning-based solutions that are faster, more reliable, and applicable for both industrial and clinical environments. This research expands understanding of the potential of ensemble methods in microbiological data classification and provides new directions for modern diagnostic technology.*

Keywords: *classification performance, decision tree, ensemble methods, escherichia coli classification, machine learning.*

Intisari—*Strain Escherichia coli (E. coli) tertentu dapat menyebabkan penyakit serius, sehingga mengidentifikasi strain berbahaya dengan akurasi tinggi merupakan prioritas dalam mendukung kesehatan masyarakat dan keamanan pangan. Namun, metode pembelajaran mesin tradisional, seperti Decision Trees, seringkali tidak cukup tangguh untuk menangani kompleksitas data biologis. Penelitian ini menyajikan solusi dengan mengevaluasi secara sistematis tujuh metode ensemble, yaitu Adaboost, Gradient Boosting, XGBoost, LightGBM, Random Forest, Bagging, dan Stacking, menggunakan dataset yang mencakup 336 sampel E. coli dengan delapan fitur biologis. Model-model ini dievaluasi berdasarkan akurasi, presisi, recall, dan skor F1, dengan optimasi parameter untuk mendapatkan hasil terbaik. Hasil penelitian menunjukkan bahwa XGBoost lebih unggul dengan akurasi, recall, dan skor F1 sebesar 88% dan presisi sebesar 87%, mengungguli metode lain. Penelitian ini memiliki keunggulan berupa pendekatan komprehensif dalam membandingkan berbagai metode ensemble secara bersamaan, disertai dengan penerapan evaluasi berbasis matriks kebingungan untuk memastikan keakuratan hasil. Selain itu, pendekatan ensemble terbukti lebih efektif dalam menangani pola data yang kompleks dan mengurangi bias dalam klasifikasi strain bakteri. Temuan ini memberikan kontribusi yang signifikan, yaitu kerangka kerja praktis untuk meningkatkan diagnostik laboratorium dan pengawasan kesehatan masyarakat, dengan solusi berbasis pembelajaran mesin yang lebih cepat, lebih andal, dan dapat*

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i3.5972

diterapkan untuk lingkungan industri dan klinis. Penelitian ini memperluas pemahaman tentang potensi metode ensemble dalam klasifikasi data mikrobiologi dan memberikan arah baru untuk teknologi diagnostik modern.

***Kata Kunci**: kinerja klasifikasi, decision tree, metode ensemble, klasifikasi escherichia coli, machine learning.*

INTRODUCTION

Escherichia coli bacteria (E. coli) are microorganisms that are very important in various fields, such as health, the food industry, and microbiology research. Most strains of E. coli live as normal inhabitants of the intestinal flora of humans and animals, but certain strains can be pathogenic and cause serious illnesses, such as diarrhea, urinary tract infections, and even life-threatening conditions, such as hemolytic uremic syndrome (HUS) [1]. Given its significant impact, accurate identification and classification of E. coli strains is crucial, not only for the prevention and treatment of diseases caused by this bacterium, but also for the development of effective control strategies in the public health and food safety sectors [2]. In recent decades, machine learning has become an increasingly important tool in the analysis of microbiological data, especially in classification tasks that involve processing large and complex data [3].

Several studies related to the use of machine learning technology in ecolimatic classification have also been carried out. The research [4], the study employs the Naïve Bayes algorithm, enhanced through the AdaBoost method, to classify *E. coli* bacteria. The findings show that, in its standalone implementation, the Naïve Bayes algorithm attains an accuracy of 76%. However, with the integration of AdaBoost, its accuracy rises substantially to 94%. This demonstrates the significant role of boosting techniques in improving the performance of weaker classifiers, such as Naïve Bayes, when applied to complex datasets like *E. coli*.

Additionally, the research evaluates the performance of other algorithms, such as Support Vector Machines (SVM) and Decision Trees, both with and without the application of AdaBoost. The results indicate that Decision Trees, when paired with AdaBoost, achieve an accuracy of 88%, while SVM achieves 92% accuracy even without the use of boosting. These results emphasize the effectiveness of ensemble methods like AdaBoost in managing biological datasets, enhancing accuracy, and stabilizing model performance.

.

The ensemble method is an approach that combines the results of several learning models to increase accuracy and reduce the possibility of prediction errors. In the context of classification, this method functions by exploiting the strengths of several different models, so that the final results are more stable and more reliable compared to using just one model [5][6][7]. Research [8] shows that the results of testing numerical data on the classification using a single classification comparison obtained accuracy results of only 63%-65%. Then further research was carried out by [9], to improve the accuracy results of diabetes classification by testing using the ensemble method and succeeded in increasing the accuracy results to 94%-96%. These findings highlight that ensemble methods consistently outperform single classifiers by leveraging the power of multiple models, thereby improving overall accuracy and reducing errors. Although various ensemble methods have been successfully applied to various types of data, research focusing on comparing the performance of ensemble methods in the context of E. coli bacterial classification is still limited. Each ensemble method has its own characteristics and advantages, which may interact with the characteristics of the microbiological data differently.

This research focuses on comparative analysis of the performance of various classification models in analyzing and classifying Escherichia coli (E. coli) bacteria, with Decision Tree as the basic model. To improve model performance, a number of ensemble methods are applied, including AdaBoost, Gradient Boosting, XGBoost, LightGBM, Bagging, Stacking, and Random Forest. The performance of these models will be evaluated using a confusion matrix to analyze the accuracy, precision, recall and F1 score of each model, then the results will be compared with each other. The main aim of this research is to provide an in-depth understanding of which ensemble methods are most effective in improving the classification performance of Escherichia coli (E. coli) bacteria. Apart from that, this research also aims to identify classification algorithm models that provide the best results by applying ensemble methods such as AdaBoost, Gradient Boosting, XGBoost, LightGBM, Bagging, Stacking, and Random Forest. Thus, it is hoped that this research can provide practical recommendations regarding the use of ensemble methods in increasing classification accuracy on E. coli datasets.

MATERIALS AND METHODS

The research process began with Data Acquisition, where a dataset of *Escherichia coli* samples was collected to serve as the foundation for classification. Following this, Data Preprocessing was carried out to ensure data quality and reliability. This step included handling outliers using the Interquartile Range (IQR) method, checking for data consistency, and addressing any issues with duplication or missing values.

Once the dataset was preprocessed, it was passed to the Classification Model stage. In this step, several ensemble methods were implemented, including Adaboost, Gradient Boosting, XGBoost, LightGBM, Random Forest, Bagging, and Stacking, using a Decision Tree as the base model. Each model was trained and tested to evaluate its ability to classify *E. coli* bacteria accurately.

Finally, in the Model Evaluation stage, the performance of the classification models was assessed using metrics such as accuracy, precision, recall, and F1-score to determine the effectiveness of each model in classifying *E. coli* bacteria.

Data Acquisition

This research uses dataset obtained from the UCI Machine Learning Repository with the link source(https://archive.ics.uci.edu/dataset/39/ecoli). Dataset was chosen because it provides relevant microbiological data for the purpose of classifying E. coli bacteria using various machine learning methods, in cluding Decision Tree models and ensemble methods.

Table 1. Sample Dataset

No	mcg	gvh	lip	chg	aac	alm1	alm2	class
1	0.49	0.29	0.48	0.5	0.56	0.24	0.35	cp
2	0.07	0.40	0.48	0.5	0.54	0.35	0.44	cp
3	0.56	0.40	0.48	0.5	0.49	0.37	0.46	cp
4	0.59	0.49	0.48	0.5	0.52	0.45	0.36	cp
5	0.23	0.32	0.48	0.5	0.55	0.25	0.35	cp

Source: (Research Result, 2024)

Preprocessing Data

In the data preprocessing stage, various important steps are taken to ensure the quality and consistency of the dataset to be used. Outlier detection is implemented using the Interquartile Range (IQR) method, where values outside the range Q1−1.5×IQR or Q3+1.5×IQR are considered as outliers. These outliers are handled to ensure that the model is not influenced by extreme values that could interfere with the analysis. In addition, missing values are handled using the data.isnull().sum() method to identify variables with missing values. If found, missing values were imputed using the mean or median to maintain data completeness and maintain an accurate distribution. Duplication checking is also carried out with the data.duplicated() function to ensure that no data is repeated, so that redundancy can be avoided and model accuracy is maintained. Proper handling of missing values and duplication is very important to improve model performance and the accuracy of research results. Although class imbalance is often a challenge in microbiological datasets, the class distribution in this Escherichia coli dataset is quite even. Therefore, techniques such as oversampling, undersampling, or weighted models are not necessary in this study. Model evaluation also did not reveal any significant errors related to class imbalance.

Classification Models

After the Data Acquisition stage, this study focuses on using classification models to analyze and classify Escherichia coli (E. coli) bacteria, with Decision Tree as the base model [10]. To enhance the model's performance, various ensemble methods are applied, including AdaBoost, Gradient Boosting, XGBoost, LightGradient, Bagging, Stacking, and Random Forest. These methods are expected to improve the accuracy and robustness of the classification of E. coli strains.

1. Decision Tree

The Decision Tree Classification Model is a very popular algorithm and one of the models most widely used in classification models [11]. Decision Trees work by dividing a dataset into smaller subsets based on decision rules generated from input features [12]. This process continues until each subset reaches a homogeneous condition or there are no features left to divide the dataset further [13]. The structure of a Decision Tree resembles a tree, with internal nodes representing decisions based on certain features and leaves representing the target class [14]. Different researchers from various fields and backgrounds have considered the problem of extending a decision tree from available data, such as machine study, pattern recognition, and statistics. In various fields such as medical disease analysis, text classification, user smartphone classification, images, and many more the employment of Decision tree classifiers has been proposed in many ways [15]. A decision tree is a tree-based technique in which any path beginning from the root is described by a data separating sequence until a Boolean outcome at the leaf node is achieved [16].

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i3.5972

2. AdaBoost
The AdaBoost (Adaptive Boosting) method improves the performance of the Decision Tree model by combining several weak models into one stronger model. Each iteration in the boosting process focuses on misclassified data from the previous step, giving greater weight to those instances so that the next model can correct the error. Key hyperparameters include *number of estimators* set to 100 and *learning rate* at 0.1, making it effective for data with high misclassification errors.

3. Gradient Boosting
Gradient Boosting is a method that combines several Decision Tree models to minimize prediction errors iteratively. Each new model is built to correct residual errors from the previous model using gradient-based optimization techniques. This method is known for its ability to improve classification performance, especially on complex datasets, by reducing errors that occur during the learning process. Important hyperparameters include number of estimators set to 100, learning rate at 0.1, and maximum depth at 3, improving classification on complex datasets by reducing errors effectively.

4. XGBost
XGBoost (Extreme Gradient Boosting) is a faster and more efficient version of Gradient Boosting, with an emphasis on performance optimization. XGBoost offers advantages in handling imbalanced data and reduces the risk of overfitting through the use of more advanced regularization techniques. This makes XGBoost one of the most powerful and popular ensemble methods for various classification tasks, including bacterial classification E. coli. Default hyperparameters include number of estimators at 100, learning rate at 0.1, maximum depth at 3, and subsample at 1.0, making XGBoost powerful for tasks like bacterial classification.

5. LightGradient
LightGradient is a Gradient Boosting method optimized for efficiency in terms of speed and memory usage. This method uses a histogram-based learning technique that allows handling large and complex datasets more efficiently. hyperparameters include number of leaves at 31, learning rate at 0.1, and maximum depth at 3, which deliver performance nearly equivalent to XGBoost, ideal for microbiological classification.

6. Bagging
Bagging (Bootstrap Aggregating) is an ensemble method that reduces model variance by combining predictions from multiple Decision Trees trained on different subsets of data. This subset is obtained through the bootstrap sampling technique, which randomly selects data with repetition. By combining prediction results from multiple models, Bagging improves the stability and reliability of predictions, thereby reducing the risk of overfitting. Key hyperparameters include number of estimators at 100 and maximum samples set to 1.0, which improve stability and reliability of predictions by reducing overfitting risks.

7. Stacking
Stacking is an ensemble method that combines the strengths of several models, including Decision Trees, using a meta-learner model. Predictions from the base model are used as additional features for the meta model, which then makes the final predictions. This approach allows combining different learning methods, thereby improving the overall performance of the model in bacterial classification E. coli. Using a meta-learner, commonly logistic regression, that takes base model predictions as inputs for final classification.

8. Random Forest
Random Forest is an ensemble method that combines predictions from many Decision Trees trained on a random subset of data and features. Random Forest is known for its stability and ability to generalize across diverse data, making it a reliable choice in a variety of classification tasks, including bacterial datasets. E. coli complex. hyperparameters include number of estimators at 100, maximum depth at 3, and minimum samples split set to 2, making it robust and reliable for complex bacterial datasets like E. coli.

The most commonly used algorithm in the Decision Tree Classification Model is CART (Classification and Regression Tree) [17]. In this algorithm, trees are built by dividing the dataset based on impurity criteria, such as Entropy. Entropy is used to measure uncertainty or impurity in a dataset and is expressed in equation 1.

$$Entropy(S) = \sum_{i=1}^{n} -Pi\ log_2\ Pi \quad (1)$$

By understanding the formula above, the data that has been obtained can be entered and processed using this algorithm for the process of creating a decision tree [18].

The use of ensemble methods in this research was motivated by the existence of class imbalance in the existing data. In order to avoid the need to adjust the dataset through techniques such as oversampling or undersampling, the ensemble approach allows the model to handle imbalanced classes naturally by combining the strengths of multiple classifiers. This method maintains the integrity of the original data distribution while improving classification accuracy and robustness, making it a very suitable approach to achieve reliable results without changing the natural structure of the dataset.

Model Evaluation

Model evaluation is an important step in the machine learning model development process, which aims to measure the model's performance in classifying data. Metrics are calculated based on the confusion matrix for binary segmentation tasks, which includes counts for true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions. [19].

Table 2. Confusion Matrix Prediction

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Source: (Research Result, 2024)

Each row in the matrix represents the actual data class, and each column represents the predicted data class or vice versa [20]. The prediction matrix is explained in Table 2.

Accuracy is used to measure the percentage of correct predictions out of all predictions made by the model. Accuracy can be written in the formula equation 2:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision is used to measure how well the model makes correct predictions for the positive class from the total positive predictions made. Precision can be written in the formula equation 3:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall is used to measure the model's ability to detect all positive data. Recall can be written in the formula equation 4:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

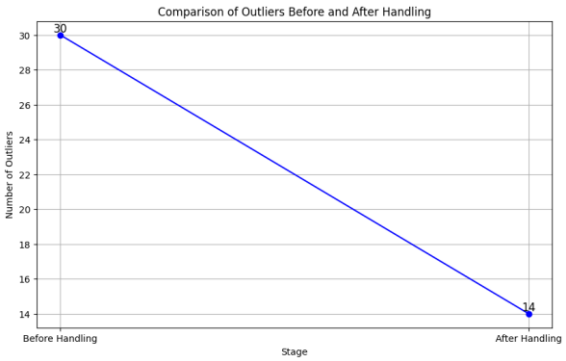
F1-Score is used to measure how well our model combines Precision and Recall capabilities, so we can understand how effective the model is in classifying data. F1-Score can be written in the formula equation 4:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (5)$$

RESULTS AND DISCUSSION

In this study, the dataset consists of 336 samples of Escherichia coli bacteria, each defined by 8 variables: mcg (McGeoch's method for signal sequence recognition), gvh (Von Heijne's method for signal sequence recognition), lip (lipoprotein prediction), chg (Signal Peptide Prediction Method), aac (discrimination of outer membrane proteins), alm1 (Score of the ALOM membrane spanning region prediction program), alm2 (Score of the ALOM membrane spanning region prediction program), and class (subcellular location of the classified protein).

In the preprocessing stage, rigorous quality checks were conducted to ensure the dataset's integrity. Missing values and duplicate entries were not present, simplifying preprocessing and allowing the focus to shift to addressing outliers. Outlier detection revealed 30 extreme values, which were reduced to 14 post-handling using the IQR method. Removing outliers reduces data variability, thereby increasing statistical power and enhancing model training. The detection of outliers was carried out using the local outlier factor, an unsupervised method for identifying anomalies [21], as shown in Figure 1. These steps are essential to ensure the quality and integrity of the data before proceeding to further analysis.



Source: (Research Result, 2024)

Figure 1. Outlier Results

The Decision Tree served as the base classification model, achieving 82% accuracy, 83% precision, 82% recall, and an F1-score of 82%.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i3.5972

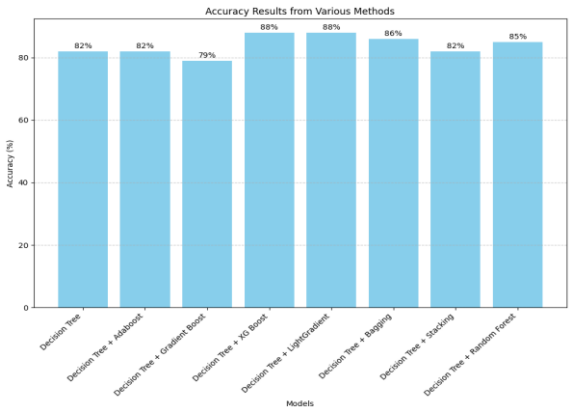
These metrics highlight its balance in performance, particularly its interpretability and ability to handle both categorical and numerical data. However, Decision Trees are prone to overfitting, particularly in datasets with complex interactions. To overcome these limitations, advanced ensemble techniques such as AdaBoost, Gradient Boosting, XGBoost, LightGradient, Bagging, Stacking, and Random Forest were applied. The performance comparison of classification between the original data and after applying ensemble methods can be seen in Tables 3, 4, 5, and 6.

Table 3. Classification Model Accuracy Results

Models	Accuracy
Decision Tree	82%
Decision Tree + adaboost	82%
Decision Tree + Gradient Boost	79%
Decision Tree + XG Boost	88%
Decision Tree + LightGradient	88%
Decision Tree + Bagging	86%
Decision Tree + Stacking	82%
Decision Tree + Random Forest	85%

Source: (Research Result, 2024)

Table 3 presents the results of comparing the performance of the basic Decision Tree model with various ensemble techniques. The combination of Decision Tree with AdaBoost achieved an accuracy of 82%, the same as the base model. Using Gradient Boosting reduced the accuracy to 79%, while XGBoost and LightGBM significantly improved the performance, reaching 88%. Bagging and Random Forest achieved accuracies of 86% and 85%, respectively. Finally, the Stacking method resulted in an accuracy of 82%. These results indicate that certain ensemble methods, such as XGBoost and LightGBM, are more effective in enhancing model performance than others. Comparison of classification accuracy values for the E.coli dataset using various models is visualized Fn figure 2.



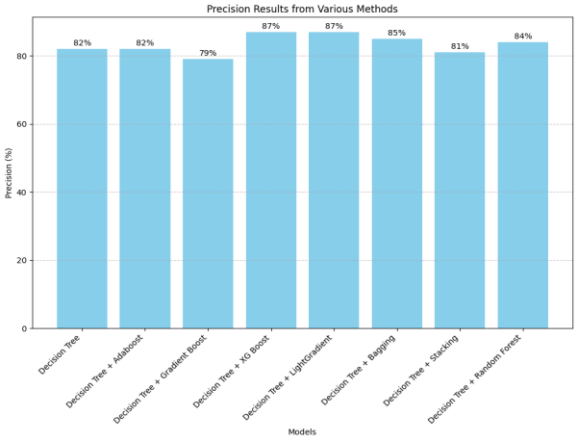
Source: (Research Result, 2024)
Figure 2. Accuracy Results

Table 4. Classification Model Precision Results

Models	Precision
Decision Tree	83%
Decision Tree + adaboost	82%
Decision Tree + Gradient Boost	79%
Decision Tree + XG Boost	87%
Decision Tree + LightGradient	87%
Decision Tree + Bagging	85%
Decision Tree + Stacking	81%
Decision Tree + Random Forest	84%

Source: (Research Result, 2024)

Table 4 presents the precision results of various classification models. The base Decision Tree model achieved a precision of 83%. When combined with AdaBoost, the precision slightly decreased to 82%. The Decision Tree with Gradient Boosting had the lowest precision at 79%, while the combination with XGBoost and LightGradient achieved the highest precision of 87%. Bagging resulted in a precision of 85%, and Stacking showed a precision of 81%. Finally, the Decision Tree with Random Forest yielded a precision of 84%. These results indicate that XGBoost and LightGradient are the most effective ensemble techniques for improving precision. Comparison of classification precision values for the E.coli dataset using various models is visualized in Figure 3.



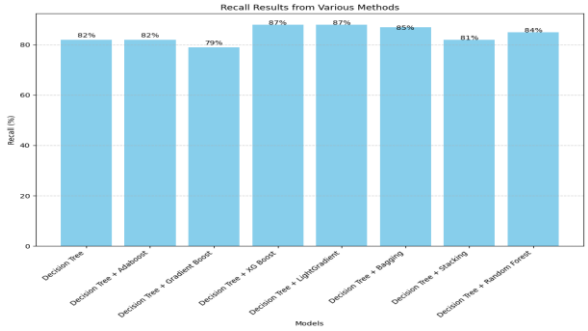
Source: (Research Result, 2024)
Figure 3. Precision Results

Table 5. Classification Model Recall Results

Models	Recall
Decision Tree	82%
Decision Tree + adaboost	82%
Decision Tree + Gradient Boost	79%
Decision Tree + XG Boost	88%
Decision Tree + LightGradient	88%
Decision Tree + Bagging	87%
Decision Tree + Stacking	82%
Decision Tree + Random Forest	85%

Source: (Research Result, 2024)

Table 5 presents the recall results of various classification models. The base Decision Tree model achieved a recall of 82%. When combined with AdaBoost, the recall remained the same at 82%. The Decision Tree with Gradient Boosting showed a recall of 79%. However, the combinations with XGBoost and LightGradient achieved the highest recall, both at 88%. Bagging resulted in a recall of 87%, while Stacking showed a recall of 82%. Finally, the Decision Tree with Random Forest achieved a recall of 85%. These results suggest that XGBoost and LightGradient are the most effective ensemble methods for improving recall. Comparison of classification recall values for the E.coli dataset using various models is visualized in Figure 4.



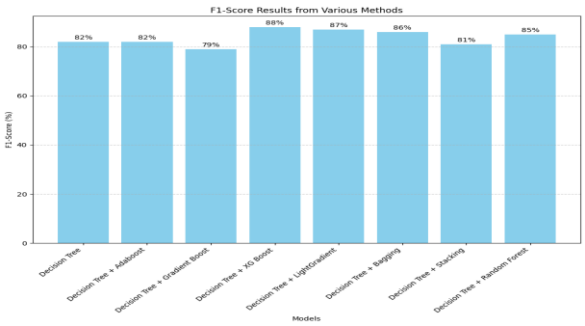
Source: (Research Result, 2024)
Figure 4. Recall Results

Table 6. Classification Model F1-Score Results

Models	F1-score
Decision Tree	82%
Decision Tree + adaboost	82%
Decision Tree + Gradient Boost	79%
Decision Tree + XG Boost	88%
Decision Tree + LightGradient	87%
Decision Tree + Bagging	86%
Decision Tree + Stacking	81%
Decision Tree + Random Forest	85%

Source: (Research Result, 2024)

Table 6 presents the F1-score results of various classification models. The base Decision Tree model achieved an F1-score of 82%. When combined with AdaBoost, the F1-score remained the same at 82%. The Decision Tree with Gradient Boosting showed the lowest F1-score at 79%. The combinations with XGBoost and LightGradient achieved the highest F1-scores, with XGBoost reaching 88% and LightGradient achieving 87%. Bagging resulted in an F1-score of 86%, while Stacking showed an F1-score of 81%. Finally, the Decision Tree with Random Forest yielded an F1-score of 85%. These results indicate that XGBoost and LightGradient are the most effective ensemble methods for improving the F1-score.



Source: (Research Result, 2024)
Figure 5. F1-Score Results

Comparison of classification recall values for the E.coli dataset using various models is visualized in Figure 5.

The performance of XGBoost and LightGradient is largely attributed to their advanced optimization techniques. XGBoost's regularization and handling of class imbalance allow it to reduce overfitting, a frequent challenge in microbiological data analysis. It also captures non-linear relationships and intricate feature interactions, making it ideal for E. coli classification tasks.

Achieving 88% accuracy with XGBoost has important real-world applications, particularly in healthcare and public health. This accuracy can improve microbial detection systems, leading to faster and more reliable identification of pathogens such as Escherichia coli, which is important for early intervention and disease prevention. Increased accuracy in bacterial classification also supports better public health monitoring of microbial threats in food and water. Deeper error analysis revealed misclassification challenges, especially between similar subcellular locations such as 'cp' (cytoplasm) and 'om' (outer membrane). This misclassification likely stems from overlapping features that obscure the distinction between categories. Future models may benefit from feature engineering or adding more biomarkers to better differentiate closely related classes.

CONCLUSION

In conclusion, this study demonstrates the significant potential of ensemble methods, particularly XGBoost and LightGradient, in improving the classification performance of Escherichia coli bacteria. These methods achieved the highest accuracy (88%) and performed exceptionally well across precision, recall, and F1-score metrics, highlighting their robustness and effectiveness in handling complex microbiological datasets. Their ability to model non-linear relationships and manage class imbalances makes

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i3.5972

them highly suitable for tasks requiring precise and reliable bacterial classification.

Moreover, the application of these models has practical implications in healthcare and public health, particularly in enhancing microbial detection and monitoring systems. While the results are promising, future research could explore advanced feature engineering and incorporate additional biomarkers to address misclassification challenges, particularly in distinguishing closely related subcellular locations. This refinement could further improve the applicability of these models in real-world scenarios, supporting faster and more accurate microbial analysis.

REFERENCE

[1] V. J. Harkins, D. A. McAllister, and B. C. Reynolds, "Shiga-Toxin E. coli Hemolytic Uremic Syndrome: Review of Management and Long-term Outcome," *Curr Pediatr Rep*, vol. 8, no. 1, pp. 16–25, Sep. 2024, doi: 10.1007/s40124-020-00208-7.

[2] A. Damena, A. Mikru, M. Adane, and B. Dobo, "Microbial Profile and Safety of Chicken Eggs from a Poultry Farm and Small-Scale Vendors in Hawassa, Southern Ethiopia," *J Food Qual*, vol. 2022, pp. 1–16, Sep. 2024, doi: 10.1155/2022/7483253.

[3] K. Qu, F. Guo, X. Liu, Y. Lin, and Q. Zou, "Application of Machine Learning in Microbiology," *Front Microbiol*, vol. 10, Sep. 2024, doi: 10.3389/fmicb.2019.00827.

[4] A. Masruro, H. Utama, and A. Triyadi, "Kolaborasi Naïve Bayes dan AdaBoost dalam Klasifikasi Bakteri E.coli," *Manajemen dan Teknologi Informasi*, vol. 2, no. 2, 2024, [Online]. Available: https://archive.ics.uci.edu/dataset/39/ecoli.

[5] R. I. Arumnisaa and A. W. Wijayanto, "Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index (HDI)," *SISTEMASI*, vol. 12, no. 1, p. 206, Sep. 2024, doi: 10.32520/stmsi.v12i1.2501.

[6] U. Indahyanti, N. L. Azizah, and H. Setiawan, "Pendekatan Ensemble Learning Untuk Meningkatkan Akurasi Prediksi Kinerja Akademik Mahasiswa," *Jurnal Sains dan Informatika*, vol. 8, no. 2, Sep. 2024, doi: 10.34128/jsi.v8i2.459.

[7] F. Churniansyah and D. W. Utomo, "Teknik Bagging pada Ensemble Learning untuk Kategorisasi Produk E-Commerce," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 10, no. 1, pp. 92–99, Sep. 2024, doi: 10.25077/teknosi.v10i1.2024.92-99.

[8] Eri Mardiani, NurRahmansyah, and Sari Ningsih, "Komparasi Metode Knn, Naive Bayes, Decision Tree, Ensemble, Linear Regression Terhadap Analisis Performa Pelajar Sma," *INNOVATIVE: Journal Of Social Science Research*, vol. 3, pp. 13880–13892, 2023.

[9] Y. Pristyanto, A. Sidauruk, and A. Nurmasani, "Klasifikasi Penyakit Diabetes Pada Imbalanced Class Dataset Menggunakan Algoritme Stacking," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 1, p. 287, Jan. 2022, doi: 10.30865/mib.v6i1.3442.

[10] B. N. Azmi, A. Hermawan, and D. Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 4, pp. 281–290, doi: 10.35746/jtim.v4i4.298.

[11] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *Informatik : Jurnal Ilmu Komputer*, vol. 18, no. 3, p. 239, doi: 10.52958/iftk.v18i3.4694.

[12] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, doi: 10.38094/jastt20165.

[13] A. K. Wardhani, E. Nugraha, and Q. Ulfiana, "Optimization of the Decision Tree Method using Pruning on Liver Disease Classification," *Journal of Applied Informatics and Computing*, vol. 6, no. 2, pp. 136–140, doi: 10.30871/jaic.v6i2.4350.

[14] E. A. Guna, M. D. D. Ghifary, E. F. Sihombing, and A. P. Datubara, "Implementasi Algoritma Decision Tree untuk Klasifikasi Data Evaluation Car Menggunakan Python," *Jusiik*, vol. 1, no. 4, [Online]. Available: https://doi.org/10.59581/jusiik-widyakarya.v1i4

[15] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[16] B. Chen, Q. Chen, and P. Ye, "Information-based massive data retrieval method based on distributed decision tree algorithm," *International Journal of Modeling, Simulation, and Scientific Computing*, vol. 14, no. 01, p. 2243002, 2023, doi: 10.1142/S1793962322430024.

[17] N. Nurussakinah and M. Faisal, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree," *Jurnal Informatika*, vol. 10, no. 2, pp. 143–149, Oct. 2023, doi: 10.31294/inf.v10i2.15989.

[18] E. Fauziningrum and E. I. Sulistyaningsih, "PENERAPAN DATA MINING METODE DECISION TREE UNTUK MENGUKUR PENGUASAAN BAHASA INGGRIS MARITIM (STUDI KASUS DI UNIVERSITAS MARITIM AMNI)," *JURNAL SAINS DAN TEKNOLOGI MARITIM*, vol. 22, no. 1, p. 41, Sep. 2021, doi: 10.33556/jstm.v22i1.285.

[19] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Res Notes*, vol. 15, no. 1, p. 210, 2022, doi: 10.1186/s13104-022-06096-y.

[20] M. M. Sugiman and H. D. Purnomo, "Prediksi Kegagalan Transformator Daya dengan Metode DGA (Dissolved Gas Analysis) Menggunakan Random Forest Berbasis TDCG," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 1, p. 441, Jan. 2024, doi: 10.30865/mib.v8i1.7036.

[21] F. Farhangi, "Investigating the role of data preprocessing, hyperparameters tuning, and type of machine learning algorithm in the improvement of drowsy EEG signal modeling," *Intelligent Systems with Applications*, vol. 15, p. 200100, 2022, doi: https://doi.org/10.1016/j.iswa.2022.200100.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri