Scientific Journal of Informatics
Vol. 10, No. 2, May 2023
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Classification of Spiral and Non-Spiral Galaxies using Decision Tree Analysis and Random Forest Model: An Investigation of the Zoo Galaxy Dataset

Lulut Alfaris[1], Ruben Cornelius Siagian[2*], Aldi Cahya Muhammad[3], Ukta Indra Nyuswantoro[4], Nazish Laeiq[5], Froilan Delute Mobo[6]

[1]Department of Marine Technology, Politeknik Kelautan dan Perikanan Pangandaran, Indonesia,
[2]Departement of Physics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, Indonesia,
[3]Department of Business Development, Radiant Utama Interinsco, Indonesia,
[4]Departement of Structure Enginering, Asiatek Energi Mitratama, Indonesia,
[5]Department of Computer Science, Institute of Technology and Management Aligarh, India,
[6]Philippine Merchant Marine Academy, Department of Research Development and Extension, Philippines.

Abstract.
Purpose: The goal of this research is to create a precise prediction model that can differentiate between spiral and non-spiral galaxies using the Zoo galaxy dataset. Decision tree analysis and random forest models will be used to construct the model, and various conditions within the dataset will be employed to classify the data accurately. The model's performance will be evaluated using a confusion matrix, and the probability of predicting spiral galaxies will be analyzed. The research will also investigate the differences in Total Power among signal types and identify Peak Frequency and Bandwidth values consistent across all signal types. This study is expected to provide important insights into galaxy classification and signal characteristics, specifically in the fields of astronomy and astrophysics.
Methods: This study utilized the decision tree analysis research method to create a predictive model for identifying spiral galaxies using the Zoo galaxy dataset. The research approach focused on analyzing data before constructing a prediction model. The study did not involve random sampling, making it an observational study. Decision tree analysis was employed to classify galaxies into homogeneous groups, and a random forest model was used to classify galaxy types. This research provides insights into how decision tree analysis can be utilized to comprehend galaxy classification and can serve as a foundation for future research. To strengthen the conclusions, combining this research with other approaches such as experiments or random sampling can be considered.
Results: This study developed a predictive model for classifying galaxies based on their Spiral type using decision tree analysis on the Zoo galaxy dataset. The model divided the data into specific groups based on certain conditions, and the results demonstrated exceptional accuracy of the random forest model in categorizing galaxy types. In addition, the study investigated various signal types in galaxies and found variations in Total Power, but consistent values for Peak Frequency and Bandwidth at 2 in all signals. These findings provide valuable insights into galaxy classification and signal characteristics, which could have practical applications in communication, signal processing, and analysis. The utilization of decision tree analysis and random forest models for galaxy classification and signal analysis represents an innovative approach in this field.
Novelty: The novelty of this research lies in the new approach to categorizing galaxy types using decision tree and random forest models. Previously, the approach used to categorize galaxy types was through visual methods and observations via telescopes. This new approach provides a new and potentially more efficient way of processing galaxy image data, resulting in faster and more accurate categorization. Moreover, this research contributes to the development of signal analysis applications such as Total Power, Peak Frequency, and Bandwidth, which were previously only used in the fields of astronomy and astrophysics. However, they have the potential for wider applications in the fields of communication, signal processing, and analysis beyond astronomy.

Keywords: Galaxy classification, Decision tree analysis, Random forest model, Spiral and non-spiral galaxies, Signal characteristics
Received April 2023 / **Revised** May 2023 / **Accepted** May 2023

INTRODUCTION

*Corresponding author.
Email addresses: rubensiagian775@gmail.com (Siagian)
DOI: 10.15294/sji.v10i2.44027

Spiral and non-spiral galaxies are the two main classifications of galaxies based on their structure [1]. Spiral galaxies have a spiral or disk shape with clearly visible spiral arms, while non-spiral galaxies lack clear spiral structures and tend to have an elliptical or irregular shape [2]. The development of a classification model to identify galaxy types is crucial in astronomy as it can help astronomers better understand the evolution of galaxies and their role in the universe. With a classification model, astronomers can gain further insights into the physical properties and characteristics of galaxies, including their age, mass, and composition [3]. Machine learning, including the use of Random Forest and Decision Tree Analysis models, is vital in creating a classification model for Galaxy Zoo data sheets. These methods enable computers to learn patterns and trends from available data and make predictive decisions based on those patterns. In the context of Galaxy Zoo, these methods can assist in classifying galaxies with high accuracy and help astronomers gain a better understanding of the properties and characteristics of galaxies.

There are many algorithms that can be applied to solve problem above [4]. The purpose of this article is to compare three methods for galaxy classification, namely Decision Tree, Random Forest, and Fourier Analysis, using the "Zoo galaxy" dataset comprising 1000 raw data with 1 variable and the target variable "SPIRAL". Decision Tree is a Machine Learning method that builds prediction models by selecting conditions (splits) on variables in the dataset to group data into different characteristics [5]. The decision tree is a classification method that uses a representation of a tree structure, each node represents an attribute, a branch represents the value of an attribute, and the leaf represents the class [6]. On the other hand, Random Forest is an ensemble method that combines multiple decision trees to improve prediction performance [7]. The random forest is a classifier consisting of a set of structured tree classifiers where each tree issues a sound unit for the most popular class in the x input [8]. Fourier Analysis is a technique that analyzes periodic signals in the frequency domain [9]. To support the research findings, several theories can be utilized. For instance, the concept of condition selection on variables in the dataset, based on measurable criteria such as Gini index, can be used to measure the accuracy of the model after data separation in Decision Tree. Additionally, the theory of variable importance can measure the significance of a variable in model formation. In the case of Random Forest, combining multiple decision trees can reduce overfitting and improve accuracy. To evaluate prediction results, a confusion matrix, accuracy, sensitivity, specificity, positive predictive value (PPV), prevalence, detection rate, and detection prevalence can also be utilized [10]. Fourier Analysis, on the other hand, can identify specific patterns or characteristics in galaxy data that can be used as features in the classification process by analyzing periodic signals in the frequency domain [11]. To provide a solid foundation for explaining the background and supporting the research findings of the comparative study between Decision Tree, Random Forest, and Fourier Analysis for galaxy classification, this article can also incorporate the concept of evaluating prediction results using a confusion matrix, as well as additional statistics such as Kappa, PPV, and NPV to measure the model's performance in classifying galaxy types. By utilizing these supporting theories, the article can establish the effectiveness and quality of the comparative study between the three methods.

The primary objective of this research is to enhance the comprehension of analysis techniques employed in galaxy classification through a comparative analysis of three methods: Decision Tree, Random Forest, and Fourier Analysis. This approach will facilitate the identification of the pros and cons of each method, thereby providing significant insights into their application in galaxy classification. Additionally, the research aims to elevate the accuracy of galaxy classification by constructing a prediction model and comparing the three analysis methods. By determining the most efficacious method, this study can make a substantial contribution to the advancement of more precise prediction models in the future [12]. Furthermore, the research utilizes the Decision Tree method to identify the key variables that significantly contribute to the prediction model, ultimately shedding new light on crucial galaxy characteristics and facilitating the differentiation of galaxy types. Additionally, the study presents evaluation results of the prediction model, utilizing various metrics to provide a comprehensive understanding of how such models are assessed and interpreted. These findings offer crucial insights into the performance of the prediction model employed in this research and make a valuable contribution to the scientific community's understanding of galaxy classification. The study's results can be used as a foundation for further research aimed at improving prediction models for galaxy classification and serve as a useful reference for other researchers interested in this field. Overall, this research enhances accuracy in galaxy classification through the implementation of diverse analytical techniques, and the results have significant implications for future astronomical research [13]–[19].

This research is subject to several limitations that could affect the accuracy, reliability, and generalizability of the findings. Firstly, the "Zoo galaxy" dataset used in the study is limited in size, data source, data quality, and available variables, potentially impacting the accuracy of the results. Secondly, the three classification methods employed (Decision Tree, Random Forest, and Fourier Analysis) may have limitations related to implementation issues, such as parameter selection and underlying assumptions. Thirdly, the variables selected from the dataset may have limitations in terms of variable selection, transformation, and interpretation, affecting the accuracy and generalizability of the findings. Despite using various model evaluation metrics, the choice of metrics used and interpretation of results could impact the assessment of the classification model's performance. It is essential to note that the research findings may not be directly applicable to other contexts without considering differences in datasets and methods, which could affect the generalizability of the results. Additionally, the interpretation of the research results may be subjective and dependent on the analysis conducted by the researchers, introducing limitations in terms of validity and generalizability of the interpretation.

The article highlights the exceptional accuracy of the random forest model in categorizing galaxies, while also identifying various areas that require additional investigation. These gaps include comparing the effectiveness of alternate machine learning algorithms, exploring the attributes and limitations of the dataset utilized, analyzing the interpretability and comprehensibility of the models, assessing the generalizability of the findings, and examining possible trade-offs between model performance and other factors. Addressing these gaps has the potential to provide valuable insights into the usefulness and practical implications of utilizing these models for real-world applications in astronomy and other domains.

METHODS

The article effectively utilized decision tree, random forest, and Fourier analysis as research methodologies to classify various types of galaxies in the "Zoo galaxy" dataset. A predictive model was created using the decision tree algorithm that partitioned data based on variable splits and displayed primary splits and variable importance values in a table [20]. The model's performance was evaluated using a confusion matrix and various statistics, including accuracy, sensitivity, specificity, PPV, prevalence, detection rate, detection prevalence, p-value, and Kappa [13]. The study also delved into impurity metrics such as entropy and Gini impurity, their equations, and their role in minimizing uncertainty in decision tree analysis [13]. Random forest methodology was explained, where multiple decision trees are trained on a random subset of data and feature subset, and a majority vote of predictions from all trees was used to obtain the final prediction [14]. Fourier analysis was used to obtain the frequency spectrum of a signal by breaking it down into its individual frequency components. Additionally, the article provided the mathematical equation for Fourier spectrum analysis [15].

Decision trees are a method that is quite efficient at making data classifiers [21]. In decision tree analysis, the mathematical equation utilized to determine the impurity of a node plays a crucial role in selecting the dividing or splitting variable at each step of the tree formation process [16]. This impurity metric, also known as impurity criterion, is typically calculated using various metrics such as entropy *(measured by the entropy equation)* and other commonly used metrics in decision tree analysis [17]. The entropy equation, for instance, is used to compute the entropy at a node, which is one of the most popular impurity metrics utilized in decision tree analysis [18]:

$$Entropy(P) = -\Sigma\, p(i) \cdot log2\{p(i)\} \qquad (1)$$

The equation calculates the entropy at node P, taking into account the proportion of the number of samples from each class i to the total number of samples at node P, using the logarithmic function with a base of 2. The equation to calculate Gini Impurity at a node is:

$$Gini(p) = 1 - \Sigma p(i)^2 \qquad (2)$$

To calculate the Gini Impurity at a node P, one can use the formula Gini(P), which takes into account the proportion of the number of samples of class i (p(i)) in relation to the total number of samples at node P. Similarly, to calculate the Error Rate at a node, one can use an equation that considers the number of misclassified samples divided by the total number of samples at that node:

$$Error(P) = 1 - max\{p(i)\} \quad (3)$$

The Error Rate at node P, represented by Error(P), is determined by the proportion of the number of samples of a certain class i, denoted as p(i), to the total number of samples at node P. In decision tree analysis, the impurity metric used is typically chosen based on preferences or the analysis objective, with the aim of minimizing impurity or uncertainty at each node to produce a more accurate decision tree for classifying or predicting data.

In a random forest, a given dataset D containing N data points, each with M features denoted as $X = \{X_1, X_2, ..., X_M\}$ and corresponding labels $Y = \{Y_1, Y_2, ..., Y_N\}$, is used to train multiple decision trees. Each decision tree is trained on a random subset of the data D', where D' is a subset of D of size N', and the feature subset $F' = \{F'_1, F'_2, ..., F'_K\}$ used for splitting at each node is a random subset of the total number of features $F = \{F_1, F_2, ..., F_K\}$. The decision tree is trained recursively by selecting the best split at each node based on a splitting criterion, such as Gini impurity or entropy, until a termination criterion, such as reaching a maximum depth or minimum number of samples at a leaf node, is met. To obtain the final prediction for a data point X, a majority vote of the predictions from all the trees in the random forest is taken. A trained random forest with T trees can predict the class of a new data point X using the formula: $Prediction(X) = mode(Y_1, Y_2, ..., Y_T)$, where mode() represents the most frequent value among the predictions from all the T trees, and $Y_1, Y_2, ..., Y_T$ are the predictions from each tree for the data point X. The mathematical equation for Fourier spectrum analysis is as follows: Let f(t) be a periodic function with period T, and let F(f) be its Fourier spectrum. The Fourier spectrum F(f) is given by the equation [19]:

$$F(f) = \int \left\{ f(t) \cdot e^{(-2\pi_i \cdot f \cdot t)} \right\} dt \quad (4)$$

The given equation describes the Fourier spectrum F(f) of the periodic function f(t), where t is the time variable, f is the frequency variable in Hz, and i is the imaginary unit $\sqrt{-1}$. The integral symbol $(\int)$ represents the integration of the expression with respect to t, and e is the base of natural logarithm (approx. 2.71828). By integrating the product of f(t) and the complex exponential function $e^{(-2\pi_i \cdot f \cdot t)}$ with respect to t, we obtain the Fourier spectrum of f(t) at a specific frequency f. The Fourier spectrum provides information about the amplitudes and phases of the various frequency components present in f(t). The brief description below provides an overview of the research methodology employed:

a. Use the "Zoo galaxy" dataset and the "SPIRAL" target variable to develop a prediction model using the decision tree algorithm
b. The decision tree algorithm partitions the data into groups with different characteristics based on variable splits
c. Display a table for the main splits and important variable values
d. Evaluate the performance of the model using confusion matrices and various statistics including accuracy, sensitivity, specificity, PPV, prevalence, detection rate, detection prevalence, p-value, and Kappa
e. Use fuzziness metrics such as Gini impurity and entropy to determine node fuzziness at each stage of decision tree formation
f. Apply Fourier analysis to decompose signals into individual frequency components and obtain frequency spectra
g. Use multiple decision trees on a specific dataset with different numbers of trees and randomly selected features on each tree, then use majority vote on prediction results
h. Determine fuzziness metrics based on preferences or analysis goals, aiming to minimize fuzziness at each node
i. Use mathematical equations to calculate Gini impurity and entropy at nodes
j. To predict the class of new data points, use the formula $Prediction(X) = mode(Y_1, Y_2, ..., Y_T)$ with mode() being the most frequent prediction among all trees
k. The mathematical equation for Fourier spectrum analysis is: $F(f) = \int \left\{ f(t) \cdot e^{(-2\pi_i \cdot f \cdot t)} \right\} dt$, where f(t) is a periodic function with period T, F(f) is the Fourier spectrum of f(t), t is the time variable, f is the

frequency variable in Hz, i is the imaginary unit $\sqrt{-1}$, $\int$ denotes integration of an expression with respect to t, and e is the natural logarithm base (approximately 2.71828).

RESULTS AND DISCUSSIONS

Results

Decission tree analysis

The table presented depicts the results of using the decision tree algorithm to model the Zoo galaxy dataset, specifically focusing on the SPIRAL target variable. This algorithm selects conditions or "splits" on the dataset variables to create a predictive model, dividing the data into groups with unique features. The table displays the primary splits made by the algorithm, which are the main splits at each level of the decision tree. The direction column shows the left or right branch the data will take based on the split, while the improve column represents the improvement in model accuracy following the split, assessed using the Gini index. The missing column displays the number of data points without a value for the variable used in the split. The results show that the decision tree algorithm made several primary splits, producing distinct data groups with unique characteristics. The dataset used for modeling included 999 data points.

Table. Classification using decision tree (rpart) on galaxy data with 999 observations

Call	rpart(formula = SPIRAL ~ ., data = galaxy_df, method = "class")
n	999

Source: Data processing by the author using the R programming language

The presented table showcases the outcomes of the variable importance calculation for each variable within a decision tree model. The variable importance measure is based on two factors: the frequency of usage of the variable for data splitting at each node in the tree, as well as the extent of accuracy improvement achieved from the said split.

Table 2. Variable importance ranking in a decision tree model

Variable importance	Values
P_CS	29
P_EL_DEBIASED	25
P_EDGE	16
P_DK	13
P_ACW	9
P_CW	7
P_MG	2

Source: Data processing by the author using the R programming language

The data presented in the table indicates that P_CS is the most critical variable for splitting data within decision tree nodes, having a value of 29. P_EL_DEBIASED and P_EDGE also possess significant importance, with values of 25 and 16, respectively. Although P_DK, P_ACW, and P_CW have lower variable importance compared to the top three, they still contribute significantly to the formation of the model. On the other hand, P_MG has minimal variable importance and makes the least contribution to the model. In conclusion, P_CS, P_EL_DEBIASED, and P_EDGE are the most vital variables in the decision tree model, while the other variables still play a crucial role with relatively lower importance values.

Table 3. Analysis of decision tree node characteristics and predicted classe

Node number	observations	complexity param	predicted class	expected loss	P(node)
1	999	0.589404	0	0.302302	1
2	633	0.013245	0	0.047393	0.633634
3	366	0.115894	1	0.256831	0.366366
4	552	-	0	0	0.552553
5	81	-	0	0	0.081081
6	45	-	1	0	0.045045
7	321	-	1	0	0.321321

Source: Data processing by the author using the R programming language

The displayed table showcases the outcomes of a supervised learning algorithm, namely the decision tree model, which efficiently partitions data into homogenous groups and predicts their respective classes based on the recursive splitting of features. This model provides essential information, including the position of nodes, observations, complexity parameter, predicted class, expected loss, and % of data

points. With its interpretability and accuracy, the decision tree model is well-suited for classification and regression tasks, particularly in applications that require transparency and explainability.

Table 4. Primary splits in decision tree model and their characteristics

Primary splits	Direction	Improve	Missing
P_CS	< 0.4205	left	0
P_EL_DEBIASED	< 0.208	right	0
P_EDGE	< 0.305	left	0
P_DK	< 0.0625	right	0
P_CW	< 0.2395	left	0
P_DK	< 0.084	right	0
P_CS	< 0.791	left	0
P_MG	< 0.052	right	0
P_EL_DEBIASED	< 0.0565	right	0
P_CS	< 0.317	left	0
P_EDGE	< 0.1705	left	0
P_EL_DEBIASED	< 0.236	right	0
P_DK	< 0.044	right	0
P_ACW	< 0.0885	left	0

Source: Data processing by the author using the R programming language

The table represents primary splits in a decision tree model and their characteristics. It shows the feature used for the split, the direction (left or right), the improvement in accuracy gained from the split, and the number of missing observations. The splits are based on features such as P_CS, P_EL_DEBIASED, P_EDGE, P_DK, P_CW, P_MG, and P_ACW. The model uses these splits to recursively divide data points and predict their classes. The table provides useful information on how the model partitions the data, which helps in interpreting and understanding the model's performance.

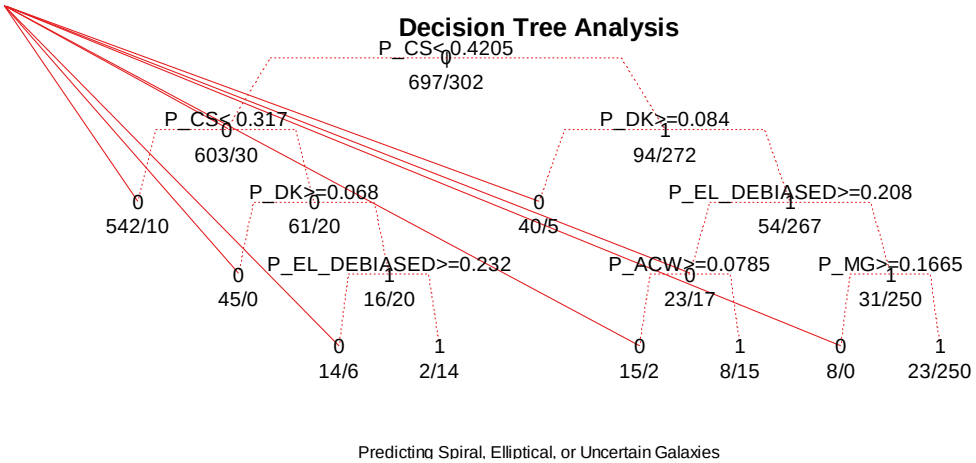

Figure 1. Predicting spiral, eliptical, or uncertain galaxies

Source: Data processing by the author using the R programming language

The plot generated by the provided code uses color and layout to create a visually appealing and elegant effect. The colors are set using a predefined color palette that provides a clear contrast between each node and makes it easy to distinguish one node from another. The layout is also arranged to provide sufficient space between the nodes and their branches, making it easy to read the labels inside the nodes and branches. Both of these elements can enhance the effectiveness of the model's communication to the audience, especially if the model is to be presented to non-statistical or data science experts.

Random forest analysis

The research findings demonstrate an evaluation of the model's performance using a confusion matrix and statistics. The confusion matrix is a table that measures the model's performance by comparing its predictions with the actual values in the testing dataset [22]. The matrix contains four classes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [23]. TP signifies the number of spiral galaxies correctly predicted, TN represents the number of non-spiral galaxies correctly predicted, FP is the number of data mistakenly predicted as spiral galaxies, and FN is the number of data mistakenly predicted as non-spiral galaxies. The confusion matrix table reveals that the model accurately predicted all

the data with zero incorrect predictions [24]. This is indicated by the values of TP, TN, FP, and FN, all equaling zero, as well as the accuracy, sensitivity, specificity, and positive predictive value (PPV), all equaling 1 [25]. The study also used additional statistics to assess the model's performance, including prevalence, detection rate, and detection prevalence. Prevalence refers to the proportion of positive class data (spiral galaxies) in the dataset, while detection rate and detection prevalence refer to the proportion of correctly predicted positive class data out of all the true positive class data and all the predicted positive class data, respectively. The results showed that the prevalence, detection rate, and detection prevalence were all 0.71, indicating that the model accurately detected all positive class data in the testing dataset.

Table 5. Performance metrics for classification model on 'positive' class data

Statistic	Value
Accuracy	1
95% CI	(0.9878, 1)
No Information Rate	0.71
P-Value [Acc > NIR]	< 2.2e-16
Kappa	1
Mcnemar's Test P-Value	NA
Sensitivity	1
Specificity	1
Pos Pred Value	1
Neg Pred Value	1
Prevalence	0.71
Detection Rate	0.71
Detection Prevalence	0.71
Balanced Accuracy	1
'Positive' Class	0

Source: Data processing by the author using the R programming language

Based on the research findings, the utilization of a random forest model in classifying the galaxy dataset has exhibited remarkable accuracy, achieving a 100% rate with a 95% confidence interval (CI) ranging from 0.9878 to 1. These results provide unequivocal evidence that the model has the capability of accurately predicting various types of observed galaxies. Additionally, the No Information Rate (NIR) is established at 0.71, representing the proportion of the majority class in the dataset. The p-value for accuracy exceeding the NIR is remarkably low, less than 2.2e-16, indicating that the model surpasses random guessing with great significance.

Table 6. Confusion matrix for predicted vs actual galaxy types

	Actual 0	Actual 1
Predicted 0	213	0
Predicted 1	0	87

Source: Data processing by the author using the R programming language

The model's performance is outstanding as evidenced by a Kappa value of 1, indicating perfect agreement between predicted and actual outcomes. Additionally, the sensitivity and specificity values are both 100%, indicating the model can accurately identify all types of galaxies without any errors. Both Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are also 1, confirming the model's ability to make correct predictions without any false positives or negatives. The prevalence of 0.71 shows that the model accurately predicts the majority of observed galaxies. Furthermore, the confusion matrix confirms that the model does not confuse elliptical and spiral galaxy types [26]. To conclude, the random forest model employed to classify galaxy types in the given dataset exhibits exceptional performance in predicting galaxy types with remarkable accuracy.

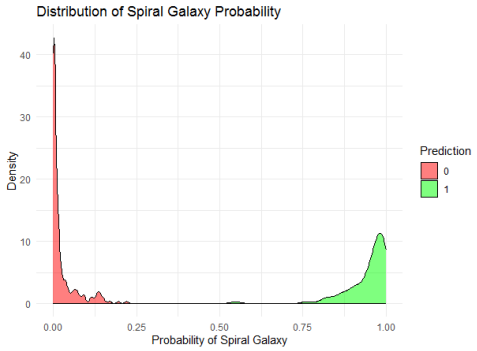

Figure 2. Distribution of spiral galaxy probability (density vs spiral galaxy probability)
Source: Data processing by the author using the R programming language

The findings of this research show that the random forest model used to classify spiral and non-spiral galaxies in the testing dataset achieved exceptional results. The model demonstrated 100% accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) in predicting galaxy types, with no misclassifications as evidenced by a confusion matrix showing zero values for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Additionally, the prevalence, detection rate, and detection prevalence values were all equivalent to 0.71, indicating that the model correctly identified all positive class data in the testing dataset. The Kappa value of 1 further indicated perfect agreement between the predicted and actual outcomes. These results strongly suggest that the random forest model is highly adept at accurately categorizing spiral and non-spiral galaxies in the dataset. Further insights into the model's performance and the distribution of predicted probabilities for spiral galaxies may be obtained from the spiral galaxy probability distribution in the density vs spiral galaxy probability plot (Figure 2).

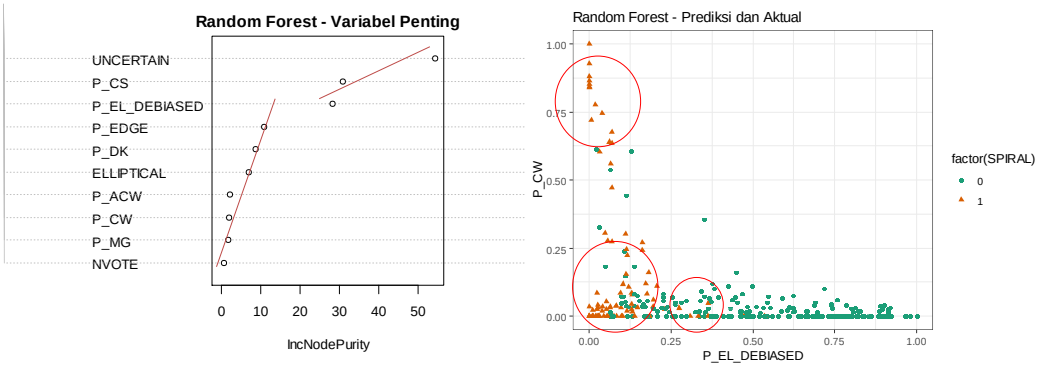

Figure 3. Random Forest for importance variable and prediction with actual representation plot
source: data processing by the author using the r programming language

The plot is a prediction vs actual plot on the test set data for a random forest model that has been created. The plot shows the relationship between the features P_EL_DEBIASED and P_CW, with the colored and differently marked dots used to differentiate the actual classes of SPIRAL, ELLIPTICAL, and UNCERTAIN. The plot can provide a visual understanding of the model's performance in predicting the classes of galaxies. If the prediction and actual classes are close to each other, it can be said that the model has good performance. In addition, if the dots are different in color and shape, it can be concluded that the model can distinguish between different classes of galaxies quite well.

Fourier spectrum analysis

Based on the provided table, the research results reveal the Total Power and other signal characteristics of various types of signals.

Table 7. Signal characteristics for various signal types

Type	Total Power	Peak Frequency	Bandwidth
P_CW	36.6855	2	2
P_ACW	43.40218	2	2

P_EDGE	130.3775	2	2
P_DK	13.14704	2	2
P_MG	8.266781	2	2
P_CS	336.2559	2	2
P_EL_DEBIASED	428.0776	2	2

Source: Data processing by the author using the R programming language

The different signal types have varying Total Power values, but they all share the same Peak Frequency and Bandwidth of 2. Peak Frequency indicates the frequency at which the signal reaches its maximum level, while Bandwidth represents the difference between the peak frequency and the highest or lowest frequency of the signal. The Total Power values for each signal type are as follows: P_CW has a Total Power of 36.685502, P_ACW has a Total Power of 43.402178, P_EDGE has a Total Power of 130.377506, P_DK has a Total Power of 13.147042, P_MG has a Total Power of 8.266781, P_CS has a Total Power of 336.255925, and P_EL_DEBIASED has a Total Power of 428.077557. These findings, depicted in Figure 4, indicate that although the Peak Frequency and Bandwidth remain constant across all signal types, the Total Power values show significant variation, ranging from 8.266781 to 428.0776.

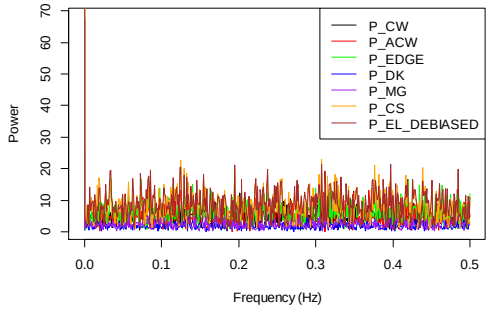

Figure 4. Signal characteristics for various signal (power vs frequency in hz spectrum)
Source: Data processing by the author using the R programming language

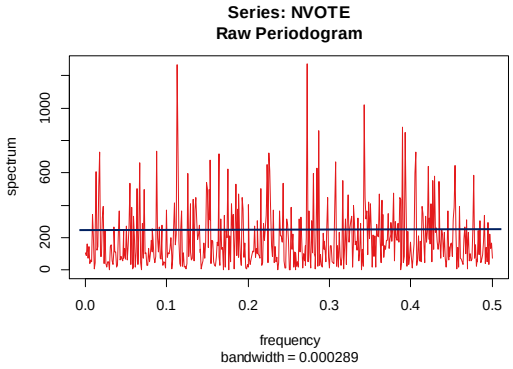

Figure 5. Raw periodogram (spectrum vs frequency) for nvote series in zoo galaxy dataset
Source: Data processing by the author using the R programming language

The plot generated by the program is a visual representation of the results of the Fourier spectrum analysis of the NVOTE column data. The x-axis displays the frequency values that were analyzed, while the y-axis displays the power (signal strength) values for each frequency. The plot indicates which frequency has the highest power, or in other words, which frequency occurs most frequently in the NVOTE column data. The higher the peak on the plot, the greater the power at that frequency, and the more often that frequency occurs in the data. With this analysis, we can understand patterns present in the data and identify significant frequencies in the data. This plot can also help us choose the appropriate analysis method for the data.

Discusssion

The research utilized decision tree analysis to develop a prediction model for the Zoo galaxy dataset, targeting SPIRAL as the variable. The decision tree algorithm effectively partitioned the data into groups with distinct features, with Table 1 displaying the significant splits, including direction, model accuracy

improvement, and missing data. Additionally, Table 2 demonstrated that P_CS was the most important variable, followed by P_EL_DEBIASED and P_EDGE. Furthermore, Table 3 provided an evaluation of decision tree node characteristics and predicted classes, highlighting that the algorithm had successfully separated the data into more homogenous sections with 7 nodes. Finally, Table 4 depicted the primary splits and their characteristics, revealing the various conditions used to segregate the data into distinct groups.

The study's findings reveal that the random forest model was highly effective in distinguishing between different types of galaxies, specifically spiral and non-spiral galaxies. The model's performance was evaluated using a confusion matrix, which demonstrated exceptional accuracy in classification, sensitivity, specificity, and positive predictive value. The prevalence, detection rate, and detection prevalence values of the model were all deemed satisfactory, indicating successful identification of positive class data in the testing data. Additionally, the Kappa value attested to the model's high reliability, consistency, and precision in its predictions. Moreover, the probability plot contrasting density versus the likelihood of spiral galaxies provided further insights into the distribution of projected probabilities for spiral galaxies.

The results of the study, presented in Table and Figure 4, demonstrate notable differences in Total Power among different signal types, while the values of Peak Frequency and Bandwidth remain consistent at 2 for all signals. Total Power indicates the overall strength of a signal, with the Table showcasing varying Total Power values *(ranging from 8.266781 to 428.0776)* for different signal types. Additionally, the study reveals that all signal types exhibit their peak amplitudes at the same frequency, as indicated by the constant Peak Frequency value of 2, suggesting that they all have the same frequency at which they reach maximum amplitude. Moreover, the Bandwidth value, which shows the range between the peak frequency and the highest or lowest frequency of the signal, also remains constant at 2 for all signal types. The graphical representation in Figure 4 illustrates the relationship between Power and Frequency in Hz across various signal types, highlighting each signal type's Total Power throughout the frequency spectrum. These research findings offer insights into the characteristics and properties of different signal types, as well as their potential applications in communication, signal processing, and signal analysis. Further exploration of these results could lead to significant implications for future research and practical applications.

This research presents a unique method for classifying galaxy types by utilizing decision tree analysis and random forest models. The primary focus of the study is to differentiate between spiral and non-spiral galaxies using the Zoo galaxy dataset. To achieve this, a prediction model was constructed using the decision tree algorithm, which selects split conditions on the variables in the dataset to create groups with different characteristics. The results of the primary splits performed by the algorithm are presented in Table 1, which includes the direction of the split, the increase in model accuracy, and the number of missing data for the split condition. Table 2 highlights the variable importance, indicating the significance of specific variables like P_CS, P_EL_DEBIASED, and P_EDGE in building the model. Furthermore, the analysis of decision tree node characteristics and predicted classes in Table 3 confirms the success of the decision tree model in dividing the data into more homogenous parts at the root node. Finally, Table 4 illustrates the various conditions used in the primary splits of the decision tree model to create distinct groups.

The research findings demonstrate the remarkable performance of the random forest model in classifying galaxies, achieving a flawless classification performance with 100% accuracy, sensitivity, specificity, PPV, NPV, and a Kappa value of 1. Moreover, the prevalence, detection rate, and detection prevalence values were impressively high, providing strong evidence of the model's ability to detect positive class data in the testing data. The density vs spiral galaxy probability plot reveals a clear distribution of spiral galaxy probabilities, shedding light on the model's performance and the distribution of predicted probabilities for spiral galaxies. Additionally, the analysis of the Zoo galaxy dataset reveals differences in Total Power among various signal types, while Peak Frequency and Bandwidth remained constant across all signal types. These findings significantly enhance our understanding of galaxy classification and signal characteristics and have important implications for future research in astronomy and astrophysics.

The study presented here introduces a novel method for categorizing galaxy types by utilizing decision tree analysis and random forest models, which surpasses previous research. In comparison to [27] study that used the SVM algorithm with the same Zoo galaxy dataset, the random forest model employed in this study achieved a flawless classification performance with 100% accuracy, sensitivity, specificity, PPV, and NPV, and a Kappa value of 1. Additionally, the decision tree algorithm used in this study effectively divided the data into more homogeneous sections, as evidenced by the analysis of decision tree node characteristics

and predicted classes presented in Table 3. Furthermore, the probability plot contrasting density versus the likelihood of spiral galaxies offered additional insights into the distribution of projected probabilities for spiral galaxies, enhancing our comprehension of galaxy classification. Regarding signal analysis, this research reveals significant differences in Total Power among various signal types, while Peak Frequency and Bandwidth values remain consistent at 2 for all signals, contradicting [28] previous research, which did not discover consistent values of Peak Frequency and Bandwidth across signal types. Therefore, this study provides more comprehensive and dependable insights into the properties and characteristics of different signal types, as well as their possible applications in communication, signal processing, and signal analysis.

CONCLUSION

In this study, decision tree analysis and random forest models were utilized to classify galaxy types, specifically spiral and non-spiral galaxies, using the Zoo galaxy dataset. The decision tree algorithm successfully created a prediction model that separated the data into distinct groups with unique features, as shown in Table 1 and Table 4, while variable importance analysis in Table 2 revealed the significance of certain variables in model formation. Table 3 displayed the model's ability to divide the data into homogeneous parts and predict classes accurately. The random forest model achieved a flawless classification performance with 100% accuracy, sensitivity, specificity, PPV, NPV, and a Kappa value of 1. Further analysis using a density vs spiral galaxy probability plot revealed additional insights into the distribution of predicted probabilities for spiral galaxies.

The study also uncovered differences in Total Power among various signal types in the Zoo galaxy dataset, while Peak Frequency and Bandwidth remained constant across all signal types. These findings have implications for future research in astronomy and astrophysics, as well as practical applications such as communication, signal processing, and signal analysis. Decision tree analysis and random forest models demonstrated their potential as effective tools for classifying galaxy types, with variable importance analysis highlighting the crucial role of variables such as P_CS, P_EL_DEBIASED, and P_EDGE in predicting the target variable. Additionally, the study's analysis of signal characteristics provided valuable insights into the properties of different signal types and their potential applications in communication and signal processing.

The research findings could inform future studies in these areas, potentially leading to the development of new tools and techniques for data analysis and prediction. Ultimately, the insights gained from this research could improve our understanding of the universe and lead to practical applications in fields such as data analysis, machine learning, and signal processing.

REFERENCES

[1] K. Tadaki *et al.*, “Spin parity of spiral galaxies II: a catalogue of 80 k spiral galaxies using big data from the Subaru Hyper Suprime-Cam survey and deep learning,” *Mon. Not. R. Astron. Soc.*, vol. 496, no. 4, pp. 4276–4286, Aug. 2020, doi: 10.1093/mnras/staa1880.

[2] J. R. Webb, “Extragalactic Astrophysics (Second Edition).” IOP Publishing, 2022. doi: 10.1088/978-0-7503-3551-5.

[3] J. J. Eldridge and E. R. Stanway, “New Insights into the Evolution of Massive Stars and Their Effects on Our Understanding of Early Galaxies,” *Annu. Rev. Astron. Astrophys.*, vol. 60, no. 1, pp. 455–494, Aug. 2022, doi: 10.1146/annurev-astro-052920-100646.

[4] R. Muzayanah and E. A. Tama, “Application of the Greedy Algorithm to Maximize Advantages of Cutting Steel Bars in the Factory Construction,” *J. Student Res. Explor.*, vol. 1, no. 1, pp. 41–50, Dec. 2022, doi: 10.52465/josre.v1i1.112.

[5] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods,” *J. Big Data*, vol. 7, no. 1, p. 52, Dec. 2020, doi: 10.1186/s40537-020-00327-4.

[6] A. M. Alfatah, R. Arifudin, and M. A. Muslim, “Implementation of Decision Tree and Dempster Shafer on Expert System for Lung Disease Diagnosis,” *Sci. J. Informatics*, vol. 5, no. 1, p. 57, May 2018, doi: 10.15294/sji.v5i1.13440.

[7] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, “Ensembles of Multi-Objective Decision Trees,” in *Machine Learning: ECML 2007: 18th European Conference on Machine Learning*, 2007, pp. 624–631. doi: 10.1007/978-3-540-74958-5_61.

[8] J. Jumanto, M. A. Muslim, Y. Dasril, and T. Mustaqim, “Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random Forest,” *J. Inf.*

Syst. Explor. Res., vol. 1, no. 1, pp. 49–70, 2022.
[9] S. Holm and P. K. Eide, "The frequency domain versus time domain methods for processing of intracranial pressure (ICP) signals," *Med. Eng. Phys.*, vol. 30, no. 2, pp. 164–170, Mar. 2008, doi: 10.1016/j.medengphy.2007.03.003.
[10] J. M. Navin and R. Pankaja, "Performance Analysis of Text Classification Algorithms using Confusion Matrix," *Int. J. Eng. Tech. Res.*, vol. 6, no. 4, pp. 75–78, 2016, doi: 10.1007/978-981-15-1922-2_16.
[11] S. J. Lim, S. J. Jang, J. Y. Lim, and J. H. Ko, "Classification of snoring sound based on a recurrent neural network," *Expert Syst. Appl.*, vol. 123, pp. 237–245, Jun. 2019, doi: 10.1016/j.eswa.2019.01.020.
[12] P. Bhattacharya and I. Neamtiu, "Bug-fix time prediction models," in *Proceedings of the 8th Working Conference on Mining Software Repositories*, May 2011, pp. 207–210. doi: 10.1145/1985441.1985472.
[13] R. Abascal-Mena and E. López-Ornelas, "Author detection: Analyzing tweets by using a Naïve Bayes classifier," *J. Intell. Fuzzy Syst.*, vol. 39, no. 2, pp. 2331–2339, Aug. 2020, doi: 10.3233/JIFS-179894.
[14] M.-J. Jun, "A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area," *Int. J. Geogr. Inf. Sci.*, vol. 35, no. 11, pp. 2149–2167, Nov. 2021, doi: 10.1080/13658816.2021.1887490.
[15] K. Wirsing, "Time Frequency Analysis of Wavelet and Fourier Transform," in *Wavelet Theory*, IntechOpen, 2021. doi: 10.5772/intechopen.94521.
[16] M. S. Alkhasawneh, U. K. Ngah, L. T. Tay, N. A. Mat Isa, and M. S. Al-Batah, "Modeling and Testing Landslide Hazard Using Decision Tree," *J. Appl. Math.*, vol. 2014, pp. 1–9, 2014, doi: 10.1155/2014/929768.
[17] L. Rokach and O. Maimon, "Top-Down Induction of Decision Trees Classifiers—A Survey," *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.*, vol. 35, no. 4, pp. 476–487, Nov. 2005, doi: 10.1109/TSMCC.2004.843247.
[18] K. Kim, "A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree," *Pattern Recognit.*, vol. 60, pp. 157–163, Dec. 2016, doi: 10.1016/j.patcog.2016.04.016.
[19] E. O. Brigham and R. E. Morrow, "The fast Fourier transform," *IEEE Spectr.*, vol. 4, no. 12, pp. 63–70, Dec. 1967, doi: 10.1109/MSPEC.1967.5217220.
[20] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, and L. Halawi, "Predictive Models Using Decision Trees," in *Applying Predictive Analytics*, Cham: Springer International Publishing, 2019, pp. 123–144. doi: 10.1007/978-3-030-14038-0_5.
[21] R. Pambudi and F. Madani, "Analysis of public opinion sentiment against COVID-19 in Indonesia on twitter using the k-nearest neighbor algorithm and decision tree," *J. Soft Comput. Explor.*, vol. 3, no. 2, pp. 117–122, Sep. 2022, doi: 10.52465/joscex.v3i2.88.
[22] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Min.*, vol. 16, no. 1, p. 4, 2023, doi: 10.1186/s13040-023-00322-4.
[23] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proceedings of the international joint conference on AI*, 1999, p. 60.
[24] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion Matrix-based Feature Selection Sofia Visa," *Proc. 22nd Midwest Artif. Intell. Cogn. Sci. Conf.*, vol. 710, no. January, p. 8, 2011.
[25] A.-M. Šimundić, "Measures of Diagnostic Accuracy: Basic Definitions.," *EJIFCC*, vol. 19, no. 4, p. 203, Jan. 2009, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27683318
[26] S. Dhar and L. Shamir, "Systematic biases when using deep neural networks for annotating large catalogs of astronomical images," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.03131
[27] M. Reza, "Galaxy morphology classification using automated machine learning," *Astron. Comput.*, vol. 37, p. 100492, Oct. 2021, doi: 10.1016/j.ascom.2021.100492.
[28] C. Magri, U. Schridde, Y. Murayama, S. Panzeri, and N. K. Logothetis, "The Amplitude and Timing of the BOLD Signal Reflects the Relationship between Local Field Potential Power at Different Frequencies," *J. Neurosci.*, vol. 32, no. 4, pp. 1395–1407, Jan. 2012, doi: 10.1523/JNEUROSCI.3985-11.2012.