Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

Evaluating the Impact of Model Complexity on the Accuracy of ID3 and Modified ID3: A Case Study of the Max_Depth Parameter

Asrianda*[1], Herman Mawengkang[2], Poltak Sihombing [3], Mahyuddin K. M. Nasution [4]

[1]Informatics, Engineering Faculty, Universitas Malikussaleh, Indonesia
[2,3,4] Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

Email: [1]asrianda@unimal.ac.id

Received : Jun 11, 2025; Revised : Jul 27, 2025; Accepted : Aug 4, 2025; Published : Oct 22, 2025

Abstract

The complexity of decision tree structures has a direct impact on the generalization capability of classification algorithms. This study investigates and evaluates the performance of the classical ID3 algorithm and its modified version in the context of tree depth. The primary objective is to identify the optimal accuracy point and assess the algorithms' robustness against overfitting. Experiments were conducted across tree depths ranging from 1 to 20, with accuracy used as the main evaluation metric. The results indicate that both algorithms achieved peak performance at depth 3, followed by a notable decline. While the classical ID3 algorithm exhibited a gradual decrease in accuracy, the modified ID3 showed a sharp drop and performance stagnation between depths 11 and 20. These findings suggest that the modified ID3 algorithm enhances sensitivity in selecting informative attributes but also increases the risk of overfitting in the absence of structural regularization mechanisms. Therefore, the study recommends the implementation of regularization strategies such as pruning and cross-validation to mitigate performance degradation caused by model complexity. This research not only contributes to the theoretical understanding of how tree depth influences classification performance but also offers practical insights for developing adaptive, stable, and accurate decision tree-based classification systems.

Keywords: *accuracy, classical ID3, decision tree, modified ID3, overfitting, performance, tree depth*

1. INTRODUCTION

The ID3 and C4.5 decision tree algorithms are widely utilized in data classification tasks. ID3 relies on information gain as the basis for selecting the most appropriate attribute at each decision node [1], while C4.5 improves upon this mechanism by employing the gain ratio to mitigate attribute selection bias [2] . Both algorithms provide flexibility in constructing efficient decision tree structures. In response to the increasing demand for advanced data analysis across various domains, several derivative methods have emerged, incorporating novel splitting criteria and intelligent search strategies. Metaheuristic-based approaches, in particular, have demonstrated improvements in both accuracy and computational efficiency [3], [4].

The development of decision tree algorithms has increasingly focused on managing high data complexity. Innovations in attribute splitting strategies, combined with metaheuristic search techniques, have enabled adaptive capabilities to identify complex and heterogeneous data patterns [5]. The integration of these approaches facilitates the construction of models that are not only accurate but also computationally efficient in handling large-scale and imbalanced datasets, making them highly applicable to modern data-driven challenges.

In constructing a decision tree, ID3 employs a greedy top-down search strategy in which the attribute with the highest information gain is selected at each step [6], [7]. This process is designed to

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

generate an optimal tree structure by efficiently mapping relationships between independent variables and the target attribute. The popularity of ID3 in rule-based classification stems from its empirical effectiveness and its ability to process numerical data reliably [8].

However, both ID3 and C4.5 exhibit several limitations that necessitate further investigation. One of the main drawbacks of ID3 is its tendency to favor attributes with a large number of distinct categories, leading to biased tree construction [9]. Although C4.5 addresses this issue through an improved splitting criterion, it does so at the cost of increased computational complexity [10]. Moreover, both algorithms are not optimally designed to handle imbalanced data, large datasets, missing values, and overfitting issues [11], [12]. Previous studies have shown that C4.5 can outperform ID3 under specific conditions, particularly when the number of attributes used is held constant [10].

To address these challenges, this study proposes a modification of the Shannon entropy formula within the ID3 framework. The proposed modification aims to reduce the bias associated with multi-valued attributes and enhance the computational efficiency of information gain calculations on large and imbalanced datasets [13]. By adapting the entropy computation to be more responsive to the underlying data distribution, the modified ID3 algorithm is expected to generate more optimal and efficient decision tree structures, thereby improving its relevance and applicability in modern data analysis scenarios.

2. METHOD

This study implements a modification of the ID3 algorithm to enhance classification effectiveness in complex and imbalanced datasets. ID3 is widely adopted due to its capability to handle nominal attributes and its attribute selection mechanism based on information gain [14]. However, the greedy approach inherent in ID3 often leads the algorithm to local optima [15] , exhibits limitations in managing large-scale datasets [16], and is prone to overfitting on training data [17].

The modification focuses on reformulating the entropy function to address challenges in attribute selection, particularly when dealing with attributes containing numerous categorical values [18], [19]. The conventional entropy function used in ID3 is considered insufficiently efficient in reducing information uncertainty, especially in datasets with complex structures [20], [21]. Therefore, this research modifies both the entropy and gain calculations to optimize node generation in the decision tree.

The methodological steps include: (1) splitting the dataset into training and testing subsets, (2) computing the initial entropy based on class proportions in the training data, (3) calculating the conditional entropy of each attribute based on its categorical values, (4) computing the gain for each attribute, and (5) selecting the attribute with the highest gain as the splitting node. This process is repeated recursively until all data instances are classified or no attributes remain.

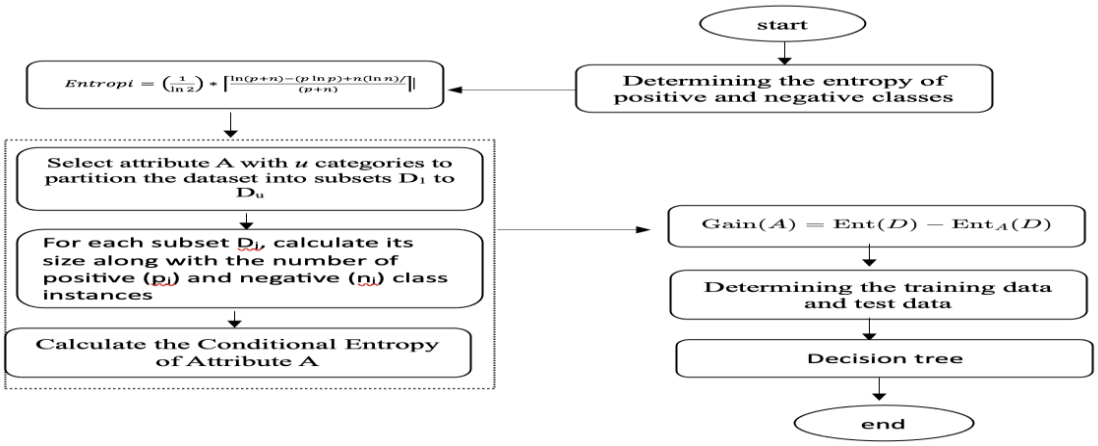

Figure. 1 Research Stages

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

Figure 1 illustrates the methodological framework for computing entropy and information gain derived from the dataset. The process begins by identifying the distribution of positive and negative classes within the dataset. The initial step involves calculating the dataset's entropy to quantify uncertainty based on the class proportions.

A categorical attribute is then selected to partition the dataset into several subsets, each representing distinct attribute values and containing a portion of the original data. Conditional entropy is computed for each subset by considering the distribution of class labels within it.

The information gain is calculated by subtracting the weighted sum of the subset entropies from the original dataset entropy. This metric serves to identify the most informative attribute for data partitioning, aiming to maximize class separation in the construction of the decision tree.

2.1. Modified entropy formula

The ID3 algorithm, introduced by Ross Quinlan, focuses on constructing a decision tree for a given set of objects [23]. The choice of tests is crucial for creating a simple decision tree, and tests are restricted to branching based on attribute values. The test selection depends on identifying the most appropriate attribute to serve as the root of the tree. For example, objects may contain p objects from class P and n objects from class N :

1) An object is assigned to class P with a probability of $\frac{p}{(p+n)}$ to class N with a probability of $\frac{n}{(p+n)}$.
2) The decision tree classifies objects and assigns them to a class based on the tree structure. The following formula describes the entropy [13], [23]:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \quad (1)$$

The second method for simplifying the entropy formula in equation (1) involves the following steps:

1) Using logarithmic identities [24]

$$\log_b(x) = \frac{\ln(x)}{\ln(b)} \quad (2)$$

With b =2

$$log_2\left(\frac{p}{p+n}\right) = \frac{ln\left(\frac{p}{p+n}\right)}{ln(2)}$$

$$log_2\left(\frac{n}{p+n}\right) = \frac{ln\left(\frac{n}{p+n}\right)}{ln(2)}$$

2) Substitute into the entropy formula:

$$\text{Ent(D)} = -\frac{\text{p}}{\text{p}+\text{n}}.\frac{\ln\left(\frac{\text{p}}{\text{p}+\text{n}}\right)}{\ln(2)} - \frac{\text{n}}{\text{p}+\text{n}}.\frac{\ln\left(\frac{\text{n}}{\text{p}+\text{n}}\right)}{\ln(2)}$$

$$Ent(D) = \frac{1}{ln(2)}\left[-\frac{p}{p+n}ln\left(\frac{p}{p+n}\right) - \frac{n}{p+n}ln\left(\frac{n}{p+n}\right)\right]$$

3) Combine terms within the logarithm:

$$-\frac{p}{p+n}\ln\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\ln\left(\frac{n}{p+n}\right)$$

4) Rewrite the entropy terms using the natural logarithm (ln):

$$-\frac{p\,ln(p) + n\,ln(n)}{p+n} + \left(\frac{p\,ln(p+n) + n\,ln(p+n)}{p+n}\right)$$

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

Simplify further:

$$-\frac{p\ln(p)+n\ln(n)}{p+n}+\ln(p+n)$$

$$Ent(D)=\frac{1}{ln(2)}\left[ln(p+n)-\frac{p\,ln(p)+n\,ln(n)}{p+n}\right]$$

5) Final simplified form:

$$Ent(D)=ln(p+n)-\frac{p}{p+n}ln(p)-\frac{n}{p+n}ln(n)$$

Final result:

The simplified entropy formula, using base-2 logarithms, is:

$$Ent(D)=\frac{1}{ln(2)}\left[ln(p+n)-\frac{p\,ln(p)+n\,ln(n)}{p+n}\right] \quad (3)$$

2.2. Modified ID3 Algorithm

The entropy formula Ent(D) is used to measure the irregularity or uncertainty [25] of a dataset D which consists of two classes: p (the number of positive instances) and n (the number of negative instances). This is a modification of Shannon's entropy, integrating the natural logarithm base standardized to base 2. The normalization factor $\frac{1}{\ln 2}$ ensures consistency. The term $\ln(p+n)$ represents the total information in the dataset, while the term $\frac{p\ln(p)+n\ln(n)}{p+n}$ contributes information based on the class proportions.

This formula is relevant for the ID3 algorithm, which uses entropy to identify the best attributes for constructing decision trees. It is particularly useful for datasets with imbalanced class distributions or those requiring smoother logarithmic scaling [11], [25].

The formula remains consistent with the standard entropy formula but differs in the logarithmic form used. Whether binary or natural logarithm (ln) is applied, the final result aligns with the standard entropy calculation.

Information Gain Formula Using Simplified Entropy [13]:

1. Calculate dataset entropy, for dataset D with p and n:

   $$Ent(D)=\frac{1}{ln(2)}\left[ln(p+n)-\frac{p\,ln(p)+n\,ln(n)}{p+n}\right]$$

Where:

- p is the number of positive instances
- n is the number of negative instances, and
- p + n is the total number of instances in the dataset

2. Calculate conditional entropy:

   Attribute A with u distinct values, the dataset D is divided into subsets D_1, D_2, .., D_n

$$Ent_A(D)=\sum_{j=1}^{u}\frac{|D_j|}{|D|}\left[ln(|D_j|)-\frac{p_j}{|D_j|}ln(p_j)-\frac{n_j}{|D_j|}1\,ln(n_j)\right] \quad (4)$$

Where:

- $|D_j|$ is the size of subset Dj

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

- p_j and n_j are the numbers of positive and negative instances in subset D_j, respectively.

3. Calculate Information Gain

   The information Gain, Gain(A) is:

$$Gain(A) = Ent(D) - Ent_A(D) \quad (5)$$

3. RESULT

As online shopping expands, contemporary society faces a "hyper choice" era, marked by an overwhelming variety of online marketplace offerings [26]. Effectively reaching customers with new offers is a significant challenge for businesses. Questions arise regarding how to attract maximum consumer attention toward new products, a concern amplified in the digital era with intense competition and information overload [27]. In the field of digital marketing, phenomena are observed through advertising campaigns aimed at achieving significant conversion rates. Given these dynamics, it becomes essential to explore consumer reactions and responses to marketing initiatives. This is valuable for understanding the mechanisms that influence engagement and decision-making processes.

In this context, the paper focuses on evaluating the impact of decision tree depth on the classification accuracy of models using the ID3 algorithm. The dataset used, Marketing Campaign, contains marketing campaign data with 23 attributes and 1 target label. The process begins with data understanding, which involves collecting and analysing marketing campaign data. The data is gathered and cleaned into a comprehensive dataset by addressing missing values and categorizing continuous data. Potential issues within the data are identified, providing an analytical foundation for subsequent research.

The dataset's target label represents consumer responses to marketing offers, with the variable distribution showing 0 (did not accept the offer) for 1,906 consumers and 1 (accepted the offer) for 334 consumers.

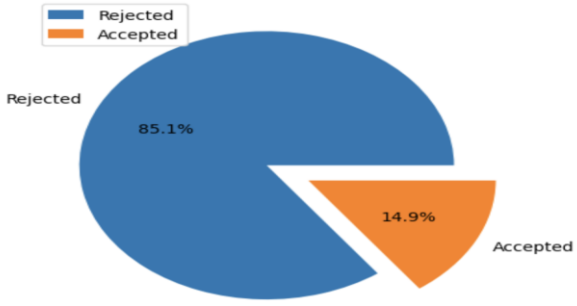

Figure 2. Consumer Response Distribution

As illustrated in Figure 2, there is a significant imbalance between consumers who accepted and those who did not accept the offer, with the proportion of non-accepting consumers being substantially higher. This imbalance must be carefully considered in data analysis and modeling, as it directly affects the performance of classification models. Class balancing approaches or the application of appropriate evaluation metrics can help mitigate this issue. Further research is warranted to explore class balancing techniques and their impact on classification accuracy.

Table 1. Metrics for ID3 and Modified ID3 with a 70:30 Data Split

Metric	ID3		Modified ID3	
	0	1	0	1
Precision	0.90	0.33	0.89	0.30
Recall	0.86	0.41	0.88	0.31
F1-score	0.88	0.37	0.88	0.30

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

The evaluation of the ID3 classification model presented in Table 1, employs precision, recall, and F1-score metrics to assess the model's performance in predicting both positive and negative classes. The precision for class 0 (did not accept the offer) is 0.90, indicating that 90% of all predictions identifying consumers as not accepting the offer were correct. In contrast, for class 1 (accepted the offer), the precision is 0.33, meaning that only 33% of predictions identifying consumers as accepting the offer were accurate. The recall for class 0 is 0.86, indicating that approximately 86% of actual instances where consumers did not accept the offer were correctly predicted by the model. For class 1, the recall is 0.41, showing that 41% of the actual instances of offer acceptance were successfully identified. The F1-score for class 0 is 0.88, reflecting a strong balance between recall and precision. However, the F1-score for class 1 is 0.37, which highlights the relatively low values of both recall and precision, resulting in a suboptimal overall performance for this class.

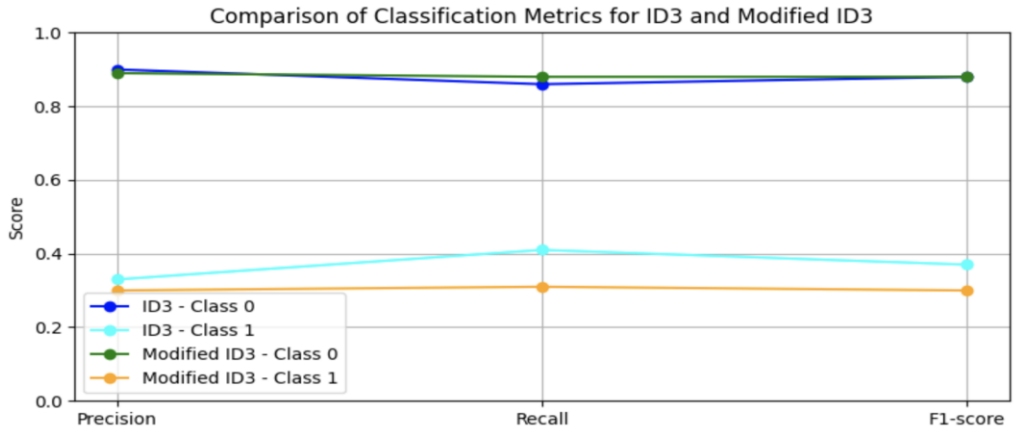

Figure 3. Comparative Evaluation of ID3 and Modified ID3 Classification Performance Metric

Figure 3, illustrates the performance evaluation of the ID3 and Modified ID3 algorithms across each class. The performance for class 0 remains consistently high in both algorithms, with precision and recall values approaching 0.9. This indicates that the models are more effective in classifying consumers who did not accept the offer compared to those who did. A slight improvement in recall and stable F1-score for class 0 in the Modified ID3 model suggests that the modification does not compromise accuracy for the majority class.

In contrast, class 1 performance (offer acceptance) remains relatively low under both approaches. The Modified ID3 model shows a decline in both precision and recall for class 1. This pattern reflects a common challenge in handling imbalanced datasets. The marginal differences in performance metrics between the two algorithms for class 1 highlight the limited improvement achieved. Therefore, future efforts should focus on addressing the minority class more effectively.

Overall, the model performs better in predicting consumers who did not accept the offer compared to those who did. This is due to the class imbalance in the dataset, where the number of consumers who did not accept the offer is significantly higher than those who did. To improve the model's performance on the minority class (class 1), techniques such as class balancing or adjusting prediction thresholds could be considered.

In evaluating the Modified ID3 model, the precision for the negative class is 0.89, meaning that 89% of predictions for class 0 were accurate. Conversely, the precision for class 1 is only 0.30, indicating that only 30% of positive predictions were correct. The recall for class 0 is 0.88, showing that 88% of actual cases of non-acceptance were identified. For class 1, the recall is 0.31, meaning that 31% of cases of acceptance were accurately detected. The F1-score, which represents the harmonic mean of precision and recall, highlights the balance between these two metrics. With a score of 0.88 for class 0 and 0.30

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

for class 1, the model is shown to be effective in classifying non-acceptance cases but struggles with acceptance cases.

These metrics provide valuable insights into model performance, particularly in the context of imbalanced datasets. The lower precision and recall values for class 1 indicate difficulties in identifying cases of offer acceptance, likely due to the smaller sample size in this class or complexity in the data patterns. Additional strategies, such as data balancing or parameter adjustments, should be considered to improve performance for the underrepresented class.

3.1.1. Adapting Model Concepts and Tree Depth

Decision tree algorithms simplify complex relationships between variables and targets by dividing original variables into more meaningful groups. The entropy parameter and max_depth (maximum tree depth) are set within a specific range to construct an optimal decision tree. The entropy parameter is used as a criterion to evaluate the quality of each node in the decision tree. Entropy measures impurity based on the level of information uncertainty and affects the structure and performance of the resulting decision tree.

The max_depth parameter determines the maximum depth of the decision tree. Setting the appropriate max_depth value is crucial to avoid overfitting. A model that is too complex and overly tailored to training data reduces its ability to generalize to new data. By optimizing this parameter, decision trees can be constructed to produce effective models that predict target variables based on input variables.

Table 2. Test Results of Max_depth on Accuracy

max_depth	Accucary	
	ID3	Modified ID3
1	0.8571	0.8571
2	0.8661	0.8661
3	0.8690	0.8690
4	0.8690	0.8542
5	0.8542	0.8408
6	0.8571	0.8289
7	0.8557	0.8199
8	0.8542	0.8140
9	0.8527	0.8095
10	0.8482	0.8065
11	0.8408	0.8021
12	0.8259	0.8021
13	0.8244	0.8021
14	0.8110	0.8021
15	0.8330	0.8021
16	0.8185	0.8021
17	0.8065	0.8021
18	0.8080	0.8021
19	0.8006	0.8021
20	0.7991	0.8021

Table 2 presents the analysis of the effect of the max_depth parameter on the accuracy of the ID3 and Modified ID3 methods shows that increasing the max_depth from 1 to 3 improves the accuracy of both methods. After reaching max_depth 3, the accuracy of ID3 remains stable, while Modified ID3 shows a decrease in accuracy. Specifically, at max_depth 3, ID3 achieves an accuracy of 89.90%, while Modified ID3 reaches 86.90%. At max_depth 4, ID3 remains at 86.90%, while Modified ID3 decreases

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

to 85.42%. This decline continues until max_depth 20, where Modified ID3 stabilizes at an accuracy of 80.21%, while ID3 decreases to 79.91%.

The results indicate that after max_depth 3, ID3 maintains its accuracy, whereas Modified ID3 experiences a decline. Selecting the optimal max_depth value is crucial for achieving the best performance in the Modified ID3 method. The decrease in accuracy for Modified ID3 after max_depth 3 may be due to the increased model complexity, which can lead to overfitting when the model adjusts too closely to the training data, resulting in poor performance on unseen test data. To address this issue, pruning techniques or further reduction of parameters can be applied to limit the model's complexity.

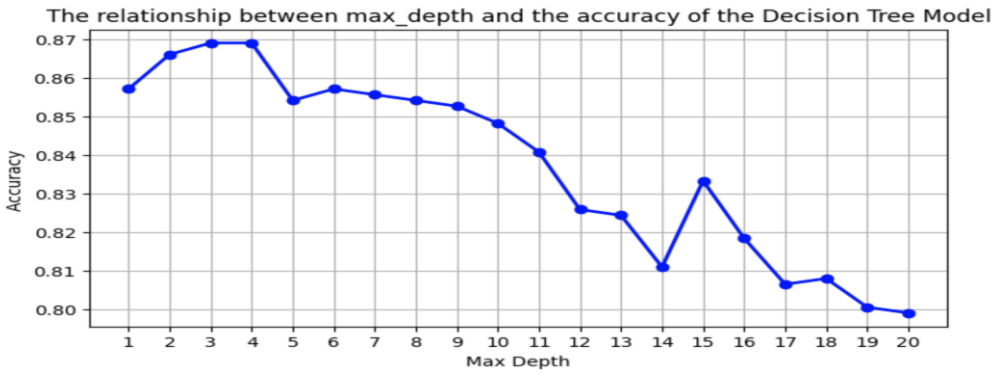

Figure 4. Tree Depth vs. Accuracy for ID3

Figure. 2 shows the relationship between the max_depth (tree depth) of the decision tree model and its accuracy. Initially, the max_depth increases from 1 to 4, with the model's accuracy significantly improving, peaking at around 0.8690. After max_depth =4, the accuracy gradually decreases, indicating that increasing the maximum depth does not further improve the model's performance. The highest accuracy occurs at max_depth=4, and accuracy slowly declines until max_depth=10. The decline becomes steeper after max_depth=10, with minor fluctuations around max_depth=14 and max_depth=16. At max_depth=20, the accuracy reaches its lowest point at around highest accuracy of 86.90%. After max_depth = 3, accuracy gradually decreases, reaching stability at max_depth = 12 with an accuracy of 80.21%. The importance of selecting the right max_depth is highlighted, and the difference in performance between ID3 and Modified ID3 shows that the Modified ID3 algorithm is more sensitive to model complexity than ID3. The significant drop in accuracy for Modified ID3 after max_depth = 3 suggests that the algorithm quickly overfits as the tree depth increases, due to additional adjustments made in Modified ID3 to handle specific features or data conditions. These adjustments can increase the burden on the model if the max_depth parameter is not optimally set. Modified ID3 requires more careful parameter tuning to ensure optimal performance.

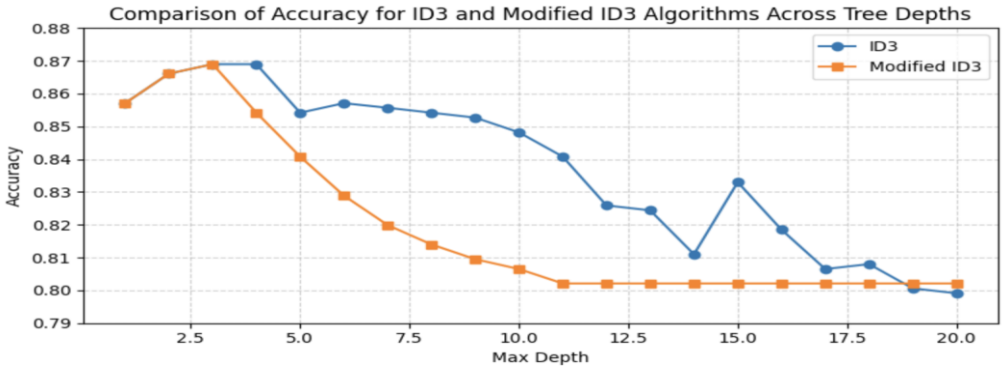

Figure 5. ID3 vs. Modified ID3 Accuracy Over Tree Depth

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

Model accuracy serves as a key indicator for evaluating the performance of classification algorithms with respect to the structural complexity of decision trees. Figure 5, illustrates the performance trajectories of the standard ID3 and the modified ID3 algorithms across varying tree depths, ranging from 1 to 20. In the initial stages, both algorithms exhibit a consistent increase in accuracy as tree depth increases. The peak accuracy for both models is observed at a depth of 3, reaching 0.8690. However, this upward trend does not persist uniformly. Notably, the modified ID3 algorithm experiences a more pronounced performance decline than the classical ID3 variant.

The classical ID3 algorithm shows a gradual decrease in performance following the accuracy peak, yet the fluctuations remain relatively stable and decelerate over deeper tree structures. This indicates a degree of resilience to overfitting, even with increased model complexity. In contrast, the modified ID3 algorithm demonstrates a stagnation in accuracy from depths 11 to 20, consistently plateauing at 0.8021. This pattern suggests that the modifications introduced into the algorithm impose structural constraints that limit model complexity, thereby preventing further increases or decreases in performance.

These findings underscore the importance of balancing flexibility and regularization in the development of tree-based classification algorithms. While the classical ID3 algorithm offers greater exploratory capacity for decision-making, it carries a higher risk of overfitting, particularly at extreme depths. At the same time, it may fail to capture optimal accuracy in shallower tree structures. Therefore, algorithm selection should be aligned with the nature of the data and the specific goals of the implementation.

Overfitting emerges as a critical concern in decision tree modeling, especially when tree depth parameters are not optimally configured. The experimental results reveal that both ID3 and its modified version exhibit accuracy improvements up to a certain depth—specifically around max_depth = 3 or 4. Beyond this threshold, accuracy gradually declines, indicating that the model begins to capture noise or non-generalizable patterns from the training data. This accuracy drop is a hallmark of classical overfitting, where increased structural complexity leads to a diminished capacity to generalize to unseen data.

In the case of the modified ID3 algorithm, the performance degradation is more abrupt than in the classical version, particularly beyond max_depth = 3. This suggests that the modifications aimed at enhancing feature sensitivity may inadvertently accelerate the onset of overfitting if not accompanied by appropriate parameter tuning. The sharper performance decline of the modified ID3 highlights its heightened sensitivity to model complexity. The imbalance between the structural complexity of the tree and the data's ability to support deeper splits emerges as a key factor that undermines the model's generalization capability.

4. DISCUSSIONS

The experimental results consistently demonstrate that decision tree depth exerts a significant impact on the performance of both the standard ID3 and the modified ID3 algorithms. The highest accuracy was achieved at a depth of 3, indicating the presence of an optimal point [28], beyond which model complexity begins to impair generalization. Beyond this point, the performance of both algorithms declined, albeit with differing degradation patterns. The classical ID3 exhibited a gradual and stable decline, suggesting its relative robustness to increasing structural complexity. In contrast, the modified ID3 showed a sharp drop in accuracy, followed by a stagnation phase between depths 11 and 20. This phenomenon reflects increased sensitivity to attribute selection and heightened risk of overfitting when complexity is not effectively controlled.

These findings underscore the critical importance of complexity control strategies in the development of decision tree algorithms, particularly when modifications are introduced to pursue

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

higher precision. An imbalance between model capacity and data structure complexity can lead to models that perform well on training data but fail to maintain accuracy on unseen data [29]. Proper tuning of the maximum depth parameter and the implementation of structural regularization techniques such as pre-pruning or post-pruning become essential to ensure generalization capability. In real-world scenarios involving imbalanced or noisy datasets, the tendency to overfit may further degrade overall classification performance.

The observed stagnation trend in the modified ID3 accuracy suggests that the algorithm enhances attribute selectivity without incorporating adequate limiting mechanisms. Consequently, the resulting decision trees expand inefficiently. Integrating algorithmic modifications with adaptive cross-validation strategies may offer a promising pathway to develop models that are not only accurate but also stable and scalable across diverse data conditions.

5. CONCLUSION

The study reveals that decision tree depth plays a pivotal role in determining the performance of both the standard ID3 and its modified counterpart. Both algorithms achieve optimal accuracy at a depth of three, beyond which a significant risk of overfitting emerges, particularly in the modified ID3. While the classical ID3 exhibits a controlled decline in performance, the modified version undergoes a rapid degradation followed by stagnation at greater depths. This indicates that the modified ID3 increases attribute selection sensitivity, thereby heightening susceptibility to overfitting if not accompanied by appropriate complexity control strategies.

The successful implementation of decision tree classification algorithms depends not only on entropy formulation or attribute selection but also on effective tree structure management and model parameterization. Recommended practices include imposing depth constraints and applying pruning techniques to prevent overfitting, along with the use of cross-validation to verify model performance consistency. This research contributes to the broader understanding of how tree structure dynamics influence generalization, providing a foundation for the development of adaptive and efficient algorithms for real-world classification tasks.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

The author expresses gratitude to all parties who contributed to the successful completion of this journal manuscript.

REFERENCES

[1] Y. Manzali, M. El Far, M. Chahhou, and M. Elmohajir, “Enhancing Weak Nodes in Decision Tree Algorithm Using Data Augmentation,” *Cybernetics and Information Technologies*, vol. 22, no. 2, pp. 50–65, 2022, doi: 10.2478/cait-2022-0016.

[2] M. Jaworski, P. Duda, and L. Rutkowski, “New Splitting Criteria for Decision Trees in Stationary Data Streams,” *IEEE Trans Neural Netw Learn Syst*, vol. 29, no. 6, pp. 2516–2529, 2018, doi: 10.1109/TNNLS.2017.2698204.

[3] R. Rivera-Lopez and J. Canul-Reich, “Construction of near-optimal axis-parallel decision trees using a differential-evolution-based approach,” *IEEE Access*, vol. 6, pp. 5548–5563, 2017, doi: 10.1109/ACCESS.2017.2788700.

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

[4] D. Saraswathi and A. Vijaya, "Search Engine Spam Detection using an Integrated Hybrid Genetic Algorithm based Decision Tree," *Int J Comput Appl*, vol. 133, no. 10, pp. 20–27, 2016, doi: 10.5120/ijca2016908027.

[5] A. Shanbhag, S. Vincent, S. B. B. Gowda, O. P. Kumar, and S. A. J. Francis, "Leveraging Metaheuristics for Feature Selection with Machine Learning Classification for Malicious Packet Detection in Computer Networks," *IEEE Access*, vol. 12, no. February, pp. 21745–21764, 2024, doi: 10.1109/ACCESS.2024.3362246.

[6] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Comput Sci*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[7] F. Umar and N. Ussiph, "Appraisal of the Classification Technique in Data Mining of Student Performance using J48 Decision Tree, K-Nearest Neighbor and Multilayer Perceptron Algorithms," *Int J Comput Appl*, vol. 179, no. 33, pp. 39–46, 2018, doi: 10.5120/ijca2018916751.

[8] W. Lee *et al.*, "Preoperative data-based deep learning model for predicting postoperative survival in pancreatic cancer patients," *International Journal of Surgery*, vol. 105, no. August, p. 106851, 2022, doi: 10.1016/j.ijsu.2022.106851.

[9] I. I. Sinam and A. Lawan, "An improved C4.5 model classification algorithm based on Taylor's series," *Jordanian Journal of Computers and Information Technology*, vol. 5, no. 1, pp. 34–42, 2019, doi: 10.5455/jjcit.71-1546551963.

[10] C. Qiu, L. Jiang, and G. Kong, "A Differential Evolution-Based Method for Class-Imbalanced Cost-Sensitive Learning," pp. 1–8, 2015.

[11] D. P. Rangasamy, S. Rajappan, A. Natarajan, R. Ramasamy, and D. Vijayakumar, "Variable population-sized particle swarm optimization for highly imbalanced dataset classification," *Comput Intell*, vol. 37, no. 2, pp. 913–930, 2021, doi: 10.1111/coin.12436.

[12] Y. Cong, J. Liu, B. Fan, P. Zeng, H. Yu, and J. Luo, "Online Similarity Learning for Big Data with Overfitting," *IEEE Trans Big Data*, vol. 4, no. 1, pp. 78–89, 2017, doi: 10.1109/tbdata.2017.2688360.

[13] Asrianda, H. Mawengkang, P. Sihombing, and M. K. M. Nasution, "OPTIMIZATION OF MARKETING CAMPAIGNS USING A MODIFIED ID3 DECISION TREE ALGORITHM," *Eastern-European Journal of Enterprise Technologies*, vol. 13, no. 2, pp. 58–70, 2025, doi: 10.15587/1729-4061.2025.327158.

[14] J. Yan, Z. Zhang, L. Xie, and Z. Zhu, "A unified framework for decision tree on continuous attributes," *IEEE Access*, vol. 7, pp. 11924–11933, 2019, doi: 10.1109/ACCESS.2019.2892083.

[15] S. Raghuwanshi and R. Ahirwal, "An Efficient Classification based Fuzzy Rough Set Theory using ID3 Algorithm," *Int J Comput Appl*, vol. 154, no. 1, pp. 31–34, 2016, doi: 10.5120/ijca2016912025.

[16] P. Rakshit, A. Ghosh, C. Chakraborty, J. Paul, and D. Das, "Skin Cancer Detection Using Deep Learning," *Lecture Notes in Electrical Engineering*, vol. 1243 LNEE, pp. 359–372, 2025, doi: 10.1007/978-981-97-6465-5_29.

[17] J. Demšar and B. Zupan, "Hands-on training about overfitting," *PLoS Comput Biol*, vol. 17, no. 3, pp. 1–19, 2021, doi: 10.1371/journal.pcbi.1008671.

[18] D. Christianti, S. Abdullah, and S. Nurrohmah, "Bayes Risk Post-Pruning in Decision Tree to Overcome Overfitting Problem on Customer Churn Classification," 2020, doi: 10.4108/eai.2-8-2019.2290487.

[19] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.

Vol. 6, No. 5, October 2025, Page. 3707-3718
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4864

[20] Y. Wang, Y. Li, Y. Song, X. Rong, and S. Zhang, "Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree," *Algorithms*, vol. 10, no. 4, pp. 1–18, 2017, doi: 10.3390/a10040124.

[21] M. Hlosta, Z. Zdrahal, and J. Zendulka, "Are we meeting a deadline? classification goal achievement in time in the presence of imbalanced data," *Knowl Based Syst*, vol. 160, no. June, pp. 278–295, 2018, doi: 10.1016/j.knosys.2018.07.021.

[22] J. R. Quinlan, "Induction of decision trees," *Mach Learn*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/bf00116251.

[23] A. Cornea, "An identity theorem for logarithmic potentials," *Osaka Journal of Mathematics*, vol. 28, no. 4, pp. 829–836, 1991.

[24] G. Maksa, "The stability of the entropy of degree alpha," *J Math Anal Appl*, vol. 346, no. 1, pp. 17–21, 2008, doi: 10.1016/j.jmaa.2008.05.034.

[25] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018*, pp. 1–11, 2018, doi: 10.1109/ICCTCT.2018.8551020.

[26] H. T. K. Le, A. L. Carrel, and H. Shah, "Impacts of online shopping on travel demand: a systematic review," *Transp Rev*, vol. 42, no. 3, pp. 273–295, 2022, doi: 10.1080/01441647.2021.1961917.

[27] C. Panico and C. Cennamo, "User preferences and strategic interactions in platform ecosystems," *Strategic Management Journal*, vol. 43, no. 3, pp. 507–529, 2022, doi: 10.1002/smj.3149.

[28] N. Aslam *et al.*, "Anomaly Detection Using Explainable Random Forest for the Prediction of Undesirable Events in Oil Wells," *Applied Computational Intelligence and Soft Computing*, vol. 2022, 2022, doi: 10.1155/2022/1558381.

[29] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, no. August, p. 100258, 2022, doi: 10.1016/j.array.2022.100258.