INSYST

Journal of Intelligent System and Computation

p-ISSN:2621-9220
e-ISSN:2722-1962

Volume 6 Number 1

April 2024

Published By **Lembaga Penelitan dan Pengabdian Masyarakat (LPPM)**
Institut Sains dan Teknologi Terpadu Surabaya (ISTTS)
formerly **Sekolah Tinggi Teknik Surabaya (STTS)**

Managed By
Departement of Informatics
Institut Sains dan Teknologi Terpadu Surabaya (ISTTS)

INSYST

Journal of Intelligent System and Computation

Volume 06 Nomor 01 April 2024

INSYST

Journal of Intelligent System and Computation

Volume 06 Nomor 01 April 2024

INSYST

Journal of Intelligent System and Computation

Volume 06 Nomor 01 April 2024

INSYST

Journal of Intelligent System and Computation

Volume 06 Nomor 01 April 2024

Daftar Isi

Prediction of Physico-Chemical Characteristics in Batu Tangerine 55 Based on Reflectance-Fluorescence Computer Vision

Safitri D. A. Ariani[1], Inggit K. Maharsih[1], and Dimas F. A. Riza[1]

[1]Bioprocess Engineering Program, Department of Biosystems Engineering, Faculty of Agricultural Technology, Brawijaya University, Malang, Indonesia

Corresponding author: Dimas F. A. Riza (e-mail: dimasfirmanda@ub.ac.id).

ABSTRACT Oranges (Citrus sp.) are one of the most abundant agricultural commodities in Indonesia. One of the popular local citruses is Batu Tangerine 55. Harvesting tangerines begins 252 days after the flowers bloom. Conventionally, we still determine the level of maturity by observing the color, shape, and hardness. The results of manual grouping tend to be subjective and less accurate. Destructive testing could be carried out and provide objective results; however, it would require sampling and damaging the fruits. Computer vision could be used to evaluate the maturity level of the fruit non-destructively. Dual imaging computer vision, i.e., reflectance-fluorescence mode, could be used to enhance the accuracy of the prediction. This study aims to develop a classification model and predict the physico-chemical characteristics of Batu Tangerine 55. Destructive testing is still being carried out to determine the value of TPT, the degree of acidity, and the firmness of the fruit. Non-destructive testing was carried out to obtain reflectance and fluorescence images. Once we obtain the destructive and non-destructive data, we will incorporate them into the classification and prediction models. The machine learning method for maturity classification uses three models, namely KNN, SVM, and Random Forest. The best results on the reflectance data (RGB) SVM model resulted in an accuracy of 1 for training data and 0.97 for testing data. The maturity parameter prediction method uses the PLS method. The best results for the predicted Brix/Acidity ratio R2 parameter are 0.81 and RMSE 3.4.

KEYWORDS Brix/Acid ratio, Machine Learning, PLS, Tangerine

I. INTRODUCTION

Indonesia stands as one of the nations blessed with abundant agricultural commodities, with oranges holding a significant place among them. Citrus fruits, particularly tangerines (*Citrus reticulata* Blanco), enjoy immense popularity among the Indonesian populace. According to data from the Central Statistics Agency (2021), Indonesia produces approximately 2,401,064 tons of tangerines annually, with East Java Province emerging as the largest contributor, accounting for 822,260 tons per year. Among the varieties under cultivation, Batu Tangerine 55 has garnered attention for its superior quality, characterized by sweet, slightly sour, and refreshing fruit flesh [1]. These tangerines typically reach harvesting maturity 252 days after flowering, necessitating careful post-harvest handling to minimize product damage during marketing.

Assessing fruit maturity holds pivotal importance in the marketing of citrus fruits in Indonesia, significantly influencing consumer preferences. Presently, the method of determining maturity levels remains predominantly conventional. Typically, farmers gauge the maturity and physical attributes of citrus fruits by visually inspecting factors such as color, shape, and hardness [2]. The process of categorizing fruit ripeness is predominantly manual, leading to subjective and often less precise results [3].

An alternative method for determining fruit maturity involves destructive testing, which entails assessing the total dissolved solids (TDS) and acidity levels in citrus fruits [4]. However, this approach has drawbacks, as it can physically damage the fruit. Hence, there is a pressing need to predict and classify tangerine maturity without causing physical harm, utilizing non-destructive testing methods, such as digital imaging [5]. Nevertheless, the accuracy of classification using conventional computer vision systems remains inadequate, necessitating the exploration of improved methodologies to enhance model accuracy [6]. Previous studies have employed computer vision in dual reflectance-fluorescence mode, enhancing predictive models by incorporating additional features from fluorescence imagery [7]. However, these studies often rely on deep learning models, which demand substantial amounts of data. Alternatively, simpler machine learning models could achieve comparable performance with smaller datasets.

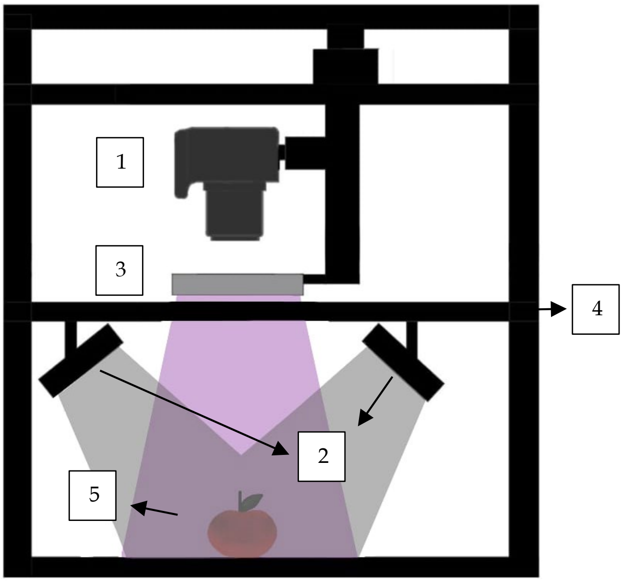
This study endeavors to develop and refine a predictive model for the physicochemical characteristics of Batu Tangerine 55 across three maturity levels. Both destructive and non-destructive testing methodologies are employed to facilitate sorting and grading processes. Destructive testing determines TDS, acidity, and fruit hardness, while non-destructive testing employs reflectance-fluorescence dual-vision computer systems. Subsequently, the data obtained from both methods are integrated into the classification and prediction model. Machine learning techniques, including KNN, SVM, and Random Forest, are applied for maturity classification. The comparative analysis of these models will reveal the most effective approach for accurately classifying Batu Tangerine 55 maturity levels.

II. RESEARCH METHODS

The experiment encompassed both non-destructive and destructive testing methodologies. Non-destructive testing involved capturing images using a mini studio setup, as depicted in Figure 1, utilizing a 700D DSLR camera. Samples comprised 55 Batu Tangerines across three ripeness levels: raw, semi-ripe, and ripe, sourced from the Research Institute of Citrus and Subtropical Plants. Image acquisition encompassed top and bottom views, facilitated by two light sources—LEDs for reflectance imaging and UV for fluorescence imaging.

Following image data collection, three non-destructive parameters were measured. First, fruit hardness was assessed using penetrometers at three equatorial points. Subsequently, the sweetness and acidity of fruit juice were gauged using the ATAGO PAL-BX|ACID 101 Brix Acidity meter.

Predictive modeling employed Partial Least-Squares Regression (PLSR) to determine the maturity parameters of Batu Tangerine 55. Widely utilized across various domains including bioinformatics, food research, medicine, and pharmacology [8]. PLSR combines principal component analysis with multiple regression analysis to predict or analyze dependent variables with multiple independent variables. This technique is particularly effective for datasets with high collinearity and numerous variables [9]. PLSR facilitated the extraction of several parameters to evaluate method accuracy, including slope values, R2 offsets, and RMSE values. R2 values indicate the proximity between real values and predictions, with higher values suggesting a stronger model relationship, ideally approaching 1. Conversely, RMSE reflects the model's error, with smaller values indicating better model performance. For predictability assessment, R2 values ≥ 0.9 are considered favorable, while values ≤ 0.64 are indicative of poorer predictability [10].



Explanation:
1. Camera
2. LED Light
3. UV LED
4. Frame
5. Sample

Figure 1. Non-destructive data retrieval scheme

Additionally, the study sought to identify an optimal machine learning model for classifying the maturity of Batu Tangerine 55. Various machine learning techniques were employed in the modeling process, including k-nearest Neighbor (k-NN), Support Vector Machine (SVM), and Random Forest. The k-NN method operates by identifying the k-nearest samples in the training dataset that closely resemble the object under consideration in the testing dataset. SVM, on the other hand, performs classification by determining an optimal hyperplane separator between different classes, particularly when the data can be linearly separated [11]. Moreover, Random Forest employs an ensemble learning approach by constructing multiple decision trees during training on the dataset. In classification tasks, it aggregates the decision outputs from individual trees, typically through a majority voting mechanism, to arrive at the final classification decision. These methods offer distinct advantages and are suitable for various types of data and classification tasks. By comparing their performance, the study aimed to identify the most effective approach for accurately classifying the maturity levels of Batu Tangerine 55 [12].

III. RESULTS AND DISCUSSION

The non-destructive testing yielded two types of images, namely reflectance images and fluorescence images, as illustrated in Figure 2. A total of 240 images were acquired for each type of light source used. These image datasets encompassed observations across three distinct maturity levels.

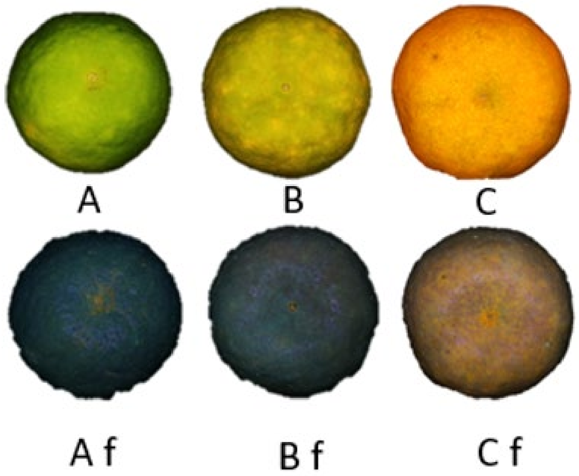


Figure 2. The acquisition result of the image of tangerine batu 55

Distinct color characteristics are observed in the images of Batu 55 tangerines across different maturity levels. The variations in color correspond to changes in the degree of maturity in the tangerine peel, attributed to the conversion of chloroplasts into chromoplasts and subsequent accumulation of carotenoids [13]. The discoloration of the fruit skin is primarily influenced by pigments such as flavonoids. Flavonoid compounds exhibit fluorescence when exposed to UV light, typically appearing yellow or blue. Notably, the use of a UV lamp as a light source results in relatively darker images compared to those obtained under white LED illumination. Nevertheless, the reflectance images capture certain colors that may not be discernible under white LED lighting. In destructive testing, measurements of firmness, Brix, and acidity values were obtained to further characterize the tangerines.

In Figure 3, it is apparent that the Brix values for partially ripe and fully ripe maturity levels exhibit insignificant differences. This suggests that tangerines classified as partially ripe are already suitable for harvesting. According to standards set by the Indonesian National Standard (SNI, 2009), tangerines are considered ready for harvest when their total dissolved solids reach 8°Brix.

Figure 4 illustrates a notable trend where the acidity value decreases as Batu Tangerine 55 progress in ripeness. This phenomenon can be attributed to the conversion of organic acids into simpler sugars, such as fructose and glucose, during the fruit ripening process [7].

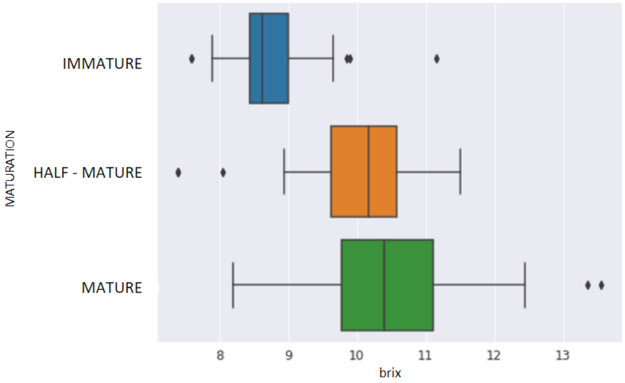


Figure 3. *Box-plot* brix

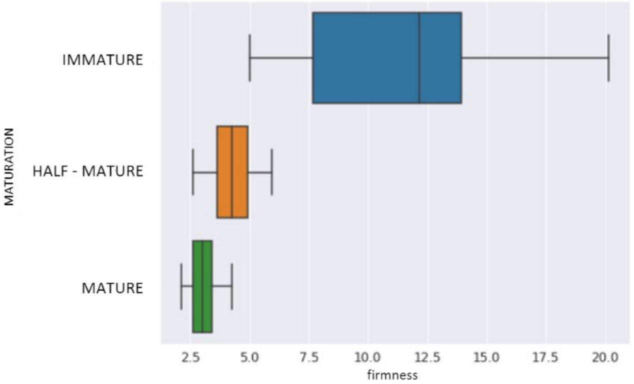


Figure 4. *Box-plot acidity*

Observing Figure 5 reveals a clear trend: as Batu Tangerine 55 reach higher levels of maturity, their firmness values decrease. This phenomenon can be elucidated by the fact that, according to [2], the softening of oranges is a hallmark of ripening, resulting from a reduction in fruit hardness. This softening process stems from changes in the chemical composition and structural integrity of carbohydrate cell walls within fruit tissues.

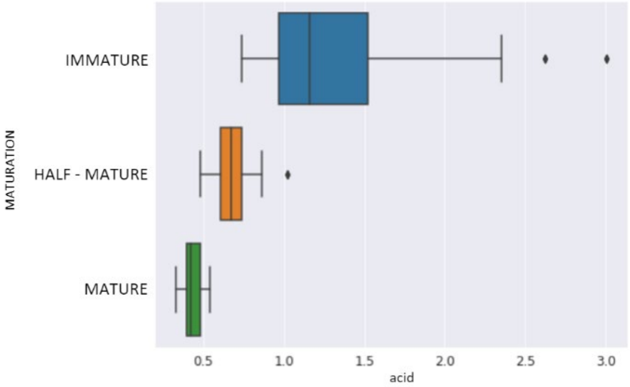


Figure 5. Box-plot firmness

Both destructive and non-destructive data were collected to create datasets for data processing. The classification task utilized three machine learning models: K-Nearest

Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, to classify three levels of maturity of Batu Tangerine 55. The data variations included RGB, UV, and combined datasets.The classification process involved two phases: the training phase, where the model was constructed, and the testing phase, where the model's accuracy was evaluated using separate data. The results are summarized in table 1. In the RGB dataset, the SVM model outperformed the other models, achieving a training accuracy of 100% and a testing accuracy of 97%. Similarly, the Random Forest model performed best in the UV dataset, with a training accuracy of 100% and a testing accuracy of 92.5%.

When considering the combined RGB and UV dataset, the SVM model once again exhibited superior performance, attaining a training accuracy of 100% and a testing accuracy of 95%. Thus, the recommended model for use with RGB image data is SVM, given its robust performance across both training and testing phases.

TABLE I
MACHINE LEARNING CLARIFICATION DATA

	Model	Scaling	*Training Accuracy*	*Test Accuracy*
RGB	KNN	MinMax	1.0	0.97
	KNN	None	0.95	0.94
	Random Forest	MinMax	1.0	0.95
	Random Forest	None	1.0	0.94
	SVM (kernel: Linear)	**MinMax**	**1.0**	**0.97**
	SVM (kernel: Linear)	**None**	**1.0**	**0.97**
UV	KNN	MinMax	0.95	0.94
	KNN	None	0.91	0.95
	Random Forest	MinMax	1.0	0.93
	Random Forest	None	1.0	0.92
	SVM (kernel: Linear)	MinMax	0.95	0.97
	SVM (kernel: Linear)	None	1.0	0.88
All	KNN	MinMax	0.95	0.92
	KNN	None	0.97	0.95
	Random Forest	MinMax	1.0	0.89
	Random Forest	None	1.0	0.89
	SVM (kernel: Linear)	MinMax	1.0	0.96
	SVM (kernel: Linear)	None	1.0	0.92

TABLE II
PREDICTION RESULTS WITH PLS COMBINED DATA

Physicochemical parameters	Factor	R^2 Calibration	RMSEC	R^2 Cross-validation	RMSECV	R^2 Prediction	RMSEP	RPD
Firmness	10	0.71	2.34	0.69	2.44	0.63	2.76	1.81
Brix	10	0.43	0.93	0.39	0.96	0.48	0.88	1.28
Acid	10	0.60	0.31	0.56	0.32	0.49	0.34	1.52
B/A	10	**0.79**	**3.58**	**0.77**	**3.70**	**0.81**	**3.48**	**2.12**

The prediction model for the maturity parameter of Batu Tangerine 55 utilized Partial Least Squares (PLS) analysis, implemented using Python software. The PLS analysis comprised two stages: calibration and validation, with 2/3 of the total 480 data points used for model training. Cross-validation, an integral part of PLS analysis, was employed to assess the accuracy of the calibration model. Subsequently, 1/3 of the total data was utilized to predict the maturity of other tangerine fruits. Table 2 presents the prediction results of the mature parameters based on physicochemical parameters derived from combined reflectance and fluorescence data. The brix/acidity ratio yielded the highest R2 value of 0.81. Statistical parameters used for model evaluation include the coefficient of determination (R2), root-mean-square error of calibration (RMSEC), and root-mean-square error of cross-validation (RMSECV). A small difference between RMSEC and RMSECV indicates model stability, with larger differences suggesting that the calibration set model does not adequately represent the validation set [14]. The accuracy achieved in this study was not superior to that of deep learning models developed in previous studies [7]. However, the machine learning model utilized herein offers advantages in terms of ease of training and implementation compared to deep learning models.

IV. CONCLUSION

A machine learning model has been developed using a reflectance-fluorescence image dataset to classify three levels of maturity of Batu Tangerine 55 fruit, employing three models: KNN, SVM, and Random Forest. The SVM model utilizing reflectance (RGB) data yielded the most favorable results, achieving a training accuracy of 100% and a testing accuracy of 97%. For the prediction of the maturity of Batu Tangerine 55 fruit using the PLS method, the brix/acidity ratio emerged as a significant parameter compared to others. Notably, the combined feature set produced the highest prediction accuracy, with an R2 value of 0.81 and an RMSE of 3.4.

ACKNOWLEDGMENTS

The author expresses gratitude to Jessica Yolanda, Abdullah Sa'id, Abdul Mulukul Ikrom, and Salsabil Lazuardi Nugroho from the Department of Biosystems Engineering, Universitas Brawijaya, for their valuable assistance during the experiment. This research was supported by funding from the 2022 Indonesian Collaborative Research through Universitas Brawijaya Rector, under decree number 6672 of 2022.

AUTHORS CONTRIBUTION

Safitri Diah Ayu Ariani: Investigation, Analysis, Visualization, Preparation of Original Drafts;
Inggit Krishna Maharsih: Validation, Review Writing and Editing;
Dimas Firmanda Al Riza: Project Administration, Software, Resources, Validation, Review Writing and Editing;

COPYRIGHT

REFERENCES

[1] A. Ashari, E. Nurcahyani, H. I. Qudus, and Zulkifli, "Analisis Kandungan Prolin Planlet Jeruk Keprok Batu 55 (*Citrus reticula blanco* Var. Crenatifolia) Setelah Diinduksi Larutan Atonik Dalam Kondisi Cekaman Kekeringan Secara In Vitro," *Analit: Analytical and Environmental Chemistry*, vol. 3, no. 1, pp. 69-78, 2018.

[2] I. Ifmalinda, K. Fahmy, and E. Fitria, "Prediction of Siam Gunung Omeh Citrus Fruit (Citrus Nobilis Var Microcarpa) Maturity Using Image Processing," Jurnal Keteknikan Pertanian, vol. 6, no. 3, pp. 335–342, Dec. 01, 2018. Doi: 10.19028/jtep.06.3.335-342.

[3] I Kadek Riastana, Ni Komang Alit Astiari, and Ni Putu Anom Sulistiawati, "Kualitas Buah Jeruk Siam (Citrus nobillis var microcarva L) Selama Penyimpanan Pada Berbagai Tingkat Kematangan Buah," GA, vol. 24, no. 1, pp. 22-28, Apr. 2020.

[4] M. G. Lieka, R. Poerwanto, and D. Efendi, "Aplikasi Ethephon dan Stiker Pascapanen untuk Perbaikan Kualitas Buah Jeruk Siam Garut (Citrus nobilis Lour)," Comm. Horticulturae Journal, vol. 2, no. 2, p. 1, Dec. 11, 2018. Doi: 10.29244/chj.2.2.1-10.

[5] R. K. Haba and K. C. Pelangi, "Pengelompokan Buah Jeruk menggunakan Naïve Bayes dan Gray Level Co-occurrence Matrix," ILKOM Jurnal Ilmiah, vol. 12, no. 1, pp. 17–24, Apr. 26, 2020. Doi: 10.33096/ilkom.v12i1.494.17-24.

[6] A. Zakiyyah , Z. Hanif , D. W. Indriani , Z. Iqbal , R. Damayanti , D. F. Al Riza , "Characterization and Classification of Citrus reticulata var. Keprok Batu 55 Using Image Processing and Artificial Intelligence," Universal Journal of Agricultural Research, Vol. 10, No. 4, pp. 397 – 404, 2022. DOI: 10.13189/ujar.2022.100409.

[7] D.F. Al Riza, A. M. Ikrom, A. A. Tulsi, Darmanto, Y. Hendrawan, "Mandarin orange (*Citrus reticulata Blanco* cv. Batu 55) ripeness parameters prediction using combined reflectance-fluorescence images and deep convolutional neural network (DCNN) regression model" Scientia Horticulturae, Volume 331, 1 May 2024, 113089, pp. 1-10, 2024

[8] A.H. Wigena "Regresi kuadrat terkecil parsial multi respon untuk statistical downscaling", Indonesian Journal of Statistics and Its Applications, Vol. 16, No. 2, pp. 12–15. 2011.

[9] V. Vijayakumar, Y. Ampatzidis, L. Costa. "Tree-level citrus yield prediction utilizing ground and aerial machine vision and machine learning", Smart Agricultural Technology, Vol. 3, February 2023, 100077, pp. 1-11. 2023.

[10] I.U. Hidayah, B. Kuswandi, L. Wulandari. "Deteksi Kemurnian Air Zamzam Menggunakan Metode Spektrofotometri Near Infra Red (NIR) dan Kemometrik" Pustaka Kesehatan, Vol. 2, No. 3, p. 439-444, sep. 2014.

[11] Y. Amrozi, D. Yuliati., A. Susilo, R. Ramadhan. "Klasifikasi Jenis Buah Pisang Berdasarkan Citra Warna dengan Metode SVM" Jurnal Sisfokom, Vol. 11, No. 3, pp. 394-399, 2022.

[12] P. Rosyani, Saprudin, R. Amalia. "Klasifikasi Citra Menggunakan Metode Random Forest dan Sequential Minimal Optimization (SMO)" Jurnal Sistem dan Teknologi Informasi, Vol. 9, No. 2, pp. 132-134, 2021

[13] A. Rahmawati and W. D. R. Putri, "Karakteristik Ekstrak Kulit Jeruk Bali Menggunakan Metode Ekstraksi Ultrasonik (Kajian Perbandingan Lama Blansing Dan Ekstraksi)," JPA, vol. 1, no. 1, pp. 26–35, Oct. 2013.

[14] A. Putri Ana, Y. A. Purwanto, and S. Widodo, “Prediksi Indeks Panen Jambu ‘Kristal’ secara Non Destruktif Menggunakan Portable Near Infrared Spectrometer,” *Jurnal Keteknikan Pertanian*, vol. 9, no. 3, pp. 103–110, Dec. 23, 2021. doi: 10.19028/jtep.09.3.103-110.

Chi-Square Histogram Analysis of Woven Fabric Images Made from Natural Dyes Due to Exposure to Sunlight

Patrisius Batarius[1] and Alfry A. J. Sinlae[1]
[1]Computer Science Program, Faculty of Engineering, Widya Mandira Catholic University, Kupang, Indonesia

Corresponding author: Patrisius Batarius (e-mail: patrisbatarius@unwira.ac.id).

ABSTRACT This research aims to conduct a Chi-square analysis on the histogram of woven fabric images dyed with natural dyes following exposure to sunlight. Woven fabrics dyed with natural dyes have attracted attention in the textile industry due to their sustainability and environmental safety. Continuous sunlight is a significant factor influencing color changes in woven fabric dyed with natural dyes. The methodology involves capturing images of woven fabric pre- and post-sunlight exposure, followed by histogram analysis using Chi-Square testing, mean, mode, and standard deviation. We utilize pre-cropped and resized grayscale images. Research findings demonstrate that sunlight significantly impacts the histogram of woven fabric images dyed with natural dyes, causing shifts in color distribution, standard deviation, and mode. These findings hold critical implications for the textile industry, particularly for manufacturers of woven fabrics dyed with natural dyes. The application of Chi-Square analysis and standard deviation provides guidelines for product design, maintenance procedures, and consumer education regarding the preservation of color quality in fabrics exposed to sunlight. Changes in the quality of woven fabric images under sunlight exposure can offer essential guidance in the care and maintenance of textile products dyed with natural dyes. This research contributes to a deeper understanding of the interplay between natural dyes, sunlight, and woven fabrics, supporting the development of sun-resistant natural dyes.

KEYWORDS: Chi-Square, natural dyes, sunlight, woven fabric

I. INTRODUCTION

The textile industry has experienced rapid development in recent decades, with various innovations aimed at improving sustainability and environmental safety. One important aspect of this effort is the use of natural dyes in the production of woven fabrics. Natural dyes have become an attractive option due to their more environmentally friendly nature compared to synthetic dyes, which often have a negative impact on the environment and human health [1][2].

However, the use of natural dyes in woven fabrics is not without its challenges. The common view is that there is a decline in the color quality of traditionally woven fabrics made from natural dyes if exposed to sunlight for a long time. One of the factors that affect the color quality of textile products that use natural dyes is sun exposure. Constant sunlight is one of the external factors that can change the color of woven fabrics that have been dyed with natural dyes [2][3]. In everyday environments, textile products exposed to sunlight often experience color changes that can affect the aesthetics and value of the product.

Therefore, it is necessary to understand the impact of sunlight on woven fabrics that use natural dyes. This study aimed to explore color changes in woven fabrics that have been dyed with natural dyes after exposure to sunlight. To achieve this goal, we used Chi-Square analysis, mean, mode, and standard deviation on the histogram of woven fabric imagery before and after exposure to sunlight. This method can provide deeper insight into changes in color distribution in sun-woven fabrics. This study limits the issue to the effect of sun exposure time on the image histogram of woven fabrics using natural dyes.

The results of this study have important implications in the context of the textile industry and the environment [2][4]. We hope that this research can provide a better understanding of the interaction between natural dyes, sunlight, and woven fabrics, as well as lay the foundation for the development of natural dyes that are more resistant to the effects of sunlight. In addition, the results of this study can be a guide for consumers and manufacturers in maintaining the color quality of textile products that use natural dyes when exposed to sunlight. Thus, this research contributes to efforts to make the textile industry more sustainable and environmentally friendly

Some research gaps or gaps that can be the basis for further research include:

1. Effect of Sun Exposure Time: This study focused on the effect of sunlight with various differences in sun exposure time on woven fabrics.
2. Natural Dyes and Types of Woven Fabrics: This study used 2 types of woven fabric images derived from 3 types of natural dyes. The first image with white color, which is made of cornmeal. The second image is a cloth consisting of 3 pieces of white, red and blue colors. The blue color is made from the bauk ulu (local language) and the red color from the bark and roots of noni wood. The process of making color with CRAdips yarn made of cotton into extra water from natural dyes.
3. Development of Sun-Resistant Natural Dyes: These findings could inspire further research in developing natural dyes that are more resistant to the effects of sunlight. This gap invites research to find solutions to maintain color quality in woven fabrics that use natural dyes, especially when exposed to sunlight.
4. Environmental and Sustainability Aspects: Although there is concern for natural dyes due to sustainability and environmental safety, this study does not explicitly explore the impact of sunlight on environmental aspects or sustainability.

II. RELATED RESEARCH

Previous studies on the effect of sunlight on textile products using natural dyes have provided valuable insights into the textile industry, but there are still drawbacks that can be improved. Some previous studies have covered diverse methods, but there is room for further research development. In this review, several previous evaluations examined the impact of sunlight on textile products using natural dyes, as well as identifying flaws in the methods used.

Other research tested the effect of sunlight on the color of woven fabrics using natural dyes with an experimental approach. They hung strips of fabric in the sun over various time intervals and measured changes in image quality at specific intervals using MSE and PSNR parameters [5][6]. The drawback was the absence of in-depth statistical analysis to support their experimental results.

Other research on computer simulations to determine the quality of the process of making natural dyes. The quality of dyeing cotton cloth using jengkol fruit peel waste (Archidendron jiringa) is calculated by calculating K/S and dE values [7]. One of the processes of testing the quality of batik coloring using natural dyes is the color difference test (L*a*b. [8]. Testing of flexing strength, rubbing effect on fabric, and light intensity of Cassia extract plant on wool fabric using CIE L*a.b* [9]. However, the drawback is strong experimental validation, as it relies solely on simulation models. The simulation results may not fully reflect real-world situations. In addition, the study did not provide a deep understanding of color change at a statistically strong level regarding the influence of sunlight on color quality. Although the process of analyzing color changes in textile products uses natural dyes with spectrophotometric methods, the drawback is that it requires expensive equipment and does not consider other external factors that can affect color change.

Another study describes the measurement with the percentage dose of each natural color extract used to show the quality of the color produced [10]. The combination of several extracts of natural dyes can be done to produce good intensity and color elasticity resistance in fabrics [11]. Utilization of bio-mordan almond peel extract as a textile dye, as an alternative that can help reduce dependence on toxic mordan metal. The amount of mordan applied in small quantities (units of g/l) affects the quality of the color used [12]. In addition, using milliliter units per gram (ml / g) as a dose of chemical use [13]. This research is more on the use of microscopy to observe changes at the microscopic level in woven fabric fibers exposed to sunlight. This research focuses more on changes at the fabric cell level. This process does not integrate color analysis of the entire fabric and focuses only on individual cell changes. The results may be less practically relevant in the textile industry.

Another study relates to color resistance tests on woven fabrics that use natural dyes. Assessment of the quality of dyeing woolen fabrics with onion dye using K/S units represented by graphic presentation [14]. The fastness of woolen fabrics with natural dyes such as madder root, chamomile, pomegranate bark, and apple tree bark is excellent. Color strength at K/S=14 [15]. Acceleration testing methods may not fully reflect the color changes that occur in woven fabrics under everyday sunlight conditions.

Although this past research has made important contributions to the understanding of the impact of sunlight on textile products that use natural dyes, there are shortcomings that need to be corrected. Therefore, future research can take a more comprehensive approach by combining various methods, such as field experiments, spectrophotometric analysis, as well as the use of microscopy to understand changes at the microscopic level. This will make it possible to gain a more complete understanding of how sunlight affects textile products that use natural dyes. In addition, the study may also consider different types of woven fabrics and different natural dyes to understand variations in response to sunlight.

Today, image processing has been developed with various methods to analyze image quality in various fields. One of them is in the textile field. Image analysis is carried out by various methods and calculations. Among them are to identify grains based on color feature extraction using RGB and HSV, shape feature extraction using Morphological Threshold, and texture feature extraction using Grey Level Co-occurrence Matrix (GLCM) and Local Binary Pattern (LBP) [16]. Likewise, in the process of classifying plants based on the shape of their leaves with edge detection methods and artificial neural networks [17] or wood fiber classifiers utilizing deep learning [18].

The image processing research above has not discussed the image from the results of statistical analysis. Accuracy in fabric degradation calculations can be predicted from the physical properties of Ultra-Violet degraded woven fabrics at various levels of exposure time [19]. Some of these physical properties can be analyzed through statistical analysis such as chi-square image histogram, mean, standard deviation, and image histogram mode. Chi-square is used for histogram matching in computer vision problems in analyzing facial images and expressions [20]. Research to reduce noise by estimating image noise levels with chi-square distribution [21].

III. RESEARCH CONTRIBUTIONS

This research has various important contributions that can have an impact on various aspects, including the textile industry, science, and the environment. The main contribution of this study is to provide a better understanding of how changes in the image histogram of woven fabrics made from natural dyes occur, over time of sun exposure. These results will provide valuable insights for producers and consumers in managing the treatment of textile products using natural dyes, as well as potentially laying the foundation for the development of natural dyes that are more resistant to the effects of sunlight, support sustainability aspects in the textile industry, and increase understanding of the interaction between natural dyes and sunlight in textile environments.
Some of the key contributions of the study include:

1. A Deeper Understanding of Natural Dyes:
   This research provides a deeper understanding of the influence of sunlight on the color quality of woven fabrics that use natural dyes. This can help textile manufacturers and researchers understand how natural dyes interact with environmental factors, particularly sunlight.
2. Textile Product Care Guide:
   The results of this study can be used as a treatment guide for textile products that use natural dyes. Consumers and manufacturers can utilize this information to maintain the color quality of woven fabrics, reducing the risk of discoloration due to sun exposure.
3. Development of Natural Dyes More Resistant to Sunlight:
   This research could encourage the development of natural dyes that are more resistant to the effects of sunlight. This can contribute to the reduction of the use of synthetic dyes that negatively impact the environment and human health.
4. Sustainability and Environmental Aspects:
   These findings support aspects of sustainability in the textile industry by promoting the use of more environmentally friendly natural dyes. This can be an important step in reducing the textile industry's impact on the environment.
5. Advanced Research:
   This research opens the door to further research that can answer more in-depth questions, such as the effect of sun exposure time, comparisons of different types of natural dyes, and practical applications in the textile industry. This has the potential to provide more comprehensive insights into the understanding of natural dyes and sunlight.

With its contributions covering practical, scientific, and environmental aspects, this research can be the basis for innovation and a better understanding of the use of natural dyes in the textile industry, with a positive impact on the industry and the environment.

IV. RESEARCH METHODS

This research method will prioritize Chi-Square analysis and histograms to understand color changes in textile products that use natural dyes after exposure to sunlight. Figure 1 shows the flow of research conducted.

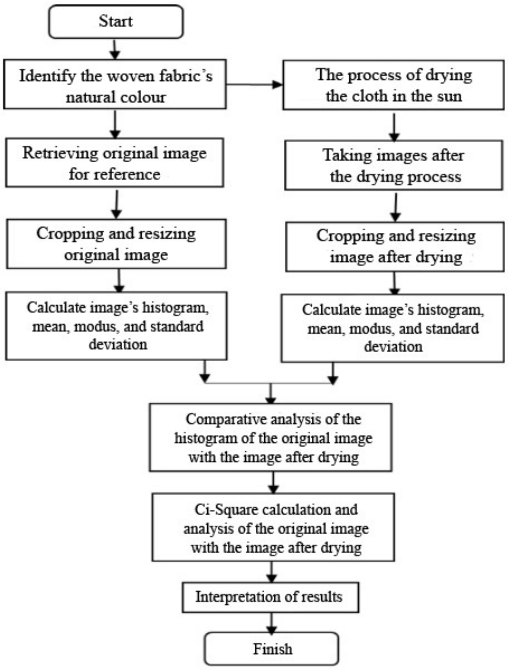

Start
Identify the woven fabric's natural colour
The process of drying the cloth in the sun
Retrieving original image for reference
Taking images after the drying process
Cropping and resizing original image
Cropping and resizing image after drying
Calculate image's histogram, mean, modus, and standard deviation
Calculate image's histogram, mean, modus, and standard deviation
Comparative analysis of the histogram of the original image with the image after drying
Ci-Square calculation and analysis of the original image with the image after drying
Interpretation of results
Finish

Figure 1. Research process flow

Here are the methodological steps to follow:

1. Identify natural dyes in woven fabrics.
   This stage is the process of collecting image samples of woven fabrics made from natural dyes. The woven fabric used is made of cotton, with natural dyes from corn powder to produce a white color. The red color is made from noni bark, the blue color is made from needle leaves.
2. The stage of taking the original image as a reference. Shooting using a mobile camera with Samsung SM-A315G F2.0 1/50s 4.60mm ISO 125 smartphone specifications with a ratio of 9:16. The shooting distance is 30 cm and the image photo process is one day after drying.

Figure 2 shows the process of taking images before and after drying.

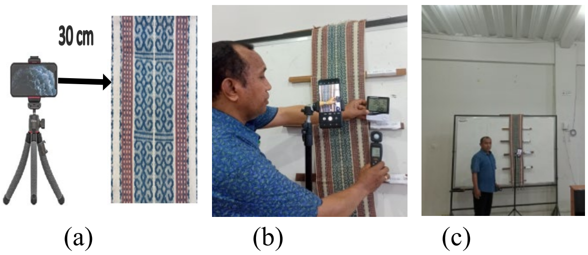


Figure 2. **The process of capturing woven fabric imagery. Camera distance with woven cloth (a), temperature measurement, room light intensity (b), lighting room conditions when taking images (c).**

The measured light intensity value during the weaving fabric image is 84 x 100 lumens. The average room temperature measured during shooting ranged from 28.30°C to 30.30°C, and the relative humidity was 43%.

3. Drying woven fabrics.
This stage is to obtain exposure to sunlight with various drying time intervals. The drying time starts at 09.00-15.00, every 1 hour the sun-dried image is taken as in step 2 above. Figure 3 shows the drying process of the fabric and the measurement of the value of light intensity, and temperature during the drying process.

Figure 3. **The drying process of woven fabrics**

TABLE I
DATA ON TEMPERATURE, AIR HUMIDITY, AND LIGHT INTENSITY AT THE TIME OF DRYING WOVEN FABRICS

Drying hour	Temperature Value	Relative Humidity	Light Intensity (x 100) Lumen
09:00 - 10:00	33,1^{0}C	42%	1710
10:00 - 11:00	33,3^{0}C	42%	1627
11:00 - 12:00	33,8^{0}C	42%	1649
12:00 - 13:00	35,8^{0}C	42%	1639
13:00 - 14:00	34,2^{0}C	45%	1550
14:00 - 15:00	34,2^{0}C	49%	1546

The value of temperature, the temperature itself, and light intensity at the time of drying woven fabrics are recorded as variables that affect the process of calculating ci-square on woven fabric images.

The data in Table 1 shows the value of temperature, temperature, and light intensity when drying fabrics.

4. The process of cropping and resizing images.
The cropping stage is carried out to take samples of the analyzed imagery. The process of image resizing to ensure the same image size between the original image and the image after sun exposure.
The process of cropping and resizing images is carried out using matlab software. Pieces of source code from the cropping and resizing process.

```
% Program cuts image
clc;
clear;
% Read image from file
namafile = 'k1_10-11.jpg';
citra = imread(namafile);
% Display image to view its contents
imshow(citra);
% Create box for crop
kotak = imrect;
%Wait until the box is completed (press Enter)
wait(kotak);
% Get the position and size of the selected box
posisi_kotak = getPosition(kotak);
% Crop the image according to the selected box
citra_crop = imcrop(citra, posisi_kotak);
% Resize image to new size
ukuran_baru = [128, 128];
citra_resize = imresize(citra_crop,
ukuran_baru);

% Displays cropped and resized image
figure;
subplot(1, 2, 1);
imshow(citra_crop);
title('Hasil Crop');
subplot(1, 2, 2);
imshow(citra_resize);
title('Hasil Resize');

% Save the cropped image and resize it to a new
file (optional)
namafile_crop_resize = 'k1_10-
111_resize128.jpg';
imwrite(citra_resize, namafile_crop_resize);
```

The results of the source code of the ‘image cropping program’, are shown in Figure 4 and Figure 5. Figure 4 is for the image of one type of color, namely white, and Figure 5 is for the image of woven fabric consisting of 3 types of colors. The cropping process that occurs (a) and the selected image crop results (b). Furthermore, the cropped image was resized again with a size of 128x128 pixels (c).

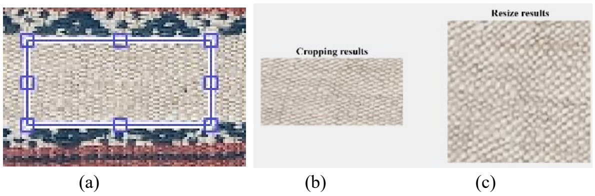
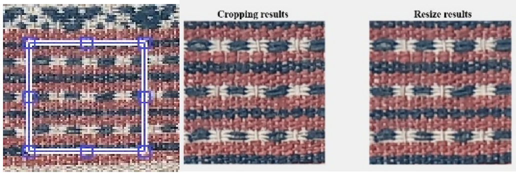


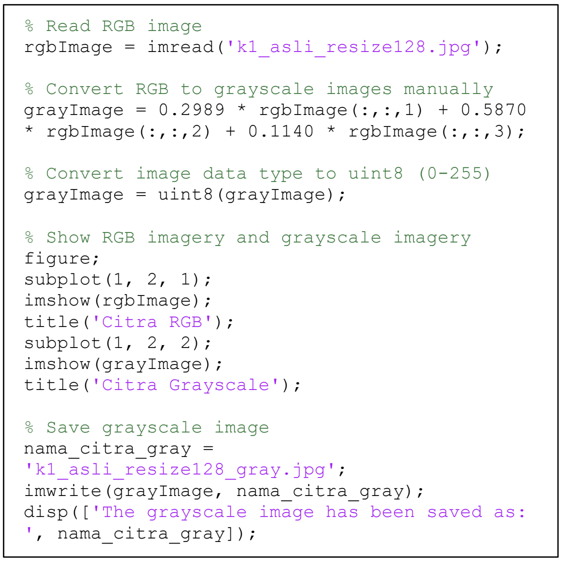
Figure 4. **The result of cropping and resizing the image of woven fabric for 1 type of color (white color). The cropped part of the woven fabric image (a), the image cropping result (b), and the resizing result (c).**



(a) (b) (c)

Figure 5. The result of cropping and resizing the image of woven fabric (for 3 types of colors). The cropped part of the woven fabric image (a), the image cropping result (b), and the resizing result (c).

The image size used for analysis is 128 x 128 pixels. The process of converting RGB images to grayscale images, with Matlab software. The source code snippet is as follows:



```
% Read RGB image
rgbImage = imread('k1_asli_resize128.jpg');

% Convert RGB to grayscale images manually
grayImage = 0.2989 * rgbImage(:,:,1) + 0.5870
* rgbImage(:,:,2) + 0.1140 * rgbImage(:,:,3);

% Convert image data type to uint8 (0-255)
grayImage = uint8(grayImage);

% Show RGB imagery and grayscale imagery
figure;
subplot(1, 2, 1);
imshow(rgbImage);
title('Citra RGB');
subplot(1, 2, 2);
imshow(grayImage);
title('Citra Grayscale');

% Save grayscale image
nama_citra_gray =
'k1_asli_resize128_gray.jpg';
imwrite(grayImage, nama_citra_gray);
disp(['The grayscale image has been saved as:
', nama_citra_gray]);
```

The results of the source code above are shown in Figure 6

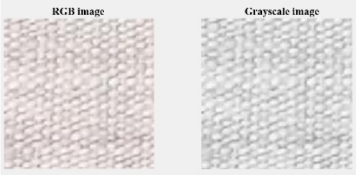


Figure 6. RGB image display and gray scale image

5. Calculation and analysis of the histogram
   The original image and the image exposed to the heat of sunlight are calculated histogram values. The calculation results analyzed several parameters such as the mean, mode, and standard deviation of each image. Image types in histogram analysis use grayscale images.
6. Calculation and analysis of ci-square original imagery with imagery after exposure to sunlight.
   Chi-square analysis is performed by calculating the per-pixel ratio between the original image and each image after sun exposure. Chi-square calculations use grayscale image types.
7. Interpretation of results.
   This stage is the process of interpreting the results of the Chi-Square analysis to determine whether the change in the histogram image of woven fabrics exposed to sunlight is significant or not. Compare results between different types of woven fabrics and natural dyes to identify differences in response to sunlight.
8. Conclusions and implications
   Summarize findings and conclusions from Chi-Square analysis and histogram. Discuss the implications of these research results in the context of the textile industry, sustainability, and the development of natural dyes that are more resistant to sunlight.

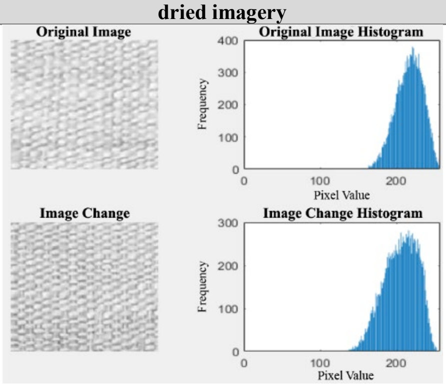
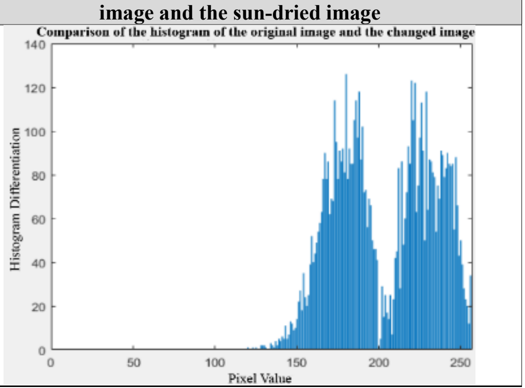
V. RESULT

The results of this study provide valuable insight into the impact of sunlight on textile products that use natural dyes. The following is a description of the results of this study:

A. WHITE COLOR

Tables 2 and 3 show the results of changes in the histogram of the original image and the image after drying. Each sun-dried image is compared to the original image. The type of image is grayscale.

TABLE II
CHANGES IN THE HISTOGRAM OF THE ORIGINAL IMAGE WITH THE IMAGE AFTER EXPOSURE TO SUNLIGHT. AS WELL AS THE DIFFERENCE IN THE HISTOGRAM OF THE TWO IMAGES

Drying hour	Histogram display of original imagery and sun-dried imagery	The difference between the histogram of the original image and the sun-dried image
09.00-10.00		

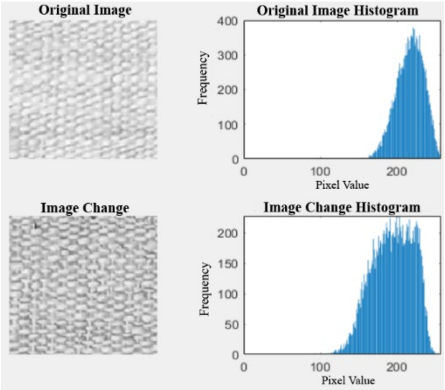
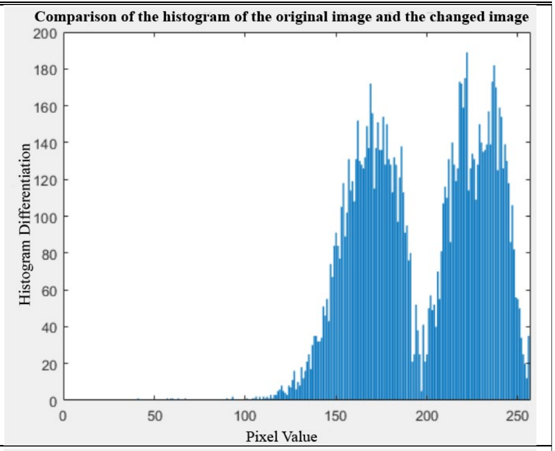
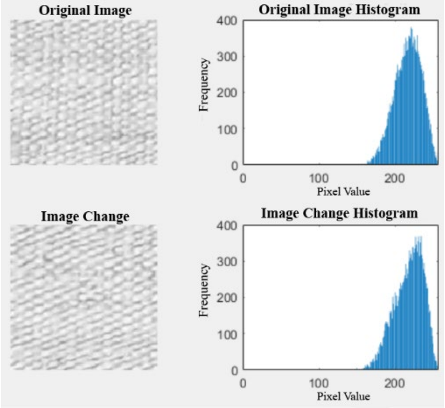
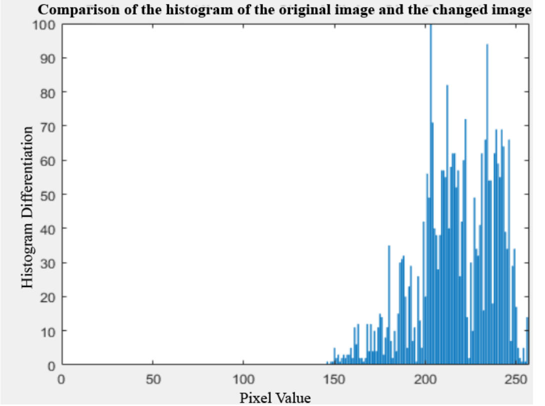
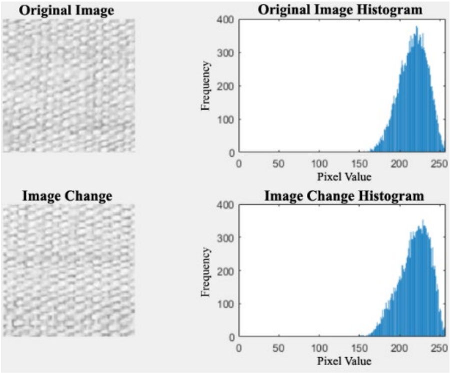
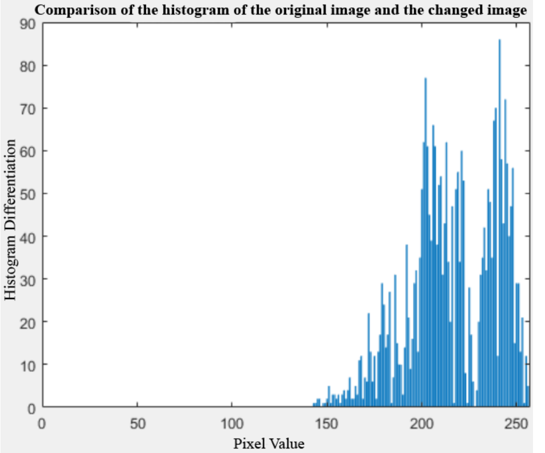
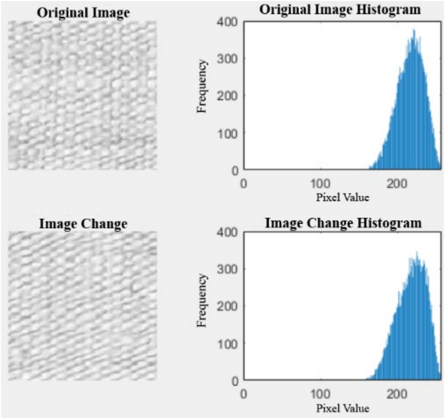
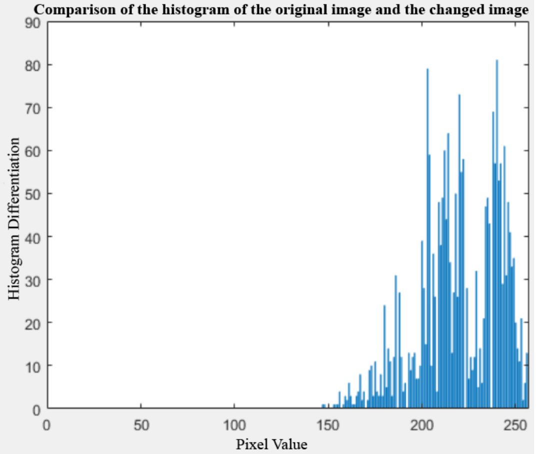
10.00-11.00		
11.00-12.00		
12.00-13.00		
13.00-14.00		

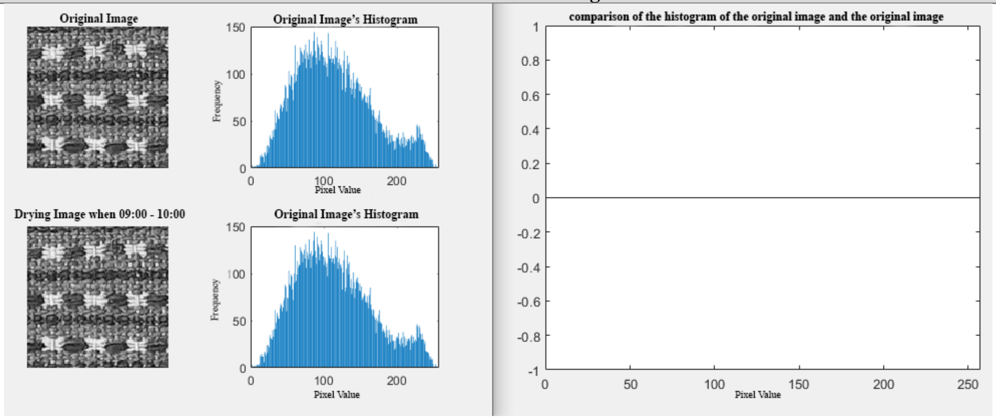
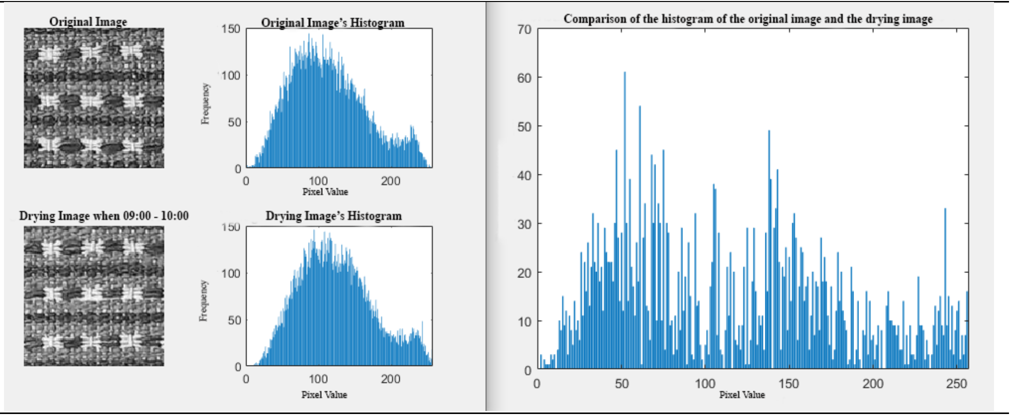
TABLE III
CHI-SQUARE VALUE, MEAN, MODE, AND STANDARD DEVIATION OF THE ORIGINAL IMAGE WITH THE IMAGE AFTER DRYING IN SUNLIGHT

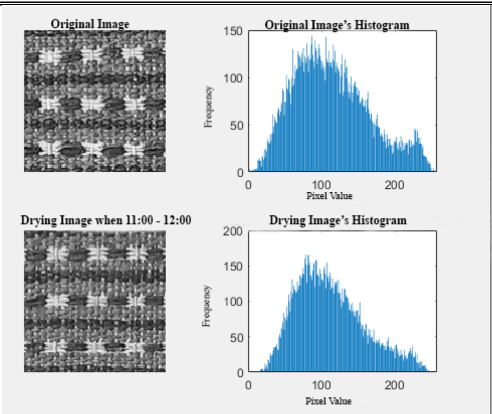
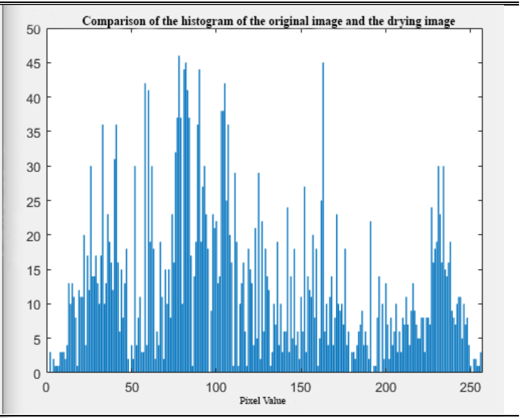
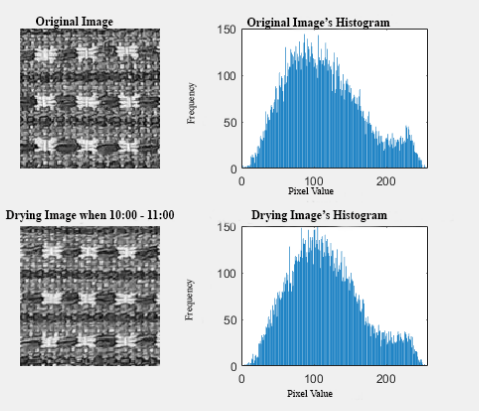
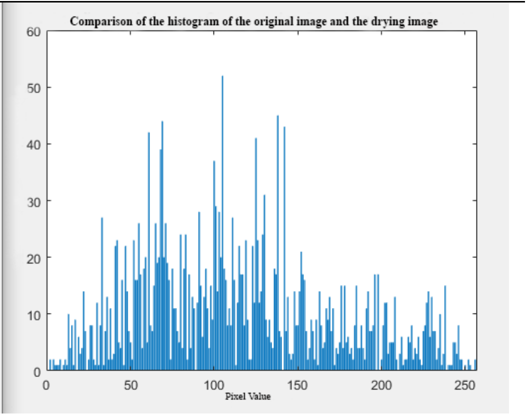
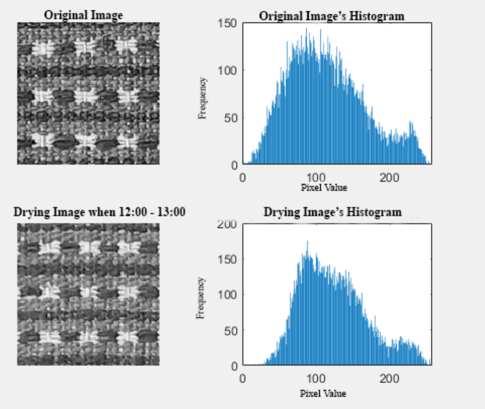
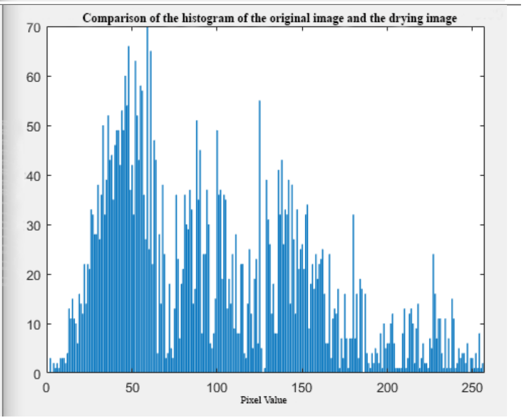
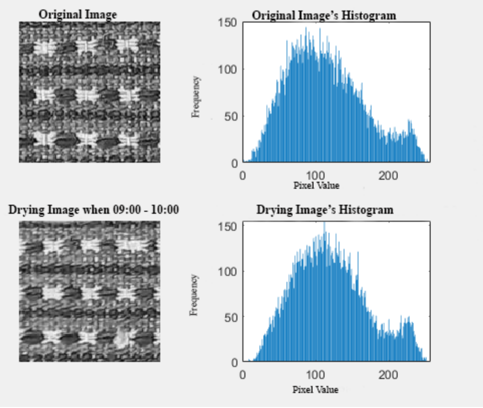
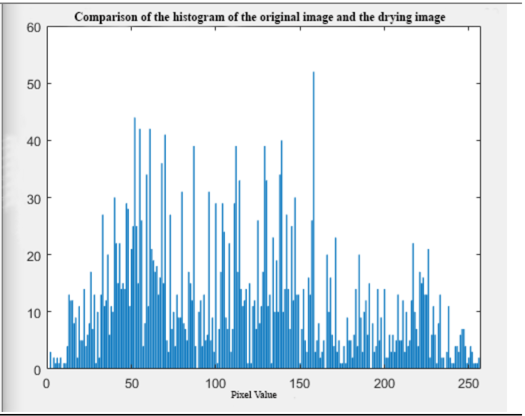
Drying hour	mean	Standard deviation	Mode	Chi-Square value of original imagery with image after drying
Original image	64	11,51	220	-
09.00-10.00	64	95,49	215	2752,2413
10.00-11.00	64	83,23	198	7160,2007
11.00-12.00	64	108,34	233	389,4071
12.00-13.00	64	107,25	228	373,4448
13.00-14.00	64	107,17	224	254,0882

B. IMAGE WITH 3 COLOR TYPES

Tables 4 and 5 show the histogram results of the original image with the image exposed to sunlight. Table 4 shows histogram values such as mean, standard deviation mode of histogram image, and chi-square value between the original image and sunlight image.

TABLE IV
CHANGES IN THE HISTOGRAM OF THE ORIGINAL IMAGE WITH THE IMAGE AFTER EXPOSURE TO SUNLIGHT, AS WELL AS THE DIFFERENCE IN THE HISTOGRAM OF THE TWO IMAGES FOR IMAGES WITH 3 TYPES OF COLORS

Drying hour	Display of the original image, the changed image and the histogram difference between the original image and the sun-dried image
Original image	
09.00-10.00	

11.00- 12.00		
10.00- 11.00		
12.00- 13.00		
13.00- 14.00		

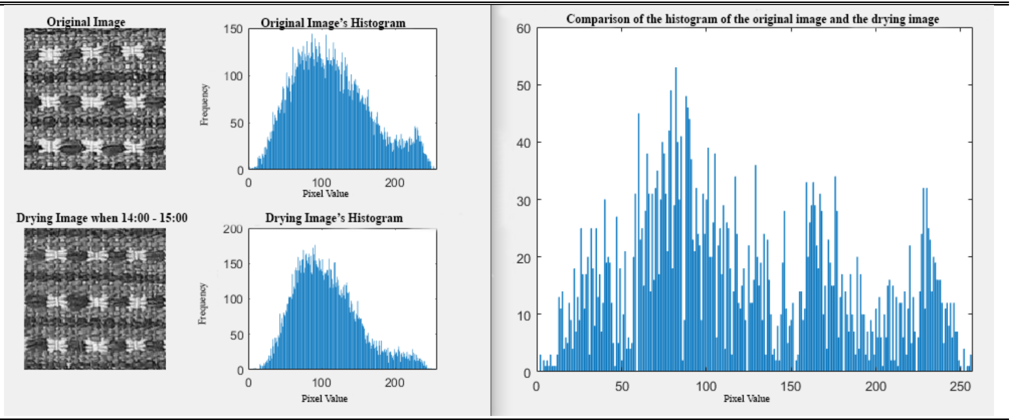
14.00- 15.00	

TABLE V
CHI-SQUARE, MEAN, MODE, AND STANDARD DEVIATION VALUES OF THE ORIGINAL IMAGE WITH THE IMAGE AFTER DRYING IN SUNLIGHT FOR IMAGES WITH 3 COLOR TYPES

Drying hour	mean	Standard deviation	Mode	Chi-Square value of original imagery with image after drying
Original image	64	41,69	86	729,5334
09.00-10.00	64	42,76	93	671,2343
10.00-11.00	64	44,26	104	322,1226
11.00-12.00	64	50,54	77	1509,5554
12.00-13.00	64	52,41	87	475,6603
13.00-14.00	64	43,23	111	821,6233
14.00-15.00	64	54,84	90	729,5334

The results of Chi-Square, standard deviation, mean and histogram mode of image tests show that changes in image histogram are statistically significant. This confirms that sunlight has a noticeable impact on the color distribution of woven fabric products that use natural dyes.

1. Standard deviation
   Standard deviations tend to rise and fall over time, indicating changes in pixel intensity variations in the image. The highest standard deviation occurred at 1 p.m., indicating the highest variation in pixel intensity at the time, while the lowest standard deviation occurred at 9 a.m. This data provides insight into how imagery intensity fluctuates during the day and may be related to changes in lighting or other factors affecting imagery during those hours. A higher standard deviation indicates a greater difference in the pixel intensity distribution of the image at a given hour
   The first image (white), drying from 11 a.m.—12:00 p.m. or 1:00 p.m. has a high standard deviation value. The difference in value from the 3rd drying time is small. The lowest standard deviation value at drying time is 09.00-10.00. While the 2nd image of drying at 14.00-15.00 has the highest and lowest values at 0900-10.00.
2. Mode Value
   This mode data can be useful for understanding how the most common pixel intensities are in an image. This value indicates a change in the nature of the image. Mode values that change over time in a drying session indicate changes in image properties. This change is the result of changes in lighting conditions during drying time. Changes in mode values reflect the difference in pixel intensity distribution between images taken each hour.
   The first image of the largest mode value during drying time is at 11:00-12:00 and lowest at 10:00-11:00. The second largest mode value in woven fabric image with a drying time of 13.00-14.00 and the lowest drying time of cloth is 11.00-12.00.

3. The chi-square value of imagery
   The Chi-Square Histogram value of the image shows a useful parameter for understanding the extent to which woven fabric images differ in terms of their pixel intensity over a period of drying time. This Chi-Square Histogram reflects the difference in pixel intensity distribution during the hours of drying. The higher the Chi-Square value, the greater the difference in the pixel intensity distribution between the original image and the sun-dried image.
   The greatest chi-square value at drying time is 10:00-11:00 for the first image 13:00-14:00 and 10:00-11:00 for the second cloth image. While the lowest value is at 14.00-15.00 for the first image at 11.00-12.00 for the second image.
   Two times compared.

The results of this study provide valuable insights into an effort to maintain the color quality of textile products that use natural dyes under sun exposure conditions. With a deeper understanding of color change and color distribution at a statistical level, the textile industry can continue to move towards sustainability, reduce environmental impact, and deliver better products to consumers.

VI. DISCUSSION

The discussion of the results of this study opens up space for a deeper understanding of the impact of sunlight on textile products that use natural dyes. Changes in the image histogram show a significant change in the image of woven fabrics made from natural dyes due to exposure to sunlight. These changes reflect shifts in color distribution that can affect the visual appearance of textile products. This is in accordance with the common observation that textile products exposed to sunlight often experience color changes.

Changes in the statistical value of woven fabric imagery caused by sunlight can vary between drying intervals and times and the type of natural dye used in woven fabrics made of cotton. Chi-Square test results, mode, and standard deviation are significant, this study provides statistical validation of the color changes observed in the histogram. This confirms that the color change is not the result of chance, but a significant effect of sun exposure.

The response to sunlight can vary between the types of natural dyes used. This is a very important finding, as it suggests that some textile products may be more resistant to the effects of sunlight than others. This can provide an opportunity for manufacturers to choose natural dyes or woven fabrics that are more suitable for specific applications.

The results of this study have practical implications for producers and consumers in the textile industry. Manufacturers can use this knowledge to develop products that are more resistant to sun-induced discoloration or provide better care guidance to consumers. Consumers, on the other hand, can use this information to maintain the quality of textile products they have under conditions of sun exposure.

This research supports aspects of sustainability in the textile industry. With a better understanding of the impact of sunlight on textile products that use natural dyes, manufacturers can reduce resource wastage, reduce the risk of unwanted discoloration, and increase efficiency in product care. The results of this study can be the foundation for the development of natural dyes that are more resistant to sunlight. This is a positive step in reducing the use of synthetic dyes that negatively impact the environment and human health, as well as creating more sustainable alternatives.

However, it's important to remember that this study still has some limitations. One is to focus on sun exposure as the only external factor affecting textile products. In real-world situations, textile products may be exposed to a variety of external factors, such as changing weather, humidity, and air pollution. Therefore, further research may consider the interaction between sunlight and these factors.

VII. CONCLUSION

This study illustrates the impact of sunlight on textile products, especially woven fabrics made from natural dyes. This impact focuses on the emphasis of the image histogram and the analysis of the Chi-Square, mode, and standard deviation of the image histogram. The contribution made to this research slightly contributed to the understanding of the textile industry, especially woven fabrics made from natural dyes. The statistical value gained in the discussion, helps manufacturers to create products that are more resistant to the effects of sunlight and supports sustainability efforts in this industry. Further research can expand the results of this study by considering other external factors and supporting the development of natural dyes that are more resistant to sunlight.

This research has not received accuracy from textile experts. This research is one way to determine the image quality of woven fabrics made from natural dyes. Further development is needed between the results of this research and the process in the textile world.

ACKNOWLEDGMENTS

Acknowledgments are conveyed to DTRPM Kemdikbudristekdikti for funding this research through Research Contract: No.1625/LL15/DT.04.91/2023 and LPPM Widya Mandira Catholic University for helping distribute the research funds.

AUTHORS CONTRIBUTION

Patrisius Batarius: Contributed to chi-square analysis and statistical analysis of histogram images.
Alfry Aristo Jansen Sinlae: Image data capture before and after sun exposure.

COPYRIGHT

REFERENCES

[1] S. Saxena and A. S. M. Raja, “Natural Dyes: Sources, Chemistry, Application and Sustainability Issues BT - Roadmap to Sustainable Textiles and Clothing: Eco-friendly Raw Materials, Technologies, and Processing Methods,” S. S. Muthu, Ed., Singapore: Springer Singapore, 2014, pp. 37–80. doi: 10.1007/978-981-287-065-0_2.

[2] T. Toprak and P. Anis, “Textile Industry’s Environmental Effects and Approaching Cleaner Production and Sustainability: an Overview,” *J. Text. Eng. Fash. Technol.*, vol. 2, no. 4, pp. 429–442, 2017, doi: 10.15406/jteft.2017.02.00066.

[3] A. K. Sarkar, “An evaluation of UV protection imparted by cotton fabrics dyed with natural colorants,” *BMC Dermatol.*, vol. 4, pp. 1–8, 2004, doi: 10.1186/1471-5945-4-15.

[4] C. R. S. de Oliveira, A. H. da Silva Júnior, J. Mulinari, and A. P. S. Immich, “Textile Re-Engineering: Eco-responsible solutions for a more sustainable industry,” *Sustain. Prod. Consum.*, vol. 28, pp. 1232–1248, 2021, doi: 10.1016/j.spc.2021.08.001.

[5] P. Batarius, A. A. Sinlae, and E. F. Fahik, “Analysis of the Quality of Natural Dyes in Weaving Exposed to Sunlight Using MSE and PSNR Parameters,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 5, pp. 797–802, 2022, doi: 10.29207/resti.v6i5.4339.

[6] M. B. Jawan, P. Batarius, and F. Tedy, “Analisis Pengaruh Sinar Matahari Terhadap Pewarna Alami pada Citra Kain Tenun,” *J. Tek. Inform. Unika ST. Thomas*, vol. 07, no. 02, pp. 141–151, 2022.

[7] S. Sofyan, F. Failisnur, and S. Silfia, “Pengaruh jenis dan metode mordan terhadap kualitas pewarnaan kain katunmenggunakan limbah kulit jengkol (Archidendron jiringa),” *J. Litbang Ind.*, vol. 8, no. 1, pp. 1–9, 2018.

[8] T. Rahayuningsih, F. S. Rejeki, E. R. Wedowati, and D. Widhowati, “Preliminary study of natural dyes application on batik,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 475, no. 1, 2020, doi: 10.1088/1755-1315/475/1/012069.

[9] T. B. Teklemedhin, “Dyeing of Wool Fabric Using Natural Dye and Natural Mordant Extracts,” *Trends in Textile Engineering & Fashion Technology*, vol. 4, no. 4. 2018. doi: 10.31031/tteft.2018.04.000593.

[10] Y. G. Naisumu, Emilia Juliyanti Bria, and Welli Herlince Kasse, “Utilization of Natural Dyes for Futus Woven Fabrics as an Alternative to Dye for Plant Tissue Preparations,” *Bioeduscience*, vol. 6, no. 1, pp. 96–107, 2022, doi: 10.22236/j.bes/618079.

[11] N. Kusumawati, A. Kristyanto, and S. Samik, “Exploration of Natural Dyes by Using a Combination of Caesalpinia sappan and Leucaena leucocephala L. Leaves.,” vol. 1, no. Snk, pp. 33–37, 2020, doi: 10.2991/snk-19.2019.9.

[12] Ö. Erdem Işmal, L. Yildirim, and E. Özdoğan, “Use of almond shell extracts plus biomordants as effective textile dye,” *J. Clean. Prod.*, vol. 70, pp. 61–67, 2014, doi: 10.1016/j.jclepro.2014.01.055.

[13] E. Rahayuningsih *et al.*, “Optimization of cotton fabrics dyeing process using various natural dye extracts,” *J. Rekayasa Proses*, vol. 16, no. 1, p. 58, 2022, doi: 10.22146/jrekpros.70397.

[14] M. Raze, F. Telegin, and S. Rahman, “Eco-friendly dyeing of wool fabric using natural dye extracted from onions outer shell by using water and organic solvents,” *ResearchGate*, vol. 3, no. 9, pp. 1–19, 2016, [Online]. Available: https://www.researchgate.net/publication/309779229%0AEco-friendly

[15] R. Mansour, S. Dhouib, and F. Sakli, “UV Protection and Dyeing Properties of Wool Fabrics Dyed with Aqueous Extracts of Madder Roots, Chamomiles, Pomegranate Peels, and Apple Tree Branches Barks,” *Journal of Natural Fibers*, vol. 19, no. 2. pp. 610–620, 2022. doi: 10.1080/15440478.2020.1758280.

[16] L. Alwi, A. T. Hermawan, and Y. Kristian, “Identifikasi Biji-Bijian Berdasarkan Ekstraksi Fitur Warna, Bentuk dan Tekstur Menggunakan Random Forest,” *J. Intell. Syst. Comput.*, vol. 1, no. 2, pp. 92–98, 2019, doi: 10.52985/insyst.v1i2.93.

[17] E. L. Santoso, E. Setyati, and Y. Kristian, “Klasifikasi Citra Daun Memanfaatkan Angular Partition, Edge Detection dan Neural Network,” *J. Intell. Syst. Comput.*, vol. 1, no. 1, pp. 27–33, 2019, doi: 10.52985/insyst.v1i1.32.

[18] S. Alamgunawan and Y. Kristian, “Klasifikasi Tekstur Serat Kayu pada Citra Mikroskopik Veneer Memanfaatkan Deep Convolutional Neural Network,” *J. Intell. Syst. Comput.*, vol. 2, no. 1, pp. 06–11, 2021, doi: 10.52985/insyst.v2i1.152.

[19] M. Moezzi, M. Ghane, and D. Semnani, “Predicting the tensile properties of UV degraded Nylon66/polyester woven fabric using regression and artificial neural network models,” *J. Eng. Fiber. Fabr.*, vol. 10, no. 1, pp. 1–11, 2015, doi: 10.1177/155892501501000101.

[20] H. Sadeghi and A.-A. Raie, “Approximated Chi-square distance for histogram matching in facial image analysis: Face and expression recognition,” in *2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP)*, 2017, pp. 188–191. doi: 10.1109/IranianMVIP.2017.8342346.

[21] Z. Wang *et al.*, “Image Noise Level Estimation by Employing Chi-Square Distribution,” in *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 2021, pp. 1158–1161. doi: 10.1109/ICCT52962.2021.9657946.

Identifying Types of Corn Leaf Diseases with Deep Learning

Rahul Firmansyah[1] **and Nur Nafiiyah**[1]
[1]Informatics Engineering Study Program, Faculty of Engineering, Lamongan Islamic University, Indonesia

Corresponding author: Nur Nafiiyah (e-mail: mynaff@unisla.ac.id).

ABSTRACT The government is trying to increase corn yields to meet the Indonesian population's food needs and for export abroad. Some farmers have yet to gain experience with the types of diseases in corn, so they need tools or systems to guide and provide information to new farmers. Many previous studies have developed automatic systems to identify corn leaf diseases, with the goal of increasing corn crop production by early recognition and control. We propose a system for identifying types of corn leaf diseases using the CNN (Convolutional Neural Network) method to be more precise in recognizing corn diseases early on. The methods used in previous research mostly used deep learning with high accuracy results above 90%. CNN is one of the deep learning methods, so we use it to identify types of leaf diseases. Our data comes from Kaggle; we process it first. The Kaggle dataset has corn plants similar to those in Indonesia, so we use this data with identification classes (Blight, Common rust, Gray leaf spot, and Healthy). The training data is 2000 images with 500 images for each class, and the testing data is 120 images with 30 images for each class. The evaluation results show that the classification process using the CNN method has an accuracy of 84.5%. The results we produced for identifying types of corn leaf disease still lack accuracy in their prediction, indicating the need to improve the CNN architecture model.

KEYWORDS CNN, Corn Leaves, Identification, Type of Disease

I. INTRODUCTION

The need for corn for food in Indonesia is increasing, and the government is trying to strengthen national food. Areas where corn is grown include North Sumatra, South Sumatra, Lampung, Central Java, East Java, Nusa Tenggara, North Sulawesi, South Sulawesi, and Maluku. And the government has developed a strategy to increase corn yields to meet Indonesia's demand and for export. Although some farmers are keen to increase rice production, several obstacles have arisen, such as disease and pest attacks on corn [1]. Farmers with experience in corn production are better equipped to handle the various diseases and pests that affect the crop. However, for novice and inexperienced farmers, recognizing the different types of corn diseases and pests requires information and guidance. Several previous studies have created a simulation system for identifying types of disease in corn [2]. An automatic system-based identification system simply inputs an image of a corn leaf and it will display information on the type of corn leaf disease.

The automatic system for identifying types of leaf disease uses machine learning methods with extraction feature methods from texture and color from RGB (Red Green Blue), HSV (Hue, Saturation, Value), L*a*b images [3]-[5]. On average, the automatic system for identifying leaf disease types using machine learning (Naive Bayes, K-Nearest Neighbor (k-NN), SVM (Support Vector Machine) has an accuracy of 70-90%. So there is a lot of research trying to increase accuracy for identification. The aim of developing a system for identifying leaf disease types is to help increase corn crop production. Because if diseases in corn can be controlled and recognized early, there is a chance of increasing crop production.

Research related to identifying types of leaf diseases using deep learning methods continues to develop, both using pretrained transfer learning architectures and creating your own architecture [1][6]-[11]. From previous research, the process of identifying types of corn leaf disease using the CNN (Convolutional Neural Network) method has an accuracy of above 90%. Therefore, to improve accuracy, we used CNN to classify the types of corn leaf diseases. We hypothesized that modifying the CNN architecture could improve the accuracy of detecting corn leaf disease types The purpose of this research is to create a system to detect types of corn leaf diseases. Differences between our research and previous ones [10], We create a CNN architecture with four times the number of convolution layers and our image size is 256x256.

II. LITERATURE REVIEW

Research related to identifying types of corn leaf diseases is included in table 1.

TABLE I
LITERATURE REVIEW

No	Method	Results
1	Feature extraction using texture (contrast value, correlation, energy, homogeneity, average, standard deviation) from L*a*b images, and classification process using k-NN [3]	Accuracy 73.3%
2	Using GLCM feature extraction from grayscale images, and HSV image feature values, then classified using k-NN [4]	70% Accuracy
3	Identifying types of leaf diseases using pretrained deep learning methods [8]	Validation data accuracy 88%
4	Identification of types of corn leaf disease from the mean features, standard deviation of RGB, HSV, and YCbCr images totaling 18 features, and 4 GLCM features (contrast, correlation, homogeneity, and energy), and the classification process with SVM [12]	99.5% Accuracy
5	Classification of types of corn leaf diseases using deep learning, input image size 32x32 [10]	94% Accuracy
6	Classification of types of corn leaf disease using ResNet50 and 224x224 image input [9]	98.3% Accuracy
7	Classifying types of corn leaf disease using HSV and GLCM (Angular Second Moment, Inverse Difference Moment, entropy and correlation) feature extraction, k-NN classification method [5]	84% Accuracy
8	Create a simulation system for identifying corn diseases, but based on 46 symptoms and 15 types of pest diseases [2]	Accuracy 73.3%
9	Identify types of corn leaf diseases using pretrained deep learningEfficientNetB0 architecture [11]	96% Accuracy
10	Identify types of corn leaf disease with CNN and 150x150 color image input [7]	94% Accuracy
11	Identify types of corn leaf disease with CNN and 50x50 image input [6]	99.9% Accuracy
12	The process of extracting the image features of corn leaves using CNN VGG-16 and 150x150 images, and then the process of classifying the types of corn leaf diseases using SVM, k-NN, and MLP [13]	SVM accuracy 93.8%, k-NN 92.1%, and MLP 94.4%
13	Classification of types of corn leaf disease with AlexNet and an input image size of 256x256 [1]	90% Accuracy

III. *METHOD*

A. DATASETS

Research data is taken from the Kaggle dataset [14]. We resize the image to 256x256. The distribution of training and testing data is presented in table 2. The number of classes in this study is four: blight, common_rust, gray_leaf_spot, and healthy, and each has the image data of corn leaves as shown in Figure 1. The data received from Kaggle was grouped by class in the form of folders. Images are stored in folders for each class.

TABLE II
DATASETS

No	Type	Training	Testing	Total
1	Blight	500	30	530
2	Common_rust	500	30	530
3	Gray_leaf_spot	500	30	530
4	Healthy	500	30	530
Total		2000	120	2120

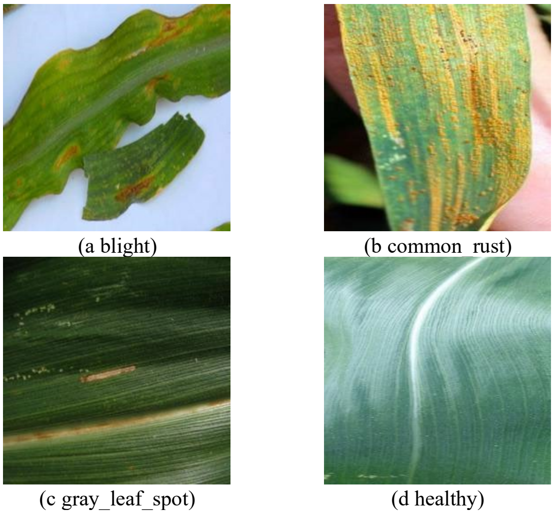

(a blight)
(b common_rust)
(c gray_leaf_spot)
(d healthy)

FIGURE 1. Example of a corn leaf dataset

B. DEEP LEARNING

Convolutional Neural Networks are very similar to standard artificial neural networks, or units arranged in the form of an acyclic graph (a graph without any cycles in it), which can be represented as a collection of neurons. The difference between CNNs is that there are hidden layers that are only connected by a subset of neurons in the previous layer. This kind of connection allows CNN to implicitly understand features. The CNN architecture produces hierarchical feature extraction through the use of filters trained for a specific purpose. In the first layer, the focus is often on recognizing edges or color changes. In the second layer, attention shifts to shape recognition. Filters in subsequent layers are generally focused on learning details from partial parts of objects, both those seen on a small scale and those seen on a larger scale. The last layer in the CNN is used to identify the object as a whole. In this feature extraction layer, an image entered into the model will be encoded into numbers. This layer consists of two elements, namely the Convolutional layer and the Polling Layer. The convolution process in image data aims to produce features from the input image using filters. These filters have weights designed to detect object characteristics, such as curved lines, edges, or color changes. The activation function is an operation for recognizing nonlinearity and improving the representation of the model. The ReLU activation function is the output value of the neuron can be expressed as 0 if the input is negative. If the input value is positive, then the output of the neuron is the activation input value itself. Pollor subsampling is the process of reducing the size of image data or matrices with the aim of overcoming unnecessary fluctuations (overfitting) by the model. At this stage, the commonly used method is Max Pooling, which is known for using the area of the pooling input feature map to get the maximum value. This method is popular because it takes a region of the input feature map and extracts its maximum value. Flatten can convert all 2-dimensional arrays smoothed by feature maps into a single linear vector to become a fully connected input layer. A fully connected layer comes from the previous process of determining the features most related to a particular class. The function of this layer is to unite all nodes into one size.

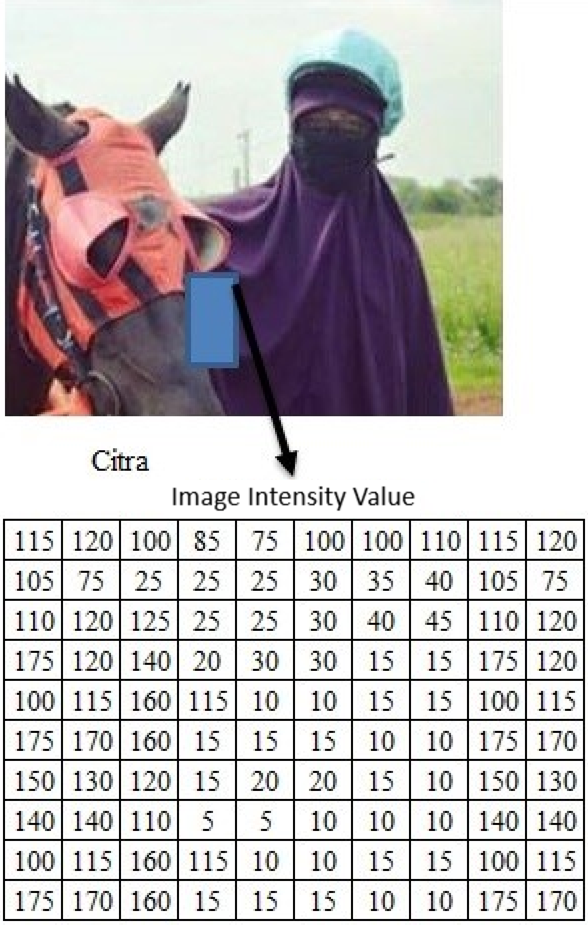


Image Intensity Value

115	120	100	85	75	100	100	110	115	120
105	75	25	25	25	30	35	40	105	75
110	120	125	25	25	30	40	45	110	120
175	120	140	20	30	30	15	15	175	120
100	115	160	115	10	10	15	15	100	115
175	170	160	15	15	15	10	10	175	170
150	130	120	15	20	20	15	10	150	130
140	140	110	5	5	10	10	10	140	140
100	115	160	115	10	10	15	15	100	115
175	170	160	15	15	15	10	10	175	170

FIGURE 2. **Image intensity value**

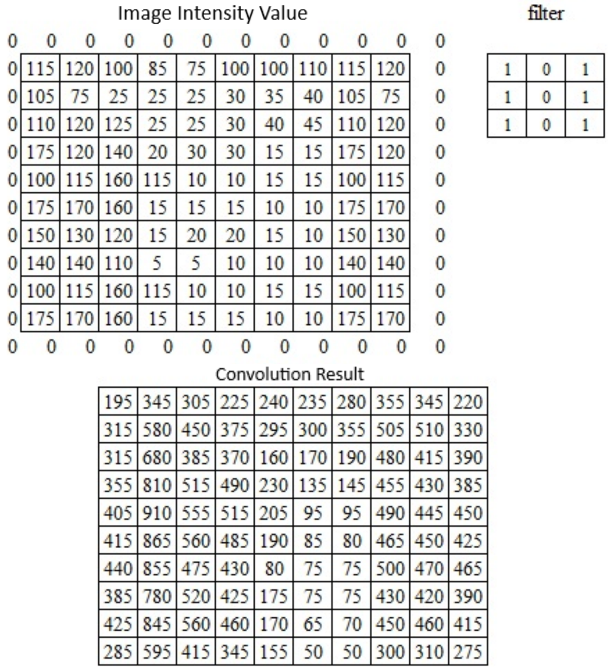


Image Intensity Value

0	0	0	0	0	0	0	0	0	0	0	0
0	115	120	100	85	75	100	100	110	115	120	0
0	105	75	25	25	25	30	35	40	105	75	0
0	110	120	125	25	25	30	40	45	110	120	0
0	175	120	140	20	30	30	15	15	175	120	0
0	100	115	160	115	10	10	15	15	100	115	0
0	175	170	160	15	15	15	10	10	175	170	0
0	150	130	120	15	20	20	15	10	150	130	0
0	140	140	110	5	5	10	10	10	140	140	0
0	100	115	160	115	10	10	15	15	100	115	0
0	175	170	160	15	15	15	10	10	175	170	0
0	0	0	0	0	0	0	0	0	0	0	0

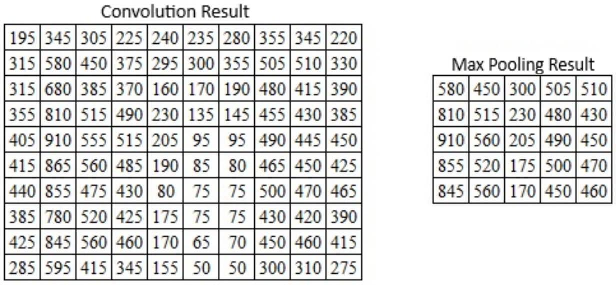
filter

1	0	1
1	0	1
1	0	1

Convolution Result

195	345	305	225	240	235	280	355	345	220
315	580	450	375	295	300	355	505	510	330
315	680	385	370	160	170	190	480	415	390
355	810	515	490	230	135	145	455	430	385
405	910	555	515	205	95	95	490	445	450
415	865	560	485	190	85	80	465	450	425
440	855	475	430	80	75	75	500	470	465
385	780	520	425	175	75	75	430	420	390
425	845	560	460	170	65	70	450	460	415
285	595	415	345	155	50	50	300	310	275

FIGURE 3. **Example of convolution**



Convolution Result

195	345	305	225	240	235	280	355	345	220
315	580	450	375	295	300	355	505	510	330
315	680	385	370	160	170	190	480	415	390
355	810	515	490	230	135	145	455	430	385
405	910	555	515	205	95	95	490	445	450
415	865	560	485	190	85	80	465	450	425
440	855	475	430	80	75	75	500	470	465
385	780	520	425	175	75	75	430	420	390
425	845	560	460	170	65	70	450	460	415
285	595	415	345	155	50	50	300	310	275

Max Pooling Result

580	450	300	505	510
810	515	230	480	430
910	560	205	490	450
855	520	175	500	470
845	560	170	450	460

FIGURE 4. **Example of max pooling**

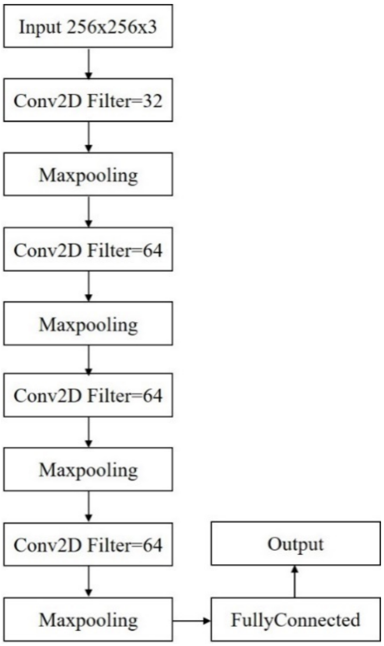
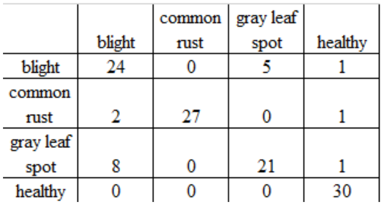


FIGURE 5. **Proposed CNN Architecture**



	blight	common rust	gray leaf spot	healthy
blight	24	0	5	1
common rust	2	27	0	1
gray leaf spot	8	0	21	1
healthy	0	0	0	30

FIGURE 6. Metric confusion results

Softmax activation transforms values from a numeric vector into a probability vector, where each possible value is proportional to the relative scale of each value in the vector. Each output value from softmax activation is interpreted as a probability in each class.

Image as in Figure 2. We took a sample of a particular part with a size of 10x10. We illustrate the convolution process of an image (Figure 3) with a size of 10x10 (Figure 2) and a filter size of 3x3. An image of size 10x10 has varying intensity values . It is then multiplied by a 3x3 filter, which results in the convolution of the same image of size 10x10, but the intensity value of each pixel is different. The convolution results take the maximum value for every 2x2 pixels to produce a max pooling process (Figure 4). The max pooling result is the best feature result from the maximum value, and the image size is reduced to 2 times smaller, for example, initially 10x10 to 5x5.

This research proposes a CNN architecture, as in Figure 5. We propose a convolution layer four times and a pooling layer four times, and the results of the feature extraction layer or convolution layer are trained. Input image of corn leaves measuring 256x256 in color. The feature map resulting from the convolution layer is 16x16 in size, meaning it has 256 feature maps.

A convolution layer is a layer that carries out the convolution process, namely multiplying each image pixel with a filter. The purpose of the convolution layer is to produce features from the image. The pooling layer is a layer that takes the best features from the convolution layer in order to represent the average image or the maximum. Our proposal uses a convolution architecture four times and pooling four times to make the extracted features more detailed. The more pooling layers, the more detailed the feature values obtained and caused the image size to decrease.

IV. RESULTS

We conducted training data experiments using optimizer={'rmsprop','sgdm'}, and learning rate={0.01;0.001}. We carried out training four times, each with 50 epoch iterations. Optimizer training results='rmsprop' with learning rate=0.001 in table 3 and optimizer training results='rmsprop' with learning rate=0.01 in table 4. Optimizer training results='sgdm' with learning rate=0.001 in table 5 and with a learning rate value = 0.01 in table 6. The results of the confusion metric evaluation are as in Figure 6. In the confusion metric evaluation results, identifying the type of leaf disease that has 100% accuracy is healthy. Moreover, the evaluation results of confusion metrics with low accuracy are gray leaf spots of only 63%.

TABLE III
TRAINING OPTIMIZER='RMSPROP' WITH LEARNING RATE 0.001

Epoch	*Iteration*	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:14	16.41%	2.0853	0.0010
4	50	00:10:52	67.19%	1.8346	0.0010
7	100	00:22:12	87.50%	0.3519	0.0010
10	150	00:33:31	77.34%	0.4690	0.0010
14	200	00:44:51	90.63%	0.2029	0.0010
17	250	00:56:11	95.31%	0.1122	0.0010
20	300	01:07:27	92.97%	0.2130	0.0010
24	350	01:18:43	96.09%	0.1001	0.0010
27	400	01:29:57	96.09%	0.0992	0.0010
30	450	01:41:13	99.22%	0.0410	0.0010
34	500	01:52:45	84.38%	0.4603	0.0010
37	550	02:04:23	100.00%	0.0211	0.0010
40	600	02:15:37	99.22%	0.0740	0.0010
44	650	02:26:51	100.00%	0.0091	0.0010
47	700	02:38:07	100.00%	0.0174	0.0010
50	750	02:49:30	99.22%	0.0430	0.0010

TABLE IV
TRAINING OPTIMIZER='RMSPROP' WITH LEARNING RATE 0.01

Epoch	*Iteration*	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:13	28.13%	2.3345	0.0100
4	50	00:11:34	54.69%	2.6889	0.0100
7	100	00:22:56	71.09%	0.6431	0.0100
10	150	00:34:28	64.84%	1.5427	0.0100
14	200	00:46:00	81.25%	0.4941	0.0100
17	250	00:57:31	85.16%	0.3802	0.0100
20	300	01:09:05	82.81%	0.3834	0.0100
24	350	01:20:37	93.75%	0.1867	0.0100
27	400	01:32:08	92.19%	0.1770	0.0100
30	450	01:43:32	85.16%	0.3998	0.0100
34	500	01:54:49	95.31%	0.1365	0.0100
37	550	02:06:08	96.09%	0.1779	0.0100
40	600	02:17:25	95.31%	0.1193	0.0100
44	650	02:28:42	95.31%	0.1163	0.0100
47	700	02:39:59	92.97%	0.1245	0.0100
50	750	02:51:17	97.66%	0.0610	0.0100

TABLE V
TRAINING OPTIMIZER='SGDM' WITH LEARNING RATE 0.001

Epoch	*Iteration*	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:13	16.41%	2.0853	0.0010
4	50	00:11:37	87.50%	0.2358	0.0010
7	100	00:23:25	90.63%	0.2112	0.0010
10	150	00:35:12	96.09%	0.1714	0.0010
14	200	00:46:39	100.00%	0.0492	0.0010
17	250	00:58:01	100.00%	0.0443	0.0010
20	300	01:09:20	100.00%	0.0279	0.0010
24	350	01:20:37	100.00%	0.0547	0.0010
27	400	01:31:54	100.00%	0.0232	0.0010
30	450	01:43:10	97.66%	0.1204	0.0010
34	500	01:54:23	100.00%	0.0401	0.0010
37	550	02:05:35	98.44%	0.0845	0.0010
40	600	02:16:48	93.75%	0.2289	0.0010
44	650	02:28:01	93.75%	0.1443	0.0010
47	700	02:39:14	96.09%	0.1229	0.0010
50	750	02:50:26	99.22%	0.0314	0.0010

TABLE VI
TRAINING OPTIMIZER='SGDM' WITH LEARNING RATE 0.01

Epoch	*Iteration*	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:12	21.09%	2.2360	0.0100
4	50	00:11:24	80.47%	0.6456	0.0100
7	100	00:22:56	86.72%	0.3969	0.0100
10	150	00:34:28	91.41%	0.1420	0.0100
14	200	00:46:02	99.22%	0.0703	0.0100
17	250	00:57:35	98.44%	0.0731	0.0100
20	300	01:09:04	99.22%	0.0532	0.0100
24	350	01:20:31	99.22%	0.0336	0.0100
27	400	01:31:57	99.22%	0.0493	0.0100
30	450	01:43:33	97.66%	0.0896	0.0100
34	500	01:55:13	95.31%	0.1032	0.0100
37	550	02:06:53	88.28%	0.2562	0.0100
40	600	02:18:33	97.66%	0.0858	0.0100
44	650	02:30:09	99.22%	0.0336	0.0100
47	700	02:41:35	100.00%	0.0172	0.0100
50	750	02:53:03	100.00%	0.0312	0.0100

Table 7 results from the average accuracy value when testing data by changing the optimizer= {'rmsprop', 'sgdm'}, and learning rate= {0.01; 0.001}. The result of changes in the optimizer that has the highest accuracy is 'sgdm', and the learning rate is 0.001.

TABLE VII
TESTING EVALUATION RESULTS

Optimizer	*Learning rate*	Testing Accuracy (%)
SGDM	0.001	87
SGDM	0.01	84
RMSPROP	0.01	82
RMSPROP	0.001	85

TABLE VIII
COMPARISON RESULTS OF RELATED RESEARCH

Method	Accuracy (%)
Our Proposal	84.5
AlexNet[1]	90
CNN[6]	99.9
CNN[7]	94
ResNet50[9]	98.3
Deep Learning[10]	94
EfficientNetB0[11]	96

Table 8 compares deep learning/CNN methods for recognizing corn leaf diseases. Our proposal has low accuracy compared with previous research.

V. CONCLUSION

We created a system for identifying leaf disease types using deep learning. Our dataset is sourced from Kaggle, and we only use 2120 images with four disease classes: blight, common rust, gray leaf spot, and healthy. The testing results for identifying types of corn leaf disease were 84.5%.

ACKNOWLEDGMENTS

Thank you to Lamongan Islamic University.

AUTHORS CONTRIBUTION

Rahul Firmasyah: reviewing, writing, and testing research; **Nur Nafiiyah:** revising and reviewing manuscripts and experiments;

COPYRIGHT

REFERENCES

[1] QN Azizah, "Classification of Corn Leaf Disease Using the AlexNet Convolutional Neural Network Method,"*sudo J. Tech. Inform.*, vol. 2, no. 1, 2023, doi: 10.56211/sudo.v2i1.227.

[2] MM Suhadi, MA Helmi, and W. Setiawan, "SIMULATION OF CLASSIFICATION OF PESTS AND DISEASES IN CORN

USING NAIVE BAYES,"*J. Simantec*, vol. 10, no. 1, 2022, doi: 10.21107/simantec.v10i1.11686.

[3] M. Khoirotul Ummah, Nur Nafi'iyah, "Identification of Corn Leaf Diseases Based on Texture Using K-Nn,"*Musamus J. Technol. Inf.*, vol. 02, no. 01, 2019, [Online]. Available: http://www.ejournal.unmus.ac.id/index.php/it/article/view/2419%0Ahttps://www.ejournal.unmus.ac.id/index.php/it/article/download/2419/1318.

[4] EH Rachmawanto and HP Hadi, "OPTIMIZATION OF FEATURE EXTRACTION ON KNN IN CLASSIFICATION OF CORN LEAF DISEASES,"*Dynamic*, vol. 26, no. 2, 2021, doi: 10.35315/dynamic.v26i2.8673.

[5] MA Setyawan, P. Kasih, M. Ayu, and D. Widyadara, "Classification of Corn Leaf Disease Based on HSV Color Space and Texture Features Using the K-NN Algorithm," in*National Seminar on Technology Innovation at UN PGRI Kediri*, 2022, pp. 67–72.

[6] AB Prakosa, Hendry, and R. Tanone, "Implementation of a Deep Learning Convolutional Neural Network (CNN) Model on Corn Leaf Disease Images for Plant Disease Classification,"*J. Educator. Technol. Inf.*, vol. 6, no. 1, 2023.

[7] D. Iswantoro and D. Handayani UN, "Classification of Corn Plant Diseases Using the Convolutional Neural Network (CNN) Method,"*J. Ilm. Univ. Batang Hari Jambi*, vol. 22, no. 2, 2022, doi: 10.33087/jiubj.v22i2.2065.

[8] MI Rosadi and M. Lutfi, "Identifying Types of Corn Leaf Disease Using Deep Learning Pre-Trained Models,"*J. Explor. IT!*, vol. 13, no. 2, 2021.

[9] IP Putra, R. Rusbandi, and D. Alamsyah, "Classification of Corn Leaf Disease Using the Convolutional Neural Network Method,"*J. Algorithm.*, vol. 2, no. 2, 2022, doi: 10.35957/algoritme.v2i2.2360.

[10] AD Nurcahyati, RM Akbar, and S. Zahara, "Classification of Disease Images on Corn Leaves Using Deep Learning with the Convolution Neural Network (CNN) Method,"*SUBMIT J. Ilm. Technol. Information and Science*, vol. 2, no. 2, 2022, doi: 10.36815/submit.v2i2.1877.

[11] F. Sarasati, F. Septia Nugraha, and U. Radiyah, "Utilization of Deep Learning Methods for Disease Classification in Corn Plants,"*J. Infortech*, vol. 1, no. 1, 2022.

[12] R. Suhendra, I. Juliwardi, and S. Sanusi, "Identification and Classification of Corn Leaf Disease Using Support Vector Machine,"*J. Technol. Inf.*, vol. 1, no. 1, 2022, doi: 10.35308/.v1i1.5520.

[13] J. Kusuma, Rubianto, R. Rosnelly, Hartono, and BH Hayadi, "Classification of Leaf Diseases in Corn Plants Using Support Vector Machine, K-Nearest Neighbors and Multilayer Perceptron Algorithms,"*J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, 2023, doi: 10.52158/jacost.v4i1.484.

[14] Kaggle, “Corn Leaf Disease Dataset.” https://www.kaggle.com/datasets/smaranjitghose/corn-or-maize-leaf-disease-dataset.

Procedural Map Generation for 'Splatted': Enhancing Player Experience through Genetic Algorithms and AI Finite State Machines in a Snowball Throwing Game

Lukky Hariyanto[1] and Hendrawan Armanto[1]

[1]Informatics Department, Faculty of Science and Technology, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia

Corresponding author: Lukky Hariyanto (e-mail: lukky.h20@mhs.istts.ac.id).

ABSTRACT Games, a now extremely prevalent form of global entertainment, have emerged as a leading industry in the entertainment media, surpassing other entertainment media such as books, films, and music. However, game development is a complex endeavor, requiring a diverse set of talents to create a decent game for people to enjoy. Some of the talents needed to create a good game is a game designer, which dictates how a player can interact with the world, a writer, which pours a meaningful story inside said world, and a composer, which uses music to elevate the emotions evoked by the game and its events. With that being said, this research aims to streamline the creation process of the game designers, specifically the level designers by focusing on procedural map generation and artificial intelligence to create a map that is in a playable state for the players to play in. Procedural map generation, facilitated by a genetic algorithm inspired by Darwin's evolutionary theory, expedites the level design process. The research explores two types of map generation—tile-based and template-based, each with distinct advantages and disadvantages. Through user acceptance tests and expert-level analysis, it is evident that the genetic algorithm performs effectively, achieving a noteworthy level of player satisfaction.

KEYWORDS Game, Genetic Algorithm, Map, Procedural Map Generation.

I. INTRODUCTION

Every human being needs entertainment in their life. Forms of entertainment may vary, one of which are games. No less than other types of entertainment such as films or books, games are a type of entertainment that requires a high level of technical complexity in its creation. Various components such as story, gameplay, system balancing, and marketing are needed in designing a game [1]–[3]. The more complex a game, the more complex its constituent components will be. Some of the important components in making games is level design and artificial intelligence for NPCs. Even though these two components are not the top priority in making a game, without a good level of design and believable AI from the NPCs, the game will feel bland.

In making a level, level design requires a lot of effort in terms of time, assets, and the endurance of the designer who designs it. Level creators must consider various things starting from the player's position, enemy position, item position, good path arrangements so that the game is interesting and balanced, and a few other considerations. For example, Live Service games such as Valorant, Apex Legends, Fortnite, or Overwatch 2 require continuous map updates so that players don't get bored of playing and switch to another game. Because of this complexity, various research was carried out to make it easier for designers to carry out level design which ultimately give birth to Procedural Map Generation [4]–[6]. Some examples of algorithms in procedural map generation are Perlin Noise [7] which is used in the game Minecraft, Fractal Terrain Generation [8] in the game Terraria, or Genetic algorithms [9]–[15] in various existing studies.

This research focuses on creating a snowball throwing game called Splatted. The creation of the level design for this game will not be done manually but automatically using a genetic algorithm. Apart from that, in this game artificial intelligence will also be developed which can influence the behavior of Non Player Characters (NPC) so that the game can be more interesting. It is expected that through this research, similar games can apply the methodology so that the games developed can become more interesting.

II. "SPLATTED" GAME

This game is developed for the purpose of research and as a case study in testing. This sub-chapter will discuss details related to gameplay and the artificial intelligence being developed.

A. GAMEPLAY

Splatted is a game in which two teams, each consisting of five players, engage in a snowball war. Each successful hit on an opposing player scores points, with the primary objective being to accumulate the highest points and emerge as the winner of the snowball war. The winning team is either the one that meets the required point target or has more points than the opposing team when the game time ends. Players can perform several actions in this game, including picking up the ball from the ground, throwing the ball with the goal of hitting an opposing player, catching the ball thrown by an opposing player, and executing a "Fakeout" to deceive an opponent attempting to catch the thrown ball. Figure 1 provides an example of Splatted game footage where the white area represent the game area (snow), and the dark gray colors represent rocks (obstacles). Players and snowball throwers are only allowed to move within the snow areas.

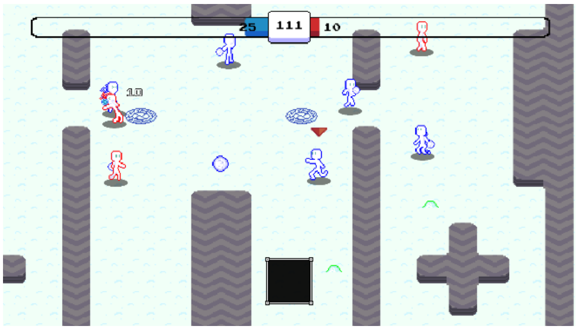
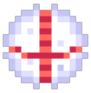
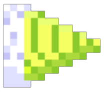


Figure 1. Screenshot of the Splatted game

B. SPECIAL BALL

Apart from the general snowballs, to add some spice into the game, several special spawners are provided in the levels. Each spawner has a special ball that can be picked up and used to attack opponents. Table 1 is a table of the special balls available and the function of these balls.

TABLE I
SPECIAL BALLS IN SPLATTED

Name	Display	Function
Ice Piercer		Pierces past opposing players and teammates
Snow-A-Rang		The ball returns to the thrower after hitting someone
Explod-o-Ball		The ball explodes after a set duration and hits all the players in the explosion area
Freezing Winter		The player hit by the thrown ball is slowed down for several seconds
Stone Auger		Penetrates players and walls, then breaks into 3 normal smaller snowballs

C. ARTIFICIAL INTELLIGENCE (NPC)

In Splatted games, the behavior of Non-Player Characters (NPCs) is governed by a Finite State Machine (FSM). Utilizing an FSM allows NPCs to have several states, each providing different behaviors. Transitions between states in the FSM are influenced by real-time conditions of the NPCs, which could be affected by other NPCs, players, or the surrounding environment. The following section will describe the various states that an NPC can possess:

1) RANDOM WALKING

If the NPC does not have a snowball, it will walk to a randomly selected location, attempting to find a snowball on the ground along the way. However, if the NPC already possess a snowball, it will start searching for opponents. If the NPC reaches the target location without finding a snowball or an opponent, a new location will be randomly selected, and the search will resume.

2) TAKE THE BALL

If the NPC in "Random Walking" state spots a snowball or a special ball, it will transition to the "Take the Ball" state. In this state, the NPC will walk towards the identified ball to pick it up. After securing the ball, or if the ball is taken by someone else, the NPC will revert to the "Random Walking" state.

3) AIM & THROW

If an NPC in the 'Random Walking' state has a ball in hand and spots a member of the opposing team, the NPC will transition to the 'Aim & Throw' state. In this state, the NPC will cease movement and aim at the opponent. After confirming the aim is accurate, the NPC will throw the ball towards the predicted future position of the opponent. Following the throw, the NPC will revert to the 'Random Walking' state.

4) FOLLOW TARGET

If the target being aimed at in the "Aim & Throw" state disappears from the NPC's view, the NPC will transition to the "Follow Target" state. In this state, the NPC will pursue the opponent with the goal of regaining vision of the target. During this pursuit, if the NPC fails to locate the opponent within a certain timeframe, the NPC will abandon the chase and revert to the "Random Walking" state.

5) CATCH BALL

Specifically, the 'Catch Ball' state can be entered from any other state when the NPC spots a ball being thrown in its direction. The purpose of this state is to enable the NPC to attempt to catch balls that are being thrown at it.

III. IMPLEMENTATION OF GENETIC ALGORITHM INTO SPLATTED GAMES

In this sub-chapter, we will delve into the details of using genetic algorithms in Splatted games. This includes everything from the representation and fitness values, to the methods employed in each operation of the genetic algorithm.

A. REPRESENTASION

In this research, two level generation models are utilized: tile-based generation and template-based generation. Each model has its own unique process and method of representation.

Tile-Based Generation

In this model, the individual representation is a 1-dimensional array with a length equal to the size of the level to be created (for instance, for a 10x10 level, the individual length would be 100). Each gene in this individual will hold a value ranging from 0 to 3, where the number 0 represents an empty area, 1 represents a rock/obstacle, 2 represents the position of a special ball spawner, and 3 represents the player's position. Figure 2 illustrates an example of a tile-based representation converted into a player-understandable level, with the level size being 3x3."

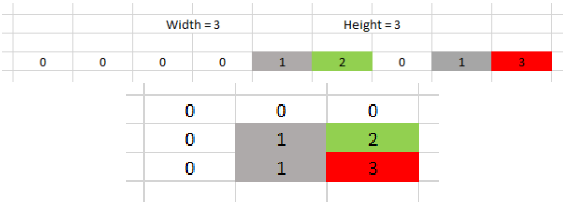


Figure 2. Example of Tile Based Representation into a map

Template-Based Generation

Similar to tile-based generation, template-based generation is also represented as a 1-dimensional array. However, unlike tile-based generation, the gene value in this model does not contain the numbers 0-3, which represent objects at the level. Instead, it has values ranging from -n to n, where n is the number of prepared templates. A negative value indicates that there will be one special ball in the middle of the level in the template with ID number x. Apart from the gene value, another difference from the tile-based representation is the length of the individual. Template-based representations have shorter individual lengths because one template consists of 5x5 tiles. For instance, if the level size is 10x10, the individual length used is 100/(5x5), or 4.

This research incorporates three types of templates, each of which has several variations. The three types of templates are as follows:

1. Oneway Template
   A Oneway template refers to a variation of the template that, when rotated by 90°, 180°, or 270°, still produces the same level. For instance, in Figure 3, the level is represented by code 12. This type of template is available in only three variations.
2. Twoway Template
   A Twoway template refers to a variation of the template that, when rotated by 90° or 270°, yields different level results. However, when rotated by 180°, it produces the same level. For instance, in Figure 3, the level is represented by code 5. This type of template is available in eight variations.
3. Fourway Template
   A 'Fourway' template refers to a variation of the template that, when rotated by 90°, 180°, or 270°, yields different level results. For instance, in Figure 3, the level is represented by either code 8 or 24. This type of template is available in five variations

In this research, a total of 39 templates were provided. These were derived from 3 oneway templates, 16 twoway templates (8 variations x 2), and 20 fourway templates (5 variations x 4). For twoway and fourway templates, a single variation will yield 2 and 4 different levels respectively when implemented. Figure 3 illustrates an example of a template-based representation converted into a player-understandable level, with the level size being 10x10.

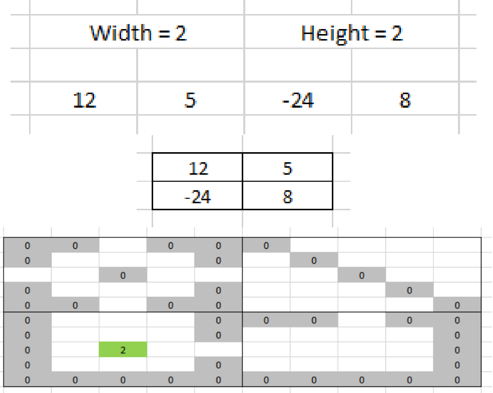
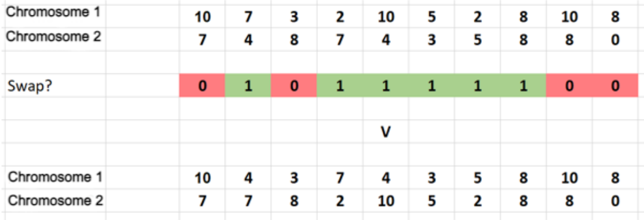


Figure 3. Example of Template-Based Representation Converted into a Map

B. GENETIC ALGORITHM OPERATOR

Genetic algorithms encompass several operators, each with a variety of algorithmic choices. This research has explored these algorithms to identify the most suitable and appropriate one for accomplishing level creation, which is the primary focus of this research. The operators and their respective choices are as follows:

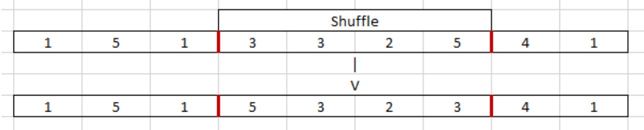
1. Parent Selection
   The selection method [16][17] used by this research is the Roulette Wheel [18]. This method is suitable for use in this research considering that the better the fitness score, the greater the chance of an individual or level candidate being selected as a parent. Where levels that are not good are generally less playable so that if they are selected it will cause the next generation to also be less playable.
2. Crossover
   This research uses the uniform crossover algorithm [19]-[22] as the operator. Where this algorithm will provide a 50% chance for each gene to be exchanged between the 2 selected parents. Figure 4 is an example of uniform crossover visualization. This algorithm is suitable to be applied because in this study 1 gene represents 1 tile/template. So that the levels produced in the next generation will have good variations.



Chromosome 1	10	7	3	2	10	5	2	8	10	8
Chromosome 2	7	4	8	7	4	3	5	8	8	0
Swap?	0	1	0	1	1	1	1	1	0	0
					V					
Chromosome 1	10	4	3	7	4	3	5	8	10	8
Chromosome 2	7	7	8	2	10	5	2	8	8	0

Figure 4. Visualization of Uniform Crossover

3. Mutation
   For the same reasons as selecting the crossover algorithm, the mutation algorithm [23][24] partial shuffle mutation or scramble mutation was chosen and used in this study. This algorithm has more opportunities for gene changes in the next generation considering that all genes between the two barriers will be randomized in order and place. Figure 5 is an example of scramble mutation visualization.



			Shuffle					
1	5	1	3	3	2	5	4	1
				|				
				V				
1	5	1	5	3	2	3	4	1

Figure 5. Visualization of Scramble Mutation

4. Elitism
   The elitism method [25][26] used in this research is to combine all offspring (children) and some parents who have the best fitness scores in the previous generation. For example, the minimum population that must be provided is 100 while the previous generation gave birth to 80 units, the remaining 20 are taken from parents who have the best fitness scores.

C. STOP CONDITION

The stopping condition for the genetic algorithm in this research is convergence. If all individuals across the last 100 generations have not shown significant development or have converged, then the iteration of the genetic algorithm is halted. This approach is adopted considering that level formation occurs during the initial game loading and needs to be efficient to prevent long waiting times for players. Although it's undeniable that level formation can sometimes take a considerable amount of time (around 20 seconds), this duration is still within the player's tolerance for waiting time.

IV. FITNESS FUNCTION

The fitness function plays a crucial role in determining the quality of an individual in the genetic algorithm. In this research, five fitness functions will be utilized for tile-based generation, and four fitness functions will be employed for template-based generation. Notably, three out of the four fitness functions for template-based generation are also used in tile-based generation.

A. FITNESS NUMBER OF STONES

This fitness function is utilized to regulate the distribution of stones within a level. The primary objective is to ensure that a level doesn't consist solely of stones, or conversely, lack of it entirely. Certain parameters are initially established, such as MinR (the minimum number of stones in a level) and MaxR (the maximum number of stones in a level).

$$m = \begin{cases} MinR - R, \ for\ R < MinR \\ R - MaxR, \ for\ R > MaxR \\ 0, \ for\ MinR < R < MaxR \end{cases} \tag{1}$$

As can be seen in (1), the value of 'm' will be made negative when the number of stones is either too few (less than MinR) or too many (greater than MaxR). However, if it falls within the range, the value of 'm' will be set to 0, indicating that there are no constraints on the number of stones in that level.

$$M = \begin{cases} A - MaxR, for\ MinR > A - MaxR \\ MinR, \ for\ MinR < A - MaxR \end{cases} \tag{2}$$

To ensure that the fitness value does not become negative, normalization is performed on the value of 'm'. Equation (2) is used to find the divisor to ensure proper normalization. There are two methods to find the divisor for normalization: subtracting the area with the maximum number of stones, or directly taking the minimum number of stones when the result of subtracting the area and the maximum number of stones is less than the minimum number of stones

$$F = \left(1 - \frac{m}{M}\right)^2 \tag{3}$$

Equation (3) represents the normalization equation utilized in this fitness function. This function ensures that the F value will never be negative. It's important to note that this fitness function is exclusively used in tile-based generation. This is because, in template-based generation, the number of stones is determined by the composition of the used template. Therefore, applying this function in a template-based context would disrupt the variation occurences in the existing templates.

B. GROUP STONE SIZE FITNESS

The fitness function is utilized to calculate the size of stone groups present within a level. The primary objective of this fitness function is to ensure that no stone group is excessively large or too small. Similar to the previous fitness function, (1) and (2) remain employed in computing this fitness function. However, unlike the previous fitness function, which was calculated for one level, in this fitness function, both equations are computed for each stone group encountered.

$$F = \left(\frac{\sum_{i=1}^{n}\left(1 - \frac{m_i}{M}\right)}{n}\right)^2 \tag{4}$$

To calculate the fitness value of a stone group size, we need to first determine the values of m and M for all stone groups. Once we have these values, we will use (4) to calculate the fitness value. It's important to note that the stone group size fitness is only applied in tile-based generation, and not in template-based generation.

C. ACCESSIBLE AREA FITNESS

This fitness function is employed to calculate the extent of the area accessible to the player. The more interconnected areas within the level, the better the level is considered. The primary objective of this function is to ensure that there are few inaccessible areas within the level, as the player character in the splatted game cannot pass through roofs or destroy obstacles.

$$F = \left(\frac{a_{terbesar}}{a_{total}}\right)^2 \tag{5}$$

Equation (5) represents the fitness function used to evaluate the quality of a level based on its area. Here, the largest value of a corresponds to the largest connected area, while the total value of a represents the total number of objects in the level that are not stones.

D. SPECIAL BALL ACCESSIBILITY FITNESS

This fitness function aims to ensure that each special ball present in the level is reachable by a player. However, this accessibility is not for all players but only for the closest player to the ball. Thus, the appearance of special balls in the level ensures that they are reachable by at least the nearest player.

$$F = \left(\frac{p_{akses}}{p_{total}}\right)^2 \tag{6}$$

Equation (6) is the fitness value equation that ensures all special balls can be reached by the nearest player. Ptotal represents the total number of special balls, while Pakses is the number of special balls that can be reached without any obstacles by the nearest player. A special ball is deemed accessible if a path search using the A* algorithm can return a path from the nearest player to the special ball without any obstacles.

E. SPECIAL BALL RATIO FITNESS

The special ball ratio fitness function is used to ensure the presence of special balls in a level. Several predetermined parameters are defined beforehand, including MinP (the minimum number of special balls in a level) and MaxP (the maximum number of special balls in a level).

$$m = \begin{cases} MinP - P, \ for\ P < MinP \\ P - MaxP, \ for\ P > MaxP \\ 0, \ for\ MinP < P < MaxP \end{cases} \tag{7}$$

As depicted in (7), it can be observed that the value of mm will be made negative when the number of special balls is insufficient (less than MinP) or excessive (greater than MaxP). However, if the number of special balls falls within the specified range, the value of m will be set to 0, indicating no constraints on the number of special balls in the level.

$$M = \begin{cases} A - MaxP, for\ MinP > A - MaxP \\ MinP, \ for\ MinP < A - MaxP \end{cases} \tag{8}$$

To prevent fitness values from becoming negative, normalization is performed on the value of m. Equation (8) presents the equation to find the divisor for proper normalization. There are two approaches to finding the divisor for normalization: subtracting the area from the maximum number of special balls or directly taking the minimum number of special balls when the difference between the area and the maximum number of special balls is less than the minimum number of stones.

$$F = \left(1 - \frac{m}{M}\right)^2 \tag{9}$$

Equation (9) represents the normalization equation used in this fitness function. Through this equation, it is ensured that the fitness value F will never be negative.

F. TEMPLATE VARIETY FITNESS

The final fitness function utilized in this research is template variety. The objective of this fitness function is to avoid repetitive occurrences of the same template within a level. Consequently, the aim is to generate levels with a high template variety, ensuring that multiple templates are used rather than repeating one or two templates multiple times.

$$x = \begin{cases} 0 & for\ t_i - T_l > 0 \\ t_i - T_l & for\ t_i - T_l < 0 \end{cases} \qquad (10)$$

As shown in (10), it can be observed that the value of x will be made negative when the number of occurrences of a template exceeds the given tolerance limit. However, if it is within the tolerance, the value of x will be set to 0. This calculation is performed for each template encountered.

$$F = \begin{cases} 0 & for\ \frac{\sum_{i=1}^{t}(1+x)}{t} < 0 \\ \left(\frac{\sum_{i=1}^{t}(1+x)}{t}\right)^2 & for\ \frac{\sum_{i=1}^{t}(1+x)}{t} > 0 \end{cases} \qquad (11)$$

After calculating the occurrences of all templates on a level using (10), (11) is employed to compute the fitness value by calculating the squared average of the x values. If the resulting value is negative, the fitness value is set to 0 to avoid interference with the values of other fitness functions. This fitness function is specifically utilized for template-based generation.

V. EXPERIMENTAL TESTING

This research employs two testing techniques. The first one involves user acceptance testing or questionnaire-based assessment. The second technique involves analyzing the levels generated by game experts.

A. USER ACCEPTANCE

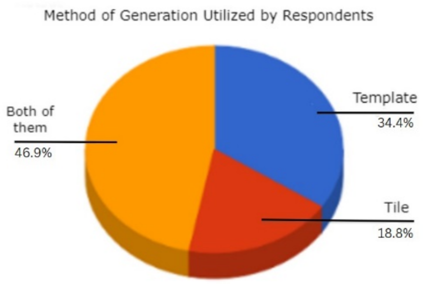


Figure 6. Level Generation Selection by Respondents

The questionnaire was completed by 32 individuals whose profile fits that of gamers aged 18–22 years, with a minimum of 2–3 hours of daily gaming, and who have previously played MOBA or 2D real-time strategy games such as Bomberman. Figure 6 depicts the percentage of respondents selecting different level generation modes. 46% opted to try both modes, while the remainder only tried one of the modes.

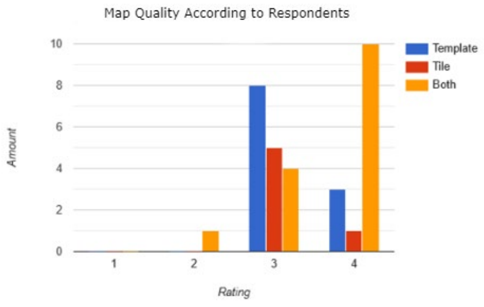


Figure 7. Respondents' Given Scores

Meanwhile, Figure 7 illustrates the ratings provided by respondents who have tried both modes as well as those who

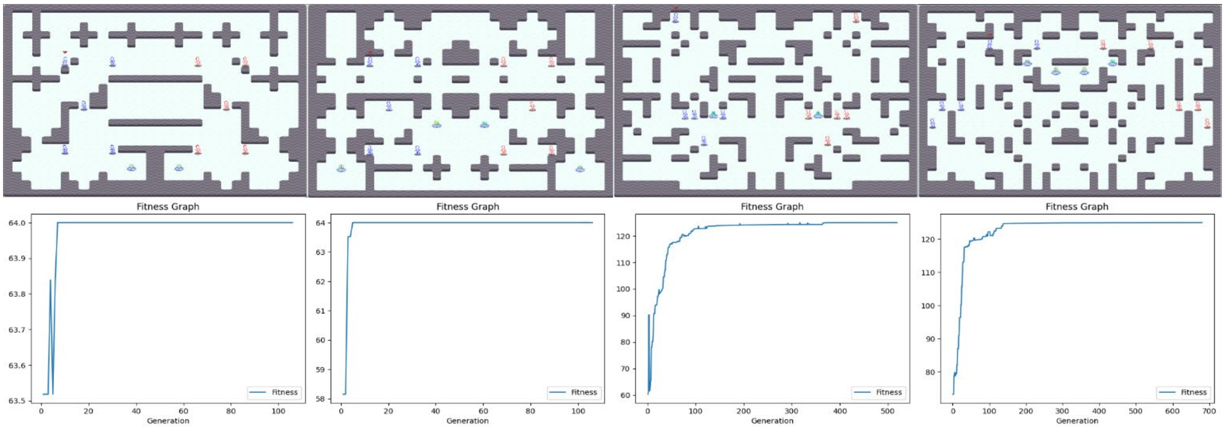


Figure 8. Visualization of Fitness Flow in Generating a Map

have tried only one mode. The questionnaire results reveal that only one respondent found the generated levels to be unsatisfactory, while the remaining 31 individuals rated the levels with a minimum score of 3. Through Figure 7, it can be inferred that the generated levels meet players' expectations.

B. ANALYSIS OF THE LEVELS

In the level analysis conducted by expert evaluation, several levels were generated using both tile-based generation and template-based generation methods. The expert, with a profile matching that of a high-rated MOBA player who spends 5-6 hours gaming daily, was asked to analyze several levels. Figure 8 (from left to right) showcases the best 20 x 30 tile-sized levels for two examples of template-based generation and two examples of tile-based generation. From the four exemplary levels, it is evident that all levels feature open areas conducive to player strategy, with no inaccessible or enclosed spaces, appropriately placed special items, and a visually appealing aesthetic suitable for gameplay.

Regarding the analysis of the levels based on their fitness values, both levels generated by template-based generation exhibit a high speed in achieving maximum fitness or generating good levels. Meanwhile, the two levels produced by tile-based generation, although capable of attaining high fitness values, require more time due to the gradual changes that occur per tile in tile-based generation.

VI. CONCLUSION

In this study utilizing the splatted game, it can be concluded that genetic algorithms perform well in generating levels that are enjoyable, playable, and meet player expectations. However, the main drawback of level generation using genetic algorithms lies in the significant dependency on the fitness function employed. The quality of generated levels improves when the fitness function aligns accurately with the desired criteria. Achieving this alignment necessitates a considerable amount of time for experimenting with and refining various fitness function variants.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to the Surabaya Institute of Integrated Science and Technology for their support and facilitation of this research.

AUTHOR CONTRIBUTIONS

Lukky Hariyanto: Application Development, Article Writing, Image and Data Provision.
Hendrawan Armanto: Article Writing, Article Copyediting, Finishing.

COPYRIGHT

REFERENCES

[1] J. Schell, *The Art of Game Design: A Book of Lenses, Third Edition*. CRC Press, 2019.

[2] E. Adams, *Fundamentals of Game Design*. Pearson Education, 2010.

[3] M. Moore, *Basics of Game Design*. CRC Press, 2016.

[4] N. A. Barriga, *A Short Introduction to Procedural Content Generation Algorithms for Videogames*. 2018.

[5] V. Kraner, I. Fister jr, and L. Brezočnik, "Procedural Content Generation of Custom Tower Defense Game Using Genetic Algorithms," 2021, pp. 493–503.

[6] S. Putra and W. Istiono, "Implementation Simple Additive Weighting in Procedural Content Generation Strategy Game," *vol*, vol. 4, pp. 9–18, 2022.

[7] E. Frank and N. Olsson, "Procedural city generation using Perlin noise." 2017.

[8] N. Sainio, "TERRAIN GENERATION ALGORITHMS," 2023.

[9] A. Lambora, K. Gupta, and K. Chopra, "Genetic Algorithm- A Literature Review," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 380–384, doi: 10.1109/COMITCon.2019.8862255.

[10] L. Haldurai, T. Madhubala, and R. Rajalakshmi, "A study on genetic algorithm and its applications," *Int. J. Comput. Sci. Eng*, vol. 4, no. 10, pp. 139–143, 2016.

[11] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimed. Tools Appl.*, vol. 80, no. 5, pp. 8091–8126, 2021, doi: 10.1007/s11042-020-10139-6.

[12] H. Armanto, H. A. Rosyid, M. Muladi, and G. Gunawan, "Evolutionary Algorithm in Game – A Systematic Review," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, May 2023, doi: 10.22219/kinetik.v8i2.1714.

[13] W. Alfonsus, A. Hendrawan, and T. J. Gunawan, "Focused Web Crawler Using Genetic Algorithms and Symbiotic Organism Search," 革新的コンピューティング・情報・制御に関する速報, vol. 15, no. 12, p. 1345, 2021.

[14] H. Armanto, K. Setiabudi, and C. Pickerling, "Komparasi Algoritma WOA, MFO dan Genetic pada Optimasi Evolutionary Neural Network dalam Menyelesaikan Permainan 2048," *J. Inov. Teknol. dan Edukasi Tek.*, vol. 1, no. 9, pp. 676–684, 2021.

[15] H. Armanto, R. D. Putra, and C. Pickerling, "MVPA and GA Comparison for State Space Optimization at Classic Tetris Game Agent Problem," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 7, no. 1, pp. 73–80, 2022.

[16] S. Prayudani, A. Hizriadi, E. B. Nababan, and S. Suwilo, "Analysis effect of tournament selection on genetic algorithm performance in traveling salesman problem (TSP)," in *Journal of Physics: Conference Series*, 2020, vol. 1566, no. 1, p. 12131.

[17] S. L. Yadav and A. Sohal, "Comparative study of different selection techniques in genetic algorithm," *Int. J. Eng. Sci. Math.*, vol. 6, no. 3, pp. 174–180, 2017.

[18] J. Y. Setiawan, D. E. Herwindiati, and T. Sutrisno, "Algoritma Genetika Dengan Roulette Wheel Selection dan Arithmetic Crossover Untuk Pengelompokan," *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 58–64, 2019.

[19] J. L. Pachuau, A. Roy, and A. Kumar Saha, "An overview of crossover techniques in genetic algorithm," *Model. Simul. Optim. Proc. CoMSO 2020*, pp. 581–598, 2021.

[20] P. Kora and P. Yadlapalli, "Crossover operators in genetic algorithms: A review," *Int. J. Comput. Appl.*, vol. 162, no. 10, 2017.

[21] A. Malik, "A study of genetic algorithm and crossover techniques," *Int. J. Comput. Sci. Mob. Comput.*, vol. 8, no. 3, pp. 335–344, 2019.

[22] L. Manzoni, L. Mariot, and E. Tuba, "Balanced crossover operators in genetic algorithms," *Swarm Evol. Comput.*, vol. 54, p. 100646, 2020.

[23] B. H. Abed-alguni, "Island-based cuckoo search with highly disruptive polynomial mutation," *Int. J. Artif. Intell.*, vol. 17, no. 1, pp. 57–82, 2019.

[24] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, and V. B. S. Prasath, “Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach,” *Information*, vol. 10, no. 12, p. 390, 2019.

[25] G. Guariso and M. Sangiorgio, “Improving the performance of multiobjective genetic algorithms: An elitism-based approach,” *Information*, vol. 11, no. 12, p. 587, 2020.

[26] H. Du, Z. Wang, W. E. I. Zhan, and J. Guo, “Elitism and distance strategy for selection of evolutionary algorithms,” *IEEE Access*, vol. 6, pp. 44531–44541, 2018.

Predictive Buyer Behavior Model as Customer Retention Optimization Strategy in E-commerce

Muhammad A. A. Hakim[1] and Terttiaavini[2]
[1] Digital Business, Faculty of Business and Social Sciences, Binawan University, Indonesia
[2] Computer Science, Faculty of Computer Science and Sciences, Indo Global Mandiri University, Indonesia

Corresponding author: Muhammad A. A. Hakim (e-mail: muhammadariv@gmail.com)

ABSTRACT Lazada is one of the rapidly growing E-commerce platforms in this digital era. One of the main challenges faced by Lazada is customer retention, where customers make purchases once or a few times before switching to other platforms. Therefore, it is important to understand buyer behavior in E-commerce through customer prediction to identify factors influencing retention. This study employs the Random Forest (RF) method to analyze Lazada customer data and formulate more effective marketing strategies. The analysis is conducted by loading preprocessed datasets into the KNIME workflow and utilizing various nodes and algorithms available in KNIME to build and evaluate predictive models. The Random Forest model is trained multiple times to achieve the highest Accuracy rate, which is 72.472%, with a fairly high level of agreement and a balanced trade-off between recall and precision. Additionally, this model successfully predicts that customers purchasing electronic equipment are potentially churning at a rate of 3.85%. Subsequently, customer strategy analysis for customer retention optimization in the E-commerce industry is conducted through data visualization using Tableau. Predictive analysis of customer behavior serves as a strong foundation for formulating effective retention strategies in the E-commerce industry. With this approach, Lazada can enhance customer experience and ensure sustainability in facing the increasingly fierce competition in the digital market.

KEYWORDS Customer Retention Optimization Strategy, E-Commerce, Predictive Behavior Model, Random Forest

I. INTRODUCTION

In the current digital era, the e-commerce industry, such as Lazada, has emerged as one of the rapidly growing digital platforms [1]. One of the main challenges faced by the Lazada platform is customer retention. Customer retention refers to the desire of customers to make repeat purchases online [2]. It has become the primary focus for Lazada due to the high cost of customer acquisition and the long-term benefits that can be derived from loyal customers. Building a strong customer base on the E-commerce platform can be achieved by increasing the Lifetime Value (LTV) of customers, leveraging recommendations and positive reviews, offering incentives for repeat purchases, and enhancing customer loyalty through optimal experiences. These strategies are key to retaining customers in the long run [3].

Although Lazada offers convenience in shopping, customer retention remains the primary focus, considering that most of them only make purchases once or a few times before eventually switching to other platforms. Therefore, it is important to understand buyer behavior in E-commerce to identify factors influencing customer retention. The factors influencing customer retention are closely related to predictive methods. In predictive methods, these factors are used as features to create models that can predict whether a customer is likely to stay or churn. By understanding the factors affecting retention, predictive models can be better trained to identify behavioral patterns that lead to customer retention, thus enabling E-commerce platforms to take more proactive steps in retaining customers.

Customer prediction allows Lazada to group buyers based on shopping loyalty patterns, devise more effective marketing strategies, and enhance the overall shopping experience.

This study employs the Random Forest (RF) method due to its ability to handle complexity and noise in data, as well as its capability to address overfitting issues. By constructing multiple decision trees randomly, RF can generate more stable and accurate predictions compared to other predictive methods [4][5]. Previous research [6] addressed the same topic, focusing on consumer review classification using Random Forest and SMOTE. They reported an Accuracy rate of 75%, which increased to 77% when utilizing 8000 max_features [6]. Another study [7] aimed to predict customer churn in a campus fashion company by identifying customers who had not made transactions for more than 365 days as 'churned'. In this research, the Random Forest Classifier model achieved the highest Accuracy compared to three other Machine Learning models. The analysis results indicated that the customer churn rate was 24.54%, while the non-churned customers accounted for 75.46%. The top five customers originated from Jakarta, West Java, Central Java, East Java, and Yogyakarta provinces, with the highest total transaction value reaching 3,997,936,774 [7]. Furthermore, research [8] performed a descriptive analysis of Shopee user data. The results revealed that more users discontinued using Shopee (54.8%) compared to those who remained active (45.2%). The classification model utilized was Random Forest due to its superior performance [8].

Based on the findings of these studies, the Random Forest method has proven to be effective in addressing the challenge of customer retention in e-commerce platforms. With its demonstrated ability to generate accurate predictions regarding customer behavior, such as churn prediction or customer segmentation [5][9], Random Forest can assist platforms like Lazada in devising more targeted marketing strategies and enhancing the shopping experience for customers [10]. Furthermore, this research also identifies customer segments vulnerable to churn, enabling proactive efforts to minimize the rate of customers leaving the platform [11]. With a better understanding of customer behavior and factors influencing churn, Lazada can take appropriate steps to improve customer retention among those vulnerable to churn and enhance the overall user experience [12].

Based on the aforementioned background, the objectives of this research are: 1) To predict customer behavior on e-commerce platforms, particularly Lazada, using the Random Forest method, and 2) To design more effective marketing strategies based on these prediction results, aiming to enhance customer retention and ensure business sustainability in this competitive digital era. Thus, this research will provide a significant contribution to managing customer relationships and improving customer retention on e-commerce platforms.

II. LITERATURE STUDY

A. RANDOM FOREST (RF)

The Random Forest method is one of the techniques in data analysis used to predict or classify data. This technique works by dividing the dataset into many small subsets and then building decision trees for each of these subsets. Subsequently, the results from all these decision trees are combined or averaged to produce more accurate predictions. This method is commonly used in machine learning due to its ability to address overfitting and provide stable results in various situations. Random Forest is an ensemble algorithm that utilizes the concept of decision trees in its model formation. Although the Random Forest algorithm itself does not directly use entropy, the decision trees it employs can utilize entropy as one of the criteria for node splitting when constructing the tree. Equation (1) represents the formula for calculating entropy where Y is the set of cases and p(c | Y) represents the proportion of class c values in Y.

$$Entropy\ (Y) = -\sum_{i} \mathrm{p(c|Y)\ log_2\ p(c|Y)} \tag{1}$$

$$Information\ Gain\ (Y, a) = Entropy\ (Y) - \sum_{ve}values\ \frac{Y_v}{Y_a}\ Entropy\ (Y_v) \tag{2}$$

Equation (2) is the formula for calculating information gain where values $\frac{Y_v}{Y_a}$ a represent all possible values in the set of cases a, Y_v is the subclass of Y with class v related to class a. Y_a is all the values corresponding to a.
The selection of attributes to be used as nodes, either as roots or internal nodes, depends on the highest information gain value possessed by the available attributes. Information Gain is used to determine the most beneficial attribute in decision-making. The value of Information Gain can be found using the formula in (3), and the gain ratio value can be observed in (4) , where Split Information $Split\ Information\ (S, A)$ is the estimated entropy value of the input variable S which has class c, while $\frac{|S_i|}{|S|}$ represents the probability of class i within that attribute.

$$Split\ Information\ (S, A) = \sum_{i}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{3}$$

$$Gain\ ratio\ (S, A) = \frac{Information\ (S, A)}{Split\ information\ (S, A)} \tag{4}$$

B. CHURN ANALYSIS

Churn analysis is the process of understanding, identifying, and managing the behavior of customers or users who cease using a product or service. It is a crucial aspect of customer relationship management (CRM) and customer retention strategies. Here are the common steps in churn analysis:

1) CUSTOMER DATA COLLECTION

The initial step in churn analysis is gathering relevant customer data. This data may include information such as customer profiles, transaction history, interactions with products or services, and more.

2) DATA EXPLORATION

Once the data is collected, the next step is to explore the data to understand patterns and trends that may be associated with churn behavior. This involves descriptive statistical analysis, data visualization, and identification of features that may influence churn.

3) PREDICTIVE MODELING

One key aspect of churn analysis is building predictive models to forecast future churn behavior. This involves using machine learning techniques such as Logistic Regression, Decision trees, Random Forests, or neural networks. These models utilize historical customer data to predict the probability of churn for new or existing customers.

4) MODEL VALIDATION

After building the predictive model, the next step is to test and validate its performance using independent data. This is important to ensure that the model has good predictive capability for accurately forecasting churn behavior.

5) CUSTOMER SEGMENTATION

One common strategy used in churn management is to divide customers into smaller groups based on similar characteristics. This is called customer segmentation. This segmentation helps companies better understand customer behavior and design more effective retention strategies.

6) IMPLEMENTATION OF RETENTION STRATEGIES

Based on churn analysis results, companies can design and implement appropriate customer retention strategies. This may involve improving the customer experience, offering incentives, or loyalty programs to encourage customers to continue using products or services.

7) MONITORING AND EVALUATION

Lastly, it is important to continually monitor and evaluate the effectiveness of customer retention strategies. This allows companies to adjust and improve their strategies over time in response to changes in customer behavior and the business environment.

Churn analysis is a continuous and iterative process. By understanding the factors influencing churn and designing appropriate strategies, companies can reduce churn rates and improve customer retention, which in turn can contribute to the long-term growth and success of the company.

III. RESEARCH METHODOLOGY

The research methodology in this study comprises several stages that outline the process of data collection and analysis, as well as the development of customer retention strategies. The research stages are described in Figure 1 below.

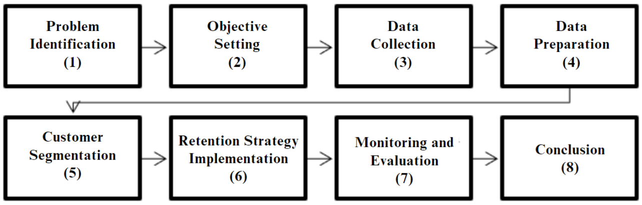

Problem Identification (1)
Objective Setting (2)
Data Collection (3)
Data Preparation (4)
Customer Segmentation (5)
Retention Strategy Implementation (6)
Monitoring and Evaluation (7)
Conclusion (8)

Figure 1. Research Stages

The stages are described in detail as follows:

1) DATA COLLECTION

The initial stage involves collecting transaction data, product preferences, and shopping behaviour from Lazada customers. This data can be obtained from Lazada's internal database or through customer survey methods. The data is obtained from sales transactions on Lazada in 2022. In this study, the focus is solely on the sales of electronic goods because the main objective is to understand customer behaviour in the context of purchasing electronic products specifically. The sales data on Lazada in 2022 is displayed in table 1.

TABLE I
ELECTRONIC SALES DATA ON LAZADA IN 2022

Dataset	Range / Frekuensi	Persentase
Catagory		
Harddisk-eksternal	4422	40,41
Laptop	701	6,41
Smart-tv	1290	11,79
China OEM	2	0,02
flash-drives	3318	30,32
televisi-digital	1211	11,07
BrandName		
No Brand	943	8,57
Merk Dll	1-20 / 920	8,37
Akari, Aoyama, Aqua, Bestrunner, Carcool, Hisense, HP COMPAQ, Ichiko, Import, Led Coocaa, Microsoft, MSI, Multi, Niko, OEM, Universal_Brand, Vandisk, Xiaomi	20-60 / 565	5,14
Apple, Changhong, Flashdisk, Ikedo, Orico, Philips, Sony,TCL,Vakind,VGen	51-100 / 707	6,43
Adata, Coocaa, Kingston, Panasonic, Transcend, Universal	101-200 / 852	7,75
Dell, LG, Seagate, Sharp, WD	201-300 / 1121	10,19
Polytron, Samsung	301-400 / 656	5,96
Acer, China OEM	401-500 / 864	7,86
HP	501-600 / 582	5,29
Toshiba	601-700 / 674	6,13
Asus, Lenovo	900-1000 / 1856	16,88
SanDisk	1000-1300 / 1258	11,44
TotalReviews		
Aoyama, AVITA, DBest, E-link, Hisense, Maxtor, Merk Lainnya, Microsoft, NYK, OEM, OneGood, OTG, PX, Robot, Sanyo, SelaluAda, SP, VANDISK, YYSL	51-100 / 1.399	0,47
Bmstore, Boneka_Nizza, Casing, Hardcase, Multi, Qflash, Redcolourful, Trisonic, Universal_Brand, Universally	101-150 / 1.186	0,40
Akari, Best CT, Bestrunner, Carcool, Hiqueen, JvGood, Sony, Vitron	151-200 / 1.424	0,48
Apple, EsoGoal, KLEVV, Max, M-Tech, Niko,Rendys chem, V-Gen	201-300 / 1.944	0,65
Ikedo, Lazada, Vakind	301-400 / 1.010	0,34
Trend's, Universal Indonesia	401-500 / 870	0,29
Aqua, Dell, Lexar, SS,Transcend	500-1000 / 3.985	1,33
Acer, Changhong, Good Shop, UGREEN	1001-1500 / 4.468	1,49
China OEM, Kingston, Panasonic, Universal, WARM	1501-2000 / 8.964	3,00
Ichiko, Orico, Seagate	2001-2500 / 6.525	2,18
Flashdisk, WD	2501-3000 / 5.471	1,83
Adata, LG	3001-3500 / 6.289	2,10
TCL	3500-4000 / 4.232	1,42

Lenovo, Philips, Polytron, Toshiba	5000-10000 / 30.079	10,06
Asus, Coocaa, Samsung, SanDisk, Sharp, Xiaomi	10000-70000 / 210.586	70,44
No brand	1.183	0,40

2) DATA PREPARATION (PRE-PROCESSING DATA)

The collected data will be prepared for further analysis, including data cleaning to remove anomalies or inconsistencies, as well as data processing to prepare it into a suitable format for analysis. In the data preparation stage, the processes include data cleaning, data processing, selection of the most relevant variables, and data validation. The data is obtained from sales on Lazada, which includes various variables such as itemId, Category, Name, BrandName, url, Price, AverageRating, TotalReviews, and RetrievedDate. From this data, the most relevant variables to be used in cluster analysis are itemId, Price, AverageRating, and TotalReviews.

3) PREDICTION ANALYSIS

Prediction analysis using the Random Forest model involves selecting, training, validating, and testing the model to make predictions on new data [13]. After validation and testing, the prediction results are evaluated using performance metrics such as Accuracy, Precision, Recall, and F1-score, while considering the interpretation of the results to understand the factors influencing the model's predictions.

In this study, the preprocessed data resulted in a cleaned dataset free from missing values, outliers, and duplicates. Additionally, relevant features have been extracted or processed, and the data has been transformed or normalized to ensure consistency and uniformity.

Prediction analysis is conducted using the KNIME application [14][15]. The analysis steps begin by loading the preprocessed dataset into the KNIME workflow. Various nodes and algorithms available in KNIME are then used to build and evaluate predictive models. The first step involves using the Excel Reader node and Column Filter to filter the data, including only relevant columns for prediction analysis, such as itemId, Price, AverageRating, and TotalReviews. The data is then partitioned using the Partitioning node, with a training dataset of 85% and a testing dataset of 15%. After that, the model is learned using the Random Forest Learner node, and tested using the Random Forest Predictor node with the test dataset. The prediction results are evaluated using appropriate evaluation metrics and can be viewed through the Score and Table View nodes. Figure 2 shows the workflow diagram of the prediction model using Random Forest (RF).

The Random Forest model was trained repeatedly with various parameter variations to achieve the highest Accuracy score. After several training sessions, the following results were obtained: Accuracy score = 72.472%; Cohen's Kappa = 0.691%; correct classified = 10524; wrong classified = 422; Recall score = 0.85; Precision score = 0.95, and F1-Score = 0.785.

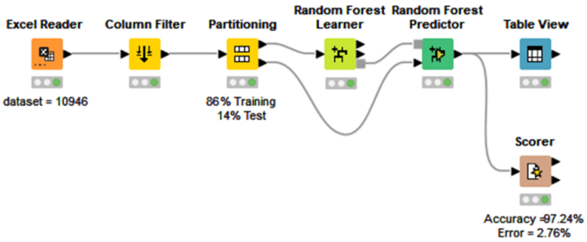


Figure 2. Workflow diagram for prediction model using Random Forest (RF)

The conclusions drawn from these testing results are as follows:

1. Accuracy score: The model has an Accuracy of 72.472%, indicating the percentage of correct predictions out of the total predictions made.
2. Cohen's Kappa: The Cohen's Kappa value is 0.691%, indicating the level of agreement between the model's predictions and the expected predictions, after correcting for chance agreement.
3. Correct classified: A total of 10524 samples were correctly classified by the model.
4. Wrong classified: There were 422 samples classified incorrectly by the model.
5. Recall score: The model's sensitivity, or recall score, is 0.85, indicating the model's ability to identify a large number of true positive cases.
6. Precision score: The model's precision rate is 0.95, depicting the proportion of true positive outcomes among all outcomes predicted positively by the model.
7. F1-Score: The harmonic mean of recall and precision, or F1-Score, is 0.785. This provides an overall overview of the model's performance by considering the balance between recall and precision.

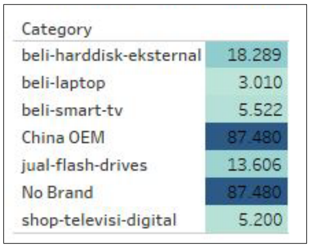
Overall, the model demonstrates a fairly good performance with decent Accuracy, a good balance between recall and precision, and a relatively high level of agreement with Cohen's Kappa at 0.691%. The prediction of churn customers is 422 customers.

4) CUSTOMER SEGMENTATION

One common strategy used in managing churn is to divide customers into smaller groups based on similar characteristics, known as customer segmentation. This segmentation helps companies better understand customer behavior and design more effective retention strategies. Analysis can be visualized using Tableau, enabling researchers to quickly observe patterns and trends in customer data [16][17]. Customer analysis is divided into several segments:

1. Sales Analysis Based on Number of Sales per Category

   Sales visualization based on the number of sales per category is presented in Figure 3.



Category	
beli-harddisk-eksternal	18.289
beli-laptop	3.010
beli-smart-tv	5.522
China OEM	87.480
jual-flash-drives	13.606
No Brand	87.480
shop-televisi-digital	5.200

Figure 3. **Sales Analysis Based on Number of Sales Per Category**

Based on product sales data, China OEM and No Brand show significant popularity with each reaching 87,480 units, while Buy External Hard Drives, Sell Flash Drives, and Buy Smart TVs have relatively high sales with 18,289, 13,606, and 5,522 units respectively. On the other hand, sales of Buy Laptops (3,010 units) and Shop Digital Televisions (5,200 units) are relatively lower. This analysis indicates that customer management strategies should emphasize maintaining the popularity of the most favored products while strengthening sales of less popular products by launching special promotions or enhancing customer service quality.

2. Sales Analysis Based on Category
Sales visualization based on category is presented in Figure 4.

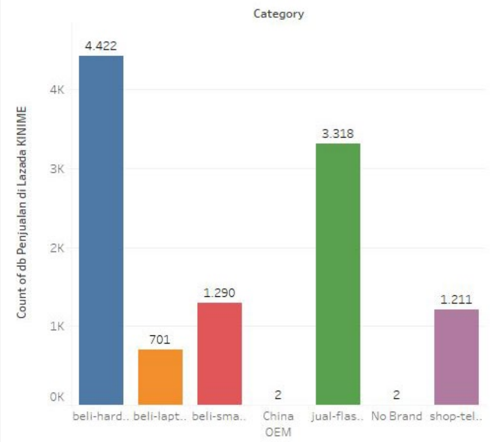


Figure 4. **Sales Analysis Based on Category**

China OEM leads the sales with 87,480 units, while Buy Laptops records the lowest sales with only 3,010 units. An effective customer management strategy should consider these differences, focusing on strengthening product availability and improving customer service quality for widely favored products, as well as developing more aggressive and innovative marketing strategies to increase interest and loyalty towards products with lower sales.

3. Sales Analysis Based on Brand Name
Sales visualization based on brand name is presented in Figure 5.

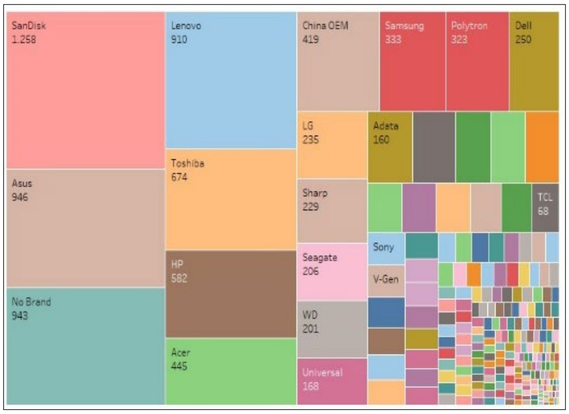


Figure 5. **Sales Analysis Based on Brand Name**

Based on sales data by BrandName, it can be seen that most sales are dominated by several major brands, such as Asus, Coocaa, Samsung, SanDisk, Sharp, and Xiaomi, which account for a total sales of 210,586 units or approximately 70.44% of total sales. Meanwhile, some other brands have lower contributions to sales, such as Apple, Dell, and LG, each having sales ranging from 201-300 to 3001-3500 units. Customer management strategies derived from this data include focusing on major brands dominating sales by improving product availability, providing special promotions, and enhancing customer service. On the other hand, attention should also be given to brands with lower sales by evaluating customer needs and preferences, as well as possibly developing more creative and targeted marketing strategies to increase brand interest and awareness.

4. Sales Analysis Based on Product Reviews
Sales visualization based on product reviews is presented in Figure 6.

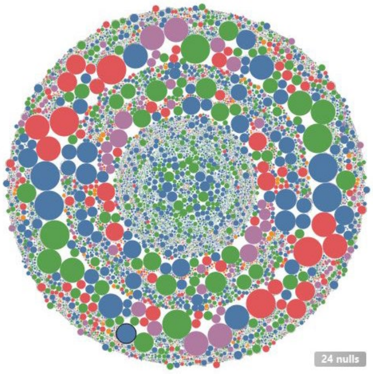


Figure 6. **Sales Analysis Based on Product Reviews**

Based on TotalReviews data, it can be observed that most brands have a diverse range of review counts. However, most brands have review counts concentrated in lower ranges, with approximately 70.44% of major brands like Asus, Coocaa, Samsung, SanDisk, Sharp, and Xiaomi having over 10,000 reviews. On the other hand, most other brands have review counts below 2,000. Customer management strategies can focus on brands with low review counts by increasing customer engagement, encouraging product reviews, and enhancing brand visibility. For brands with high review counts, it is important to maintain and strengthen relationships with loyal customers, encourage positive interactions, and provide prompt and satisfactory responses to reviews given. This will help increase customer trust and expand the loyal customer base.

5) IMPLEMENTATION OF RETENTION STRATEGIES

Based on the analysis of the four segments, the implementation of customer retention strategies can be carried out with a comprehensive approach. First, to improve customer retention based on sales per category, strategies can focus on maintaining the popularity of the most favored products such as China OEM and No Brand, while also strengthening the sales of less popular products like Buy Laptops and Shop Digital Televisions by launching special promotions or enhancing customer service quality. Second, considering the analysis of sales based on categories, retention strategies can focus on strengthening product availability and improving customer service quality for widely favored products, as well as developing more aggressive and innovative marketing strategies for products with low sales. Third, considering sales based on brands, retention strategies can focus on major brands dominating sales by improving product availability, providing special promotions, and enhancing customer service, while also paying special attention to brands with lower sales through evaluation of customer needs and preferences and development of more creative and targeted marketing strategies. Fourth, considering the analysis of sales based on product reviews, retention strategies can focus on increasing customer engagement and encouraging product reviews for brands with low review counts, while also maintaining and strengthening relationships with loyal customers and providing satisfactory responses to reviews for brands with high review counts. With a holistic and focused approach, the implementation of these retention strategies is expected to increase customer loyalty and strengthen the company's position in the e-commerce market. Monitoring and Evaluation. Finally, it is important to continuously monitor and evaluate the effectiveness of customer retention strategies. This allows the company to adjust and improve its strategies over time according to changes in customer behavior and the business environment.

IV. CONCLUSION

The findings of this research are as follows:

1. Predictive Analysis with Random Forest resulted in an accuracy rate of 72.472%, as well as a good balance between recall and precision. The test results indicate that the model performs well in predicting customer behavior based on factors influencing retention. The Random Forest method is capable of identifying customer segments vulnerable to churn, enabling proactive efforts to minimize the rate of customer attrition from the platform.
2. The Retention/Churn analysis shows that Lazada still faces challenges in maintaining the popularity of electronic products, where a large percentage of loyal customers accounts for 96.15%, while customers indicated as churning are 3.85%.
3. In enhancing customer retention on E-commerce platforms like Lazada, several key strategies can be employed. Firstly, by segmenting customers based on their shopping behavior, companies can better understand their needs to devise more effective marketing strategies. Then, by providing personalized experiences through product recommendations and relevant promotional offers, relationships with customers can be strengthened. Additionally, improving product availability and quality, as well as innovation in marketing, are also important to attract customer interest and maintain their loyalty. Continuous monitoring and evaluation of retention strategies are also necessary to adjust the company's approach to changes in customer behavior and market trends. By implementing these strategies comprehensively, Lazada can strengthen its position in the E-commerce market and sustain business growth.

AUTHORS CONTRIBUTION

Muhammad Arif Abdul Hakim: Investigation, data collection, analysis, review writing, and editing.
Terttiaavini: Data analysis, validation, Interpretation of results, visualization.

COPYRIGHT

REFERENCES

[1] M. A. A. Hakim, “Analysis Of The Influence Of Content Personalisation On Consumer Purchasing Decisions In The E-commerce Industry,” *J. MIRTE*, vol. 2, no. 2, pp. 10–11, 2023.

[2] M. I. Eid, “Determinants of E-commerce customer satisfaction, trust, and loyalty in Saudi Arabia,” *J. Electron. Commer. Res.*, vol. 12, no. 1, pp. 78–93, 2011.

[3] M. Harahap, Y. Lubis, and Z. Situmorang, “Data Science bidang Pemasaran : Analisis Prilaku Pelanggan,” *Data Sci. Indonesai*, 2021.

[4] N. F. Sahamony, T. Terttiaavini, and H. Rianto, “Analisis Perbandingan Kinerja Model Machine Learning untuk Memprediksi Risiko Stunting pada Pertumbuhan Anak,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 413–422, 2024, [Online]. Available: https://journal.irpi.or.id/index.php/malcom/article/view/1210

[5] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, “Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 257–265, 2024, doi: 10.57152/malcom.v4i1.1078.

[6] N. Istiqamah and M. Rijal, “Klasifikasi Ulasan Konsumen Menggunakan *Random Forest* dan SMOTE,” *J. Syst. Comput. Eng.*, vol. 5, no. 1, pp. 66–77, 2024, doi: 10.61628/jsce.v5i1.1061.

[7] T. N. Muthmainnah and A. Voutama, "Pendekatan Data Science Untuk Menemukan Customer *Churn* Pada Perusahaan Fashion Dengan Metode Machine Learning," *J. Teknol. Sist. Inf. dan Sist. Komput. TGD*, vol. 6, no. 2, pp. 463–471, 2023, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsk/index

[8] W. Fajrin Mustafa, S. Hidayat, and D. Hatta Fudholi, "Prediksi Retensi Pengguna Baru Shopee Menggunakan Machine Learning," *J. Media Inform. Budidarma* , vol. 8, no. 1, pp. 612–623, 2024, doi: 10.30865/mib.v8i1.7074.

[9] N. F. Sahamony, T. Terttiaavini, and H. Rianto, "Analysis of Performance Comparison of Machine Learning Models for Predicting Stunting Risk in Children ' s Growth," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. April, pp. 413–422, 2024.

[10] A. A. S. Bratan, H. Taan, and Y. L. Ismail, "Economics and Digital Business Review Meningkatkan Retensi Pelanggan Paket Data Internet Telkomsel di Kota Gorontalo melalui Strategi Customer Relationship Management (CRM)," *Econ. Digit. Bus. Rev.*, vol. 5, no. 1, pp. 47–56, 2024.

[11] N. Suryana, "Prediksi *Churn* Dan Segmentasi Pelanggan TV Berlangganan (Studi Kasus Transvision Jawa Barat)," *J. TEDC*, vol. 11, no. 2, pp. 185–191, 2019, [Online]. Available: https://ejournal.poltektedc.ac.id/index.php/tedc/article/view/77

[12] A. S. Wibowo, "Analisis *Churn* Nasabah Bank Dengan Pendekatan Machine Learning dan Pengelompokan Profil Nasabah dengan Pendekatan Clustering," *Konstr. Publ. Ilmu Tek. Perenc. Tata Ruang dan Tek. Sipil*, vol. 2, no. 1, pp. 30–41, 2024, doi: https://doi.org/10.61132/konstruksi.v2i1.43.

[13] A. Primajaya *et al.*, "*Random Forest* Algorithm for Prediction of Precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, pp. 27–31, 2018.

[14] D. Marcelina, A. Kurnia, and T. Terttiaavini, "Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 293–301, 2023, doi: 10.57152/malcom.v3i2.952.

[15] KNIME Official, "KNIME Analytics *Platform*," 2023. https://www.knime.com/knime-analytics-*platform*

[16] J. Kurniawan *et al.*, *Analisa dan Visualisasi data*, 2023rd ed. Bandung: Widina Bhakti Persada Bandung, 2023.

[17] M. Ariandi *et al.*, "Analisis Visualisasi Data Kecamatan Kertapati menggunakan," *J. Jupiter,* vol. 14, no. 2, pp. 366–373, 2022.

Comparison of CNN Transfer Learning in Detecting Superior Local Fruit Types in Bali

Nyoman Purnama[1]

[1]Information System Department, Faculty of Information Technology and Design, Primakara University, Bali, Indonesia

Corresponding author: Nyoman Purnama (e-mail: purnama@primakara.ac.id).

ABSTRACT Bali Province is an island that has unique geographical conditions, as well as the diversity of fruit it has. The specialty of local fruit is not only of economic value for food needs but also for religious ceremonial needs. Bali provincial government is currently actively promoting local fruit so that it can be used as consumption for Bali's increasingly rapid tourism. Several superior fruits were developed as an effort to raise the potential of local fruit in the tourism sector. Some of the superior fruits are Balinese snake fruit and sapodilla. However, snake fruit is one of the superior local fruits in Bali which has not experienced degradation over time. This research aims to detect the types of snake fruit in Indonesia. This fruit is not popular compared to imported fruit. Therefore, an application is needed to recognize this type of snake fruit automatically. This research uses a deep learning method with the CNN (Convolutional Neural Network) algorithm. This algorithm is able to recognize and classify an image well. The fruit images used were 400 fruits for 4 types of snake fruit. Where the training data for snake fruit is special because it has different skin and fruit contents. In this research, 2 transfer learning models from the CNN algorithm were also compared, namely mobilenetv2 and ResNet152. Based on the test results, it was found that the best level of accuracy was obtained using the ResNet152 model with an accuracy value of 92% in identifying images of Balinese snake fruit.

KEYWORDS CNN, Local Bali Fruits, ResNet152, VGG16

I. INTRODUCTION

Indonesia is known for its diverse range of fruit plants, facilitated by its archipelagic geography where each region boasts unique fruit varieties. The diversity of native Indonesian fruits plays a crucial role not only in meeting nutritional needs due to their high vitamin content, beneficial for health [1], but also holds religious significance. In religious ceremonies, fruits are often used as offerings alongside leaves and flowers. This practice is prominently observed in Hindu religious rituals, where various fruits are utilized, each carrying profound philosophical meanings [2].

Bali, as one of Indonesia's provinces with distinctive geographical features, indirectly fosters unique diversity in its fruit varieties. The uniqueness of fruits in Bali extends beyond their nutritional benefits to their role as ceremonial complements in religious events. However, local Balinese fruit faces marketing challenges due to the presence of fruits from outside the region and imported fruits. To address this, the Bali provincial government promptly issued Regional Regulation No. 99 of 2018, mandating all tourism components to use local Balinese fruits [3]. Some local fruits from Bali have even been designated as superior varieties that can compete with fruits from other regions. These superior local fruits include Siam Orange, mangosteen, bananas, and Balinese Snake Fruit (Salak Bali). Based on research by Made Tamba, snake fruit and mangosteen will remain superior fruits and will not degrade into non-superior fruits in Bali. Snake fruit reflects the diversity of flavors and textures found across various regions in Indonesia, making it a popular and highly regarded fruit nationwide. Specific areas are known for distinctive varieties of snake fruit, such as Pondoh Snake Fruit (Salak Pondoh) from Yogyakarta, Balinese Snake Fruit from Bali, Condet Snake Fruit (Salak Condet) from Jakarta, and Sidempuan Snake Fruit (Salak Sidempuan) from Magelang.

The similarity in texture and shape among different varieties of snake fruit poses a significant challenge in identifying the specific type originating from Bali compared to other regions. In this study, researchers employed a classification method. Classifying fruit types is a task that requires time and expertise [4]. The advancement of computer vision allows for efficient and accurate automation of fruit-type classification. One effective method for automated classification is through deep learning, a rapidly evolving field within machine learning [5]. A prominent deep learning method capable of processing image information is the Convolutional Neural Network.

Convolutional Neural Network (CNN) is an extension of the Multilayer Perceptron (MLP) designed specifically for processing two-dimensional data. Unlike MLP, where each neuron is one-dimensional, CNN represents neurons in a two-dimensional form [6]. CNN is a deep learning architecture tailored for structured data arrays. Widely employed in computer vision, CNNs have become pivotal in various visual applications such as image classification and have demonstrated effectiveness in natural language processing for text classification as well. CNN is renowned for its automatic feature extraction capabilities. In contrast to traditional machine learning methods that rely on manual feature extraction, CNNs automatically extract features in layers like convolutional, pooling, and Rectified Linear Unit (ReLU) activation. Following feature extraction, classification tasks are performed in the Fully Connected Layer (FCL) with softmax activation [6].

Transfer learning is a technique in neural networks where a model trained on one task is utilized to solve a different task. It serves as a foundational concept behind many popular machine learning applications such as speech recognition, object detection, and natural language processing [7]. Several examples of transfer learning in CNN algorithms include VGG16, AlexNet, MobileNet, and ResNet. MobileNet is a Convolutional Neural Network architecture specifically designed to address excessive computing resource requirements. As implied by its name, Mobile, researchers at Google developed this CNN architecture for mobile devices [8]. MobileNet released its second version in April 2017. Similar to MobileNetV1, MobileNetV2 still utilizes depthwise and pointwise convolutions but introduces two new features: 1) linear bottleneck, and 2) shortcut connections between bottlenecks. On the other hand, ResNet, or Residual Network, is another artificial neural network architecture that introduces shortcut connections across layers and applies activation functions to the preceding layers [9]. There are several variants of ResNet, with ResNet152 being one of them. ResNet152 comprises 152 layers in its network architecture. Due to its complexity, this model achieved success in the ILSVRC competition in 2015 for its minimal error rate [10].

Research on fruit classification using the YOLOv3-based CNN method has been conducted by Mr. Wibi Bagas et al. [11]. In their study, they classified 10 types of fruits using 2333 images. The training process involved 5000 iterations and achieved an accuracy of 90% in the first test of each fruit and 70% for the second test on out-of-test data images. Fruit classification using the fruit-360 dataset has been conducted by Febian Fitra Maulana. They utilized 15 out of 111 classes available in the dataset, achieving an accuracy of 91.42%. From these studies, it is evident that CNN is an effective method for image classification. Fruit classification research has also been undertaken by Myongkyoon Yang [12], titled "Fruit Classification using Convolutional Neural Network." They classified 7 categories of fruits with a dataset of 1000 images. Based on their findings, CNN demonstrates strong classification capability with an error rate of 10%. Research on the accuracy of CNN algorithms using various architectures has been conducted by Wahyudi Setiawan in the paper titled “Perbandingan Arsitektur Convolutional Neural Network Untuk Klasifikasi Fundus” [13].

Research on the performance of ResNet152 and AlexNet in classifying types of skin cancer has been conducted by Tommy Saputra [10]. In their study, an accuracy of 87.85% in skin cancer classification was achieved using ResNet152. Another study on classification using the RESNET model was conducted by Vijay Gadre, titled "Waste Classification using ResNet152" [9]. In their research, waste was classified based on various characteristics using ResNet152. The results of the study indicate that the success of waste classification using ResNet152 depends on the quality and diversity of the training data. Research on classification using MobileNetV2 for butterfly image classification has been conducted by Desi Ramayanti. The research dataset consisted of 4955 images labeled with 50 butterfly species, each sized 224 x 224 x 3. The best accuracy achieved by MobileNetV2 without fine-tuning reached 96% [14].

Based on the background and previous research, this study develops a CNN architecture using fruit images as test data, focusing on a prominent local fruit from Bali province, Balinese Snake Fruit. Four types of snake fruit will be classified based on their origin: Balinese Snake Fruit, Pondoh Snake Fruit, Condet Snake Fruit, and Sidempuan Snake Fruit. Transfer learning CNN models ResNet152 and MobileNetV2 will be compared. The diverse textures and shapes of various snake fruit in Indonesia pose a challenge in distinguishing their origins. Detection of the prominent local fruit, Balinese Snake Fruit, will be conducted using CNN classification methods. By comparing transfer learning CNN models ResNet152 and MobileNetV2, the study aims to effectively classify these local Bali fruits and achieve the highest accuracy possible.

II. RESEARCH METHODS

In this study, an experimental research method was employed with the following stages:

A. DATA COLLECTION

The data utilized in this research consists of digital images. These images were collected from various sources and collectively referred to as the Dataset. The Dataset was obtained from search engine datasets such as Google and Bing. Selected images were chosen based on adequate lighting conditions, backgrounds with minimal noise, and intact or peeled snake fruit images. The collected Dataset includes images of four types of snake fruit: Balinese Snake Fruit, Pondoh Snake Fruit, Condet Snake Fruit, and Sidempuan Snake Fruit. Before using this data, preprocessing was conducted by categorizing each image into respective folders according to its category. The collected image Dataset remains in RGB color format with JPG extension.

Before using this image Dataset in the classification process using transfer learning CNN models, the data underwent preprocessing. Pixel size standardization was applied to all images, followed by data normalization. To increase the Dataset size, augmentation of these snake fruit images was performed initially. The subsequent process involved converting pixel values of the images into array form, aiming to standardize input sizes for the CNN network.

B. METHOD USED

The method employed in this study for fruit classification is one of the Deep Learning methods. Deep Learning can address problems with large amounts of data [15]. By utilizing Deep Learning, it allows us to create systems capable of learning at desired speeds and accuracies. One example of a Deep Learning method used in this research is Convolutional Neural Network.

This algorithm is efficient in image processing and widely used in image recognition [16]. CNN is not fundamentally different from other neural networks like artificial neural networks, as they all have biases, weights, and activation functions. However, what sets CNN apart from other neural networks is its specialized layer called the Convolutional Layer [17]. Image processing for leaf classification is performed using kernel filters. These filters are used to obtain fragments (strides) from an image. The process of obtaining these fragments/strides is called convolution. The process of this convolution is depicted in Figure 1 below.

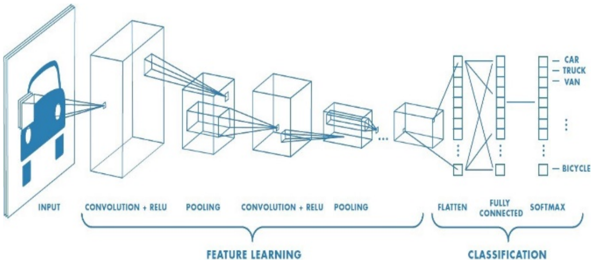


FIGURE 1. Process in Convolutional Neural Network.

MobilenetV2 and ResNet152 are transfer learning CNN architectures used in image classification applications [18]. Each architecture has its own strengths and weaknesses. Based on previous research, MobilenetV2 is a transfer learning model suitable for devices with limited resources. On the other hand, ResNet152 achieves high accuracy due to its deep convolutional layers [10]. In this study, each architecture will be used to train the same dataset. There are 500 fruit image data points, which will be split into training, validation, and test sets with a ratio of 80:10:10.

C. EXPERIMENT, EVALUATION, AND VALIDATION OF RESULTS

In this study, the Tensorflow framework developed by Google is utilized for developing fruit image classification using CNN with transfer learning models ResNet152 and MobilenetV2. Tensorflow offers numerous features related to image classification, utilizing Keras as a high-level interface for machine learning development. The classification results will be evaluated using precision, recall, and F1 Score metrics.

Comparison is applied to accuracy, precision, recall, and F1-score values in prediction results during testing, calculated using a confusion matrix. True Positive (TP) indicates instances where actual and predicted values are both positive, while True Negative (TN) indicates instances where both are negative. False Positive (FP) shows instances where actual values are negative but predicted as positive, and False Negative (FN) indicates instances where actual values are positive but predicted as negative [19].

Precision measures the accuracy of the system in providing requested information compared to the system's responses. Recall, on the other hand, measures the system's success in retrieving information [20]. Broadly, the process of detecting prominent local Bali fruit types is depicted in Figure 2.

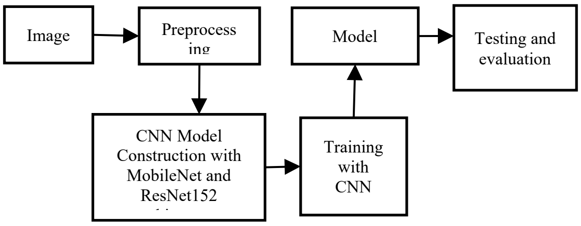


FIGURE 2. Research diagram

III. RESULT AND DISCUSSION

A. DATA PREPARATION

This study utilizes 4 images of local snake fruit: Balinese Snake Fruit, Pondoh Snake Fruit, Condet Snake Fruit, and Sidempuan Snake Fruit. These image data are collectively referred to as the dataset. Classification processes are conducted using transfer learning convolutional neural networks, namely ResNet152 and MobileNetV2. Evaluation of test results employs precision, F2-score, and recall metrics. In the initial stage, data collection involved gathering 250 images from search engines for these 4 types of snake fruit. These image data underwent augmentation to increase the training set to 400 images. Augmentation methods included adjustments for brightness, rotation, and vertical flips.

The composition of training, validation, and test data is set at 80:10:10 ratio. Thus, each fruit category comprises 80 training, 10 validation, and 10 test data points. Each type of snake fruit has an equal number of 100 training images. After collection, the data was categorized accordingly, with both training and test data placed into folders named after each fruit. Examples of these images for each type of snake fruit used in this study are shown in Figure 3.

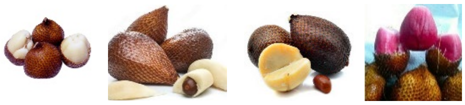


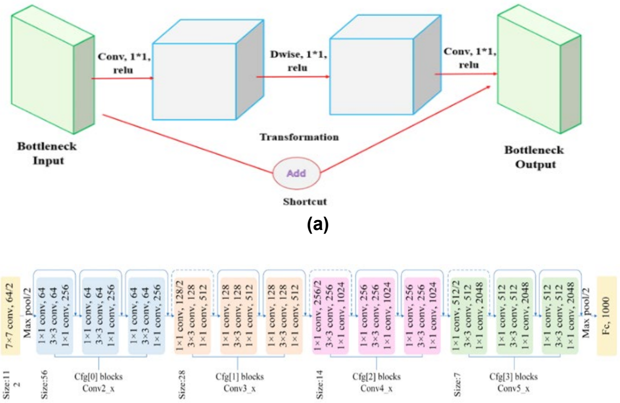
FIGURE 3. Initial images of Balinese Snake Fruit, Pondoh Snake Fruit, Condet Snake Fruit, and Sidempuan Snake Fruit

The next step involves image processing. In this stage, Python programming language is utilized within the Google Colab IDE environment to aid in processing. The image processing library used is OpenCV, a free library designed for image processing in Python. Since OpenCV uses the BGR (Blue, Green, Red) mode, the images collected in each folder are initially converted to the RGB (Red, Green, Blue) mode using the `cv2.cvtColor` function. The resulting images are then resized to 224x224 pixels. This step is necessary due to the varying sizes of the collected images, and resizing to 224x224 pixels accelerates the training process. Similarly, images for testing undergo the same image processing as those for training. Each test set consists of 20 images per fruit category. Before input into the CNN network, pixel values of both training and test images are normalized.

B. CONSTRUCTION OF RESNET152 AND MOBILENETV2 TRANSFER LEARNING MODELS

The next process involves constructing transfer learning models using CNN architectures, specifically MobileNetV2 and ResNet152. MobileNetV2 offers the advantage of being deployable on devices with low resources, especially mobile devices. It utilizes convolutional layers with filter thickness tailored to the input image thickness. On the other hand, the ResNet152 architecture is based on the concept of residual learning, comprising convolutional, normalization, and ReLU activation layers followed by residual blocks. In addition to transfer learning, this study employs the Sequential model in the model creation process. The Sequential model is a type of deep learning architecture that enables sequential layer-by-layer model building. This approach is commonly used for constructing deep learning models with Keras, particularly in TensorFlow.

Figure 4 illustrates the comparison of the MobileNetV2 and ResNet152 models utilized in this research. The models were constructed using the TensorFlow and Keras libraries in Python, with programming conducted in the Google Colab IDE environment.

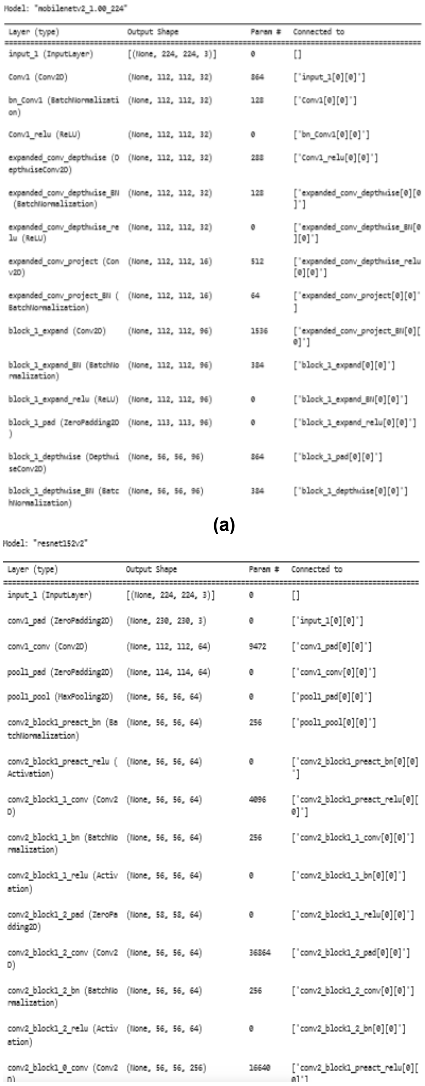


(a)

(b)

FIGURE 4. Comparison of MobileNetV2 (a) and ResNet152 (b) architecture models.

Regarding the parameters used in this model creation process, the learning rate was set to 0.001, with 100 epochs and a batch size of 32. The activation function employed was ReLU, optimized using Adam, and the loss function used was categorical_crossentropy, suitable for datasets with more than one label. The training was conducted in two phases: initially, by freezing or maintaining the pre-trained layers that had previously learned general features from classification tasks. Only the last few layers designated for fruit class classification were trained. This approach leverages the knowledge already captured in the lower layers of the transfer learning pre-trained model.



```
Model: "mobilenetv2_1.00_224"
______________________________________________________________________________________________
 Layer (type)                   Output Shape          Param #   Connected to
==============================================================================================
 input_1 (InputLayer)           [(None, 224, 224, 3)]  0         []

 Conv1 (Conv2D)                 (None, 112, 112, 32)   864       ['input_1[0][0]']

 bn_Conv1 (BatchNormalizati     (None, 112, 112, 32)   128       ['Conv1[0][0]']
 on)

 Conv1_relu (ReLU)              (None, 112, 112, 32)   0         ['bn_Conv1[0][0]']

 expanded_conv_depthwise (D     (None, 112, 112, 32)   288       ['Conv1_relu[0][0]']
 epthwiseConv2D)

 expanded_conv_depthwise_BN     (None, 112, 112, 32)   128       ['expanded_conv_depthwise[0][0
  (BatchNormalization)                                           ]']

 expanded_conv_depthwise_re     (None, 112, 112, 32)   0         ['expanded_conv_depthwise_BN[0
 lu (ReLU)                                                       ][0]']

 expanded_conv_project (Con     (None, 112, 112, 16)   512       ['expanded_conv_depthwise_relu
 v2D)                                                            [0][0]']

 expanded_conv_project_BN (     (None, 112, 112, 16)   64        ['expanded_conv_project[0][0]'
 BatchNormalization)                                             ]

 block_1_expand (Conv2D)        (None, 112, 112, 96)   1536      ['expanded_conv_project_BN[0][
                                                                 0]']

 block_1_expand_BN (BatchNo     (None, 112, 112, 96)   384       ['block_1_expand[0][0]']
 rmalization)

 block_1_expand_relu (ReLU)     (None, 112, 112, 96)   0         ['block_1_expand_BN[0][0]']

 block_1_pad (ZeroPadding2D     (None, 113, 113, 96)   0         ['block_1_expand_relu[0][0]']
 )

 block_1_depthwise (Depthwi     (None, 56, 56, 96)     864       ['block_1_pad[0][0]']
 seConv2D)

 block_1_depthwise_BN (Batc     (None, 56, 56, 96)     384       ['block_1_depthwise[0][0]']
 hNormalization)
```

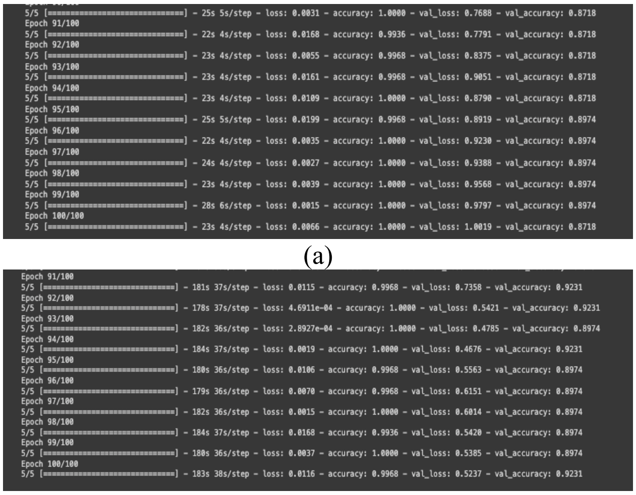
(a)

```
Model: "resnet152v2"
______________________________________________________________________________________________
 Layer (type)                   Output Shape          Param #   Connected to
==============================================================================================
 input_1 (InputLayer)           [(None, 224, 224, 3)]  0         []

 conv1_pad (ZeroPadding2D)      (None, 230, 230, 3)    0         ['input_1[0][0]']

 conv1_conv (Conv2D)            (None, 112, 112, 64)   9472      ['conv1_pad[0][0]']

 pool1_pad (ZeroPadding2D)      (None, 114, 114, 64)   0         ['conv1_conv[0][0]']

 pool1_pool (MaxPooling2D)      (None, 56, 56, 64)     0         ['pool1_pad[0][0]']

 conv2_block1_preact_bn (Ba     (None, 56, 56, 64)     256       ['pool1_pool[0][0]']
 tchNormalization)

 conv2_block1_preact_relu (     (None, 56, 56, 64)     0         ['conv2_block1_preact_bn[0][0]
 Activation)                                                     ']

 conv2_block1_1_conv (Conv2     (None, 56, 56, 64)     4096      ['conv2_block1_preact_relu[0][
 D)                                                              0]']

 conv2_block1_1_bn (BatchNo     (None, 56, 56, 64)     256       ['conv2_block1_1_conv[0][0]']
 rmalization)

 conv2_block1_1_relu (Activ     (None, 56, 56, 64)     0         ['conv2_block1_1_bn[0][0]']
 ation)

 conv2_block1_2_pad (ZeroPa     (None, 58, 58, 64)     0         ['conv2_block1_1_relu[0][0]']
 dding2D)

 conv2_block1_2_conv (Conv2     (None, 56, 56, 64)     36864     ['conv2_block1_2_pad[0][0]']
 D)

 conv2_block1_2_bn (BatchNo     (None, 56, 56, 64)     256       ['conv2_block1_2_conv[0][0]']
 rmalization)

 conv2_block1_2_relu (Activ     (None, 56, 56, 64)     0         ['conv2_block1_2_bn[0][0]']
 ation)

 conv2_block1_0_conv (Conv2     (None, 56, 56, 256)    16640     ['conv2_block1_preact_relu[0][
 D)                                                              0]']
```

(b)

FIGURE 5. Summary of MobileNetV2 (a) and ResNet152 (b) models.

To mitigate overfitting, the training process implemented early stopping, a technique used to halt training early if signs of overfitting or a lack of performance improvement on validation data occured. The objective of early stopping is to prevent the model from excessively memorizing the training data. Figure 5 presents a summary of the model results generated by the Keras library in Python, used in this fruit classification process.



```
5/5 [==============================] - 25s 5s/step - loss: 0.0031 - accuracy: 1.0000 - val_loss: 0.7688 - val_accuracy: 0.8718
Epoch 91/100
5/5 [==============================] - 22s 4s/step - loss: 0.0168 - accuracy: 0.9936 - val_loss: 0.7791 - val_accuracy: 0.8718
Epoch 92/100
5/5 [==============================] - 23s 4s/step - loss: 0.0055 - accuracy: 0.9968 - val_loss: 0.8375 - val_accuracy: 0.8718
Epoch 93/100
5/5 [==============================] - 23s 4s/step - loss: 0.0161 - accuracy: 0.9968 - val_loss: 0.9051 - val_accuracy: 0.8718
Epoch 94/100
5/5 [==============================] - 23s 4s/step - loss: 0.0109 - accuracy: 1.0000 - val_loss: 0.8790 - val_accuracy: 0.8718
Epoch 95/100
5/5 [==============================] - 25s 5s/step - loss: 0.0199 - accuracy: 0.9968 - val_loss: 0.8919 - val_accuracy: 0.8974
Epoch 96/100
5/5 [==============================] - 22s 4s/step - loss: 0.0035 - accuracy: 1.0000 - val_loss: 0.9230 - val_accuracy: 0.8974
Epoch 97/100
5/5 [==============================] - 24s 4s/step - loss: 0.0027 - accuracy: 1.0000 - val_loss: 0.9388 - val_accuracy: 0.8974
Epoch 98/100
5/5 [==============================] - 23s 4s/step - loss: 0.0039 - accuracy: 1.0000 - val_loss: 0.9568 - val_accuracy: 0.8974
Epoch 99/100
5/5 [==============================] - 28s 6s/step - loss: 0.0015 - accuracy: 1.0000 - val_loss: 0.9797 - val_accuracy: 0.8974
Epoch 100/100
5/5 [==============================] - 23s 4s/step - loss: 0.0066 - accuracy: 1.0000 - val_loss: 1.0019 - val_accuracy: 0.8718
```

(a)

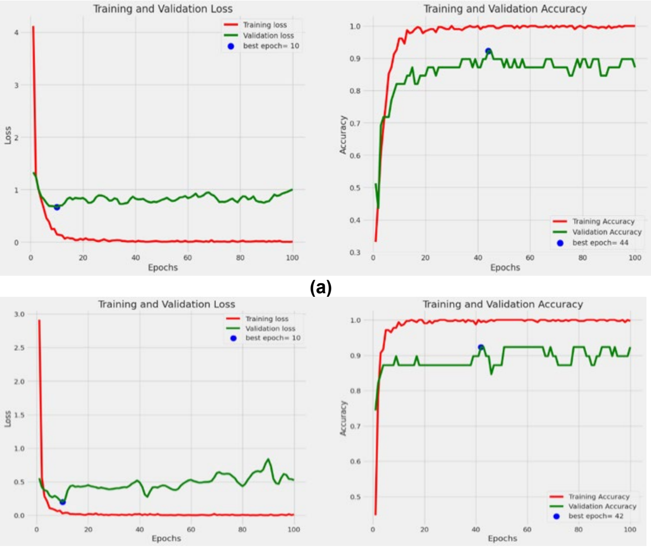
```
Epoch 91/100
5/5 [==============================] - 181s 37s/step - loss: 0.0115 - accuracy: 0.9968 - val_loss: 0.7358 - val_accuracy: 0.9231
Epoch 92/100
5/5 [==============================] - 178s 37s/step - loss: 4.6911e-04 - accuracy: 1.0000 - val_loss: 0.5421 - val_accuracy: 0.9231
Epoch 93/100
5/5 [==============================] - 182s 36s/step - loss: 2.8927e-04 - accuracy: 1.0000 - val_loss: 0.4785 - val_accuracy: 0.8974
Epoch 94/100
5/5 [==============================] - 184s 37s/step - loss: 0.0019 - accuracy: 1.0000 - val_loss: 0.4676 - val_accuracy: 0.9231
Epoch 95/100
5/5 [==============================] - 180s 36s/step - loss: 0.0106 - accuracy: 0.9968 - val_loss: 0.5563 - val_accuracy: 0.8974
Epoch 96/100
5/5 [==============================] - 179s 36s/step - loss: 0.0070 - accuracy: 0.9968 - val_loss: 0.6151 - val_accuracy: 0.8974
Epoch 97/100
5/5 [==============================] - 182s 36s/step - loss: 0.0015 - accuracy: 1.0000 - val_loss: 0.6014 - val_accuracy: 0.8974
Epoch 98/100
5/5 [==============================] - 184s 37s/step - loss: 0.0168 - accuracy: 0.9936 - val_loss: 0.5420 - val_accuracy: 0.8974
Epoch 99/100
5/5 [==============================] - 180s 36s/step - loss: 0.0037 - accuracy: 1.0000 - val_loss: 0.5385 - val_accuracy: 0.8974
Epoch 100/100
5/5 [==============================] - 183s 38s/step - loss: 0.0116 - accuracy: 0.9968 - val_loss: 0.5237 - val_accuracy: 0.9231
```

(b)

FIGURE 6. Comparison of training results for MobileNetV2 (a) and ResNet152 (b) models.

C. TESTING RESULTS AND EVALUATION

Following the creation of two models using the same parameters, the networks were trained for 100 epochs. The initial training process utilized the MobileNetV2 architecture. As shown in Figure 6 Part B, the training results using the MobileNetV2 model yielded commendable accuracy, aligning with its intended use on resource-constrained devices. In part c of Figure 6, the training process in the final epoch with the ResNet152 model is depicted. The performance results of ResNet152 demonstrated superior accuracy compared to MobileNetV2.



(a)

(b)

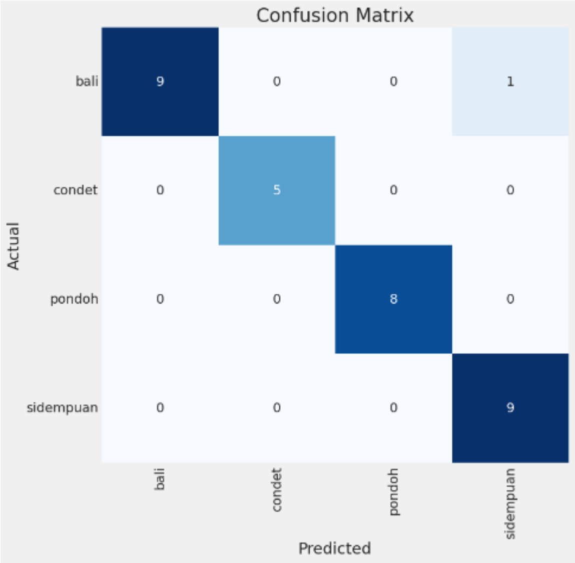
FIGURE 7. Comparison of training and validation loss data for MobileNetV2 (b) and ResNet152 (c) models.

In Figure 6, the performance results of each model after the training process are shown. In the figure 6 a is the result of training for mobilenetv2 and figure 6 b is for resnet152. To evaluate the models on the training dataset, the evaluate() function available in the Keras library is used. This function takes the same input and output as used to train the model. It generates predictions for each input-output pair and collects scores, including the average loss and any configured metrics such as accuracy. The evaluate() function returns a list with two values: the first value is the model's loss on the dataset, and the second value is the model's accuracy on the dataset. From the figure, it is evident that the best epoch for achieving the highest accuracy was obtained in the ResNet152 transfer learning model.

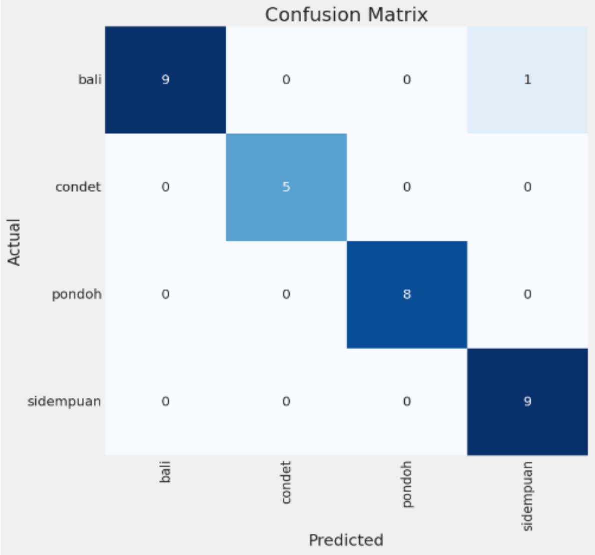
TABLE I
COMPARISON OF MODEL PERFORMANCE BETWEEN RESNET152 AND MOBILENETV2.

	Resnet152	MobilenetV2
Accuraccy	94.1%	84.72%
F1-Score	90%	87%
Recall	91%	89%
Precission	92%	92%

In Figure 7, the graph illustrates the accuracy and loss results for both training and validation data during the model creation process. Accuracy represents the ratio of correct predictions (both positive and negative) to the total number of instances for each class. Meanwhile, the loss function indicates how well the model's predictions match the actual results. In model creation, the goal is to minimize the loss value. The lowest accuracy value was observed in the MobileNetV2 model. The predictions are then evaluated to determine accuracy, recall, precision, and F1-score. Below, Table I compares the training accuracy performance results of the three models tested in this study.



(a)



(b)

FIGURE 8. Comparison of confusion matrices for MobileNetV2 (a) and ResNet152 (b) models.

In Figure 8, the confusion matrix for the MobileNet and ResNet152 models is depicted. Based on the confusion matrix, MobileNetV2 correctly classified 33 images, while ResNet152 correctly classified 46 images. The confusion matrix indicates that ResNet152 outperformed the other model in image classification. This outcome also reflects the longer training time required for ResNet152 compared to other models. MobileNetV2 had an average training time of 0.04 ms, whereas ResNet152 required 1.4 ms. This difference is due to ResNet152's more numerous and complex layers compared to MobileNetV2. However, MobileNetV2 strikes a balance between accuracy and the time required for training and testing.

Based on the research, evaluation scores for each fruit image were obtained using ResNet152, as shown in Table II below:

TABLE II
EVALUATION RESULTS OF THE TRAINING DATA FOR THE RESNET152 MODEL.

Snake Fruit Type	Precission	Recall	F1-score
Pondoh Snake Fruit	1.00	0.86	0.87
Balinese Snake Fruit	0.7	1.00	0.85
Condet Snake Fruit	1.00	0.87	0.93
Sidempuan Snake Fruit	0.55	0.84	0.68

Meanwhile, the evaluation results using the MobileNetV2 model yielded the following outcomes, as shown in Table III below:

TABLE III
EVALUATION RESULTS OF THE TRAINING DATA FOR THE MOBILENETV2 MODEL.

Class	Precission	Recall	F1-score
Pondoh Snake Fruit	1.00	0.75	0.86
Balinese Snake Fruit	0.75	1.00	0.86
Condet Snake Fruit	0.70	1.00	0.82
Sidempuan Snake Fruit	0.83	0.91	0.87

Based on the above evaluation results, it is evident that Balinese Snake Fruit exhibits higher precision values in both the ResNet152 and MobileNetV2 models. On the other hand, Pondoh Snake Fruit demonstrates higher recall values in both models. Both types of snake fruit are characterized by distinct shapes and skin colors compared to other fruit varieties.

IV. CONCLUSION

Based on the research conducted with a dataset of 400 images of 4 types of snake fruit, the system successfully identified fruit types based on their categories. The dataset comprised 80 images for training, 10 for validation, and 10 for testing. The training process utilized the TensorFlow and Keras libraries in the Google Colab IDE. With the same parameters, the MobileNetV2 model achieved an accuracy of 84.62%, while ResNet152 achieved 92.31%. ResNet152 demonstrated the highest accuracy in identifying local superior snake fruit. However, ResNet152 is disadvantaged by longer training times compared to MobileNetV2. MobileNetV2 exhibited a good accuracy result with faster training and testing processes. The accuracy difference between the two models was not substantial. This study has limitations, including the manual classification process. Future developments could involve mobile-based applications for real-time fruit classification. The snake fruit classification process presents its challenges, such as using training data that includes both whole fruits and peeled fruits, impacting the accuracy achieved.

AUTHORS CONTRIBUTION

Nyoman Purnama: Investigation, data collection, analysis, review writing, coding and editing.

COPYRIGHT

REFERENCES

[1] N. Adiputra, “Fungsi Buah Dan Daun Tanaman Dalam Budaya Bali Sebuah Kajian Terhadap Tanaman Upacara.”

[2] N. Widya Utami, N. Purnama, I. Putu, And R. Prajna, “Klasifikasi Tanaman Upakara Adat Hindu Di Kebun Raya Eka Karya Bali Menggunakan Algoritma Convolutional Neural Network,” 2023.

[3] I. M. Tamba, “Kajian Buah-Buahan Lokal Unggulan Provinsi Bali dan Potensi Dinamisnya,” *JIA (Jurnal Ilmiah Agribisnis) : Jurnal Agribisnis dan Ilmu Sosial Ekonomi Pertanian*, vol. 9, no. 2, pp. 126–132, Apr. 2023, doi: 10.37149/jia.v9i2.1117.

[4] F. Fitra Maulana and N. Rochmawati, “Klasifikasi Citra Buah Menggunakan Convolutional Neural Network”.

[5] N. E. A. Mimma, S. Ahmed, T. Rahman, and R. Khan, “Fruits Classification and Detection Application Using Deep Learning,” *Sci Program*, vol. 2022, 2022, doi: 10.1155/2022/4194874.

[6] S. Juliansyah *et al.*, “Klasifikasi Citra Buah Pir Menggunakan Convolutional Neural Networks,” *Jurnal Infra Petra*, vol. 7, no. 1, pp. 489–495, 2021, doi: 10.22441/incomtech.v11i1.10185.

[7] R. Pathak, “CLASSIFICATION OF FRUITS USING CONVOLUTIONAL NEURAL NETWORK AND TRANSFER

LEARNING MODELS.” [Online]. Available: https://www.researchgate.net/publication/364254116

[8] Y. Miftahuddin and F. Zaelani, “Perbandingan Metode Efficientnet-B3 dan Mobilenet-V2 Untuk Identifikasi Jenis Buah-buahan Menggunakan Fitur Daun,” 2022.

[9] V. Gadre, S. Sashte, and A. Sarnaik, “WASTE CLASSIFICATION USING RESNET-152,” *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, no. 01, Jan. 2023, doi: 10.55041/ijsrem17421.

[10] T. Saputra, M. Ezar, and A. Rivan, “STRING (Satuan Tulisan Riset dan Inovasi Teknologi) Analisis Performa Resnet-152 Dan Alexnet Dalam Klasifikasi Jenis Kanker Kulit.” [Online]. Available: https://challenge.isic-

[11] M. C. Wujaya and L. W. Santoso, “Klasifikasi Pakaian Berdasarkan Gambar Menggunakan Metode YOLOv3 dan CNN.”

[12] “Fruit Classification using Convolutional Neural Network (CNN),” *Precision Agriculture Science and Technology*, vol. 3, no. 1, 2021, doi: 10.12972/pastj.20210001.

[13] W. Setiawan, “Perbandingan Arsitektur Convolutional Neural Network Untuk Klasifikasi Fundus,” *Jurnal Simantec*, vol. 7, no. 2, pp. 48–53, 2020, doi: 10.21107/simantec.v7i2.6551.

[14] D. Ramayanti, D. Asri, and L. Lionie, “Implementasi Model Arsitektur VGG16 dan MobileNetV2 Untuk Klasifikasi Citra Kupu-Kupu Article Info ABSTRAK,” *JSAI: Journal Scientific and Applied Informatics*, vol. 5, no. 3, 2022, doi: 10.36085.

[15] I. P. W. Prasetia and I Made Gede Sunarya, “Image Classification of Balinese Seasoning Base Genep Based on Deep Learning,” *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 13, no. 1, Mar. 2024, doi: 10.23887/janapati.v13i1.67967.

[16] S. A. Maulana *et al.*, “Penerapan Metode CNN (Convolutional Neural Network) Dalam Mengklasifikasi Jenis Ubur-Ubur,” *Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)*, vol. 2, no. 4, pp. 122–130, 2023, doi: 10.55606/juprit.v2i4.3084.

[17] A. Prima, “Rancang Bangun Sistem Pendeteksi Aneka Ragam Buah Menggunakan MobileNetv2,” *Jurnal Sistim Informasi dan Teknologi*, pp. 208–215, Jul. 2023, doi: 10.60083/jsisfotek.v5i2.217.

[18] M. Sanjaya and E. Nurraharjo, “Deteksi Jenis Rempah-Rempah Menggunakan Metode Convolutional Neural Network Secara Real Time,” 2023.

[19] A. Eka *et al.*, “JEPIN (Jurnal Edukasi dan Penelitian Informatika) Klasifikasi Jenis Rempah Menggunakan Convolutional Neural Network dan Transfer Learning,” 2023.

[20] C. Z. Basha, B. N. L. Pravallika, and E. B. Shankar, “An efficient face mask detector with pytorch and deep learning,” *EAI Endorsed Trans Pervasive Health Technol*, vol. 7, no. 25, pp. 1–8, 2021, doi: 10.4108/eai.8-1-2021.167843.

A Hybrid Approach Using K-Means Clustering and the SAW Method for Evaluating and Determining the Priority of SMEs in Palembang City

Terttiaavini[1]
[1]Master of Computer Science, Faculty of Computer Science and Sciences, Indo Global Mandiri University, Indonesia

Corresponding author: Terttiaavini (avini.saputra@uigm.ac.id)

ABSTRACT The current efforts to develop Small and Medium Enterprises (SMEs) are still facing challenges in setting appropriate targets. Although the Palembang City Cooperative and SME Agency has launched various programs and initiatives to support SME development, they have not yet successfully identified the SMEs that should be given priority for development. This study aims to apply a hybrid approach that combines the K-Means Clustering method and Simple Additive Weighting (SAW) to evaluate and prioritize SME development in Palembang City. The K-Means Clustering method is used to group SMEs based on their characteristics, while SAW provides preference values (V_i). The SME data was obtained from the Palembang City Cooperative and SME Agency, covering 362 SME units. The K-Means Clustering results yielded two clusters: Cluster 0 as the High Growth Cluster and Cluster 1 as the Stability and Improvement Cluster. Validation using cross-validation showed that this model achieved an accuracy of 99.72%. The SAW analysis on Cluster 0 indicated that the Kopi Kaljo SME received the highest priority with a preference value of 45.71. This study confirms that this hybrid approach is effective in grouping SMEs based on their characteristics and prioritizing them based on data-driven evaluation. The research results are expected to help the Palembang City Cooperative and SME Agency design more effective and targeted assistance programs to optimize the contribution of SMEs to local economic growth to the maximum extent.

KEYWORDS Hybrid approach, K-Means Clustering, Simple Additive Weighting, SMEs

I. INTRODUCTION

Small and Medium Enterprises (SMEs) play a crucial role in the local and national economy, especially in developing cities like Palembang. SMEs not only create employment opportunities but also contribute to overall economic growth. In Palembang, SMEs serve as the main driver in improving community welfare and reducing unemployment rates. They also act as a pillar in strengthening the local economic structure, providing stability in times of economic crisis [1].

The Cooperative and SMEs Office of Palembang City has launched various programs and initiatives to support the development of SMEs in the region. However, the main challenge is how to classify SMEs based on certain characteristics and how to effectively prioritize their development. The data from the Department of Cooperatives and SMEs of Palembang City in 2022 recorded 1,103 SMEs, while more than 160,000 SMEs remain unregistered [2]

Traditional methods often fall short in handling the complexity of diverse SMEs data. With various types of SMEs having different characteristics and performances, an appropriate approach is needed to evaluate and determine their development priorities effectively [3].

One of the main issues faced by policymakers and SME managers is how to classify SMEs based on certain characteristics and how to effectively prioritize their development [4]. A more accurate and data-driven approach is needed to evaluate the performance and potential of SMEs more precisely, thereby providing targeted and appropriate support according to the needs of each SME [5].

This study aims to develop and implement a hybrid approach combining K-Means Clustering and Simple Additive Weighting (SAW) methods to evaluate and determine the development priorities of SMEs in Palembang City [6]. The K-Means Clustering method is used to group SMEs based on their characteristics, while the SAW method is employed to assign preference values to each clustered

SME, thus identifying those with the highest performance and potential [7].

The method used in this study is a hybrid approach combining two data mining techniques. First, K-Means Clustering is used to group SMEs based on their characteristics [8]. This approach helps identify underlying patterns among different SME groups. Second, Simple Additive Weighting (SAW) is used to assign preference values to each clustered SME, allowing the identification of SMEs with the highest performance and potential within each cluster. The combination of these two techniques is expected to provide a more comprehensive and accurate overview of the SME conditions in Palembang City and support better decision-making in the development and support of SMEs [9].

The benefits of this study include a deeper understanding of SMEs in Palembang City by identifying the underlying patterns and characteristics of various SME groups. By using the K-Means Clustering and Simple Additive Weighting (SAW) approaches, this study provides development priority recommendations for SMEs based on more accurate data-driven evaluations. This is expected to enhance the effectiveness of support for SMEs through the design of more adaptive and sustainable assistance programs, thereby maximizing the contribution of SMEs to local economic growth [10][11].

By implementing this hybrid model, it is expected to provide strategic recommendations to the Department of Cooperatives and SMEs of Palembang City in designing more effective and targeted assistance programs for priority SMEs. Through comprehensive and data-driven evaluation, this model enables more accurate identification of the needs and potentials of SMEs within each cluster. This is anticipated to enhance the efficiency of resource allocation and assistance, and strengthen the contribution of SMEs to sustainable local economic growth [12].

II. LITERATURE REVIEW

A literature review of the K-Means Clustering method, Simple Additive Weighting (SAW), and the application of RapidMiner can provide an in-depth understanding of the concepts, applications, and relevance of each method in the context of the development and evaluation of Small, and Medium Enterprises (SMEs). Here is an overview of the literature review used:

A. K-MEANS CLUSTERING METHOD

The K-Means Clustering method is a data analysis technique used to group data into different clusters based on certain similarities in characteristics [13]. This technique is widely applied in various studies due to its effectiveness in clustering data without prior labels or supervision [14]. The stages in the K-Means Clustering method are as follows:

1. Centroid Initialization
   Randomly select K initial centroids from the data points as the initial cluster centers.
2. Data Point Allocation to Clusters
   Assign each data point to the nearest cluster based on the Euclidean distance between the data point and the centroid.
3. New Centroid Calculation
   Recalculate the position of the new centroid in each cluster by taking the average of all data points that belong to the cluster.
4. Iteration
   Repeat steps 2 and 3 until a stopping condition is met, such as no significant changes in the centroid positions or the maximum number of iterations is reached.

Formulas Used in K-Means Clustering

1. Euclidean Distance
   To calculate the distance between two points in n-dimensional space, Equation (1) is the Euclidean formula [15].

$$distance(X_i, C_j) = \sqrt{\sum_{k=1}^{n} (X_{ik} - C_{jk})^2} \qquad (1)$$

   X_i is the i-th data point, C_j is the j-th centroid, and n is the number of dimensions.
2. Centroid Update
   After all data points are allocated to clusters, the new centroid C_j is calculated as the average of all data points X_i that belong to the j-th cluster. Equation (2) represents the formula for calculating the new centroid C_j.

$$C_j = \frac{1}{|S_j|} \sum_{X_i \in S_j} X_i \qquad (2)$$

   S_j is the number of data points in the j-th cluster.

The use of the K-Means algorithm for clustering SMEs offers advantages in identifying patterns in data without the need for prior class labels. This algorithm is scalable for large datasets, easy to interpret, and aids in determining the optimal clusters using the Elbow method [16]. By utilizing the Elbow method, the K-Means algorithm can automatically determine the optimal number of clusters based on significant drops in the Sum of Squared Errors (SSE) values. This enables researchers or practitioners to efficiently and accurately group SMEs based on data characteristics.

B. THE SIMPLE ADDITIVE WEIGHTING (SAW) METHOD

SAW is a multi-criteria decision-making technique used to evaluate alternatives based on the relative weights of each criterion [17]. SAW is employed to assign preferences to SMEs that have been clustered using the K-Means Clustering method. With SAW, each SME is assessed according to several predetermined criteria. The steps involved in the SAW method are as follows:

1. Determining Criteria (C_i)
   Criteria selected should be relevant and representative of the evaluation goals. These criteria are typically chosen based on an analysis of the needs and characteristics of the SMEs being evaluated.
2. Determining Suitability Ratings (R) and Weights (W_i)
   Each alternative A_i is assessed using suitability ratings R_{ij} for each criterion C_j. Weights W_j are assigned to indicate the relative importance of each criterion $\boldsymbol{C_j}$.
3. Creating the Decision Matrix (X) and Normalization
   The decision matrix X has dimensions $m\ x\ n$, where m is the number of alternatives and n is the number of criteria. Each element X_{ij} of matrix $\boldsymbol{X}$ represents the suitability rating R_{ij} of alternative A_i for criterion C_j.

$$X = \begin{bmatrix} X_{11} & \dots & X_{1n} \\ \dots & \dots & \dots \\ X_{m1} & \dots & X_{mn} \end{bmatrix}$$

   This structured approach allows for a systematic evaluation of SMEs based on weighted criteria, facilitating informed decision-making in developmental and support programs.
4. Normalization of the Decision Matrix (R)
   Normalization is performed to transform each element X_{ij} into the same range, based on whether the attribute is a benefit or a cost attribute. The normalization of the decision matrix is done using (3) and (4):

   For benefit attributes

$$r_{ij} = \frac{X_{ij}}{\max X_{ij}} \tag{3}$$

   For cost attributes

$$r_{ij} = \frac{\min X_{ij}}{X_{ij}} \tag{4}$$

   r_{ij} is the normalized value of element i on attribute j, where X_{ij} is the original value of element i on attribute j. $\max X_{ij}$ represents the maximum value of attribute j, and $\min X_{ij}$ represents the minimum value of attribute j. Normalization can be performed using various methods, such as min-max normalization or z-score normalization, depending on the nature of the data.
5. The calculation of Preference Value (V_i)
   After obtaining the normalized matrix R, the preference value V_i for each alternative A_i is calculated by summing the products of matrix R with the weight vector W using (5).

$$V_i = \sum_{j=1}^{n} W_j \times R_{ij} \tag{5}$$

   V_i is the preference value or score for alternative A_i

   The final result of the SAW process is the ranking of alternatives based on the value of V_i. Alternatives with higher V_i values are considered the best solutions or highest priorities

C. RAPIDMINER APPLICATION

RapidMiner is an open-source platform that provides various tools for data analysis, including data mining processes, predictive modeling, and business analytics. RapidMiner can help optimize the evaluation and decision-making processes related to SMEs by leveraging its visualization tools, data processing capabilities, and modeling functionalities offered by the platform [18].

By integrating literature on this topic, the research can develop a holistic approach to evaluating and developing SMEs using K-Means Clustering and SAW with the assistance of RapidMiner. This literature review will provide a strong theoretical foundation and practical insights to design effective and applicable research methodologies in the context of SMEs in Palembang or other regions.

D. HYBRID APPROACH FOR EVALUATING AND PRIORITIZING SMEs

This research aims to address existing research gaps by introducing a novel hybrid approach that combines the K-Means Clustering and Simple Additive Weighting (SAW) methods. The existing research gap lies in the challenge of prioritizing and developing Small and Medium Enterprises (SMEs) based on their diverse characteristics and needs.

Previous studies have contributed by categorizing SMEs into various clusters such as high, medium, and low [19], independent, developing, and assisted [20], as well as micro and small businesses [21], and strong and weak sustainability groups [22]. However, their weakness lies in their limited focus solely on classification and categorization. The approaches used tend to be descriptive and lack the utilization of objective data to determine development priorities. This limitation restricts the ability to provide specific and strategic recommendations for SMEs.

Traditional methods often struggle to manage the complexity and variation present in SME data, making it difficult to determine which SMEs should receive priority support and development.

Here is a detailed explanation of how this approach innovates and adds value compared to existing methods:

1. Hybrid Approach: The integration of K-Means Clustering allows segmentation of SMEs into different groups based on their characteristics such as income, number of employees, and business scale. This clustering provides a fundamental understanding of SME clusters, identifying groups like "High Growth" and "Stability and Improvement," which represent SMEs with different development needs and potentials.
2. SAW Method: After clustering, the SAW method is used to assign preference values (V_i) to each SME within

the identified clusters. This method evaluates SMEs based on predefined criteria to objectively measure development priorities.

3. Comprehensive Evaluation: Unlike traditional subjective approaches, this hybrid model ensures comprehensive, data-driven evaluation of SMEs. It leverages statistical analysis and machine learning techniques to gain insights from a dataset encompassing 362 SMEs from the Cooperative and SME Agency of Palembang City.
4. Value Proposition: The innovation lies in seamlessly integrating clustering for segmentation and SAW for prioritization, enabling policymakers and stakeholders to design assistance programs tailored to the identified needs of SME clusters. This approach optimizes resource allocation and enhances the effectiveness of support programs, thereby maximizing SME contributions to sustainable local economic growth.
5. Comparison with Existing Methods: Unlike single-method approaches that may overlook nuanced differences among SMEs or rely solely on subjective evaluations, the hybrid model in this study offers a structured and objective framework. It combines the strengths of clustering (for grouping similar SMEs) and SAW (for prioritizing based on criteria) to provide a holistic view that traditional methods may lack.

Overall, this research aims to bridge gaps by introducing a hybrid approach that is effective not only in categorizing SMEs but also in prioritizing them based on objective criteria. This innovation is expected to improve the accuracy and effectiveness of policy formulation and strategic planning for SME development in Palembang, offering a model that can be applied and adapted in similar contexts.

III. RESEARCH METHODOLOGY

In this section, a detailed explanation will be provided regarding the steps and approaches used to implement the K-Means Clustering and Simple Additive Weighting (SAW) methods in evaluating and prioritizing the development of Small, and Medium Enterprises (SMEs) in Palembang City [23]. Figure 1 illustrates the research stages, covering the process from start to finish in implementing the hybrid approach using K-Means Clustering and Simple Additive Weighting (SAW).

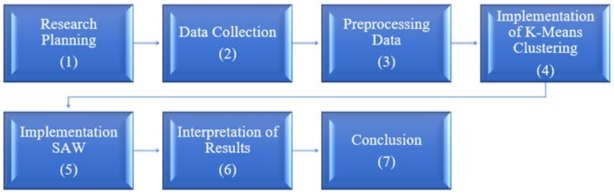


FIGURE 1. Research Stages

To achieve the objectives of this research, several stages will be detailed comprehensively. These stages are designed to ensure that the research is conducted systematically and comprehensively, so that the results obtained can significantly contribute to the evaluation and development of SMEs in Palembang. The following are the research stages to be implemented

A. DATA COLLECTION

Data collection for this research utilizes information provided by the Department of Cooperatives and SMEs of Palembang City. This data includes details from approximately 362 Small, and Medium Enterprises (SMEs) operating in Palembang. Sourcing data from this department is considered highly relevant as it provides direct access to information on characteristics, financial performance, and other factors influencing SMEs in the region [24]. The acquired data includes the SME name, owner's name, education, ownership status, business location status, district, business scale, business type, number of employees, revenue, operational costs, profit, average production quantity, buyer category, target customers, products, monthly sales volume, sales method, and transaction method. The total dataset consists of 362 items.

B. DATA PREPROCESSING

Data preprocessing is a crucial stage in the data analysis process aimed at cleaning, organizing, and preparing raw data for further analysis. In this stage, RapidMiner application is used for data preprocessing. The explanation of the data preprocessing stage is as follows:

1. Data Cleaning This process involves examining the data, determining attributes, and handling missing or incomplete values. Based on statistical analysis using RapidMiner, all data is complete, without outliers, and ready for further analysis.
2. Feature Selection Next, relevant and significant features are selected for clustering analysis and evaluation using SAW. The selected features include Education, District, Business Scale, Number of Employees, Revenue, Operational Costs, Profit, Average Sales, Number of Products Sold, and Transaction Method. Table I shows the features in Palembang City SMEs
3. Data Transformation
   In this stage, data is converted and adjusted to prepare it for further analysis. The data transformation process involves converting categorical data into numerical values, specifically for Education, Business Scale, and Transaction Method.
4. Data Normalization
   In the context of SME data, the range of values for each criterion can vary significantly. Data normalization aims to standardize the scale of input variables so that different ranges of values do not distort the results of clustering. Through normalization, variables with larger scales do not dominate the distance calculation between

data points, thus preventing bias in cluster formation based on Euclidean distance or other metrics.

TABLE I
FEATURES OF SMEs IN PALEMBANG CITY (*IN THOUSANDS)

Features	data
Business owner's education	High School = 174; Diploma = 42; Bachelor's Degree= 146
District	Alang-alang Lebar = 27; Bukit Kecil = 3; Gandus = 6; Ilir Barat I = 32; Ilir Barat II = 24; Ilir Timur I = 24; Ilir Timur II = 15; Ilir Timur III = 16; Jakabaring = 18; Kalidoni = 24; Kemuning = 14; Kertapati = 3; Plaju = 5; Sako = 38; Seberang Ulu I = 10; Seberang Ulu II = 18; Sematang Borang = 8; Sukajadi Timur = 1; Sukarami = 37; Luar Palembang = 40;
Business Scale	Micro Enterprises = 343; Small Business = 16; Medium Business = 1
Number of Employees*	$m = 0 : 55$; $1 \leq m \leq 3 = 279$; $4 < m \leq 6 = 24$; $7 < m \leq 10 = 3$; $m = 50 : 1$
Revenue	$n \leq 1.000 = 74$; $1.000 < n \leq 5.000 =168$; $5.000 < n \leq 10.000 = 69$; $10.000 < n \leq 50.000 = 50$; $n > 100.000 = 1$
Operational Costs*	$r \leq 1.000 =184$; $1.000 < r \leq 5.000 = 141$; $5.000 < r \leq 10.000 =27$; $r > 10.000 =10$
Profit*	$p \leq 1.000 = 23$; $1.000 < p \leq 5.000 =194$; $5.000 < p \leq 10.000 = 34$; $10.000 < p \leq 50.000 = 11$
Average Sales*	$x \leq 10 =117$; $10 < x \leq 50 = 88$; $50 < x \leq 100 = 42$; $100 < x \leq 500 = 46$; $500 < x \leq 1000 = 28$; $x > 1000 = 41$
Number of Products Sold*	$y \leq 10 = 120$; $10 < y \leq 50 = 22$; $50 < y \leq 100 = 59$; $100 < y \leq 500 = 87$; $500 < y \leq 1000 = 39$; $y > 1000 = 35$
Transaction Method*	Online = 53; Offline = 51; Both = 258

C. K-MEANS CLUSTERING IMPLEMENTATION

1. Implementing K-Means Clustering using RapidMiner involves several operators: read excel, select attributes, normalize, k-means clustering, and cluster distance performance. Each operator plays a role in preparing and analysing the data to cluster SMEs based on their characteristics. Figure 2 shows the clustering workflow for building a K-Means model using RapidMiner.

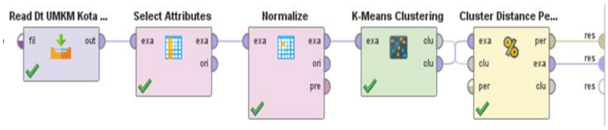


FIGURE 2. Clustering workflow

2. The first step involves setting the number of clusters and The first stage involves setting the number of clusters and relevant attributes using the Select Attributes, Set Parameters, and Normalize operators. Next, the K-Means algorithm will iterate to find the centroid for each cluster and group the data based on its proximity to the centroid using the K-Means operator. The clustering results are evaluated to measure their quality, often by considering the inertia value of the clusters using the Cluster Distance Performance operator, as well as visualizing the patterns formed using the Scatter Plot operator, thus providing valuable insights for decision-making related to the development strategy of SMEs in Palembang City
3. Elbow Method. The Elbow Method is used to determine the optimal number of clusters in cluster analysis. This method involves plotting the Sum of Squared Errors (SSE) values for various numbers of clusters (K) and then identifying the point where the decrease in SSE starts to slow significantly. This point resembles an elbow shape on the plot and indicates the optimal number of clusters for analysis. Table II presents a comparison of centroid distance values for each cluster.

TABLE II
CALCULATE THE ELBOW PLOT IN CLUSTERING

K	Elbow
2	4.925
3	3.927
4	3.234
5	2.262
6	1.694
7	1.239

To visualize the centroid distance values to determine the optimal number of clusters using the elbow method, a line diagram can be displayed as shown in Figure 3.

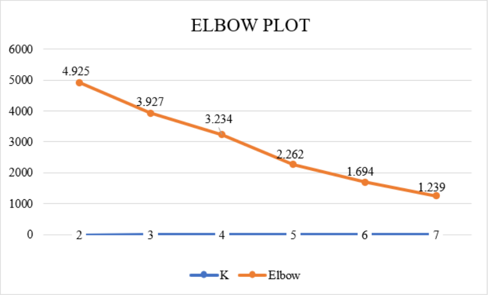


FIGURE 3. Elbow Plot Visualization

The results of the Elbow Plot calculation indicate that five clusters (K=5) are the optimal choice. However, for this analysis, it was decided to use two clusters (K=2).

This decision is based on the research objectives and the consideration of a clearer and more coherent interpretation of the results. The cluster names that represent these two clusters reflect their main characteristics and purposes. The clustering results with K=2 show that Cluster 0, which is the High Growth Cluster, has 314 items, while Cluster 1, which is the Stability and Improvement Cluster, has 48 items, making a total of 362 items. The clustering results with K=2 can be visualized in a Scatter Plot diagram. A Scatter Plot diagram is used to display the relationship between two variables in bivariate data. This patterns, correlations, and trends between these variables diagram helps in analyzing and visualizing data to find. Figure 4 shows the Scatter Plot trends between these variables diagram helps in analyzing and visualizing data to find. Figure 4 shows the Scatter Plot diagram for K=2, which displays the distribution of data within the two formed clusters.

.

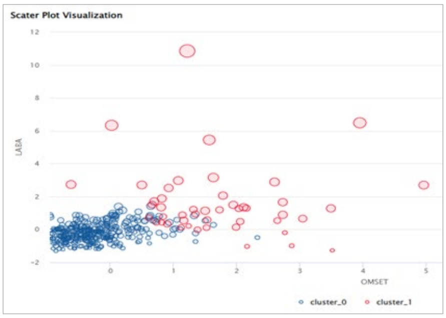


FIGURE 4. **Scatter Plot Diagram for K=2**

A clearer understanding is needed to measure how well the model can generalize to new data not seen during training. This technique divides the data into several subsets, training the model on one subset and testing it on another subset in turn. This approach provides a more accurate evaluation of the model's performance when faced with unseen data. Cross-validation calculations can be used to determine how well the model can generalize to new data not seen during training, thus providing a more accurate assessment of model performance.

The results of the cross-validation calculation generate a Performance Vector Table. The Performance Vector Table displays performance metrics for each fold used in the cross-validation, along with the mean of these metrics. Performance metrics can include Accuracy, Precision, and Recall values. Table III shows the Performance Vector Table resulting from the cross-validation calculation.

TABLE III
PERFORMANCE VECTOR TABLE

	true cluster_0	**true cluster_1**	**class precision**
pred. cluster_0	335	1	99.70%
pred. cluster_1	0	26	100.00%
class recall	100.00%	96.30%	

The cross-validation test results are detailed as follows:

1. Overall Accuracy: 99.72% This accuracy value indicates that the model can cluster SMEs with a success rate of 99.72% of the total data tested. This very high accuracy level demonstrates the model's excellent ability to distinguish between clusters.
2. Pred. Cluster_0**:** A total of 335 data points that belong to Cluster_0 was correctly grouped into Pred. Cluster_0. Only 1 data point from True Cluster_1 was incorrectly grouped into Pred. Cluster_0. The precision for Pred. Cluster_0 is 99.70%, meaning that of all the data predicted as Cluster_0, 99.70% actually belong to Cluster_0.
3. Pred. Cluster_1: All data points that belong to Cluster_1 (26 data points) was correctly grouped into Pred. Cluster_1. No data from True Cluster_0 was incorrectly grouped into Pred. Cluster_1. The precision for Pred. Cluster_1 is 100.00%, meaning that all data predicted as Cluster_1 actually belong to Cluster_1.
4. Class Recall**:** The recall for Cluster_0 is 100.00%, meaning all data points that should belong to Cluster_0 was correctly grouped. The recall for Cluster_1 is 96.30%, meaning that of all data points that should belong to Cluster_1, 96.30% were correctly grouped, with an error rate of only 3.70%.

Overall, these test results show that the implemented clustering model is highly reliable and can be used with a high degree of confidence to cluster SMEs in Palembang according to the specified characteristics. This provides confidence that the clustering results can be used as a reference in designing more targeted assistance and development programs.

D. IMPLEMENTATION OF SIMPLE ADDITIVE WEIGHTING (SAW)

After conducting cross-validation testing and identifying that Cluster 0 has relevant results for further processing, the data will be ranked to help determine the most optimal SMEs. Ranking is performed using the Simple Additive Weighting (SAW) method. The criteria are determined based on the previously established criteria. The steps in the SAW method are explained as follows:

1. Determination of Criteria (C_i)
   Based on the data in Cluster 0 and the SME data, relevant criteria are established to determine the development priorities of the SMEs. These criteria include various aspects such as business owner's education (C1), number of employees (C2), revenue (C3), operational costs (C4), profit (C5), average sales (C6), number of products sold per month (C7), and lending method (C8).
2. Determining Suitability Ratings (R) and Weights (W_i)
   Each alternative (SME) is evaluated or given suitability ratings based on the established criteria. Next, the relative weight of each criterion is determined to establish the importance of each criterion in the decision-making process. The criterion weights are determined based on mathematical analysis. These factors help establish the importance of each criterion in the context of decision-making. The determination of criterion weights (W_i) is explained in Table IV.
3. Normalization of Decision Matrix
   Normalization of the Decision Matrix is performed to ensure that attribute values within the decision matrix are on a uniform scale. This is crucial to ensure fair comparison of each attribute, avoiding bias due to differences in scale and units across different attributes. Through normalization, each attribute is evaluated within the same range, typically [0, 1], thereby making the total score calculation in the SAW method more

accurate and representative. Normalization of the Decision Matrix is conducted using (3) and (4).

TABLE IV
DETERMINATION OF ALTERNATIVES, CRITERIA AND SUITABILITY CHAIN
(*IN THOUSANDS)

Criteria (C_i)	Alternatives (A_i)	Suitability Ratings	Weights (W_i)
C1	Business owner's education	High School Diploma Bachelor's Degree	1 2 3
C2	Number of Employees	$0 \leq n < 10$ $10 \leq n < 50$ $n \geq 50$	1 2 3
C3	Revenue*	$0 \leq m < 1$ $1.000 \leq m < 5.000$ $5.000 \leq m < 10.000$ $m \geq 10.000$	1 2 3 4
C4	Operational Costs*	$0 \leq c < 1.000$ $1.000 \leq n < 5.000$ $5.000 \leq n < 10.000$ $n \geq 10.000$	1 2 3 4
C5	Profit*	$0 \leq l < 1.000$ $1.000 \leq l < 5.000$ $5.000 \leq l < 10.000$ $l \geq 10.000$	1 2 3 4
C6	Average Sales	$0 \leq x < 100$ $100 \leq x < 500$ $500 \leq x < 1.000$ $x \geq 1.000$	1 2 3 4
C7	Number of Products Sold	$0 \leq x < 100$ $100 \leq x < 500$ $500 \leq x < 1.000$ $x \geq 1.000$	1 2 3 4
C8	Transaction Method	online offline both	1 2 3

4. Calculation of Preference Value (V_i)
   Calculation of the preference value (V_i) is used to determine the ranking of each alternative based on the predefined criteria. The preference value is computed using equation (5). In this study, the results of the preference value calculation are displayed for the top 10 rankings only. Table V presents the Calculation of Preference Value (V_i) with the top 10 entries.

TABLE V
RESULTS OF PREFERENCE VALUE CALCULATION (V_I)

Initial SMEs	SMEs Name	V_i	Ranking
X208	Kopi kaljo	45,71	1
X187	Warung Neknang	41,31	2
X126	Pempek Ce' Anie	40,79	3
X125	Habar Jumputan	39,45	4
X181	Ikan bakar gegana	37,23	5
X242	Tiara bakery	35,10	6
X313	Benawa Coffee Roastery	35,05	7
X90	Dewul	33,78	8
X150	Rusnani	33,30	9
X225	Kemcum	32,92	10

The result of using the Simple Additive Weighting (SAW) method to determine the most optimal UKM shows that Kopi Kaljo has the highest preference value with a value of 45.71. Kopi Kaljo ranks first in the ranking list, followed by Warung Neknang with a preference value of 41.31. Meanwhile, the lowest value is Arasshop with a value of 10.41.

IV. CONCLUSION

Based on the analysis using the hybrid approach of K-Means Clustering and Simple Additive Weighting (SAW) on SME data in Palembang City, several key conclusions can be drawn:

1. The use of the K-Means Clustering model successfully grouped SMEs into two main clusters: the High Growth Cluster dominated by 314 SMEs, and the Stability with Improvement Cluster consisting of 48 SMEs. This result provides a clear picture of SME distribution based on their characteristics and performance in this region.
2. Validation results of the model showed a very high accuracy rate of 99.72%. The Performance Vector Table confirms that the model effectively classifies SMEs into the appropriate clusters. The High Growth Cluster has a precision of 99.70% and recall of 100.00%, while the Stability with Improvement Cluster has a precision of 100.00% and recall of 96.30%. This indicates that this clustering model is reliable for decision-making related to SME development strategies.
3. The application of the SAW method on clustered SMEs can identify the most optimal SMEs to prioritize in development programs. For instance, SMEs like Kopi Kaljo received the highest preference value with Vi of 45.71, placing it as the top priority for development. This approach allows for a more focused and comprehensive assessment of each SME, ensuring more effective and strategic resource allocation.
4. This research not only provides deep insights into the conditions of SMEs in Palembang City, but also establishes a strong foundation for better decision-making to support local economic growth through more measured and sustainable assistance programs. The implementation of this hybrid model is expected to serve as a valuable guide for stakeholders in designing more effective and supportive policies for SMEs amidst complex economic dynamics.
5. This hybrid method can be further developed by considering the integration of other clustering methods or applying more advanced weighting methods for priority evaluation. Future research could explore how the use of other machine learning techniques like random forest or neural networks could enhance the accuracy and relevance of evaluation results.
6. The implications of these findings are that stakeholders can optimize the type and amount of support provided to SMEs, including training, working capital, and other supportive infrastructure, by understanding their clusters.

AUTHORS CONTRIBUTION

Terttiaavini: Research Design and Conceptualization, Data Collection. Data Preprocessing, Implementation of K-Means Clustering, Implementation of SAW Method, Analysis and Interpretation, Writing and Documentation,Visualization, Review and Editing, Validation and Cross-Validation

COPYRIGHT

REFERENCES

[1] T. Tambunan, “Micro, small and medium enterprises in times of crisis: Evidence from Indonesia,” *J. Int. Counc. Small Bus.*, vol. 2, no. 4, pp. 278–302, 2021, doi: 10.1080/26437015.2021.1934754.

[2] Y. Abdullah and I. Gultom, “Dinas Koperasi dan UMKM Palembang bentuk tim data usaha kecil,” *Antara News*. https://sumsel.antaranews.com/berita/643913/dinas-koperasi-dan-umkm-palembang-bentuk-tim-data-usaha-kecil (accessed Apr. 12, 2023).

[3] A. K. M. H. Islam, M. R. Sarker, M. I. Hossain, K. Ali, and K. M. A. Noor, “Challenges of Small- and Medium-Sized Enterprises (SMEs) in Business Growth: A Case of Footwear Industry,” *J. Oper. Strateg. Plan.*, vol. 4, no. 1, pp. 119–143, 2021, doi: 10.1177/2516600x20974121.

[4] T. Terttiaavini and T. S. Saputra, “Pendampingan Manajemen Pengelolaan Pasar Sekojo Dengan Membangun Market Management One Gate System (MMOGS),” *Selaparang J. Pengabdi. Masy. Berkemajuan*, vol. 7, no. 4, pp. 2543–2551, 2023, [Online]. Available: https://journal.ummat.ac.id/index.php/jpmb/article/view/19594/8211

[5] D. Marcelina, A. Kurnia, and T. Terttiaavini, “Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 293–301, 2023, doi: 10.57152/malcom.v3i2.952.

[6] T. Terttiaavini, Y. Hartono, E. Ermatita, and D. P. Rini, “Comparison of Simple Additive Weighting Method and Weighted Performance Indicator Method for Lecturer Performance Assessment,” *Mod. Educ. Comput. Sci.*, vol. 15, no. 2, pp. 1–11, 2023, doi: 10.5815/ijmecs.2023.02.01.

[7] Y. Kustiyahningsih *et al.*, “Integration K-Means clustering and AHP for recommendations batik MSMEs,” *E3S Web Conf.*, vol. 499, pp. 1–7, 2024, doi: 10.1051/e3sconf/202449901006.

[8] M. S. Kaiser, J. Xie, and V. S. Rathore, *Information and Communication Strategies for Competitive Technology (ICTCS 2021)*, Lecture No., vol. 401, no. Ictcs. Springer, 2023. doi: 10.1007/978-981-19-0098-3_14.

[9] Terttiaavini and T. S. Saputra, “Analisa Pelatihan Strategi Manajemen Penjualan Produk Umkm Menggunakan Digital Marketing Bagi Masyarakat Terdampak Covid-19 Di Kampung Keluarga Berhasil (Kb) Layang-Layang Palembang,” in *Seminar Nasional AVoER XII 2020*, Palembang: Fakultas Teknik Universitas Sriwijaya, 2020, pp. 18–19. [Online]. Available: http://ejournal.ft.unsri.ac.id/index.php/avoer/article/view/251

[10] T. Terttiaavini and T. S. Saputra, “Analisa Pelatihan Strategi Manajemen Penjualan Produk UMKM Menggunakan Digital Marketing Bagi Masyarakat Terdampak Covid-19 Di Kampung Keluarga Berhasil (Kb),” in *Seminar Nasional AVoER XII 2020*, 2020, pp. 18–19.

[11] T. Terttiavini, L. Hertati, Y. Yulius, and T. S. Saputra, “Strategi Digital Marketing dan Inovasi produk untuk meningkatkan daya saing UMKM Ikan Pedo Serbuk di Kabupaten Muratara,” Palembang, 2023. [Online]. Available: https://drive.google.com/file/d/1KlK9ef0ZALIe65CWnqi49cMNBIEF8wFT/view?usp=sharing

[12] S. Rochayatun, . S., and R. Bidin, “Mode of Entry Strategy on SMEs Internationalization in East Java: A Review of Literature,” *Asian J. Econ. Bus. Account.*, vol. 22, no. 15, pp. 20–32, 2022, doi: 10.9734/ajeba/2022/v22i1530626.

[13] T. Terttiaavini *et al.*, “Clustering Analysis of Premier Research Fields,” *Int. J. Eng. Technol.*, vol. 7, no. 4.44, p. 43, 2018, doi: 10.14419/ijet.v7i4.44.26860.

[14] K. P. Sinaga and M. Yang, “Unsupervised K-Means Clustering Algorithm,” *IEEE Access*, vol. 8, no. May, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796 Unsupervised.

[15] R. Suwanda, Z. Syahputra, and E. M. Zamzami, “Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K,” *J. Phys. Conf. Ser.*, vol. 1566, no. 1, 2020, doi: 10.1088/1742-6596/1566/1/012058.

[16] F. Liu and Y. Deng, “Determine the Number of Unknown Targets in Open World Based on Elbow Method,” *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 5, pp. 986–995, 2021, doi: 10.1109/TFUZZ.2020.2966182.

[17] Terttiaavini, Y. Hartono, Ermatita, and D. P. Rini, “Comparison of Simple Additive Weighting Method and Weighted Performance Indicator Method for Lecturer Performance Assessment,” *Int. J. Mod. Educ. Comput. Sci.*, vol. 15, no. 2, pp. 1–11, 2023, doi: 10.5815/ijmecs.2023.02.01.

[18] L. Kovács and H. Ghous, “Efficiency comparison of Python and RapidMiner,” *Multidiszcip. Tudományok*, vol. 10, no. 3, pp. 212–220, 2020, doi: 10.35925/j.multi.2020.3.26.

[19] L. Magdalena and R. Fahrudin, “Penerapan Data Mining Untuk Koperasi Se-Jawa Barat Menggunakan Metode Clustering pada Kementerian Koperasi dan UKM,” *J. Digit*, vol. 9, no. 2, p. 190, 2020, doi: 10.51920/jd.v9i2.120.

[20] D. Marcelina, A. Kurnia, and T. Terttiaavini, “Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. October, pp. 293–301, 2023.

[21] S. A. Mustaniroh, L. S. Jauhari, and J. M. Maligan, “Strategi Pengembangan Klaster Ukm Produksi Bandeng Asap dengan Menggunakan Metode K-Means Clustering dan Fuzzy Ahp,” *J. Keteknikan Pertan. Trop. dan Biosist.*, vol. 8, no. 1, pp. 101–106, 2020, doi: 10.21776/ub.jkptb.2020.008.01.10.

[22] P. Dauni, P. Pratiwi, and R. T. Prasetio, “Pemetaan Keberlangsungan Hidup Umkm Guna Optimalisasi Bantuan Kredit Menggunakan Algoritma Fuzzy C-Means,” *J. Responsif Ris. Sains dan Inform.*, vol. 5, no. 1, pp. 61–69, 2023, doi: 10.51977/jti.v5i1.1051.

[23] D. Anton, T. Avini, A. Heryati, and H. Saputra, *Business Process Reengineering*, 1st ed. Tasikmalaya: Perkumpulan Rumah Cemerlang Indonesia, 2023.

[24] Hartatik *et al.*, *Data Science For Business (Pengantar & Penerapan berbagai Sektor)*, no. September 2016. 2023. [Online]. Available: https://www.data-science.ruhr/about_us/

INSYST

Journal of Intelligent System and Computation

Volume 06 Nomor 01 April 2024

Author Guidelines

- Manuscript should be written in Indonesia and be submitted online via journal website. Online Submission will be charged at no Cost
- Manuscript should not exceed 15 pages including embedded figures and tables, without any appendix, and the file should be in Microsoft Office (.doc/.docx). download template
- Title should be less than 15 words
- Abstracts consists of no more than 200 words, contains the essence of the article and includes a brief background, objectives, methods and results or findings of the study. Abstract is written in one paragraph.
- Keywords are written in Indonesia three to five words/phrases, separated with coma and consist of important words/phrases from the article.
- Author's name, affiliation, affiliation address and email. State clearly and include country's name on your affiliation address.
- The main text of the writing should be consists of: Introduction, Method, Result and Discussion, and Conclusion; followed by Acknowledgment and Reference
- Introduction State adequate background, issues and objectives, avoiding a detailed literature survey or a summary of the results. Explain how you addressed the problem and clearly state the aims of your study.
- Used method is the scientific in the form of study of literature, observation, surveys, interviews, Focus Group Discussion, system testing or simulation and other techniques commonly used in the world of research. It is also recommended to describe analysis techniques used briefly and clearly, so that the reader can easily understand.
- Results should be clear, concise and not in the form of raw data. Discussion should explore the significance of the results of the work, not repeat them. Avoid extensive citations and discussion of published literature. INSYST will do the final formatting of your paper.
- Conclusion should lead the reader to important matter of the paper. Authors are allowed to include suggestion or recommendation in this section. Write conclusion, suggestion and/or recommendation in narrative form (avoid of using bulleting and numbering)
- Acknowledgments. It is highly recommended to acknowledge a person and/or organizations helping author(s) in many ways. Sponsor and financial support acknowledgments should be included in this section. Should you have lots of parties

to be acknowledged, state your acknowledgments only in one paragraph. Avoid of using bulleting and numbering in this section

- The number of references are not less than 10 with at least 8 primary references. Primary references are include journal, thesis, disertation and all kinds of research reports. All refferences must come from source published in last 7 years.
- Figure and table should be in black and white, and if it is made in color, it should be readable when it is later printed in black and white.
- Figure and table should be clearly readable and in a proportional measure to the overall page.

Tim Redaksi
Journal of Intelligent System and Computation
Departement of Informatics
Institut Sains dan Teknologi Terpadu Surabaya
Jl. Ngagel Jaya Tengah 73-77 Surabaya
Email: insyst@istts.ac.id
Website: https://jurnal.stts.edu/index.php/INSYST/index