# PILAR NUSA MANDIRI

## Journal of Computing and Information System

**Vol. 20 No. 1 March 2024**

**P-ISSN: 1978-1946 (Print)**

**E-ISSN: 2527-6514 (Online)**

**Publisher:**
Lembaga Penelitian dan Pengabdian Masyarakat Universitas Nusa Mandiri
Jl. Jatiwaringin Raya No. 02 RT 08 RW 013 Kelurahan
Cipinang Melayu Kecamatan Makassar Jakarta Timur 13620
Phone: 021 28534471
https://ejournal.nusamandiri.ac.id/index.php/pilar

PILAR NUSA MANDIRI: JOURNAL OF COMPUTING AND INFORMATION SYSTEM

# EDITORIAL BOARD

Vol. 20 No 1 March 2024
E-ISSN: 2527-6514
P-ISSN: 1978-1946

Editorial Address : Kampus Universitas Nusa Mandiri Tower Jatiwaringin
Jl. Jatiwaringin Raya No. 02 RT 08 RW 013 Kelurahan Cipinang Melayu
Kecamatan Makassar Jakarta Timur 13620

Website : https://ejournal.nusamandiri.ac.id/index.php/pilar/index

Editorial Email : jurnal.pilar@nusamandiri.ac.id

Vol. 20 No 1 March 2024
E-ISSN: 2527-6514
P-ISSN: 1978-1946

# PREFACE

Welcome to the Pilar Nusa Mandiri: Journal of Computing and Information System. This journal, with e-ISSN 2527-6514 and p-ISSN 1978-1946, is published biannually in March and September. The current issue, March 2024 (Vol. 20 No. 1), exemplifies our steadfast commitment to advancing knowledge in pivotal fields such as Artificial Intelligence System, Genetic Algorithms, Information Systems, Technology Management, Designing Information Systems, Data Mining, Business Intelligence, Image Processing, Database System, and Decision Support System.

In this issue, we present a collection of scholarly articles that delve into cutting-edge research and innovative practices across these domains. Each contribution reflects rigorous inquiry and thoughtful exploration, aiming to address contemporary challenges and contribute to the evolving landscape of computing and information systems.

We extend our gratitude to the authors, reviewers, and editorial team whose dedication has made this publication possible. Their expertise and commitment ensure that Pilar Nusa Mandiri remains a vital platform for disseminating impactful research and fostering intellectual discourse.

We invite readers, researchers, and practitioners alike to explore the insights and findings presented in this issue, confident that they will find inspiration and knowledge that enriches their understanding and informs their practice.

Thank you for your continued support and interest in Pilar Nusa Mandiri: Journal of Computing and Information System.

Warm regards,

Editor-in-Chief

Vol. 20 No. 1 March 2024
E-ISSN: 2527-6514
P-ISSN: 1978-1946



# TABLE OF CONTENTS

Vol. 20 No 1 March 2024
E-ISSN: 2527-6514
P-ISSN: 1978-1946

# ANALYSIS OF WHISPER AUTOMATIC SPEECH RECOGNITION PERFORMANCE ON LOW RESOURCE LANGUAGE

**Riefkyanov Surya Adia Pratama[1]; Agit Amrullah[2*]**

Informatika[1,2]
Universitas AMIKOM Yogyakarta, Sleman, Indonesia[1,2]
www.amikom.ac.id[1,2]
rsurya@students.amikom.ac.id[1], agit@amikom.ac.id[2*]

(*) Corresponding Author



***Abstract****—Implementing Automatic Speech Recognition Technology in daily life could give convenience to its users. However, speeches that can be recognized accurately by the ASR model right now are in languages considered high resources, like English. In previous research, a few regional languages like Javanese, Sundanese, Balinese and Btaknese are used in automatic speech recognition. This research aim is to improve speech recognition using the ASR model on low-resource language. The dataset used in this research is the Javanese dataset specifically because there is a high-quality Javanese speech dataset provided by previous research. The method used is fine-tuning the Whisper model which has been trained on 680,000 hours of multilingual voice data using a Javanese speech dataset. To reduce computation requirements, parameter efficient fine-tuning (PEFT) implemented in the fine-tuning process. The trainable parameter is reduced to <1% because the implementation of PEFT reduces the computation required by the model for fine-tuning. The best WER evaluation result is 13.77%, achieved by the fine-tuned Whisper large-v2 model compared to the base model of Whisper large-v2, which achieves 89.40% in WER evaluation. Performance improvement in WER evaluation showed that fine-tuning effectively improves the performance of the Whisper automatic speech recognition model on recognizing speeches in low-resource languages like the Javanese language compared to the Original Whisper model performance with minimal computational cost needed for fine-tuning large model.*

***Keywords:*** *automatic speech recognition, low-resources language, whisper fine-tuning.*

**Abstrak—***Implementasi dari teknologi Automatic Speech Recognition (ASR) pada kehidupan sehari-hari dapat memberikan kemudahan bagi para penggunanya. Namun, suara ucapan yang dapat dikenali dengan akurat oleh model ASR saat ini adalah suara ucapan dengan bahasa-bahasa sumber daya besar seperti bahasa inggris. Pada penelitian sebelumnya pengenalan suara telah dipergunakan pada beberapa bahasa daerah baik Jawa, Sunda, Bali dan Batak. Penelitian ini bertujuan untuk melakukan peningkatan penengenalan suara ucapan pada model ASR pada bahasa bersumber daya rendah. Dataset yang digunakan pada penelian ini secara spesifik adalah dataset Bahasa jawa karena terdapat dataset ucapan berbahasa jawa yang berkualitas tinggi yang disediakan oleh sebuah penelitian sebelumnya. Metode yang digunakan adalah fine-tuning pada model Whisper yang telah dilatih pada 680,000 jam data suara multilingual dengan menggunakan dataset ucapan berbahasa Jawa. Untuk mengurangi kebutuhan sumber daya komputasi, diimplementasikan parameter efficient fine-tuning (PEFT) pada proses fine-tuning. Trainable parameter dari model berkurang menjadi <1% sebagai hasil dari implementasi PEFT yang mana mengurangi kebutuhan sumber daya komputasi untuk fine-tuning model. Hasil evaluasi Word Error Rate (WER) terbaik adalah 13.77% pada model Whisper large-v2 yang telah dilakukan fine-tuning dibandingkan pada model Whisper large-v2 tanpa fine-tuning yang mana WER adalah 89.40%. Peningkatkan performa pada evaluasi WER menunjukan bahwa fine-tuning efektif untuk meningkatkan pengenalan suara ucapan otomatis model Whisper pada bahasa besumber daya rendah seperti bahasa Jawa dibandingkan dengan performa* *model Whisper aslinya dengan kebutuhan komputasi seminimal mungkin untuk model yang besar.*

***Kata Kunci****: pengenalan ucapan otomatis, bahasa bersumber daya rendah, whisper fine-tuning.*

## INTRODUCTION

In recent years, Artificial intelligence technology has been developing and helping in many life activities since it was first known in 1956 (Zhang & Lu, 2021), and based on a report by (Zhang et al., 2022), publication related to AI research topic growing from 200000 in 2010 to 496010 total research publications in 2021. The Automatic Speech Recognition is one of many fields that have become the focus of various research (Alharbi et al., 2021). Automatic speech recognition (ASR) is converting speech directly from speech to word sequence using a specific algorithm using a computer (Kumar & Mittal, 2019). Several implementations of ASR technology in everyday life are internet surfing or browsing using speech voice, speech recognition, operating IoT devices using speech voice, and general human-machine interaction based on user speech voice, biometric media, *et cetera* (Zhang et al., 2019).

Two major ASR models were used and researched: the hybrid and end-to-end (E2E) models (Li, 2022). There is a transition from a widely used hybrid model to E2E because E2E models have proved more efficient than the others. The E2E model works by directly mapping input sequences to word sequences. There is three technique that is considered successful in the implementation of the E2E model: Connectionist Temporal Classification (CTC), Attention-based Encoder-Decoder (AED), and Recurrent Neural Network Transducer (RNN-T). One State-Of-Art model that implements those E2E techniques is Whisper by OpenAI.

Web-scale supervised Pre-Training For Speech Recognition (Whisper) is a model that builds based on transformer encoder-decoder architecture (Vaswani et al., 2023) and is trained in weakly supervised on 680,000 hours of multilingual and multitasking (speech recognition, translation, and language identification) speech data (Radford et al., 2022). However, as mentioned in (Radford et al., 2022), Whisper's performance could be better in recognizing language categorized as low-resource. Low-resource language is a language that has little data in the form of digital, or that can be processed by a computer directly. One example of this language category is the Javanese language (Butryna et al., 2020). Whisper performance evaluated using Character Error Rate (CER)/Word Error Rate (WER) on this type of language is remarkably low.

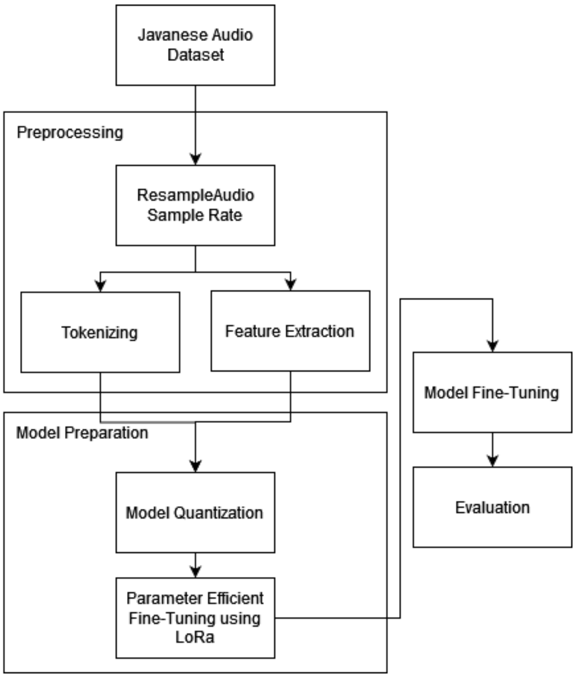
Research by (Rouditchenko et al., 2023) stated a comparison of performance between the XSL-R and Whisper model in zero-shot conditions ( without fine-tuning) where the evaluation of model performance is lower in less seen or unseen language, which can categorized as low-resource language. There is research (Novitasari et al., 2020) to build an ASR model for ethnic languages in Indonesia. One of the best results is evaluating ASR model performance in recognizing speech in the Javanese language, which is 20.20% in CER evaluation.

The solution proposed to improve Whisper performance in low-resource language, mainly Javanese, is fine-tuning the Whisper model. Fine-tuning improves the AI model in specific downstream tasks (Liu et al., 2022; Min et al., 2022; Wei et al., 2022; Yang et al., 2023). Fine-tune is done by training the model using a specific dataset, which, in this case, is a dataset of speech audio that is spoken using the Javanese language, so the model is more familiar with the pattern of data. There are some problems in fine-tuning large language models like Whisper, one of which is that fine-tuning requires immense computation power. To reduce computation costs in the fine-tuning process, the Parameter Efficient Fine-Tuning (PEFT) (Fu et al., 2022) method will reduce trainable parameters trained during the fine-tuning process.

This research scope focuses on fine-tuning the Whisper ASR model in one language categorized as a low-resource language, i.e., the Javanese language. Therefore, the research aims to improve Whisper ASR's performance that WER evaluated in the Javanese language, categorized as a low-resource language. Hopefully, this research could be used as a reference by other researchers to improve ASR performance in low-resource language.

## MATERIALS AND METHODS

This research aims to improve the Whisper ASR model for the Javanese language through fine-tuning using low computation. The first step is to convert the audio sample rate in the dataset from 48kHz to 16kHz to achieve maximal performance in the fine-tuning process. Then, preprocessing data is split into two stages: tokenization and feature extraction. After preprocessing, the model was prepared to compute minimal computation cost using PEFT Low-Rank Adaptation (LoRA) to reduce its trainable parameter. After all preparation is set, the Fine-tuning process could be commenced. After fine-tuning is complete, Model performance will evaluated using WER evaluation and compared with the zero shot (without fine-tuning) model. The details of the research flow can be seen in Figure 1.




Source: (Research Result, 2023)
Figure 1. Research Flow

**Data Collection**

The dataset will be used from (Butryna et al., 2020), which contains compilations of speech audio with .wav extension, file ID, and transcription of each audio file. There is a total of 5822 audio files with their ID and transcription, with an average duration of each audio being 3 seconds, and the total duration of all audio files is plus minus 4.85 hours. The dataset will be split into two parts, one for fine-tuning or training models and the other for evaluating model performance. Based on (Joseph, 2022), the most optimal ratio for dataset splitting is 8:2, with 8 being the training dataset and 2 being the evaluation dataset.

**Audio Resampling**

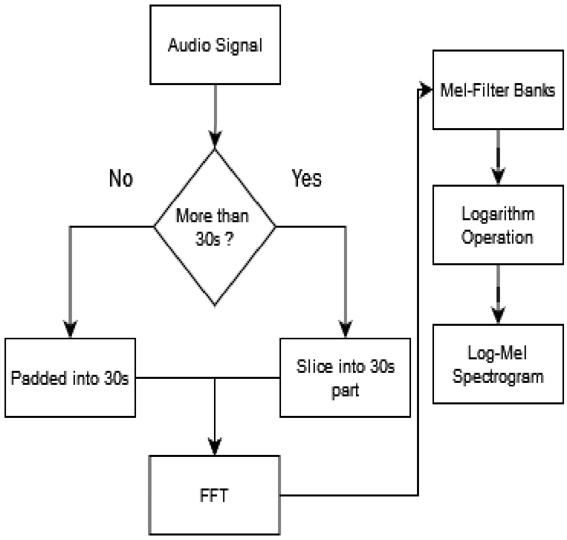
The sample rate audio in the dataset that will be used is 48kHz. Based on (Radford et al., 2022), Whisper, especially its feature extractor, works best at a 16kHz sample rate. So, to achieve maximal performance in the fine-tuning process, Audio needs to be resampled from 48kHz to 16kHz. Audio resample done using bandlimited sinc interpolation with the help of torch audio library.

**Tokenization**

Tokenization significantly affects language model performance even more in low-resource languages (Toraman et al., 2023). Whisper's built-in Byte-level BPE tokenizer will be used, separating sentences into words and words into tokens (Radford et al., 2019; Wang et al., 2019). There is 96 language that are recognized by Whisper's built-in tokenizer, one of which is Javanese.

**Feature Extraction**

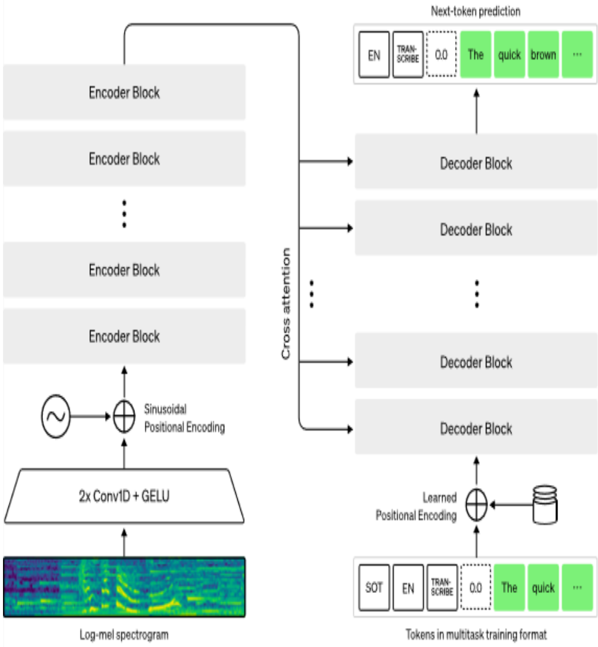
Feature extraction will be done using Whisper's built-in feature extractor. The whisper feature extractor works in two steps. First, the whisper feature extractor pads the audio signal to match the 30-second duration. If audio exceeds the 30-second limit, then audio will be sliced up into two parts. If the audio signal is less than the 30-second limit, then the audio signal will padded with zero or silence into 30 seconds duration. The next step is converting audio into an 80-channel log-mel spectrogram visual. The audio signal will first be converted into a Fast Fourier Transform (FFT) measurement. Then, this measurement will be inserted into the mel-filter bank and apply the logarithm operation. The result is a visualization of the log-mel spectrogram. The workflow of Whisper Feature Extraction can be seen in Figure 2.




Source: (Research Result, 2023)
Figure 2. Whisper Feature Extractor Process

**Model Quantization**

Whisper is a model built based on an encoder-decoder transformer (Vaswani et al., 2023) or could be called a seq-2-seq model because it works by mapping input sequences into word sequences. Weak supervision training on 680,000 hours of multilingual audio has been done prior. As can be seen in Figure 3, the Whisper model will first encode the log-mel spectrogram from the Whisper feature extractor to form a sequence from the encoder's hidden state and then insert it into the decoder to predict the text token autoregressively based on the previous token condition and then encoder hidden state.




Source: (Radford et al., 2022)

Figure 3. Whisper Architecture

Whisper variant model can be seen in Table 1. For this research, models that will be used are Whisper base, Small, Medium, and Large V2.

Table 1. Whisper Model Comparison

| Model | Layer | Width | Heads | Size |
|---|---|---|---|---|
| Tiny | 4 | 384 | 6 | 39 M |
| Base | 6 | 512 | 8 | 74 M |
| Small | 12 | 768 | 12 | 244 M |
| Medium | 24 | 1024 | 16 | 769 M |
| Large-V2 | 32 | 1280 | 20 | 1550 M |

Source: (Radford et al., 2022)

Fine-tuning models with parameters that exceed 200 M requires immense computational cost. The Free computational power Google Colaboratory provides is 14.7 GB T4 GPU and 12.7 GB RAM. To address this problem, model quantization is proposed. Model quantization first proposed by (Dettmers et al., 2022) is 8-bit quantization. The full model works in full precision using the fp32 data type. The model will be loaded in quarter precision or 8-bit data type to reduce memory and computational requirements and proved that only ~5% loss in model accuracy compared to the full precision model.

### Parameter-Efficient Fine-Tuning

To reduce even more computational and memory requirements, Parameter-Efficient Fine-Tuning (PEFT) is proposed. PEFT works by freezing model parameters, so only required parameters related to specific tasks will be trained (Liu et al., 2022). This reduction is possible because of the LoRA method. Low-rank Adaptation (LoRA) is a method proposed by (Hu et al., 2021) that works by adding a decomposition matrix ( Update matrix) to model weights and training only newly added weights, resulting in the reduction of trainable parameters and consequently reducing memory and computational cost.

### Model Fine-Tuning

Fine-tune done in Google colaboratory environment with system specification 14.7GB T4 GPU and 12.7GB RAM. The hyperparameter that used for fine-tuning can be seen in Table 2.

Table 2. Hyperparameter

| Hyperparameter | Value |
|---|---|
| Learning rate | 1e-3 |
| Epoch | 10 |
| Batch_size | 8 |
| Optimizer | AdamW |
| Evaluation Strategy | Epoch |
| Gradient_Checkpointing | True |

Source: (Research Result, 2023)

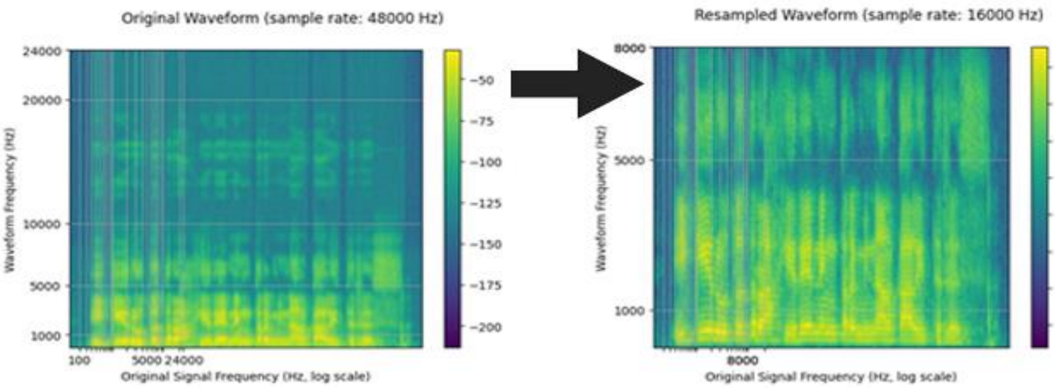
### Evaluation

Word Error Rate (WER) is the evaluation matrix used to evaluate the model. WER calculation is the same as the Character Error Rate (CER). Nevertheless, WER represented ASR performance more accurately because the error rate was evaluated based on words instead of characters. Therefore, it can be said that when ASR is evaluated from WER, its error rate is higher than one thst evaluated with CER. In previous research (Novitasari et al., 2020), the evaluation used to evaluate the ASR model is CER. This method is used because the language contains some characters outside the standard alphabet. But in Javanese, it can be written in the standard alphabet. So, the WER evaluation will be used to provide a more accurate evaluation of ASR performance in recognizing Javanese speech. Mathematically, WER can be calculated with the equation (1) below where S is substitution, D is deletion, I is insertion, and C is correct.

$$WER = \frac{(S + D + I)}{(S + D + C)} \quad \text{..........................................} (1)$$

## RESULTS AND DISCUSSION

Whisper, especially its feature extractor, works best at a 16 kHz sample rate after the original sample rate of 48 kHz resampling to 16 kHz; the result of audio resampling can be seen in Figure 4.




Source: (Research Result, 2023)

Figure 4. Sample Rate Waveform Comparison

As shown in Figure 4, the original waveform visualization with a 48kHz sample rate is represented in 24kHz waveform frequency on the Y-axis and 24kHz original signal frequency on the X-axis. Compared to the original audio sample rate, the 16kHz resampled waveform in Figure 4 is represented with 8kHz waveform frequency and 8kHz original Signal Frequency, a zoomed version of Figure 4. The difference between the two sample rate visualizations is an artifact seen in the upper 8kHz of the original sample rate in Figure 4. This artifact removal does not affect speech in audio because the main speech audio is represented in the 8kHz Y-axis and X-axis range.

At the data splitting, the ratio of splitting used is 8:2, resulting in 4657 training data and 1165 evaluation data in a random state, which means that every epoch where 4657 data have been trained will be evaluated with 1165 data.

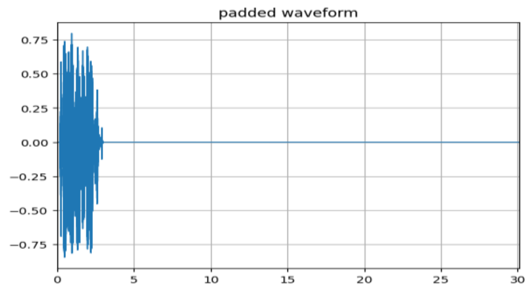
Example of the result from tokenizing with Whisper tokenizer can be seen in Table 3.

Table 3. Result of Tokenizing

| Transcription | Tokenized |
|---|---|
| bar ngepeki sayuran banjur ditawake neng bandungan | {bar, ngepeki, sayuran, banjur, ditawakake, ning, bandungan} |
| nyebrang menyang ketapang adoh | {nyebrang, menyang, ketapang, adoh} |
| Puding ingkang didamel purimas radi mambet | {pudding, ingkang, didamel, purimas, radi, mambet} |

Source: (Research Result, 2023)

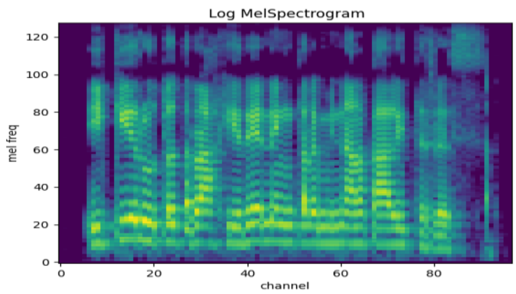
In Table 3, the Whisper tokenizer successfully tokenizes transcription text in the Javanese language into tokens per word. Because all audio data duration is less than 30 seconds, the Whisper feature extractor will pad audio waveform with 27 seconds of 0 signal or silence. The audio waveform after padding can be seen in Figure 5. In Figure 5, it can be seen that 27 seconds of silence was successfully added after the speech ended in 3 seconds.




Source: (Research Result, 2023)

Figure 5. Padded Audio Waveform

The next step is to convert this waveform into an 80-channel log-mel spectrogram visualization. The result of visualization can be seen in Figure 6.




Source: (Research Result, 2023)

Figure 6. Log -mel Spectrogram

The log-mel spectrogram was successfully visualized, and as can be seen, the log-mel spectrogram was divided into 80 channels on the X-axis. The silence is automatically omitted, and detected audio is converted into a log-mel spectrogram; as can be seen, visualization is almost identical to the waveform spectrogram in Figure 5.

After all preprocessing steps are complete, the research will proceed into model preparation. First, we load the model in 8-bit quantization, then apply PEFT using LoRa to reduce trainable parameters. The result of the reduction can be seen in Table 4.

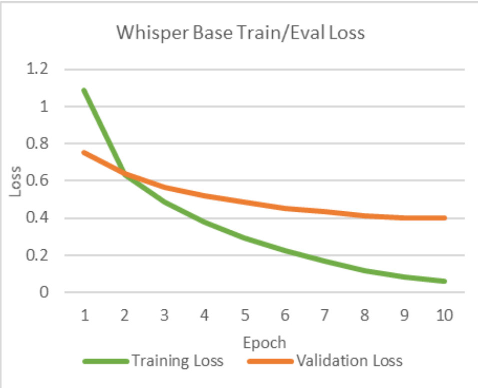
PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 1 March 2022 | DOI: 10.33480/pilar.v20i1.4633

Table 4. Quantization and PEFT result

| Model | Base Model Param. | LoRa Param. | Percentage % |
|---|---|---|---|
| Base | 73,183,744 | 589,824 | 0.80 |
| Small | 243,504,384 | 1,769,472 | 0.72 |
| Medium | 768,576,512 | 4,718,592 | 0.62 |
| Large-V2 | 1,551,169,280 | 7,864,320 | 0.51 |

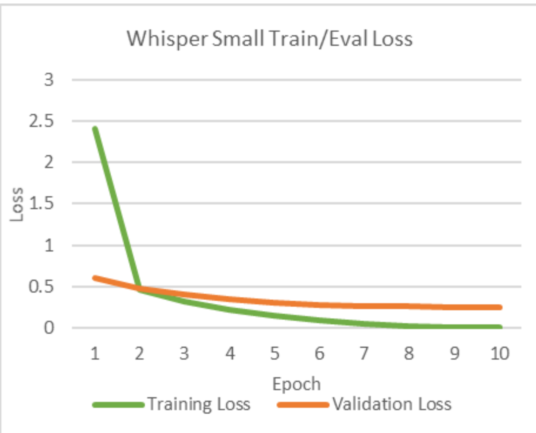
Source: (Research Result, 2023)

After the model is prepared, the fine-tuning process is ready to start. Figures 7, Figure 8, Figure 9, and Figure 10 presented a graph of training and evaluation loss of each model during the fine-tuning process.



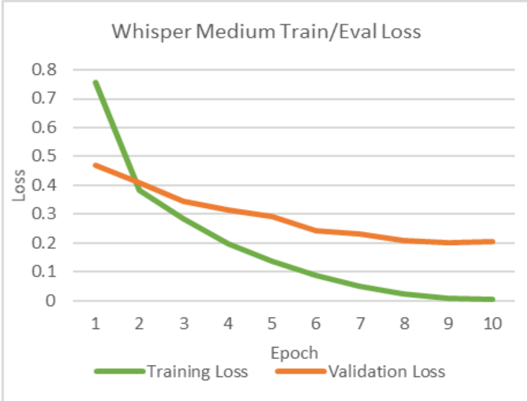

Source: (Research Result, 2023)
Figure 7. Whisper Base Training/Eval Loss

As shown in Figure 7, training and evaluation loss is reduced as the epoch continues for the Whisper-Base model. Significant improvement happened in epoch two as the model became more familiar with the pattern of Javanese speech data. There is no indication of overfitting or underfitting based on the gap between training and evaluation loss value.



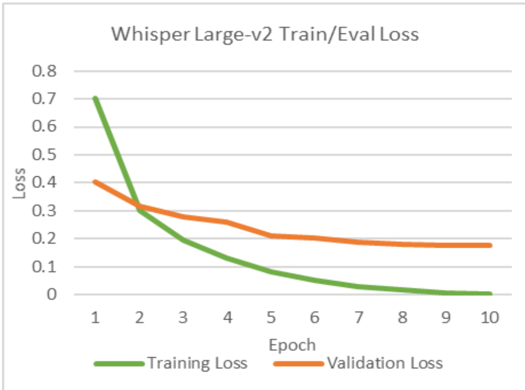

Source: (Research Result, 2023)
Figure 8. Whisper Small Training/Eval Loss

Figure 8 shows a drastic reduction in training loss in epoch two and measures up to evaluation loss. There is no significant reduction from epoch three to ten. Otherwise, there is no indication of overfitting or underfitting in the Whisper-Small model fine-tuning process.




Source: (Research Result, 2023)
Figure 9. Whisper Medium Training/Eval Loss

In Figure 9, a significant reduction of training loss happened in epoch two. The evaluation loss reduction goes down smoothly, but in epoch ten, evaluation loss is increased slightly by 0.001065%. To that end, the fine-tuned model will be used on epoch nine. The training and evaluation loss graph shows no indication of overfitting or underfitting.




Source: (Research Result, 2023)
Figure 10. Whisper Large-v2 Training/Eval Loss

Like the other model, Whisper-Large-v2 training and evaluation loss during fine-tuning goes down for each epoch, as shown in Figure 10. Training loss starts to measure up with evaluation loss in epoch two. There is a slight increase of 0.001422% in evaluation loss in epoch ten. Because of that, the end model that will be used is the model on epoch nine. There is no overfitting or underfitting indication in Whisper-Large-v2 during the fine-tuning process, as shown in the graph in Figure 10.

Comparing the results from Figures 7,8,9, and 10, there is one similarity between all models: the most significant loss value reduction is at epoch 2. This happens because models start familiarizing themselves with data patterns after two epochs. All models stop improving at epoch 9. The spotted difference is on the smaller model ( base and small model) loss value, as shown in Figures 7 and 8, which goes down smoothly each epoch after epoch 2. On the other hand, in Figures 9 and 10, the medium and large model shows a visible drastic loss value reduction in epochs 3, 4, 5, and 6. This proves that larger models are learning better than smaller ones.

Table 5 shows the WER evaluation result and comparison between the base whisper model (model before fine-tuned) and the fine-tuned whisper model.

Table 5. WER Comparison

| Model | Word Error Rate(WER) % | |
|---|---|---|
| | Base Model | Fine-tuned Model |
| Base | 116.87 | **28.57** |
| Small | 109.63 | **18.84** |
| Medium | 110.62 | **15.97** |
| Large-V2 | 89.40 | **13.77** |

Source: (Research Result, 2023)

In Table 5, the result of the WER evaluation is consistently better for each fine-tuned model. The improvement in WER evaluation of the performance of each Whisper ASR model is up to 85% reduction in error rate compared to the model without fine-tuning. Thus, the fine-tuning process improved WER evaluation results for low-resource language, in this case, Javanese language speech recognition.

## CONCLUSION

Based on the experiment conducted above, the research could be concluded that fine-tuning the whisper model on a low-resource language dataset could improve Whisper ASR model performance in recognizing speech spoken in low-resource language, which in this research case, Javanese language that measured with Word Error Rate(WER) evaluation. The improvement in the WER result is significantly better for every Whisper model tested in this research.

## REFERENCE

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., ... Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, *9*, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). Llm. int8 (): 8-bit matrix multiplication for transformers at scale. arXiv preprint arXiv:2208.07339.

Fu, Z., Yang, H., So, A. M. C., Lam, W., Bing, L., & Collier, N. (2023, June). On the effectiveness of parameter-efficient fine-tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 11, pp. 12799-12807).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(4), 531-538.

Kumar, A., & Mittal, V. (2019). Speech recognition: A complete perspective. International Journal of Recent Technology and Engineering (IJRTE), 7(6), 78-83.

Li, J. (2022). Recent advances in end-to-end automatic speech recognition. APSIPA Transactions on Signal and Information Processing, 11(1).

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. A. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Advances in Neural Information Processing Systems, 35, 1950-1965.

Min, S., Lewis, M., Zettlemoyer, L., & Hajishirzi, H. (2021). Metaicl: Learning to learn in context. arXiv preprint arXiv:2110.15943.

Novitasari, S., Tjandra, A., Sakti, S., & Nakamura, S. (2020). Cross-lingual machine speech chain for javanese, sundanese, balinese, and bataks speech recognition and synthesis. arXiv preprint arXiv:2011.02128.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Rouditchenko, A., Khurana, S., Thomas, S., Feris, R., Karlinsky, L., Kuehne, H., ... & Glass, J. (2023). Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages. arXiv preprint arXiv:2305.12606.

Vol. 18, No. 1 March 2022 | DOI: 10.33480/pilar.v20i1.4633

Wang, C., Cho, K., & Gu, J. (2020, April). Neural machine translation with byte-level subwords. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 05, pp. 9154-9160).

Butryna, A., Chu, S. H. C., Demirsahin, I., Gutkin, A., Ha, L., He, F., ... & Wibawa, J. A. E. (2020). Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: an overview. arXiv preprint arXiv:2010.06778.

Toraman, C., Yilmaz, E. H., Şahinuç, F., & Ozcelik, O. (2023). Impact of tokenization on language models: An analysis for turkish. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 1-21.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.

Yang, H., Zhang, M., Tao, S., Ma, M., & Qin, Y. (2023, February). Chinese ASR and NER Improvement Based on Whisper Fine-Tuning. In 2023 25th International Conference on Advanced Communication Technology (ICACT) (pp. 213-217). IEEE.

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., ... & Perrault, R. (2021). The AI index 2021 annual report. arXiv preprint arXiv:2103.06312.

Zhang, X., Peng, Y., & Xu, X. (2019, September). An overview of speech recognition technology. In 2019 4th International Conference on Control, Robotics and Cybernetics (CRC) (pp. 81-85). IEEE.

# DESIGN AND DEVELOPMENT OF AN INTERNAL QUALITY AUDIT INFORMATION SYSTEM BASED PPEPP CYCLE

**Ahmad Yani[1*]; Lalu Darmawan Bakti[2]; Ardiyallah Akbar[3]; Bahtiar Imran[4]**

Teknologi Informasi[1], Rekayasa Perangkat Lunak[2], Teknik Komputer[3], Rekayasa Sistem Komputer[4]
Universitas Teknologi Mataram, Mataram, Indonesia[1,2,3,4]
https://utmmataram.ac.id[1,2,3,4]
m4dy45@gmail.com[1*], laludarmawanbakti.utm@gmail.com[2], ardiyallah_akbar@ymail.com[3], bahtiarimranlombok@gmail.com[4]
(*) Corresponding Author



***Abstract***—*The Mataram University of Technology Quality Assurance Institute already has and has established national education standards plus the standards set by universities following Permendikbud number 3 of 2020. However, there are problems with the implementation of Internal Quality Audits, where the implementation of internal quality audits is very less effective and efficient, good in terms of time, cost, and energy. This is because the Mataram University of Technology Quality Assurance Institute only has 3 auditors to audit 12 study programs in one year and even spends two months in a row. This is an important concern for researchers to build and produce an internal quality audit information system application program that can help implement the internal quality audit process carried out by the Mataram University of Technology Quality Assurance Institute. The design of the internal quality audit information system was carried out using the prototyping method. The application of the prototyping method in system design will make information system builders better and more structured. The internal quality audit information system was built using the PHP programming language with the CodeIgniter framework and MySQL as the database and implementing Code-View-Controller (MVC). The main objective of this research is to produce an internal quality audit information system so that it can assist the Mataram University of Technology Quality Assurance Institute in documenting and optimizing higher education quality management in a planned and sustainable manner following the PPEPP cycle.*

***Keywords:*** *information systems, internal quality audit, PPEPP cycle, quality.*

***Abstrak***—*Lembaga Penjaminan Mutu Universitas Teknologi Mataram sudah memiliki dan menetapkan standar nasional pendidikan ditambah dengan standar yang ditetapkan oleh perguruan tinggi sesuai dengan permendikbud nomor 3 tahun 2020. Namun terdapat permasalahan pada pelaksanaan Audit Mutu Internal, dimana pelaksanaan audit mutu internal sangat kurang efektif dan efesien, baik dari segi waktu, biaya dan tenaga. Hal ini disebabkan karena Lembaga Penjaminan Mutu Universitas Teknologi Mataram hanya memiliki 3 orang Auditor untuk mengaudit 12 program studi dalam satu tahun berjalan dan bahkan sampai menghabiskan waktu dua bulan berturut-turut. Hal inilah yang menjadi perhatian penting peneliti untuk membagun dan menghasilkan program aplikasi sistem informasi audit mutu internal yang dapat membantu pelaksanaan proses audit mutu internal yang di lakukan oleh Lembaga Penjaminan Mutu Universitas Teknologi Mataram. Perancangan sistem informasi audit mutu internal dilakukan dengan menggunakan metode prototyping. Penerapan metode prototyping dalam perancangan sistem akan membuat pembangun sistem informasi menjadi lebih baik dan terstruktur. Sistem informasi audit mutu internal dibangun menggunakan bahasa pemrograman PHP dengan framework CodeIgniter dan MySQL sebagai databasenya serta menerapkan Code-View-Controller (MVC). Adapun tujuan utama dari penelitian ini adalah menghasilkan sistem informasi audit mutu internal sehingga mampu membantu Lembaga Penjaminan Mutu Univeristas Teknologi Mataram dalam mendokumentasikan dan mengoptimalisasikan manajemen mutu perguruan tinggi secara berencana dan berkelanjutan sesuai dengan siklus PPEPP.*

***Kata Kunci****: sistem informasi, audit mutu internal, siklus PPEPP, mutu.*

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4729

## INTRODUCTION

Quality higher education is a big task that has been mandated by the Government through Law Number 12 of 2012 concerning Higher Education, where all universities in Indonesia must implement a Quality Assurance System to produce and create quality education. Quality higher education is higher education that produces graduates who are able to actively develop their potential and produce science and/or technology that is useful for society, nation, and state (Kebudayaan, 2020).

To produce quality education, universities must implement a Quality Assurance System in a planned and sustainable manner in accordance with the Cycle of Determining, Implementing, Evaluating, Controlling, and Improving Higher Education Standards. The evaluation as intended in the PPEPP cycle is carried out through an Internal Quality Audit (Direktorat Penjaminan Mutu, 2018).

Internal Quality Audit is a systematic, independent, and documented testing process to ensure the implementation of activities in higher education in accordance with established procedures and standards to achieve institutional goals. Thus, AMI is a very strategic stage in developing the quality of higher education, especially to improve quality on an ongoing basis (Direktorat Penjaminan Mutu, 2018).

The quality of higher education is the level of conformity between the implementation of higher education and Higher Education Standards consisting of National Higher Education Standards and Standards set by Higher Education Institutions (Kebudayaan, 2020) . To achieve this suitability, evaluation must be carried out.

Evaluation is a comparison activity between the output of activities that have been implemented by institutions and study programs with the fulfillment of national higher education standards and established higher education standards (Kementerian Riset, Teknologi, 2018). Without a good and planned evaluation, universities will not be able to carry out control and improve their quality standards.

The Mataram University of Technology Quality Assurance Institute has so far established and implemented standards, both national higher education standards and standards set by universities in accordance with Permendikbud No. 3 of 2020 (Kebudayaan, 2020). However, there are weaknesses in the Internal Quality Audit process carried out, where the Internal Quality Audit is still carried out manually and has not been systemized. Implementing an Internal Quality Audit which is carried out manually will take a very long time, even months, and require a lot of energy and costs, as well as other facilities, so, to overcome these problems, an application system is needed that able to help and provide information quickly and efficiently. appropriate, related to the data and information needed in the process of implementing the Internal Quality Audit (Muslim et al., 2021).

An information system is a technique that has the task of forming, processing, storing, analyzing, and disseminating information to achieve agreed goals (Komalasari et al., 2023).

Internal Quality Audit is a routine activity carried out by the Quality Assurance Agency repeatedly to ensure the implementation of the standards that have been set so that the quality of education can be achieved very well (Febriyanti & Irawan, 2020).

An internal quality audit information system is the application of information technology to help processes or activities carried out by a group of people become better and easier to produce the required information (Agus et al., 2023).

The Prototype method is a software development method that allows interaction between system developers and system users, so as to overcome incompatibility between developers and user (Hasanah & Untari, 2020). By applying this prototype method, system development becomes better because it suits user needs (Erkamim et al., 2022).

Meaning, that when internal auditors are efficient, they will be able to extend the scope of their tasks and carry them out effectively. Ultimately, when internal auditors become more effective and efficient, it is expected that stakeholders will be more likely to be persuaded and trust the work of the electronic internal audit (Alqudah et al., 2023).

There have been several previous studies related to the research that will be carried out, including The results of research by Suryo Widiantoro and Yodi with the title "Design and Build an Internal Quality Audit Information System Based on IAPS 4.0" in 2020. Where the research only produced a design and framework from the Internal Quality Audit information system which cannot yet be implemented (Widiantoro & Yodi, 2020), while in this research we will design and build an Internal Quality Audit information system application that can immediately be implemented to help the performance of the Mataram University of Technology Quality Assurance Institute in improving the quality of education..

Research carried out by Dwi Rani Febriyanti and Hendri Irawan with the title “Implementation of a Web-Based Internal Quality Audit Information System to Increase Work Efficiency Case Study: Budi Luhur University Quality Assurance Institute” in 2020. In this research, the quality management implemented or used still uses the old management, namely PDCA (Febriyanti & Irawan, 2020). Meanwhile, the research that will be carried out will apply the latest quality management based on the PPEPP Cycle which is in accordance with Permenristekdikti No. 62 of 2016 Article 5 (Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 62 Tahun 2016 tentang Sistem Penjaminan Mutu Pendidikan Tinggi, 2016).

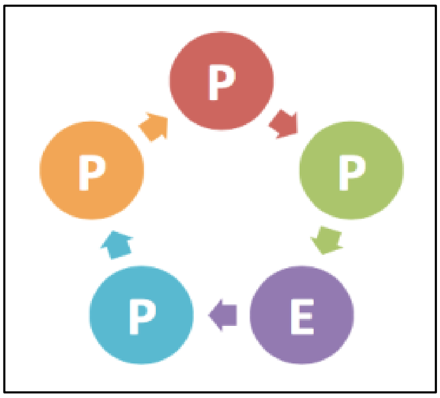
Research carried out by Andie, Muhammad Hasbi, and Hasanuddin with the title "Internal Quality Audit Information System (SIAMI)" in 2021. This application system produces an Internal Quality Audit report based on input questions entered into the application system and there is no visible control and improvement process, apart from that, this application was built using the waterfall method, PHP programming language, Dreamweaver 5.5 as a text editor (Andie et al., 2021). Meanwhile, the system to be built will be up to date in terms of output produced by a system that follows the PPEPP cycle in the audit process carried out so that the system will not only produce a quality audit report but will produce control over the results of the internal quality audit that has been carried out. Apart from that, this system will be built using the Prototyping method and the CodeIgniter framework which adopts the Model View Controller (MVC) design pattern (Norfifah et al., 2023).

Research carried out by Eva Faja Ripanti and H. A. Oramahi with the title "Design of an Information System for Management of Internal Quality Audits (AMI) in Higher Education" in 2021. The designed application cannot be fully implemented or finished, because some of the corrections or improvements that will be given, both by the auditee and the auditor, are still done manually or outside the system (Ripanti & Oramahi, 2021). Meanwhile, the research carried out will produce an internal quality audit information system that is capable of providing corrections or corrective actions through the information system that is built.

Based on the research that has been carried out, several differences or updates can be drawn from the research, including those related to cycles, software development methods, and outputs produced by information systems. The aim of this research is to produce an internal quality audit information system application that is able to assist universities in improving the quality of education based on the Determination, Implementation, Evaluation, Control, and Improvement (PPEPP) cycle which can be implemented every year continuously and sustainably in order to create a culture quality of higher education.

## MATERIALS AND METHODS

The prototyping method is a software development method that can be applied to the development of small and large systems with the hope that the development process can run well. This prototyping aims to collect information and design and build a system based on the needs of users (Kustanto & Chernovita, 2021), namely the Mataram University of Technology Quality Assurance Institute following the Determination (Penetapan), Implementation (Penerapan), Evaluation (Evaluasi), Control (Pengendalian), and Improvement (Peningkatan) or abbreviated as (PPEPP) cycle as shown in Figure 1.




Source: (Directorate of Quality Assurance, 2018)
Figure 1. PPEPP Cycle

In the internal quality audit information system program, the determination menu will contain an input menu for documents and instruments created based on the quality standards held by the Mataram University of Technology Quality Assurance Institute. For the implementation menu, each unit will upload all files or evidence of standard implementation that have been implemented in the current year, while at the evaluation stage, the auditor will provide responses or comments on files or evidence of standard implementation that have been uploaded by each unit or study program. At the control stage, all study programs or units within the Mataram University of Technology will provide feedback on the responses or comments given by the auditor, while in the improvement menu, new indicators will be entered into each standard document in accordance with the results that have been achieved which will then be determined and implemented in the following year, so that the cycle can continue to be implemented in a planned and sustainable manner. The stages in creating this internal quality audit information system are as follows:

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4729

**Requirements Collection**

In gathering the design requirements for an internal quality audit information system based on the PPEPP cycle, of course, researchers must involve application system users, namely the Mataram University of Technology Quality Assurance Institute team, in order to find out the problems and obstacles that have occurred so far. So by knowing these constraints, the need for data and information to design and build an internal quality audit information system can be easily determined. The data that can be collected include:

**a. Study Program Data**

The Mataram University of Technology has 12 study programs consisting of 7 academic education programs and 5 vocational education programs, as shown in Table 1.

Table 1. Study Program Data

| No | Study Program | Education Programs |
|---|---|---|
| 1 | Informatics Engineering | |
| 2 | Information Systems | |
| 3 | Information Technology | |
| 4 | Computer Systems Engineering | |
| 5 | Software Engineering | |
| 6 | Management | |
| 7 | Law | Academic (Undergraduate Program) |
| 8 | Computer Engineering | |
| 9 | Informatics Management | |
| 10 | Computerized Accounting | Vocational (Diploma Program) |
| 11 | Administrative Management | |
| 12 | Secretary | |

Source: (Universitas Teknologi Mataram, 2022)

**b. Auditor Data**

The Mataram University of Technology Quality Assurance Institute has 3 internal auditors who have carried out internal quality audits, including Lalu Delsi Samsumar, M.Eng, Ahmad Yani, M.Kom and Karina Nurwijayanti, M.Pd.

**c. Instrument Data along with Internal Quality Audit Report**

Internal Quality Audit instrument data and output are used as samples in creating an internal quality audit information system as shown in Table 2.

Table 2. Internal Quality Audit information system

| No | Criteria | Question | Audit Notes | findings | Document |
|---|---|---|---|---|---|
| 1 | Cooperation | Has the institution implemented collaboration in the 3 fields, namely Education, Research, and Community Service? | Yes. Collaboration has been carried out in these 3 fields. | Achieved | Cooperation Agreement Letter (MoU) Proof of Collaboration |
| 2 | | Has the number of local/regional level collaborations reached 4 per year? | Yes. Local level cooperation has reached 9 collaborations. | Achieved | Cooperation Agreement Letter (MoU) Proof of Collaboration Collaboration Monitoring and Evaluation Report |
| 3 | | Has the number of international level collaborations reached 1 per year? | There is no international cooperation yet. | Major | |
| 4 | | Has the number of collaborations in the education sector reached 2 per year? | Yes. Collaboration in the education sector has reached 4 collaborations and has exceed | Achieved | Cooperation Agreement Letter (MoU) Proof of Collaboration |

| No | Criteria | Question | Audit Notes | findings | Document |
|---|---|---|---|---|---|
| | | | ed the standards that have been set | | |

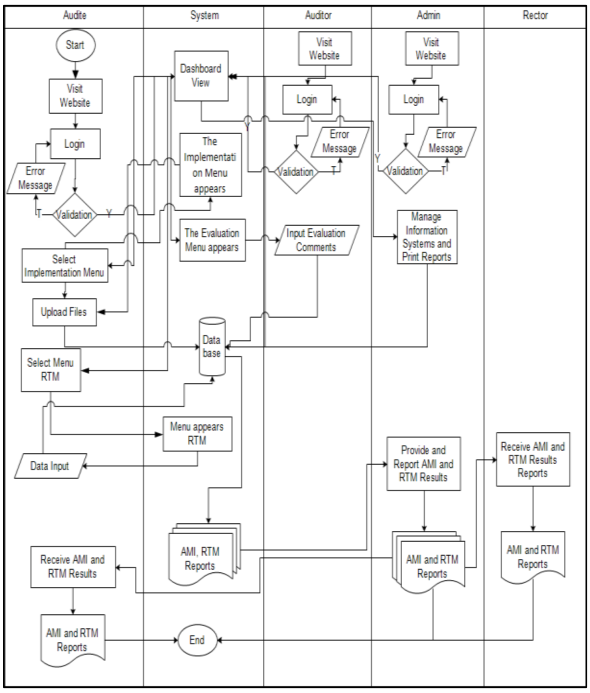
Source: (Universitas Teknologi Mataram, 2022)

## Design Process

Based on data and information from the LPM TEAM, the researcher then carried out a system design process which included:

### a. System Flowchart

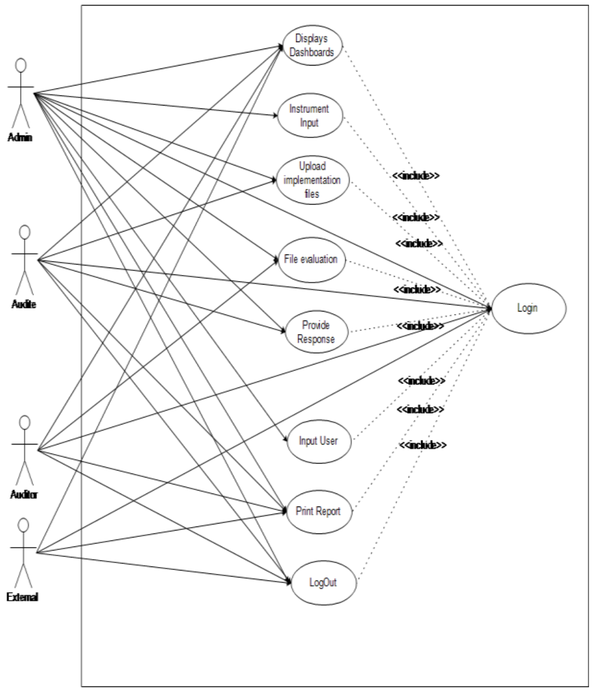
This system flowchart describes how the flow of the internal quality audit information system works, both by the Audite (Study Program) as the auditee, the Auditor (Evaluation TEAM) as the assessor of the data that has been uploaded or attached by the auditee to the system, and the Admin as the manager full of the information system created. The system flowchart image looks like in Figure 2.




Source: (Research Result, 2023)
Figure 2. System Flowchart

### b. Use Case Diagrams

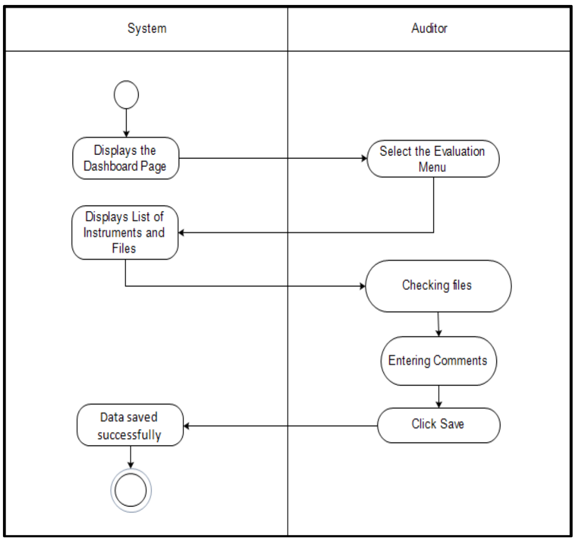
The Use Case Diagram will describe the activities that can be carried out by each actor when accessing the internal quality audit information system. The use case diagram that was built is shown in Figure 3.




Source: (Research Result, 2023)
Figure 3. Use Case Diagram

### c. Activity Diagrams

The activity diagram depicts the activities that will be carried out by the auditor, where the auditor will examine the files that have been attached or uploaded by the auditee to the internal quality audit information system, to ensure that the attached documents comply with the instrument items. The following is an example of an activity diagram carried out by an auditor after logging in to the internal quality audit information system as shown in Figure 4.

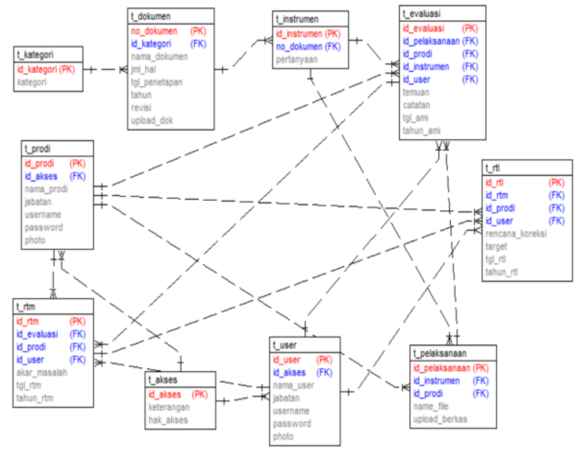



Source: (Research Result, 2023)
Figure 4. Activity Diagram

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4729

**d. Entity Relationship Diagram (ERD)**

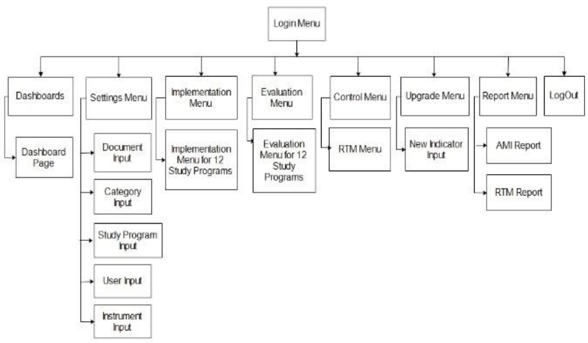
This entity relationship diagram depicts the relationships formed between one table and another table in the internal quality audit information system database. The design of the ERD internal quality audit information system can be seen in Figure 5.



Source: (Research Result, 2023)
Figure 5. Entity Relationship Diagram (ERD)

**e. Program Architecture**

The program architecture will describe how the menus are arranged in the quality audit information system that will be built. The architecture of the internal quality audit information system program is shown in Figure 6.




Source: (Research Result, 2023)
Figure 6. Program Architecture

**Building Prototypes**

At this stage, an internal quality audit information system will be created, using the CodeIgniter framework, PHP programming language, and MySQL as a database.

**Prototype Evaluation**

Prototype evaluation is carried out through testing carried out using black box testing with the aim of finding out whether the system being built is able to run well, both in terms of input and output produced (Yani et al., 2022).

## RESULTS AND DISCUSSION

This research produces an internal quality audit information system that can help the Mataram University of Technology Quality Assurance Institute overcome the problems faced when carrying out the internal quality audit process. Apart from that, this research also has the latest results from previous research, both in terms of software development methods. , interface design, and resulting output. The results of the implementation of the internal quality audit information system can be explained as follows:

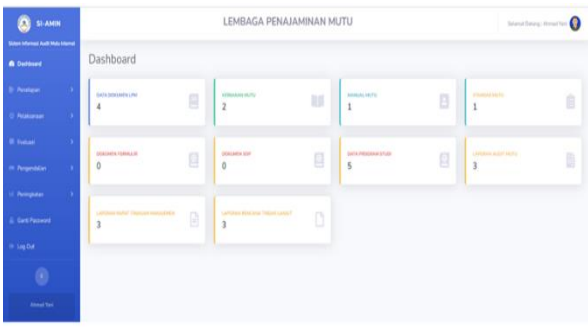
**1. System Login Page**

The login page is used for users to enter the internal quality audit information system, where the user must enter their user and password to be able to enter the dashboard page. The login page menu looks like Figure 7


Source: (Research Result, 2023)
Figure 7. Login page

**2. Dashboard page**

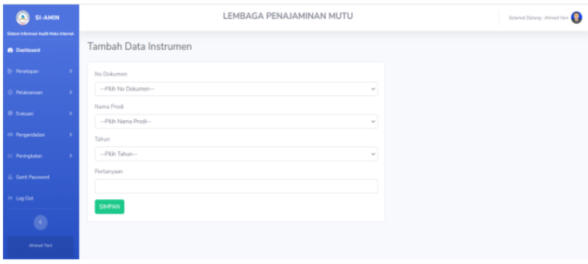
The dashboard page is used to view all the information in the internal quality audit information system. The appearance of the dashboard page is shown in Figure 8.




Source: (Research Result, 2023)
Figure 8. Dashboard page

**3. Instrument Input Form**

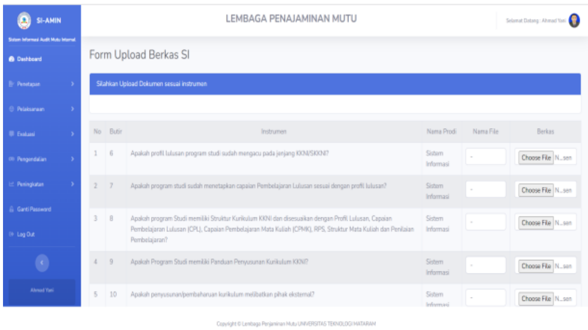
The instrument data input page is used to enter questions that are used when conducting internal quality audits, where the questions are taken from the standard documents of the Mataram University of Technology Quality Assurance Institute. The display of the instrument data input form is shown in Figure 9.




Source: (Research Result, 2023)
Figure 9. Instrument Input Page

## 4. Study Program Implementation Page

The standard implementation page is a page used by each study program to upload standard implementation files based on questions that have been entered into the internal quality audit information system. The display of the standard implementation document upload page is shown in Figure 10.




Source: (Research Result, 2023)
Figure 10. Implementation of the Study Program

On this page each study program will go to each page according to its study program, on this page the study program will write the name of the document and upload the document according to the available questions, which will later be evaluated by the auditor team on the evaluation page.

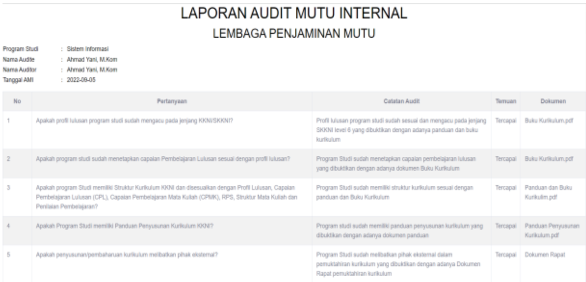
## 5. Evaluation Page

The evaluation page is used by auditors to provide an assessment of documents that have been uploaded by the study program in implementing the standards that have been set so that with these documents the auditor will provide an assessment of the conformity of the document with the instrument. If the documents comply or do not comply with or exceed the established standards, the auditor will assign a category to the findings and provide notes on the results found. The appearance of the evaluation page looks like Figure 11.


Source: (Research Result, 2023)
Figure 11. Evaluation page

## 6. Report Page

The report page is used to print the results of internal quality audit reports that have been carried out or filled in by the auditee (study program) or by the auditor. The appearance of the Internal Quality Audit Report is shown in Figure 12.




Source: (Research Result, 2023)
Figure 12. Internal Quality Audit Report

Based on the results of information system testing using the black box testing method, it can be ensured that all processes in the system can run well and smoothly without any errors. The results of the trial results on this internal quality audit information system can be seen in Table 3.

Table 3. Evaluation Results

| No | Module | Scenario | Result | Description |
|---|---|---|---|---|
| 1 | Instrument Data Set Menu | Instrument Data Input | Instruments can be added | Valid |
| 2 | Study Program Implementation Menu | Input Document Name | Document Name can be input | Valid |
| | | Upload Implementation Documents according to the instrument | Documents can be Uploaded | Valid |
| 3 | | View implement | Displays implemen | Valid |

| No | Module | Scenario | Result | Description |
|---|---|---|---|---|
| | Evaluation Menu (Auditor) | ation documents | tation documents | |
| | | Provide information on evaluation results | Information on the evaluation results can be input | Valid |
| | | Select a finding category | Can select the finding category | Valid |
| 4 | Control Menu | Input the Root of the problem | Successfully input the root of the problem | Valid |
| | | Input corrective action | Successfully input corrective action | Valid |
| | | Select a repair deadline | Successfully selected the repair deadline | Valid |
| 5 | Upgrade Menu | Input improvement indicator | Successfully input standard improvement indicators | Valid |
| 6 | Report Menu | Select Audit Report Based on Study Program | Displays the Study Program Quality Audit Report | Valid |
| | | Select Management Review Meeting Report | Display Management Review Meeting Report | Valid |

Source: (Research Result, 2023)

Based on the results of the trial implementation of the internal quality audit information system using the black box testing method, all menus and data input have been confirmed to run well follow what the user expects, and are free from errors.

## CONCLUSION

With the internal quality audit information system produced in this research, the Mataram University of Technology Quality Assurance Institute can overcome and resolve every problem and obstacle that has been encountered when carrying out internal quality audits, because all the activities carried out are structured and programmed appropriately well in a system. Apart from that, this internal quality audit information system can provide convenience for quality assurance institutions in making reports and documentation as well as improving quality periodically and continuously.

## REFERENCE

Agus, R. S., Winarni, A., & Enjinering Indorama, P. (2023). Sistem Audit Mutu Internal Politeknik Enjinering Indorama. *Jurnal Manajemen dan Bisnis Jayakarta*, *5*(1), 45–64. https://doi.org/https://doi.org/10.53825/jmbjayakarta.v5i01.188

Alqudah, H., Lutfi, A., Abualoush, S. habis, Al Qudah, M. Z., Alshira'h, A. F., Almaiah, M. A., Alrawad, M., & Tork, M. (2023). The Impact Of Empowering Internal Auditors On The Quality Of Electronic Internal Audits: A case Of Jordanian Listed Services Companies. *International Journal of Information Management Data Insights*, *3*(2), 1–8. https://doi.org/10.1016/j.jjimei.2023.100183

Andie, A., Hasbi, M., & Hasanuddin, H. (2021). Sistem Informasi Audit Mutu Internal (Siami). *Technologia: Jurnal Ilmiah*, *12*(2), 110–120. https://doi.org/10.31602/tji.v12i2.4758

Direktorat Penjaminan Mutu. (2018). *Pedoman Audit Mutu Internal (AMI) Kementerian Riset, Teknologi, dan Pendidikan Tinggi Direktorat Penjaminan Mutu Direktorat Jenderal Pembelajaran dan Kemahasiswaan* (Edisi ke-1). Direktorat Penjaminan Mutu.

Erkamim, M., Wati, V., Wartono, W., & Fitriyadi, F. (2022). Rancang Bangun Sistem Informasi Pengelolaan Penelitian Dosen Menggunakan Metode Prototyping. *Tekinfo: Jurnal Ilmiah Teknik Industri dan Informasi*, *10*(2), 117–128. https://doi.org/https://doi.org/10.31001/tekinfo.v10i2.1567

Febriyanti, D. R., & Irawan, H. (2020). Penerapan Sistem Informasi Audit Mutu Internal Berbasis Web Guna Meningkatkan Efisiensi Kerja Studi Kasus: Lembaga Penjaminan Mutu Universitas Budi Luhur. IDEALIS: InDonEsiA journaL Information System, 3(1), 474-480.

Hasanah, F. N., & Untari, R. S. (2020). Buku Ajar Rekayasa Perangkat Lunak. Umsida Press, 1-119.

Kebudayaan, M. P. dan. (2020). Peraturan Menteri Pendidikan Dan Kebudayaan Nomor 03 Tahun 2020 Tentang Standar Nasional Perguruan Tinggi.

Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 62 Tahun 2016 tentang Sistem Penjaminan Mutu

Pendidikan Tinggi, Kemenristekdikti (2016).

Kementerian Riset, Teknologi, dan P. T. (2018). Pedoman Sistem Penjaminan Mutu Internal (Edisi ke-2). Direktorat Penjaminan Mutu.

Komalasari, R., Akhriana, A., Mandias, G. F. dkk. (2023). Pengantar Teknologi Informasi (D. Gustian (ed.); Cetakan Ke). Indie Press.

Kustanto, G. E. A., & Chernovita, H. P. (2021). Perancangan Sistem Informasi Manajemen Berbasis Web Studi Kasus : PT Unicorn Intertranz. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *8*(4), 719. https://doi.org/10.25126/jtiik.2021844849

Muslim, I., Maksum, S. A. R., Rika, S. P., & Henin Rasio Silvana. (2021). Rancang Bangun Sistem Audit Mutu Internal Guna Optimalisasi Kinerja Penjaminan Mutu Perguruan Tinggi. *Sistemasi: Jurnal Sistem Informasi*, *10*(2), 490–501. https://doi.org/https://doi.org/10.32520/stmsi.v10i2.1374

Norfifah, Veri, J., & Yunita, P. (2023). Rancang Bangun Sistem Informasi Audit Mutu Internal. *Journal of Applied Computer Science and Technology (JACOST)*, *4*(2), 108–117. https://doi.org/https://doi.org/10.52158/jacost.v4i2.539

Ripanti, E. F., & Oramahi, H. A. (2021). Rancangan Sistem Informasi Pengelolaan Audit Mutu Internal (AMI) Perguruan Tinggi. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, *7*(1), 93–99. https://doi.org/10.26418/jp.v7i1.44330

Widiantoro, S., & Yodi, Y. (2020). Rancang Bangun Sistem Informasi Audit Mutu Internal Berbasis IAPS 4.0. Jurnal Ilmu Komputer Dan Bisnis, 11(2), 2446-2454. https://doi.org/https://doi.org/10.47927/jikb.v11i2.6

Universitas Teknologi Mataram. (2022). *Kebijakan Mutu* (A. Yani (ed.); 03 ed.). Universitas Teknologi Mataram.

Yani, A., Zenuddin, Z., Hambali, H., Muslim, R., & Imran, B. (2022). Decision Support System Of Rewarding On Lecturer Performance Using Fuzzy Tsukamoto Method Case Study At Mataram University Of Technology. *Jurnal Pilar Nusa Mandiri*, *18*(2), 131–138. https://doi.org/10.33480/pilar.v18i2.3548

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4937

# CLUSTERING OF POPULAR SPOTIFY SONGS IN 2023 USING K-MEANS METHOD AND SILHOUETTE COEFFICIENT

**Nur Rohman[1*]; Arief Wibowo[2]**

Magister Ilmu Komputer[1], Sistem Informasi[2]
Universitas Budi Luhur, Jakarta, Indonesia[1,2]
https://www.budiluhur.ac.id/[1,2]
2211601667@student.budiluhur.ac.id[1*], arief.wibowo@budiluhur.ac.id[2]
(*) Corresponding Author



**Abstract**—*The rapid advancement of technology and globalization in this era has brought about comprehensive and easily accessible music streaming services, one of which is Spotify. According to Kompas.com, Spotify has experienced a rise in subscribers up to 130 million, as a platform that offers various features besides music streaming. Spotify also provides a better user experience and has the ability to compete with other music streaming platforms. The mission of this research is to classify popular Spotify song data in 2023, which can aid in a deeper understanding of listener preferences or music trends. Based on the test results, there were 2 clusters obtained with cluster 0 containing 863 data and cluster 1 containing 90 data. From the testing results conducted in the K-Means analysis, a Silhouette Coefficient of 0.81 was obtained, which falls into the category of Strong Structure. From these results, it can be suggested that cluster formation was done very well to provide more personalized and relevant music recommendations to Spotify platform users. By understanding the preferences and patterns of listeners revealed through clustering, streaming services can enhance user experience by providing more tailored content.*

***Keywords:*** *clustering, data mining, k-means, silhouette coefficient, spotify.*

**Abstrak**—*Pesatnya perkembangan teknologi dan globalisasi pada era ini menghadirkan layanan streaming music yang lengkap dan mudah diakses, salah sastunya adalah Spotify. Dilansir dari Kompas.com spotify mengalami kenaikan pelanggan hingga 130juta pelanggan, sebagai platform yang mempunyai fitur berbagai macam selain music streaming, Spotify juga memberikan pengalaman pengguna yang lebih baik dan memiliki kemampuan untuk bersaing dengan platform streaming musik lainnya. Misi dari penelitian ini adalah untuk menggolongkan data lagu-lagu pupuler Spotify pada tahun 2023 yang dapat membantu dalam pemahaman lebih lanjut tentang preferensi pendengar atau tren musik. Berdasarkan hasil pengujian yang dilakukan, cluster yang didapatkan sebanyak 2 cluster dengan cluster 0 berisi 863 data sedangkan cluster 1 berisi 90 data. Dari hasil pengujian yang dilakukan dalam analisis K-Means, diperoleh Silhouette Coefficient sebesar 0,81 yang masuk dalam kategori Struktur Kuat, Dari hasil ini, dapat disarankan bahwa pembentukan cluster dilakukan dengan sangat baik untuk memberikan rekomendasi musik yang lebih personal dan relevan kepada pengguna platform Spotify. Dengan memahami preferensi dan pola pendengar yang terungkap melalui pengelompokan, layanan streaming dapat meningkatkan pengalaman pengguna dengan memberikan konten yang lebih disesuaikan.*

***Kata Kunci****: klasterisasi, data mining, k-means, silhouette coefficient, spotify.*

## INTRODUCTION

The rapid development of technology and globalization has brought about increasingly comprehensive and accessible music streaming services across various platforms. One of them is Spotify. Spotify offers various services such as digitally listening to music using an internet connection. With Spotify, we can easily enjoy music by connecting to the digital internet, listening to music anywhere and anytime, and of course, using portable devices. The activity of listening to music seems to have become a habit that is difficult to ignore or forget (Navisa, Hakim, & Nabilah, 2021). Spotify is a large platform with a substantial user base. Certainly, analysis is needed to enhance and strengthen competitiveness with other platforms (Privandhani, 2022). In this research process, the researcher utilized public data by conducting clustering of popular songs in 2023 based on the artist and frequently played tracks. This study was conducted because users tend to listen to music based on their favorite artists.

Clustering method, is an unsupervised approach where the nature of each cluster is not predetermined. This process is based on the similarity of attributes within a group. (Ramadhani et al., 2022). According by (Nisa & Yustanti, 2021) Clustering or clustering is a data analysis method that aims to form data groups (clusters) based on similarity of characteristics among the members of each group. The main goal of clustering is to produce groups that have high similarity within the group and, at the same time, have significant differences between the groups. This process can help understand patterns that may exist in the data and facilitate further analysis.

In this research, the researcher utilizes the K-Means algorithm to process data, and this process is carried out using Google Colab and the Python programming language. According by (Aji et al., 2023) The K-Means algorithm is one of the clustering algorithms. The K-Means algorithm is used to cluster data, observations, or cases based on similarities in what is being studied. K-means clustering is a non-hierarchical cluster analysis technique that attempts to divide existing objects into one or more clusters or groups based on their properties. In the K-Means method, accuracy towards object sizes is very high; this algorithm is relatively concise and efficient when handling a large number of objects.

In the previous research conducted by (Hasyim & Muafi, 2022) in determining the promotion strategy for a program implemented with the K-Means clustering technique, the most effective promotion strategy to target new family planning participants is to focus on the most popular family planning programs It is concluded that the solution is to have a certified BPPKB team with associated costs. Additionally, it might also be appropriate to implement field-based promotions for new family planning participants by adapting the use of the promotional mix. In the research (Wahyudi et al., 2023), utilizing the K-Means method and Davies-Bouldin Index, three attributes and 793 data formed three clusters. Further validation was conducted using the Davies-Bouldin Index (DBI), which resulted in a DBI value of 0.679 for the first cluster, 0.816 for the second cluster, and 0.837 for the third cluster. The lower the DBI value, the higher the quality. Therefore, from the DBI values obtained, it can be concluded that the first cluster with 128 products and a DBI value of 0.679 falls into the category of very popular. In the previous research, the K-Means method was also employed to classify the top 50 Spotify songs from 2010 to 2019 (Musyarofah et al., 2022). In another study conducted by (Triyandana et al., 2022), he K-Means method and DBI (Davies-Bouldin Index) method were used. The case study involved clustering food and beverage menus based on sales levels, resulting in 3 clusters with a model accuracy value of -0.457. The study conducted by (Ramadhani et al., 2022) involves categorizing information to identify several disaster-prone areas in the Purbalingga region. Based on data analysis, there are five groups of disaster-prone areas in Purbalingga Regency with different risk levels: very high, high, moderate, low, and very low. The classification of these areas can serve as a proactive measure against potential disasters, enabling effective ongoing prevention efforts to minimize the impact of disasters on the community. In this research, the analysis results were obtained from the circular cluster diagram, bar graph cluster, and coordinate points related to the valence of songs.

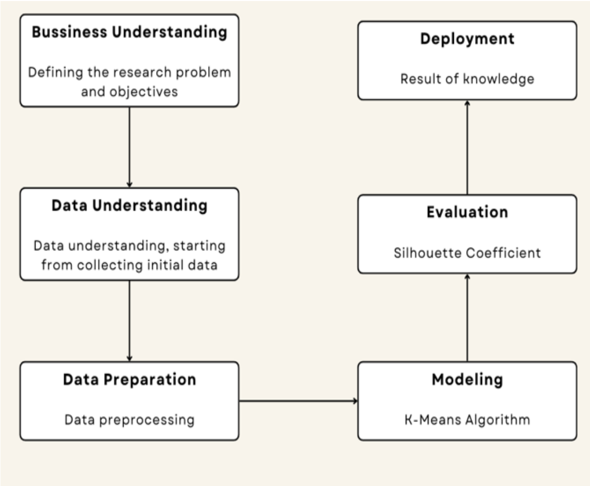
This research aims to determine the optimal number of clusters for grouping songs on the Spotify application in 2023 using the silhouette coefficient as an evaluation method. Additionally, this research will conduct the grouping of popular songs in the Spotify application in 2023 to help further understand listener preferences or music trends, with clusters of popular and unpopular songs. By analyzing groups of popular songs, we can identify emerging music trends and patterns. This information can be used by music producers, record labels, and DJs to make better decisions regarding marketing strategies and music production. Furthermore, based on the results of this research, it is hoped that stakeholders in the music industry can make better decisions regarding music development, promotion, and distribution. They can use the information they obtain to optimize their music catalogs and maximize the impact of their promotions. Based on the good accuracy values, the method used employs clustering techniques with the K-Means method. The data processing steps carried out in this research are conducted using the Python programming language.

## MATERIALS AND METHODS

In this research, the researcher employs the data mining technique known as CRISP-DM (Cross-Industry Standard Process for Data Mining). This is a technique that utilizes a data mining process model commonly used by researchers to solve problems. The research process follows the six

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4937

phases of CRISP-DM, as described by (Fransiska et al., 2022), as shown in Figure 1.




Source: (Fransiska et al., 2022)
Figure 1. Schema Research CRISP-DM

**Business Understanding**

Business Understanding is the first level of CRISP-DM and is a crucial component. In this stage, the researcher defines the problem of the data mining object and determines the research goals (Fransiska et al., 2022). Business understanding can also be interpreted as a comprehensive understanding of the research scope to meet the overall project and business needs or specific objectives and goals. According by (Dhewayani et al., 2022) Business Understanding is a process such as setting business objectives, understanding the situation and conditions being investigated, and establishing research goals with the aim of solving problems through data mining.

**Data Understanding**

Data Understanding is the initial stage in understanding data, which involves studying and describing the data, identifying constraints related to data quality, and searching for the required data as initial hypotheses. According by (Dhewayani et al., 2022) Data understanding is the initial step where you consider the data you will use, gather initial data, and assess the quality of your data. When understanding the data, each feature is analyzed through a descriptive process(Nasari & Am, 2023).

**Data Preparation**

In this stage, data preparation includes selecting, integrating, and cleaning the data to be used. For example, selecting tables, records, and attributes that have been collected to perform data grouping and selection into pre-determined clusters, as well as processing the data for use in the modeling stage. According by (Dhewayani et al., 2022) Data preparation is a process that occurs after data collection.

In this stage, data goes through a series of processes, including identification, data selection, data cleaning, and data transformation. The activities performed in the Data Preparation stage include Data Selection, Data Cleaning, and Data Transformation, according to (Fahmi et al., 2021).

A. Data Selection

   Data Selection is the stage where data is chosen, and attributes are selected based on the goals of data mining.

B. Data Preprocessing

   This process involves cleaning the data by handling outliers, noisy data, and missing values. The goal of this phase is to ensure data quality.

C. Data Transformation

   This process involves grouping attributes into new data, followed by data integration and transformation according to its purpose, then processed in the data mining process.

**Modelling**

This process begins with the application of modeling techniques and data mining algorithms suitable for the research. K-Means is used in the modeling stage. K-Means is a relatively simple and fast non-hierarchical clustering technique.

In the K-Means method, K-Means divides the existing data into one or more groups (clusters) and places data with similar characteristics in the same cluster, while data with different characteristics are placed in different clusters. (Pratama et al., 2022).

**Evaluation**

In this step, the focus is on carrying out a quality modeling stage. This step is also conducted to assess the effectiveness of the model used and whether it aligns with the existing K-Means standards (Fransiska et al., 2022). In the Evaluation stage, the researcher uses the Silhouette Coefficient method, which serves to measure the accuracy and quality of the obtained clusters in the modeling stage.

The Silhouette Coefficient is useful for assisting in choosing the optimal number of clusters in the K-Means algorithm. The Silhouette Coefficient values range from -1 to 1, with the guideline that if the positive value approaches 1, it can be concluded that objects in the data are in suitable clusters and have good distance from other clusters. Conversely, if the value is -1, the data objects have been placed in less suitable clusters. The data in Table 1 reflects the accuracy level in measuring the Silhouette Coefficient.

Table 1. The Standard Value of Silhouette Coefficient

| Silhouette Coefficient | Standard value |
|---|---|
| 0.7 < Silhouette Coefficient ≤ 1.0 | Strong Structure |
| 0.5 < Silhouette Coefficient ≤ 0.7 | Medium Structure |
| 0.25 < Silhouette ≤ 0.5 | Weak Structure |
| Silhouette ≤ 0.25 | No Structure |

Source: (Fransiska et al., 2022)

This evaluation stage assesses whether the applied modeling is appropriate and suitable for this research case and whether it achieves the desired goals.

The results of the evaluation are used to decide whether to proceed to the next steps or to restart if the goals are not achieved.

***Deployment***

The final stage in the CRISP-DM method is the Deployment stage, which is conducted to generate knowledge or information that can be presented in the form of creating applications or simple reports.

## RESULT AND DISCUSSION

Based on the results of the clustering research using the CRISP-DM (Cross-Industry Process for Data Mining) method. The CRISP-DM Data Mining research method is a combination of qualitative and quantitative methods to describe information from the research subjects and then provide prescriptive recommendations (Asyraf & Prasetya, 2023). And using the Python programming language, supported by Google Colab tools, and covering the discussions as follows:

**Bussiness Understanding**

Spotify is one of the music streaming services that can be listened to and played anywhere. Currently, Spotify has approximately 130 million users who use it as a music streaming service. However, this streaming service faces numerous competitors, and there is a likelihood that the user base of around 130 million could decline if the service is not improved. Spotify can conduct an analysis based on popular songs or music by utilizing data mining.

**Data Understanding**

The first phase of data understanding involves recognizing and comprehending the data we possess, as well as analyzing it to discover potential information and actions that can be taken from the data. The dataset used in this research was obtained from the website Kaggle.com. This dataset consists of 953 entries collected in the year 2023 and is available in CSV (comma-separated values) format.

This dataset contains 24 relevant attributes used in the research. Refer to Figure 2 for details.

```
spotify.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 953 entries, 0 to 952
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   track_name            953 non-null    object
 1   artist_name           953 non-null    object
 2   artist_count          953 non-null    int64
 3   released_year         953 non-null    int64
 4   released_month        953 non-null    int64
 5   released_day          953 non-null    int64
 6   in_spotify_playlists  953 non-null    int64
 7   in_spotify_charts     953 non-null    int64
 8   streams               953 non-null    object
 9   in_apple_playlists    953 non-null    int64
 10  in_apple_charts       953 non-null    int64
 11  in_deezer_playlists   953 non-null    object
 12  in_deezer_charts      953 non-null    int64
 13  in_shazam_charts      903 non-null    object
 14  bpm                   953 non-null    int64
 15  key                   858 non-null    object
 16  mode                  953 non-null    object
 17  danceability_%        953 non-null    int64
 18  valence_%             953 non-null    int64
 19  energy_%              953 non-null    int64
 20  acousticness_%        953 non-null    int64
 21  instrumentalness_%    953 non-null    int64
 22  liveness_%            953 non-null    int64
 23  speechiness_%         953 non-null    int64
dtypes: int64(17), object(7)
memory usage: 178.8+ KB
```

Source: (Research Results, 2024)

Figure 2. Information Data Spotify

**Data Preparation**

In this phase, the researcher carries out data processing from the Knowledge Discovery in Database (KDD) phase, such as data cleaning, data integration, data selection, and data transformation. In this stage, the 953 data from the Kaggle.com website will be processed through several steps before entering the clustering phase. The steps are as follows:

A. The first stage of data preparation is to determine which data will be processed from the obtained dataset. Not all data will be processed; columns that will not be used in the Modeling stage are removed, as shown in Figure 3.

```
# Droping columns that aren't require
df = data.drop(columns=['track_name', 'artist_name', 'released_month', 'released_day','released_year','in_shazam_charts','key'])
df.any()

artist_count          True
in_spotify_playlists  True
in_spotify_charts     True
streams               True
in_apple_playlists    True
in_apple_charts       True
in_deezer_playlists   True
in_deezer_charts      True
bpm                   True
mode                  True
danceability_%        True
valence_%             True
energy_%              True
acousticness_%        True
instrumentalness_%    True
liveness_%            True
speechiness_%         True
dtype: bool
```

Source: (Research Results, 2024)

Figure 3. Information Attribute Data

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4937

B. The second stage involves processing data with missing values by removing empty values in attributes where such gaps exist. However, based on Figure 4, it is apparent that the data to be used does not have any missing or null values. Therefore, no values are deleted in this process.

```
[5] df.isnull().sum()

    artist_count            0
    in_spotify_playlists    0
    in_spotify_charts       0
    streams                 0
    in_apple_playlists      0
    in_apple_charts         0
    in_deezer_playlists     0
    in_deezer_charts        0
    bpm                     0
    mode                    0
    danceability_%          0
    valence_%               0
    energy_%                0
    acousticness_%          0
    instrumentalness_%      0
    liveness_%              0
    speechiness_%           0
    dtype: int64
```

Source: (Research Results, 2024)
Figure 4. Finding Empty Or Null Values

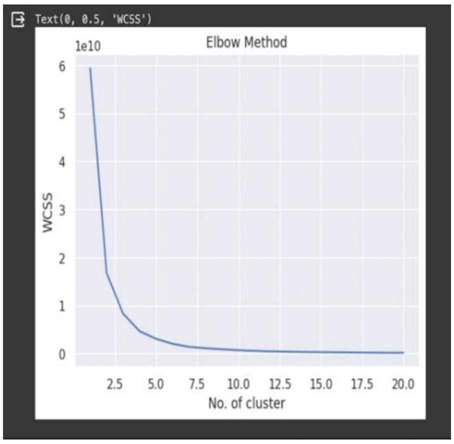
C. The third stage involves data transformation. In this stage, the researcher performs data transformation through data normalization or Min-Max scaling. Figure 5 shows the results of the normalization process.

```
[ ] # Normalization of data
    from sklearn.preprocessing import MinMaxScaler
    datatypes = ['int16', 'int32', 'int64', 'float16', 'float32', 'float64']
    normarization = data.select_dtypes(include=datatypes)
    for col in normarization.columns:
        MinMaxScaler(col)

[ ] x=normarization.values
    print(normarization)
    # x

        artist_count  released_year  released_month  released_day  \
    0              2           2023               7            14
    1              1           2023               3            23
    2              1           2023               6            30
    3              1           2019               8            23
    4              1           2023               5            18
    ..           ...            ...             ...           ...
    948            1           2022              11             3
    949            1           2022              10            21
    950            2           2022              11             3
    951            3           2022              10            20
    952            1           2022              11             4

        in_spotify_playlists  in_spotify_charts  in_apple_playlists  \
    0                    553                147                  43
    1                   1474                 48                  48
    2                   1397                113                  94
    3                   7858                100                 116
    4                   3133                 50                  84
```

Source: (Research Results, 2024)
Figure 5. Information Normalization Data

## Modelling

The next step is the modeling stage, where the researcher determines the number of clusters to be used in the data clustering process using the K-Means method. The researcher uses the Elbow method to determine the number of clusters, and the results can be seen in Figure 6.




Source: (Research Results, 2024)
Figure 6. Grapich Cluster Elbow Method

From the results of the Elbow method documented in Figure 6, it can be concluded that from the processed and analyzed data, 2 clusters are optimal. In this method, the optimal cluster is the point forming the elbow.

## Evaluation

In this stage, the researcher employs an approach with the Silhouette Coefficient technique, which is used to measure or test the quality of the previously obtained clusters in the Modeling stage. The researcher measures and tests 2-5 clusters. Table 2 shows the accuracy results for each cluster.

Table 2. Results of the Silhouette Coefficient Values

| No | Cluster | Accuration Silhouette Coefficient |
|---|---|---|
| 1 | 2 | 0.8113997870257713 |
| 2 | 3 | 0.7431462113579907 |
| 3 | 4 | 0.6924926306048924 |
| 4 | 5 | 0.6701590066697268 |

Source: (Research Results, 2024)

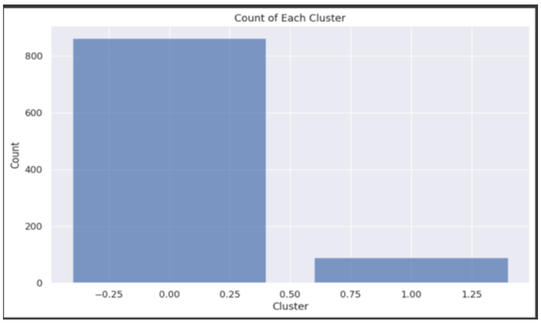
From Table 2, it can be concluded that cluster 2 from the K-Means process obtains the highest silhouette value compared to clusters 3, 4, or 5, which is 0.81. Based on Table 1, cluster 2 falls into the Strong Structure criteria. Silhouette Coefficient results greater than 0.7 indicate that the quality assessment of the Spotify popular song clustering in 2023 using the K-Means algorithm shows excellent quality for cluster 2.

## Deployment

The deployment stage is carried out to generate information or knowledge related to the research that has been conducted, based on the results of the previous stages. Based on the K-Means clustering results, out of 953 Spotify songs data in the year 2023, cluster 0 contains 863 data, while cluster 1 contains 90 data as shown in Figure 8 for cluster visualization. It can be analyzed that cluster 0 has a better distribution compared to cluster 1. Validation testing was then conducted using the Silhouette Coefficient method, and the testing resulted in a value of 0.81, as seen in Figure 7.

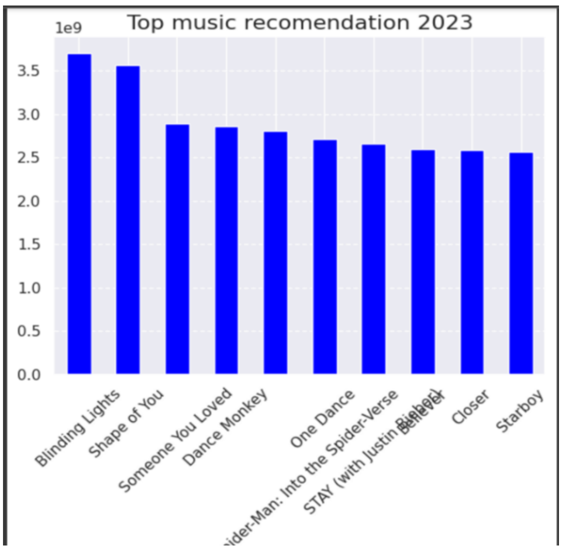
```
[ ] from sklearn.metrics import silhouette_score

    silhouette_score(x, features)

    0.8113997870257713
```

Source: (Research Results, 2024)
Figure 7. Display of the Cluster Model Results




Source: (Research Results, 2024)
Figure 8. Cluster Visualization Display

The bar graph in Figure 9 shows that the songs "Blinding Lights" by The Weeknd and "Shape of You" by Ed Sheeran dominated the index of the most played songs in 2023. Meanwhile, the songs "Dance Monkey" by The Weeknd and "One Dance" by Drake, WizKid, Kyla also ranked in the top five positions. Additionally, "STAY" (with Justin Bieber), "Believer" by Imagine Dragons, "Closer" by The Chainsmokers, Halsey, and "Starboy" by The Weeknd, Daft Punk were in the last positions, making them among the songs that were difficult to play in 2023.




Source: (Research Results, 2024)
Figure 9. Chart of Popular Songs

## CONCLUSION

In this study, the researcher successfully applied the K-means clustering method and the CRISP-DM method to test the data. This process involves six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. It then classified Spotify's popular songs data for the year 2023 using Google Colab as the development environment and Python as the programming language. The study used 17 attributes from the dataset acquired from Kaggle.com. Based on the Elbow method analysis, the optimal number of clusters (K) found was two. Cluster formation with the K-Means method resulted in cluster 0 with 863 popular songs and cluster 1 with 90 less popular songs. The evaluation of clustering quality with the Silhouette Coefficient method produced a value of 0.81, which approaches the maximum value of 1. This indicates that the clustering performed is of very high quality. The conclusion of this study shows the effectiveness of the K-Means method in classifying popular songs on Spotify and its potential to provide more targeted recommendations to users. The suggestions for this research have significant potential, such as utilizing the latest Spotify data, clustering based on genres or sub-genres, and implementing Spotify data using the K-Medoids algorithm. This is because the K-Medoids algorithm is suitable for addressing the weaknesses of K-Means. K-Medoids is sensitive to noise and outliers, making it a better choice for certain datasets. Therefore, incorporating these suggestions could enhance the robustness and accuracy of the research findings.

## REFERENCE

Aji, N. S., Natsir, F., & Istianah, S. (2023). *Penentuan Penjualan Barang Berdasarkan Pengelompokan Produk dengan K-Means Clustering Metode CRISP-DM*.

Asyraf, H., & Prasetya, E. (2023). Implementasi Metode CRISP DM dan Algoritma Decision Tree Untuk Strategi Produksi Kerajinan Tangan pada UMKM A. *Jurnal Media Informatika Budidarma*, *8*, 94–105. https://doi.org/10.30865/mib.v8i1.7050

Dhewayani, F. N., Amelia, D., Alifah, D. N., Sari, B. N., & Jajuli, M. (2022). Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM. *Jurnal Teknologi Dan Informasi*, *12*. https://doi.org/10.34010/jati.v12i1

Fahmi, R. N., Jajuli, M., & Sulistiyowati, N. (2021). Analisa Pemetaan Tingkat Kriminalitas Di Kabupaten Karawang Menggunakan

Algoritma K-Means. *Journal of Information Technology and Computer Science (INTECOMS)*, *4*(1). Retrieved from www.pasundanekspres.co

Fransiska, N., Anggraeni, D., & Enri, U. (2022). Pengelompokkan Data Kemiskinan Provinsi Jawa Barat Menggunakan Algoritma K-Means dengan Silhouette Coefficient. *Jurnal Teknologi Informasi Komunikasi*, *9*, 29–35. https://doi.org/10.38204/tematik.v9i1.921

Hasyim, F., & Muafi. (2022). *Implementasi Data Mining Dalam Menentukan Strategi Promosi Program KB Menggunakan Algoritma K-Means Clustering*. *3*(1). Retrieved from https://ejournal.unuja.ac.id/index.php/core

Musyarofah, U. L., Alima, S. N., & Kartika, D. S. Y. (2022, September). KLASIFIKASI TOP 50 SPOTIFY TAHUN 2010-2019 MENGGUNAKAN METODE K-MEANS CLUSTERING. In Prosiding Seminar Nasional Teknologi dan Sistem Informasi (Vol. 2, No. 1, pp. 215-220).

Nasari, F., & Am, A. N. (2023). Implementasi K-Medoids Clustering Dalam Pengelompokkan Harga 8 Jenis Minyak Goreng. *SINTECH JOURNAL*, *6*. Retrieved from https://doi.org/10.31598

Navisa, S., Hakim, L., & Nabilah, A. (2021). Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM. In *Jurnal Sistem Cerdas*.

Nisa, C., & Yustanti, W. (2021). Studi Perbandingan Algoritma Klastering Dalam Pengelompokan Persediaan Produk (Studi Kasus : Subdirektorat Perencanaan Sarana Prasarana Dan Logistik PTN X). *JEISBI*, *02*.

Privandhani, N. A. (2022). Clustering Pop Songs Based On Spotify Data Using K-Means And K-Medoids Algorithm. Jurnal Mantik, 6(2), 1542-1550.

Ramadhani, D. I., Damayanti, O., Thaushiyah, O., & Kadafi, A. R. (2022). Penerapan Metode K-Means Untuk Clustering Desa Rawan Bencana Berdasarkan Data Kejadian Terjadinya Bencana Alam. *JURIKOM (Jurnal Riset Komputer)*, *9*(3), 749. https://doi.org/10.30865/jurikom.v9i3.4326

Pratama, E. F. A., Khairil, K., & Jumadi, J. (2022). Implementasi Metode K-Means Clustering Pada Segmentasi Citra Digital. Jurnal Media Infotama, 18(2), 291-301.

Triyandana, G., Putri, L. A., & Umaidah, Y. (2022). Penerapan Data Mining Pengelompokan Menu Makanan dan Minuman Berdasarkan Tingkat Penjualan Menggunakan Metode K-Means. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 6). Retrieved from http://jurnal.polibatam.ac.id/index.php/JAIC

Wahyudi, T., Sa'adah, N., & Puspitasari, D. (2023). Penerapan Metode K-Means Pada Data Penjualan Untuk Mendapatkan Produk Terlaris di PT. Titian Nusantara Boga. *Jurnal Sains Dan Teknologi*, *5*(1), 228–236. https://doi.org/10.55338/saintek.v5i1.1379

# ABILITY CONVOLUTIONAL FEATURE EXTRACTION FOR CHILI LEAF DISEASE USING SUPPORT VECTOR MACHINE CLASSIFICATION

**Rizal Amegia Saputra[1*]; Toto Haryanto[2]; Sri Wasyianti[3]**

Sistem Informasi Akuntansi Kampus Kota Sukabumi[1]
Universitas Bina Sarana Informatika, Sukabumi, Indonesia[1]
https://www.bsi.ac.id/[1]

Departemen Ilmu Komputer[2]
IPB University, Bogor, Indonesia[2]
https://www.ipb.ac.id/[2]

Sistem Informasi Akuntansi Kampus Kota Bogor[3]
Universitas Bina Sarana Informatika, Bogor, Indonesia[3]
https://www.bsi.ac.id/[3]

rizal.rga@bsi.ac.id[1*], totoharyanto@apps.ipb.ac.id[2], sri.siw@bsi.ac.id[3]
(*) Corresponding Author



***Abstract—****Chili plants are among the most commonly used food ingredients in various dishes in Indonesia. Leaves on chili plants are often affected by disease; if the disease is not treated immediately, it can damage the plant and cause crop failure. Early detection of chili plant diseases is important to reduce the risk of crop failure. The development of technology and the application of machine-learning algorithms can automatically monitor chili plants using a computer system. Using this algorithm, the system analyzes and identifies diseases that a camera can observe and record. In this study, the proposed method for feature extraction uses a convolutional neural network (CNN) algorithm with transfer learning using VGG19. For classification using SVM for training data, accuracy generated 95%, precision 95%, recall 95%, and F1-Score 95%, and testing data accuracy generated 90%, precision 89%, recall 90%, and F1-Score 89%, proving that the convolutional process with architecture VGG19 and SVM algorithm is acceptable for classification. In future research, other architectures or extraction fusions can be used to maximize the results.*

***Keywords:*** *leaf chilli diseases, SVM, transfer learning.*

**Abstrak**—*Tanaman cabai merupakan salah satu bahan makanan yang paling sering digunakan dalam berbagai masakan di Indonesia. Daun pada tanaman cabai yang sering terkena penyakit, jika penyakitnya tidak segera ditangani, maka penyakit tersebut dapat merusak tanaman dan mengakibatkan gagal panen, deteksi penyakit tanaman cabai secara dini sangat penting dilakukan, untuk mengurangi resiko gagal panen. Perkembangan teknologi dan penerapan algoritma machine learning dapat melakukan pengawasan terhadap tanaman cabai secara otomatis menggunakan sistem komputer. Dengan menggunakan algoritma ini penyakit yang dapat dilihat dan direkam oleh kamera akan dapat di analisis dan diidentifikasi kan oleh sistem. Pada penelitian ini metode yang diusulkan untuk ekstraksi fitur menggunakan konvolusi dari algoritma CNN dengan transfer learning menggunakan VGG19 dan untuk klasifikasi menggunakan SVM untuk data training Accuracy yang dihasilkan 95%, Precision 95%, Recall 95% dan F1-Score 95%, dan data testing Accuracy yang dihasilkan 90%, Precision 89%, Recall 90% dan F1-Score 89%, hal ini membuktikan bahwa proses convolutional dengan arsitektur VGG19 dan algoritma SVM sangat baik untuk proses klasifikasi.* Untuk penelitian lanjutan dapat menggunakan arsitektur lainnya atau menggunakan ekstraksi gabungan agar lebih maksimal.

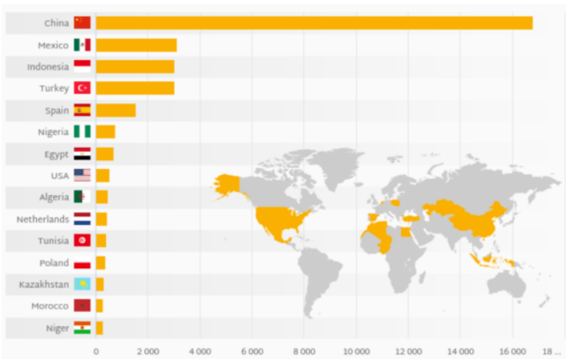
***Kata Kunci****: penyakit daun cabai, SVM transfer learning.*

## INTRODUCTION

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4961

Indonesia is among the ten largest producers of chili plants in the world; nationally, chili plants are the most produced crop (Rahman et al., 2023). The need for chili in the Indonesian region continues to increase every year, in 2022 as much as 636.56 thousand tons based on horticultural statistical data, an increase of 6.78% compared to 2021 (Santika, 2023).

The chili plant is not a staple food crop but a complementary seasoning for Indonesian cuisine, with prices always fluctuating, making chili peppers a contributor to inflation in the Indonesian economy (Rosalina & Wijaya, 2020). To maintain price stability, production must be increased, and the quality of chili plants maintained (Barus et al., 2022).



Source: (Helgi Library, 2023)
Figure 1. World's Largest Chili Producing Countries in 2021

Harvest failure is one of the causes of price instability in chili crops (Polii et al., 2019), and several factors, such as pests and diseases, cause crop failure in chili plants (Firmansyah et al.,2020). Pests and diseases pose a serious threat to farmers as they can lead to a decrease in the quality and quantity of crops produced (Islam et al., 2020), so the worst impact that farmers will experience is a big loss.

Chili plants have diseases that often attack and are difficult to control when infected with leaf curl and yellow viruses (Renfiyeni et al., 2023). The disease can be visually identified by the symptoms that appear; however, visual identification has similarities, so errors can occur. This is because each person has a different assessment of the visual identification results (Rozlan & Hanafi, 2022). To solve this problem visually, computer technology that can recognize digital images has been rapidly developed (Susim & Darujati, 2021). The analysis of a digital image is a regular or random pattern (Stoilov et al., 2012), and features are the most important part of the analysis of an image. The analysis results can provide information about the structure of the surface, changes in intensity, or brightness of the color (Juandri & Anwar, 2023). features are the most important parts in the analysis of an image, where the analysis results can provide information about the structure of the surface, changes in intensity, or brightness of the color (Muzahid et al., 2020).

CNN have many architectures, one of which is VGG-19. VGG-19 is an architecture consisting of 19 layers: 16 convolutional layers, five max pooling layers, three fully connected layers, and one SoftMax layer (Mascarenhas & Agarwal, 2021). The input image size of this architecture is 224 × 224, and this architecture has been used to train more than 1 million images obtained from the ImageNet database. In addition, this architecture has a 3x3 kernel and has 5 blocks with various sizes of convolutional layers in each block, which then adds a max pooling layer as a separator for each block (Marcella et al., 2022).

Machine Learning (ML) is a type of artificial intelligence (AI) that is widely used in various fields (Suradiradja, 2021), especially in agriculture. Many studies have used these methods for automatic decision-making with computing systems (Yana & Nafi'iyah, 2021). The ML method has many models, one of which is often used and has very good accuracy in solving classification problems, namely the Support Vector Machine algorithm (Abdullah & Abdulazeez, 2021). SVMs have advantages over other ML models, such as the problem of overfitting (Nusinovici et al., 2020), the ability to work well on relatively small datasets (Rahayu et al., 2022), and the ability to overcome the problem of unbalanced or unevenly distributed data in the class (Wahab et al., 2019).

Previous research on chili leaf disease, namely: Research Araujo et al., 2019. Identification of chili leaf disease with CNN and YoloV2 algorithms. The research results were 61.49% (Das Chagas Silva Araujo et al., 2021). Wahab et al research, 2019. Detecting chili leaf disease with the K-means algorithm for segmentation and SVM for classification. The results of the study obtained an accuracy of 57.1% (Wahab et al., 2019). Karuna DKK research, 2019. Using the CNN algorithm with several architectures, the result is that the application built can classify chili leaf disease (Karuna et al., 2019). Research Windarningsih 2019, identification of viruses causing yellow leaf curl disease in chili using PCR-RFLP The results obtained by PCR can detect the disease (Windarningsih, 2019). Nuanmeesri and Sriurai 2021, using the Multi-Layer Perceptron Neural Network (MLPNN) algorithm by comparing Feature Selection Filters (IG, GR) and Wrapper Feature Selection. The results obtained by the MLPNN and Wrapper have the highest accuracy of 98.91% (Nuanmeesri & Sriurai, 2021). A study by Zikra et al.

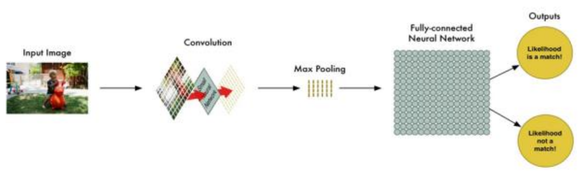
2021, using 4 angle type GLCM texture extraction and SVM Method. Their results obtained an accuracy of 95% (Zikra et al., 2021). Research by Patil and Lad 2021, using SVM and KNN algorithms using 4 GLCM texture extraction angles. The results showed that KNN had greater accuracy, namely 93% compared to the SVM algorithm of 83.33% (Patil & Lad, 2021). Dzaky, 2021. The CNN algorithm was applied using AlexNet architecture. Their results obtained an accuracy of 90%. Rozlan and Hanafi 2022, using the Deep Learning algorithm (VGG16, InceptionV3, and EfficientNetB0). The results showed that InceptionV3 had the highest accuracy of 98.83% (Rozlan & Hanafi, 2022). Kelikualiq et al. 2022. Using the CNN algorithm with AlexNet architecture on the chili plant health monitoring system based on the IOT. The results of this study obtained an accuracy of 63% (Kelikualiq et al., 2022). Research Anggraeni, Widayana, Rahayu, & Rozikin, 2022, the CNN algorithm was used for the classification of the three types of chili leaf disease. Their research obtained an accuracy of 60% (Anggraeni et al, 2022). Hafidhoh Research 2022, using GLCM feature extraction with pixel distance d = 1 to 5 and the SVM algorithm with Gaussian and polynomial kernel functions. The results showed an accuracy of 83% for the polynomial kernel (Hafidhoh, 2023).

Based on the explanation of previous research, this research will apply the SVM algorithm for classification such as research (Zikra et al., 2021; Patil & Lad, 2019; Wahab et al., 2019; Hafidhoh, 2023). However, the difference is that in this study, feature extraction uses the CNN algorithm with VGG19 transfer learning.

## MATERIALS AND METHODS

### 1. CNN

A Convolutional Neural Network (CNN) is a type of deep learning specifically designed to recognize patterns in grid-structured data, such as images. CNN utilizes convolution operations to automatically extract image features, both low and high dimensional images (Muzahid et al., 2020).




Source: (Lina, 2019)
Figure 2. Illustration of CNN Architecture

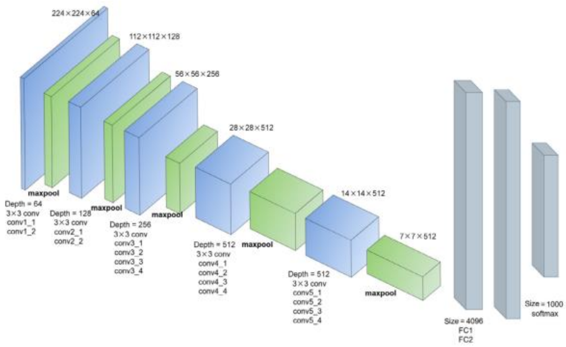
There are several layers in the CNN algorithm, one of which is the Convolutional Layers: this layer is the main operation in the CNN convolution layer. In this context, the 2D convolution between the input (image) and filter (kernel) can be explained in formula (1).

$$(I*K)\ (x,y) = \Sigma_i \Sigma_j I(x+i,y+j)\cdot K(i,j) \ldots\ldots\ldots\ldots\ldots (1)$$

Where $I$ is the input image, $K$ is the kernel, $x$ and $y$ are the positions in the image, and $i$ serta $j$ are the indices in the kernel. The convolution result $(I*K)$ $(x, y)$ is the value at position (x, y) in the convolved feature map.

### 2. VGG19

The VGG19 architecture, which is part of the VGG family of models, is known to have significant advantages in the field of image recognition. The model showed good performance in image recognition.




Source: (Mohbey et al., 2022)
Figure 3. VGG19 Architecture

### 3. SVM

SVM is a machine learning algorithm used for classification and regression tasks (Ibrahim & Abdulazeez, 2021). The goal is to determine the best hyperplane that separates the two data classes as much as possible with the maximum margin.

The SVM equation for multiclass classification problems depends on the approach used. In the One-Versus-Rest (OVR) approach, each class has its own SVM model. In the One-Versus-One (OVO) approach, there is a binary SVM model for each possible class pair (Setiawan et al., 2022). In this study, using the One-Versus-Rest (OVR) approach, if we have N classes in a multi-class problem, then there will be N binary SVM models. Each model separates one class from another. The general equation for prediction with the OVR approach is shown in formula (2).

$$f\ (x) = argmax_i(\omega_i . x + b_i) \ldots\ldots\ldots\ldots\ldots (2)$$

Description: $f\ (x)$ is the predicted class for the test data $x$. $\omega_i$ is the weight vector of the SVM model for class $i$. $b_i$ is the bias constant of the SVM model for class $i$.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4961

Each SVM OVR model will provide a score or value:

$$\omega_i . x + b_i \text{ ......................................... (3)}$$

For test data $x$, and the class with the highest score will be selected as the final prediction.

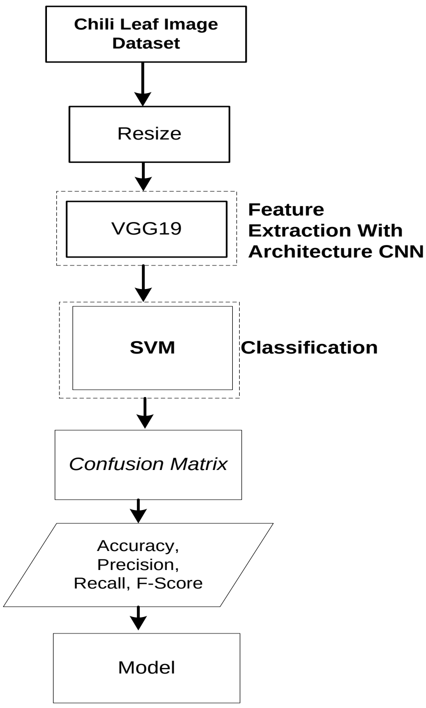
### 4. Confusion Matrix

The performance evaluation in this study used a Confusion Matrix. This technique is a prediction matrix measured based on the values of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) (Liyantoko, Candradewi, & Harjoko, 2019). There are 3 confusion matrices were used in this study.

Akurasi = (TP + TN) / (TP+FP+FN+TN) ............. (3)
Presisi = (TP) / (TP+FP) ...................................... (4)
Recall = (TP) / (TP + FN) ...................................... (5)
F1 Score = 2 * (Recall*Precision) / (Recall + Precision) .............................................................. (6)

### 5. Research method

Figure 4 shows the stages of the research method planned in this research.




Source: (Research Results, 2024)
Figure 4. Research Method

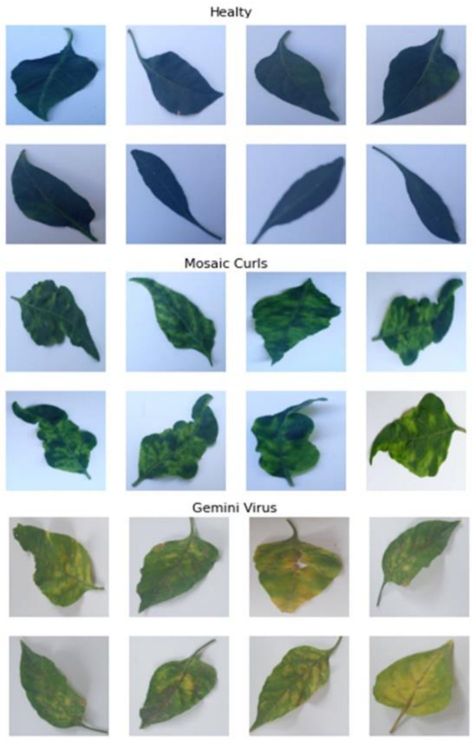
Figure 4 is the flow of the research method that will be planned, the first step is to collect image data on the types of chili leaf diseases from the farmers' gardens, after collecting all the images are resized in the python application automatically, after that step three perform feature learning using CNN architecture, namely VGG19, then the SVM algorithm classification process, then evaluate the results using a confusion matrix with the evaluation compared, namely accuracy, precision, recall and F-Score.

## RESULTS AND DISCUSSION

### 1. Data

Image data were collected for as many as 300 data points, consisting of 100 data points on healthy leaves, 100 data points on mosaic curls, and 100 data points on gemini virus, using a smartphone, and the data were recorded one by one based on the type of disease.

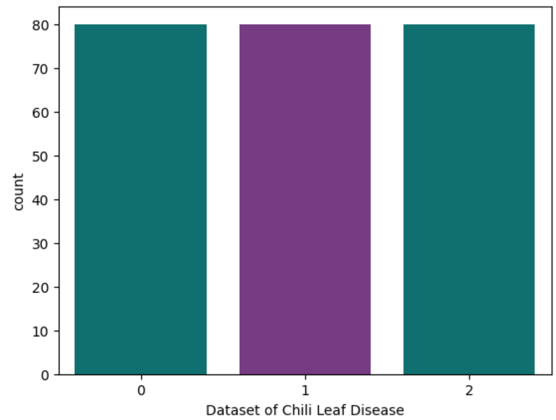
In this study, dara was processed using the Jupyter notebook with Microsoft Visual Studio Code. Figure 5 is a display of the three types of chili leaf data.




Source: (Research Results, 2024)
Figure 5. Chili Leaf Data

Of the three types of data, as much as 300 were collected, and the data were divided into 80% training data and 20% testing data. Then the data label is converted into numeric numbers 0 = Healthy Leaves, 1 = Mosaic Curls and 2 = Gemini Virus. Figure 6 is a display of the three types of chili leaves that have been converted into numeric numbers.

Figure 6 shows an even distribution of data from three types of chili leaves, as much as 80 data from one type of data.




Source: (Research Results, 2024)

Figure 6: Number of images by disease type

2. **Feature Extraction Process**

The next stage is the image resizing process carried out as an input with a size of 244 × 244 × 3. At this convolutional stage, image features are extracted using transfer learning VGG19 using the image set as. Initialize the initial weight and extractor fixed features. Process results for Jupiter notebook applying VGG19 transfer learning.

Table 1. Model "vgg19"

| Layer (type) | Output Shape | Param # |
|---|---|---|
| input_5 (Input Layer) | [(None, 224, 224, 3)] | 0 |
| block1_conv1 (Conv2D) | (None, 224, 224, 64) | 1792 |
| block1_conv2 (Conv2D) | (None, 224, 224, 64) | 36928 |
| block1_pool (MaxPooling2D) | (None, 112, 112, 64) | 0 |
| block2_conv1 (Conv2D) | (None, 112, 112, 128) | 73856 |
| block2_conv2 (Conv2D) | (None, 112, 112, 128) | 147584 |
| block2_pool (MaxPooling2D) | (None, 56, 56, 128) | 0 |
| block3_conv1 (Conv2D) | (None, 56, 56, 256) | 295168 |
| block3_conv2 (Conv2D) | (None, 56, 56, 256) | 590080 |
| block3_conv3 (Conv2D) | (None, 56, 56, 256) | 590080 |
| block3_conv4 (Conv2D) | (None, 56, 56, 256) | 590080 |
| …… | | |
| Total params: 20,024,384 | | |
| Trainable params: 0 | | |
| Non-trainable params: 20,024,384 | | |

Source: (Research Results, 2024)

3. **Process Evaluation**

After the resize process and feature extraction process, then the next classification process uses a machine learning algorithm, namely SVM. Table 2 presents the results of the training and testing processes.

Table 2. Training and Testing Data Results

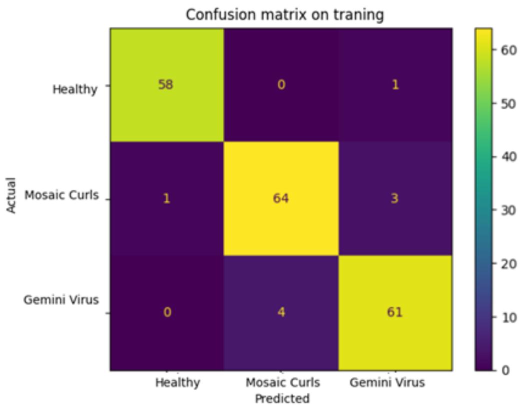
| No | Algorithm CSVM | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| 1 | Training | 95% | 95% | 95% | 95% |
| 2 | Testing | 90% | 89% | 90% | 89% |

Source: (Research Results, 2024)

In Table 2, it can be explained that by applying transfer learning VGG19 and classification using SVM for training data, the resulting accuracy is 95%, precision 95%, recall 95% and F1-Score 95%, and the resulting testing data accuracy is 90%, precision 89%, recall 90% and F1-Score 89%.

4. **Confusion Matrix**

The confusion matrix is a step to evaluate the performance of a classification model. The confusion matrix provides information about the correct and incorrect predictions made by the model compared to the actual label or class. In this confusion matrix, three classes are considered: Healthy Leaf, Mosaic Curl and Gemini Virus. The confusion matrix results of the training process are shown in Figure 7.

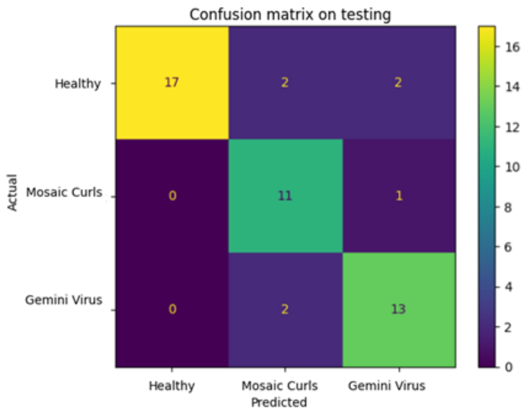



Source: (Research Results, 2024)

Figure 7. Confusion Matrix of Training

Figure 7 shows the Confusion Matrix of the training results, where the vertical axis (y-axis) shows the true class (True Labels), and the horizontal axis (x-axis) shows the predictions made by the model (Predicted Labels).

The classification results show that for 'Healthy Leaf', the model has successfully identified 60 cases correctly, and only one of the 'Healthy Leaf' cases was mistakenly classified as 'Gemini Virus'. In the 'Mosaic Curly' category, one case was misclassified as 'Healthy Leaf' and 64 cases were correctly classified, while the other three cases were misclassified as 'Gemini Virus.' For 'Gemini Virus, the model made the mistake of not recognizing four cases as 'Mosaic Curl, but managed to classify 61 cases correctly. The Confusion Matrix of the training results showed a fairly good performance, with the majority of cases correctly classified in each category. However, there were some misclassifications that occurred, particularly in the recognition of 'Gemini Virus,' which may require further investigation to determine the cause of the errors and improve the accuracy of the model.




Source: (Research Results, 2024)

Figure 8. Confusion Matrix Testing

Figure 8 shows the confusion matrix of the test results of the model identifying the 'Healthy Leaf' class with a total of four errors, although the majority of 'Healthy Leaf' cases 15 out of 21 were correctly classified. In the case of 'Mosaic Curly,' the model performed better, correctly identifying 11 out of 12 cases. Only one 'Mosaic Curly' case was misclassified as 'Gemini Virus. For 'Gemini Virus,' the model accurately classified 13 out of 15 cases, with two cases misclassified as 'Mosaic Curls.'

Overall, the confusion matrix shows that the model has a good ability to classify the three categories, but there is still room for improvement, especially in terms of minimizing misclassification between categories. These errors can come from a variety of factors, such as features that are not sufficiently discriminative and class imbalance in the training data.

## CONCLUSION

This research successfully demonstrated the applicability of convolutional feature extraction techniques using the transfer learning architecture VGG19 for image classification tasks. By combining the features extracted by VGG19 and using SVM as a classifier, the model achieved very high performance on the training data, with Accuracy, Precision, Recall, and F1-Score values all reaching 95%. In the testing data process, there was a decrease in performance, and the results showed a good accuracy of 90%, precision of 89%, recall of 90%, and F1-Score of 89%. This decrease in performance is common, considering that models tend to perform better on training data than on testing data. The use of VGG19 transfer learning for feature extraction with SVM classification proved to be an excellent combination for image classification. Although there were errors, they need to be further analyzed to understand the basis of each misclassification. Future research can apply improvement strategies, such as collecting more training data, applying data augmentation techniques, exploring different architectural models, or using fusion extraction.

## REFERENCE

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification a review. Qubahan Academic Journal, 1(2), 81-90.

Anggraeni, D. S., Widayana, A., Rahayu, P. D., & Rozikin, C. (2022). Metode Algoritma Convolutional Neural Network pada Klasifikasi Penyakit Tanaman Cabai. *STRING (Satuan Tulisan Riset Dan Inovasi Teknologi)*, *7*(1), 73. https://doi.org/10.30998/string.v7i1.13304

Barus, M. D. B., Mustafa, M., & Thahirah, F. S. (2022). Analisis Trend Produksi Dan Harga Komoditas Cabai Untuk Meningkatkan Produktivitas Desa Lau Gumba Kabupaten Karo. *NUSANTARA: Jurnal Ilmu Pengetahuan Sosial*, *9*(2), 527–531.

Das Chagas Silva Araujo, S., Malemath, V. S., & Sundaram, K. M. (2021). Symptom-Based Identification of G-4 Chili Leaf Diseases Based on Rotation Invariant. *Frontiers in Robotics and AI*, *8*(May), 1–8. https://doi.org/10.3389/frobt.2021.650134

Firmansyah, E., Hidayat, C., Husna, R., Yuliani, S., & Utomo, R. (2020). Application of the Bayes Theorem to the Expert System for Diagnosing Big Red Chili Plants. *Proceedings of the 1st International Conference on Islam, Science and Technology, ICONISTECH 2019, 11-12 July 2019, Bandung, Indonesia*. EAI. https://doi.org/10.4108/eai.11-7-2019.2297526

Hafidhoh, N. (2023). Identifikasi Penyakit Daun Tanaman Cabai Merah Dengan Ekstraksi Fitur Dan Klasifikasi Support Vector Machine. *Hasil Penelitian Dan Pengabdian Kepada Masyarakat (SEHATI ABDIMAS) 2022*, *5*(1), 64–74.

Helgi Library. (2023). Which Country Produces the Most Green Chilli and Peppers? Retrieved October 29, 2023, from https://www.helgilibrary.com/ website: https://www.helgilibrary.com/charts/which-country-produces-the-most-green-chilli-and-peppers/

Ibrahim, I., & Abdulazeez, A. (2021). The Role of Machine Learning Algorithms for Diagnosing Diseases. *Journal of Applied Science and Technology Trends*, *2*(01), 10–19. https://doi.org/10.38094/jastt20179

Islam, A. H. M. S., Schreinemachers, P., & Kumar, S. (2020). Farmers' knowledge, perceptions and management of chili pepper anthracnose disease in Bangladesh. *Crop Protection*, *133*(February), 105139. https://doi.org/10.1016/j.cropro.2020.105139

Juandri, & Anwar, N. (2023). BULLET : Pengenalan Warna Terhadap Objek Dengan Model Analisis Elemen Data Warna Gambar Berbasis Deep Neural Network. *Jurnal Multidisiplin Ilmu*, *2*(01), 23–31. Retrieved from https://journal.mediapublikasi.id/index.php/bullet

Karuna, M., Varsha, B. S., M, S. R., Meghana, G. K., & Student, B. (2019). Early Detection of Chili Plant Leaf Diseases using Machine Learning. *International Journal of Engineering Science and Computing*, *9*(5), 22328–22335. Retrieved from http://ijesc.org/

Kelikualiq, R., Elektro, F. T., Telkom, U., Bogi, N., Karna, A., Elektro, F. T., … Cabai, D. (2022). *Sistem Monitoring Kesehatan Tanaman Cabai Pada Smart Greenbox Menggunakan Algoritma Convolutional Neural Network ( CNN ) Berbasis Internet Of Things ( IOT ) Dan Website*. *8*(6), 3962–3964.

Lina, Q. (2019). Apa itu Convolutional Neural Network? Retrieved from medium.com website: https://medium.com/@16611110/apa-itu-convolutional-neural-network-836f70b193a4

Marcella, D., Yohannes, Y., & Devella, S. (2022). Klasifikasi Penyakit Mata Menggunakan Convolutional Neural Network Dengan Arsitektur VGG-19. *Jurnal Algoritme*, *3*(1), 60–70. https://doi.org/10.35957/algoritme.v3i1.3331

Mascarenhas, S., & Agarwal, M. (2021). A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, 96–99. IEEE. https://doi.org/10.1109/CENTCON52345.2021.9687944

Mohbey, K. K., Sharma, S., Kumar, S., & Sharma, M. (2022). COVID-19 identification and analysis using CT scan images: Deep transfer learning-based approach. In *Blockchain Applications for Healthcare Informatics* (pp. 447–470). Elsevier. https://doi.org/10.1016/B978-0-323-90615-9.00011-6

Muzahid, A. A. M., Wan, W., & Hou, L. (2020). A new volumetric CNN for 3D object classification based on joint multiscale feature and subvolume supervised learning approaches. Computational Intelligence and Neuroscience, 2020, 1-17. https://doi.org/10.1155/2020/5851465

Nuanmeesri, S., & Sriurai, W. (2021). Multi-Layer Perceptron Neural Network Model Development for Chili Pepper Disease Diagnosis Using Filter and Wrapper Feature Selection Methods. *Engineering, Technology and Applied Science Research*, *11*(5), 7714–7719. https://doi.org/10.48084/etasr.4383

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., … Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Patil, A., & Lad, K. (2019). Review of Dieases Detection and Classification for Chilli Leaf using Various Algorithams. *Proceedings of 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019*, 1–4. https://doi.org/10.1109/ICECCT.2019.8869080

Patil, A., & Lad, K. (2021). Chili Plant Leaf Disease Detection Using SVM and KNN Classification. In *Advances in Intelligent Systems and Computing* (Vol. 1187, pp. 223–231). Singapore: Springer. https://doi.org/10.1007/978-981-15-6014-9_26

Polii, M. G. M., Sondakh, T. D., Raintung, J. S. M., Doodoh, B., & Titah, T. (2019). Kajian Teknik Budidaya Tanaman Cabai ( Capsicum annuum L .) Kabupaten Minahasa Tenggara. *Eugenia*, *25*(3), 73–77.

Rahayu, A. S., Fauzi, A., & Rahmat, R. (2022). Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Spotify. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(2), 349. https://doi.org/10.30865/json.v4i2.5398

Rahman, E., Rizko, R. A., Sujiono, S., & Taufik, T. (2023). Segmentasi Wilayah Indonesia Yang Berpotensi Budidaya Cabai Dengan Hierarki Clustering. *JUTIN : Jurnal Teknik Industri Terintegrasi*, *6*(4), 1223–1235. https://doi.org/10.31004/jutin.v6i4.20010

Renfiyeni, Afrini, D., Mahmud, Nelvi, Y., Harissatria, Surtina6, D., & Elinda, F. (2023). Pengendalian Hama Dan Penyakit Tanaman Cabai. *Communnity Development Journal*, *4*(2), 4952–4961.

Rosalina, R., & Wijaya, A. (2020). Pendeteksian Penyakit pada Daun Cabai dengan Menggunakan Metode Deep Learning. *Jurnal*

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4961

*Teknik Informatika Dan Sistem Informasi*, *6*(3), 452–461. https://doi.org/10.28932/jutisi.v6i3.2857

Rozlan, S., & Hanafi, M. (2022). Efficacy of chili plant diseases classification using deep learning: A preliminary study. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*(3), 1442–1449. https://doi.org/10.11591/ijeecs.v25.i3.pp1442-1449

Santika, E. F. (2023). Konsumsi Cabai Besar dan Rawit 2022 Sudah Melampaui Capaian Sebelum Pandemi. Retrieved from databoks.katadata.co.id website: https://databoks.katadata.co.id/datapublish/2023/06/22/konsumsi-cabai-besar-dan-rawit-2022-sudah-melampaui-capaian-sebelum-pandemi

Setiawan, M. H., Gunadi, I. G. A., & Indrawan, G. (2022). Klasifikasi Pelayanan Kesehatan Berdasarkan Data Sentimen Pelayanan Kesehatan menggunakan Multiclass Support Vector Machine. *Jurnal Sistem Dan Informatika (JSI)*, *17*(1), 47–54. Retrieved from https://www.jsi.stikom-bali.ac.id/index.php/jsi/article/view/512

Stoilov, G., Kavardzhikov, V., & Pashkouleva, D. (2012). A Comparative Study of Random Patterns for Digital Image Correlation. *Journal of Theoretical and Applied Mechanics*, *42*(2), 55–66. https://doi.org/10.2478/v10254-012-0008-x

Suradiradja, K. H. (2022). Algoritme Machine Learning Multi-Layer Perceptron dan Recurrent Neural Network untuk Prediksi Harga Cabai Merah Besar di Kota Tangerang. Faktor Exacta, 14(4), 194-205.

Susim, T., & Darujati, C. (2021). Pengolahan Citra untuk Pengenalan Wajah (Face Recognition) Menggunakan OpenCV. *Jurnal Syntax Admiration*, *2*(3), 534–545. https://doi.org/10.46799/jsa.v2i3.202

Wahab, A. H. B. A., Zahari, R., & Lim, T. H. (2019, July). Detecting diseases in chilli plants using K-means segmented support vector machine. In 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC) (pp. 57-61). IEEE.

Windarningsih, M. (2019). Identification of virus causing the yellow leaf curl diseases on chili pepper in Lombok Island by PCR-RFLP technique. *AIP Conference Proceedings*, *2199*, 1–6. https://doi.org/10.1063/1.5141297

Yana, Y. E., & Nafi'iyah, N. (2021). Klasifikasi Jenis Pisang Berdasarkan Fitur Warna, Tekstur, Bentuk Citra Menggunakan SVM dan KNN. *RESEARCH : Journal of Computer, Information System & Technology Management*, *4*(1), 28. https://doi.org/10.25273/research.v4i1.6687

Zikra, F., Usman, K., & Patmasari, R. (2021). Deteksi Penyakit Cabai Berdasarkan Citra Daun Menggunakan Metode Gray Level Co-Occurence Matrix Dan Support Vector Machine. *Seminar Nasional Hasil Penelitian Dan Pengabdian Masyarakat*, *ISSN: 2598*(E-ISSN: 2598-0238), 105.

# OPTIMIZATION OF POTATO LEAF DISEASE IDENTIFICATION WITH TRANSFER LEARNING APPROACH USING MOBILENETV1 ARCHITECTURE

**Herlambang Brawijaya[1*]; Eva Rahmawati[2]; Toto Haryanto[3]**

Department of Computer Science, Faculty of Mathematics and Natural Sciences[1,3]
Institut Pertanian Bogor, Bogor, Indonesia[1,3]
https://ipb.ac.id[1,3]

Informatika[2]
Universitas Nusa Mandiri, Jakarta, Indonesia[2]
https://www.nusamandiri.ac.id/[2]

lambangbrawijaya@apps.ipb.ac.id[1*]; eva.ehw@nusamandiri.ac.id[2]; totoharyanto@apps.ipb.ac.id[3]
(*) Corresponding Author



***Abstract***—*Diseases affecting potato leaves frequently lead to significant setbacks for farmers, reducing the overall harvest and the quality of the potatoes. Given the critical need for prompt disease detection, this research introduces the use of the MobileNet framework grounded in the Convolutional Neural Network (CNN) for adept detection of potato leaf ailments. During the research, potato leaf images undergo processing, and their distinct features are gleaned using CNN. Then, harnessing the MobileNet framework, these images undergo classification to ascertain the existence of diseases. The aspiration is that the formulated model can pinpoint diseases with notable precision, rapid feedback, and enhanced computational adeptness. Initial findings underscore the potential of this methodology in discerning potato leaf diseases, providing renewed optimism for farmers grappling with plant health issues. Experiments using the Transfer Learning approach showed good performance in classification and displayed a high accuracy rate of 99.2%.*

***Keywords:*** *CNN, disease classification, mobilenet architecture, potato leaf disease, transfer learning.*

**Abstrak**—*Penyakit yang menyerang daun kentang kerap menyebabkan dampak negatif bagi petani, menurunkan jumlah dan mutu hasil panen. Menyadari urgensi deteksi dini terhadap penyakit ini, studi ini merekomendasikan penggunaan arsitektur MobileNetV1 yang berlandaskan Convolutional Neural Network (CNN) sebagai solusi cekatan dalam identifikasi penyakit daun kentang. Selama proses penelitian, citra daun kentang diolah dan karakteristiknya dianalisa melalui CNN. Berikutnya, dengan bantuan arsitektur MobileNetV1, citra-citra tersebut dikelompokkan untuk konfirmasi keberadaan penyakit. Tujuan dari penelitian ini adalah menciptakan model yang dapat mengidentifikasi penyakit dengan presisi yang tinggi, kecepatan tanggap, dan keefisienan komputasi. Temuan pendahuluan menegaskan bahwa metode ini berpotensi dalam identifikasi penyakit daun kentang, memberikan optimisme bagi petani yang berhadapan dengan masalah penyakit pada tanamannya. Dengan percobaan menggunakan pendekatan Transfer Learning menunjukkan hasil performa yang baik dalam klasifikasi dan menampilkan nilai akurasi yang cukup tinggi, yaitu 99.2%.*

***Kata Kunci****:* CNN, klasifikasi penyakit, mobilenet arsitektur, penyakit daun kentang, *transfer learning.*

## INTRODUCTION

Potatoes are known as a type of tuber rich in carbohydrates. In addition, potatoes also contain good nutrients for the body, ranging from antioxidants, vitamins, and minerals (Lesmana et al., 2022). In terms of growth, potatoes are one of the food crops that grow well in the mountainous areas of Indonesia. These factors drive the high production of potatoes in the country, reaching millions of tons. In fact, according to data from the Central Bureau of Statistics (BPS), potato production in Indonesia continues to increase year by year. This was shown in 2022, with potato production in Indonesia amounting to 1.42 million tons. This number increased by 4.21% compared to the previous year's production of 1.36 million tons.

In the cultivation process, there are several aspects that need to be considered so that crop production can run smoothly, including when planting potatoes (Gaikwad & Musande, 2023; Jaya & Sahlinal, 2022). However, in the cultivation process, there are often problems with disease attacks on potato plants, one of which is diseases on potato leaves. If farmers are not careful in monitoring the symptoms of disease on potato leaves, the disease can be a major factor affecting the decline in the quality and quantity of potato crop production.

Early pattern recognition of potato leaf diseases is crucial; an effective approach is needed to detect these diseases to improve production and plant quality. Farmers often face difficulties in identifying diseases in the early stages through traditional methods, especially novice farmers. The ability to detect diseases in the early stage will give farmers an advantage in increasing their harvest. The success of the search and identification process highly depends on the appropriate and accurate keywords in assisting disease recognition through digital searches. Keyword mismatches will result in inappropriate search results. Diagnosing potato leaf diseases in traditional ways is not only at risk of error but also requires slow time. On the other hand, systems supported by technology tend to be faster and more cost-efficient. The development of the industrial revolution 4.0 has had many impacts on various sectors, including agriculture. In the agricultural sector, the existence of digital technology offers various solutions, such as automatic detection of diseases on potato leaves.

Deep learning methods are currently a good solution in the ability to recognize complex patterns and automate the process of image classification (Wani et al., 2022). Using image data from an object to be recognized, it is very possible to use deep learning methods in the training process of pattern recognition on potato leaves. The system will be able to recognize diseases on potato leaves based on the given image data. The Convolutional Neural Network (CNN) is the most efficient part of the Deep Learning method in its ability to select features in images (Anim-Ayeko et al., 2023; Gao et al., 2021; Sharma et al., 2020).

Research on potato leaf diseases has been widely carried out before, and this research is aimed at classifying diseases on potato leaves. Currently, deep learning methods are considered the most appropriate method to develop algorithms capable of categorizing various diseases on an object, including potato leaves (Dasgupta et al., 2020; Saputra et al., 2021). This adopted approach is expected to provide high-accuracy classification results for potato leaf diseases. This CNN model applies multi-layered convolution in extracting and integrating a very large dataset, which distinguishes the CNN model from conventional image classification methods (Khan et al., 2023; Rashid et al., 2021). One of the challenges with using this CNN model is that the data available for image classification is not always abundant. There will be a possibility of imbalance in the number of samples from each class, which can affect accuracy in classification. During the classification process, the application of the transfer learning model can also be used as the basis for a previously trained model, which will facilitate the classification process compared to doing the process raw or from a model that has not been trained before (Akther et al., 2021; Islam et al., 2019; Sagar & Jacob, 2020).

Regarding the research that has been conducted by various researchers, studies related to potato leaf diseases have been conducted by several researchers using different methods and results presented. A study by (M & Kristiyanti, 2023) researched potato leaves using the CNN method and the MobileNetV2 architecture. In their study, due to the unbalanced and limited data amounting to 2,152 colored potato leaf images obtained from the Kaggle repository, the researchers applied a data augmentation technique to address this issue. Thus, in the classification process, the researchers compared results without performing data augmentation and with data augmentation, and also conducted experiments using several Transfer Learning methods, namely InceptionV3, VGG16, InceptionResNetV2, and MobileNetV2. From the results using MobilenetV2 without data augmentation, an accuracy rate of 97.6% was achieved, and by performing data augmentation, an accuracy of 99.6% was achieved. A study by (Arshad et al., 2023) on various objects such as tomatoes, apples, and potato leaves from PlantVillage (Kaggle) yielded an accuracy rate of 94.25% using the PDDPNet method. Research by (Nishad et al., 2022) on potato leaf objects used the CNN algorithm with the VGG16 architecture and a dataset obtained from PlantVillage (Kaggle) and Mendeley totaling 2,580 images, split in an 80:20 ratio for training and testing data. By performing data augmentation, an accuracy rate of 97% was achieved. A study by (Rozaqi et al., 2021a) classified potato leaves with a dataset of 450 images obtained from PlantVillage (Kaggle). By testing several methods, accuracy results were obtained as follows: InceptionV3 78%, VGG16 95%, and ResNet-50 at 78%. Research by (Tiwari et al., 2020) used potato leaf images totaling 2,152 images obtained from PlantVillage (Kaggle).

Using the VGG19 architecture, they achieved an accuracy rate of 97.8%.

After reviewing several previous studies, it was decided that this research would use the Convolutional Neural Network model with the MobileNetV1 architecture to classify potato leaf diseases. The reason for using the Convolutional Neural Network (CNN) method in this research is its efficiency in recognizing complex patterns and automating the process of image classification. Deep learning methods, particularly CNN, are highly effective in utilizing image data for the training process of pattern recognition. In the context of potato leaf disease detection, the CNN model is capable of accurately recognizing diseases based on the provided image data, making it a suitable and powerful tool for this application. In our research, the choice of MobileNetV1 architecture as the foundation of our deep learning model for identifying potato leaf diseases was driven by considerations of computational efficiency, inference speed, and high accuracy (Sharma et al., 2021). With its lightweight design, MobileNetV1 offers significant advantages in terms of computational efficiency and inference spee factors that are crucial for field applications where computational resources are limited and response speed is critical (Howard et al., 2017). Furthermore, in our evaluation of various CNN architectures using the same dataset, we observed that research with MobileNetV2 achieved an accuracy of 99.6% after data augmentation (M & Kristiyanti, 2023). Although this result is similar to our achievement, we assessed that MobileNetV1 provides an optimal balance between accuracy and computational efficiency a crucial quality for implementation on edge devices. The use of the MobileNetV1 architecture in image classification can also be explored in terms of its efficiency in extracting features and optimizing the model size for environments with limited resources, especially when compared to other CNN-based architectures such as VGG16, AlexNet, and others (Suganthi & Sathiaseelan, 2020). This underscores the strategic thinking behind the selection of MobileNetV1, affirming its superiority for field applications of potato leaf disease detection.

## MATERIALS AND METHODS

The research method carried out in this study includes several stages, namely, the preparation of the dataset used, the division of training data, test data, and validation data, the design of the classification process architecture, followed by the creation of a model for the classification process.

### Datasets

The data used in this study come from a public dataset sourced from Kaggle (https://www.kaggle.com/datasets/rizwan123456789/potato-disease-leaf-datasetpld). The image data utilized consists of 4,072 colored leaf images with a resolution of 256x256 pixels, encompassing three disease classes: EarlyBlight, Healthy, and LateBlight.

### Data Split

Once the data was compiled, it was divided into two groups for classification: training data and testing data. The data division ratio adopted is 90:10, where 90% is used for training and 10% for testing. The data breakdown can be seen in table 1.

Table 1. Dataset Distribution

| No | Class | Training (90%) | Testing (10%) | Total |
|---|---|---|---|---|
| 1 | EarlyBlight | 1.465 | 163 | 1.628 |
| 2 | Healthy | 918 | 102 | 1.020 |
| 3 | LateBlight | 1.282 | 142 | 1.424 |
| | Total | 3.665 | 407 | 4.072 |

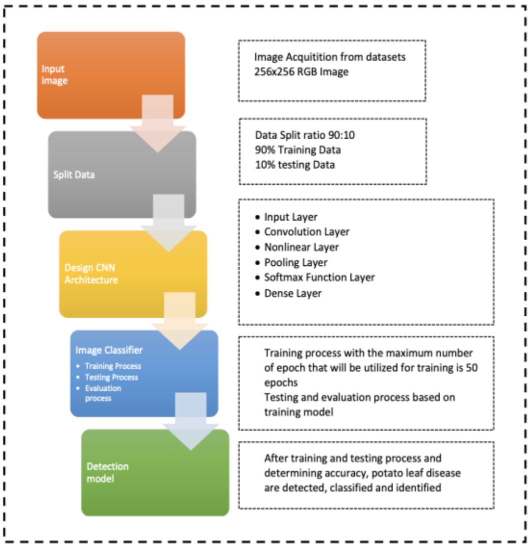
Source :(Rashid et al., 2021)

For validation data, 100 images per class were used or 7.3% of the total images. In this study, data was used without any modifications or augmentation. This decision was made to maintain the integrity of the original data, facilitate the replication of research by other researchers using the same dataset, and evaluate the model under conditions that resemble real-world scenarios, including facing the imbalance in the number of samples among classes that often occurs in application settings. We chose not to implement techniques such as oversampling or undersampling because we wanted to assess the efficiency and effectiveness of the MobileNetV1 architecture in classifying potato leaf diseases under as-is conditions.

### Classification Design

After the data division phase, the next step involves the design of the architecture for image classification. The design process begins with the preparation of the dataset as input data, the division of training and test data, followed by the creation of a model using a transfer learning scheme up to the classification results. The scheme of the process can be viewed in figure 1.

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4718



Source : (Research Results, 2023)
Figure 1. General Scheme for Image Classification

The CNN architecture was created using the MobileNetV1 design. Using an image input resolution of 224x224 pixels (Saputra et al., 2021). This resolution choice aligns with the standard resolution of MobileNetV1. The model was developed by running tests using a Conv2D with 32 Kernels and 3 classes. The activation function employed is the Rectified Linear Unit (ReLU), with the Adam compiler, and a loss function using categorical_crossentropy. The results of the classification process will be saved as a model checkpoint, as illustrated in Table 2.

Table 2. MobileNetV1 and Sequential model

| Model: "sequential" | | |
|---|---|---|
| Layer (type) | Output Shape | Param # |
| Mobilenet_1.00_224 (Functional) | (None, 7, 7, 1024) | 3228864 |
| conv2d (Conv2D) | (None, 5, 5, 32) | 294944 |
| Global_average_pooling2d (GlobalAveragePooling2D) | (None, 32) | 0 |
| dense (Dense) | (None, 3) | 99 |
| Total params: 3523907 (13.44 MB) | | |
| Trainable params: 295043 (1.13 MB) | | |
| Non-trainable params: 3228864 (12.32 MB) | | |

Source : (Research Results, 2023)

## RESULTS AND DISCUSSION

The experiments in this study were conducted using Python programming and the Tensorflow framework, supported by the Apple Silicon M1 Pro processor and 16 GB RAM. The training and testing data processes were conducted in two trials. The first trial used an epoch setting of 50, while the second involved hyperparameter tuning.

### Training and Testing Results

In the first trial, an epoch of 50 was set. An epoch is a specific parameter that describes how many times the entire dataset runs through the CNN model. The classification process's performance on training and validation data using the transfer learning model and an epoch setting of 50 was quite satisfactory.

```
Epoch 1/50
26/26 [==============================] - 22s 861ms/step - loss: 0.0174 - accuracy: 0.9997 - val_loss: 0.2196 - val_accuracy: 0.9213
Epoch 2/50
26/26 [==============================] - 21s 820ms/step - loss: 0.0168 - accuracy: 0.9994 - val_loss: 0.2076 - val_accuracy: 0.9287
Epoch 3/50
26/26 [==============================] - 21s 823ms/step - loss: 0.0145 - accuracy: 0.9997 - val_loss: 0.2258 - val_accuracy: 0.9225
Epoch 4/50
26/26 [==============================] - 21s 808ms/step - loss: 0.0137 - accuracy: 0.9997 - val_loss: 0.2102 - val_accuracy: 0.9262
Epoch 5/50
26/26 [==============================] - 21s 809ms/step - loss: 0.0126 - accuracy: 0.9997 - val_loss: 0.2313 - val_accuracy: 0.9200
Epoch 6/50
26/26 [==============================] - 21s 807ms/step - loss: 0.0110 - accuracy: 1.0000 - val_loss: 0.2168 - val_accuracy: 0.9311
Epoch 7/50
26/26 [==============================] - 21s 808ms/step - loss: 0.0104 - accuracy: 1.0000 - val_loss: 0.2154 - val_accuracy: 0.9274
Epoch 8/50
26/26 [==============================] - 21s 824ms/step - loss: 0.0091 - accuracy: 1.0000 - val_loss: 0.2236 - val_accuracy: 0.9237
Epoch 9/50
26/26 [==============================] - 21s 821ms/step - loss: 0.0085 - accuracy: 1.0000 - val_loss: 0.2234 - val_accuracy: 0.9262
Epoch 10/50
26/26 [==============================] - 22s 860ms/step - loss: 0.0079 - accuracy: 1.0000 - val_loss: 0.2369 - val_accuracy: 0.9225
Epoch 49/50
26/26 [==============================] - 21s 809ms/step - loss: 0.0012 - accuracy: 1.0000 - val_loss: 0.2650 - val_accuracy: 0.9299
Epoch 50/50
26/26 [==============================] - 21s 809ms/step - loss: 0.0011 - accuracy: 1.0000 - val_loss: 0.2642 - val_accuracy: 0.9287
```

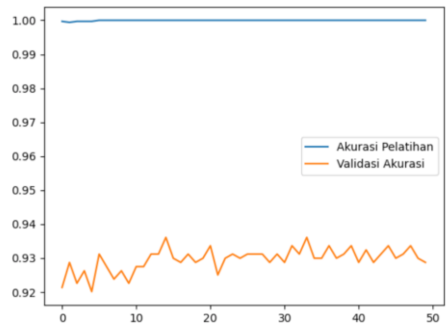
Source : (Research Results, 2023)
Figure 2. First Experiment Process

Referring to figure 2, the first trial using an epoch of 50 displayed good training and validation accuracy values. Training accuracy increased to 1.0000 and validation accuracy rose to 0.9287. Similarly, training error dropped to 0.0011, and validation error reached 0.2642. The model evaluation process in Figure 3 showed a final accuracy value of 0.9926 or 99.26%.

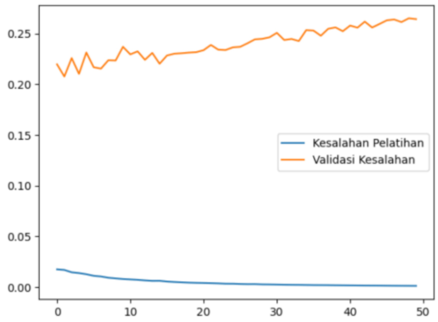
```
['loss', 'accuracy']
26/26 [==============================] - 17s 648ms/step - loss: 0.0011 - accuracy: 1.0000
[0.0010915168095380068, 1.0]
7/7 [==============================] - 4s 582ms/step - loss: 0.2642 - accuracy: 0.9287
[0.2641914486885071, 0.9286592602729797]
13/13 [==============================] - 2s 151ms/step - loss: 0.0229 - accuracy: 0.9926
[0.022865256294608116, 0.9926289916038513]
```

Source : (Research Results, 2023)
Figure 3. Model Experiment Result (First Experiment)

In Figure 4, the obtained accuracy values showed regular training and testing accuracy values, although the increase wasn't significant. In Figure 5, validation loss appeared stable without significant fluctuations.

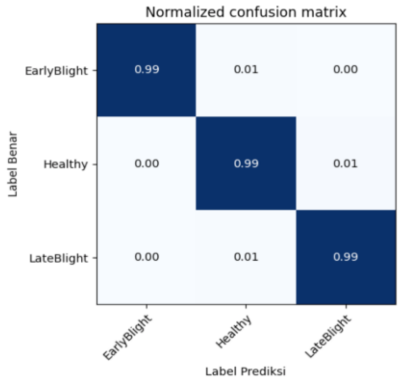


Source : (Research Results, 2023)
Figure 4. Accuracy Value Graph (First Experiment)




Source : (Research Results, 2023)
Figure 5. Loss Value Graph (First Experiment)

Additionally, the evaluation results and confusion matrix from the first trial can be seen in figure 6.




Source : (Research Results, 2023)
Figure 6. Confussion Matrix First Experiment

Considering the confusion matrix in figure 6, an accuracy of 99.2% was achieved. Predicted and actual values can be seen in table 3.

Table 3. First Experiment Classification Performance

| | EarlyBlight | Healthy | LateBlight |
|---|---|---|---|
| EarlyBlight | 162 | 1 | 0 |
| Healthy | 0 | 101 | 1 |
| LateBlight | 0 | 1 | 141 |

Source : (Research Results, 2023)

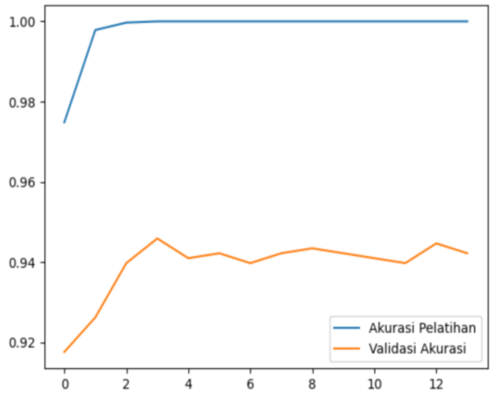
From table 3, it's inferred that the classification results from the first trial were satisfactory, with one error in each class, totaling three errors.

In the second trial, hyperparameter tuning was performed with settings of epoch 50, batch size 64, optimization using the Adam optimizer, setting Categorical Cross Entropy as the loss function, and using "save best" in model checkpoint and early stopping focusing on validation loss, patience 10, and verbose 1. The results are shown in figure 8, where there was no improvement in accuracy and loss on epoch 14, with training accuracy reaching 1.000 and validation accuracy 0.9422, training error 0.0013, and validation error 0.2443.

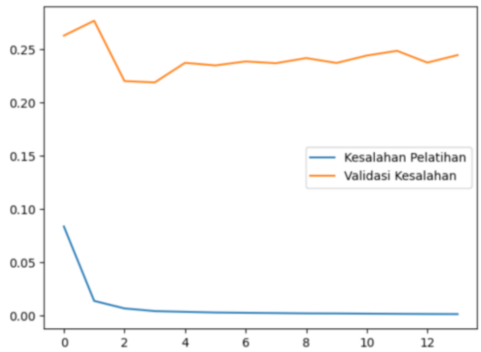
```
Epoch 1/50
26/26 [==============================] - ETA: 0s - loss: 0.0835 - accuracy: 0.9748
Epoch 1: val_loss improved from inf to 0.26252, saving model to best_model.h5
26/26 [==============================] - 22s 830ms/step - loss: 0.0835 - accuracy: 0.9748 - val_loss: 0.2625 - val_accuracy: 0.9176
Epoch 2/50
26/26 [==============================] - ETA: 0s - loss: 0.0136 - accuracy: 0.9979
Epoch 2: val_loss did not improve from 0.26252
26/26 [==============================] - 21s 826ms/step - loss: 0.0136 - accuracy: 0.9979 - val_loss: 0.2765 - val_accuracy: 0.9262
Epoch 3/50
26/26 [==============================] - ETA: 0s - loss: 0.0066 - accuracy: 0.9997
Epoch 3: val_loss improved from 0.26252 to 0.21991, saving model to best_model.h5
26/26 [==============================] - 21s 810ms/step - loss: 0.0066 - accuracy: 0.9997 - val_loss: 0.2199 - val_accuracy: 0.9397
Epoch 4/50
26/26 [==============================] - ETA: 0s - loss: 0.0041 - accuracy: 1.0000
Epoch 4: val_loss improved from 0.21991 to 0.21859, saving model to best_model.h5
26/26 [==============================] - 21s 831ms/step - loss: 0.0041 - accuracy: 1.0000 - val_loss: 0.2186 - val_accuracy: 0.9459
Epoch 5/50
26/26 [==============================] - ETA: 0s - loss: 0.0034 - accuracy: 1.0000
Epoch 5: val_loss did not improve from 0.21859
26/26 [==============================] - 22s 829ms/step - loss: 0.0034 - accuracy: 1.0000 - val_loss: 0.2370 - val_accuracy: 0.9410
Epoch 6/50
26/26 [==============================] - ETA: 0s - loss: 0.0027 - accuracy: 1.0000
Epoch 6: val_loss did not improve from 0.21859
26/26 [==============================] - 21s 820ms/step - loss: 0.0027 - accuracy: 1.0000 - val_loss: 0.2346 - val_accuracy: 0.9422
Epoch 7/50
26/26 [==============================] - ETA: 0s - loss: 0.0025 - accuracy: 1.0000
Epoch 7: val_loss did not improve from 0.21859
26/26 [==============================] - 21s 809ms/step - loss: 0.0025 - accuracy: 1.0000 - val_loss: 0.2383 - val_accuracy: 0.9397

Epoch 13/50
26/26 [==============================] - ETA: 0s - loss: 0.0013 - accuracy: 1.0000
Epoch 13: val_loss did not improve from 0.21859
26/26 [==============================] - 21s 810ms/step - loss: 0.0013 - accuracy: 1.0000 - val_loss: 0.2372 - val_accuracy: 0.9446
Epoch 14/50
26/26 [==============================] - ETA: 0s - loss: 0.0013 - accuracy: 1.0000
Epoch 14: val_loss did not improve from 0.21859
26/26 [==============================] - 21s 812ms/step - loss: 0.0013 - accuracy: 1.0000 - val_loss: 0.2443 - val_accuracy: 0.9422
Epoch 14: early stopping
```

Source : (Research Results, 2023)
Figure 7. Second Experiment Process

In this trial, error values and accuracy can be seen in Figures 8 and 9. Figures 8 and 9 indicate a slight increase in accuracy and a decrease in error values, although not significantly.




Source : (Research Results, 2023)
Figure 8.Accuracy Value Graph (Second Experiment)




Source : (Research Results, 2023)
Figure 9. Loss Value Graph (Second Experiment)

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4718

The confusion matrix in the second trial showed improved results compared to the first trial with an accuracy of 99.51%, as seen in Figure 10.

```
['loss', 'accuracy']
26/26 [==============================] - 17s 642ms/step - loss: 0.0011 - accuracy: 1.0000
[0.0011103596771135926, 1.0]
7/7 [==============================] - 4s 568ms/step - loss: 0.2443 - accuracy: 0.9422
[0.2442903220653534, 0.9421893954277039]
13/13 [==============================] - 2s 158ms/step - loss: 0.0281 - accuracy: 0.9951
[0.028144290670752525, 0.9950860142707825]
```

Source : (Research Results, 2023)

Figure 10. Model Experiment Result (Second Experiment)

Predicted and actual values are shown in table 4.

Table 4. Second Experiment Classification Performance

| | EarlyBlight | Healthy | LateBlight |
|---|---|---|---|
| EarlyBlight | 162 | 1 | 0 |
| Healthy | 0 | 102 | 0 |
| LateBlight | 0 | 1 | 141 |

Source : (Research Results, 2023)

In table 4, the confusion matrix classification performance showed reduced errors, with two mistakes found in the EarlyBlight and LateBlight classes.

From the two trials conducted, performance metrics like precision, recall, and f1-score can also be observed in table 5.

Table 5. Classification Validation Performance Metrics

| | Class | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Experiment 1 | EarlyBlight | 1.00 | 0.99 | 1.00 |
| | Healthy | 0.98 | 0.99 | 0.99 |
| | LateBlight | 0.99 | 0,99 | 0.99 |
| Experiment 2 | EarlyBlight | 1.00 | 0.99 | 1.00 |
| | Healthy | 0.98 | 1.00 | 0.99 |
| | LateBlight | 1.00 | 0.99 | 1,00 |

Source : (Research Results, 2023)

**Performance Comparison**

Table 6 displays a performance results comparison between the proposed CNN model and models from previous studies.

Table 6. Comparison between the proposed model and previous studies

| Author(s), Year | Algoritma | Dataset | Accuracy |
|---|---|---|---|
| (M & Kristiyanti, 2023) | CNN – MobileNetV2 | Kaggle (2.152 Images) | 97.6% (without Augmentation), 99.6% (with augmentation) |
| (Arshad et al., 2023) | PLDPNet | Kaggle (2.152 Images) | 94.25% |
| (Nishad et al., 2022) | CNN - VGG16 | Kaggle (2.580 Images) | 97% |
| (Rozaqi at al., 2021b) | CNN – VGG16 | Kaggle (450 Images) | 95% |
| (Tiwari et al., 2020) | CNN _ VGG19 | Kaggle (2.152 Images) | 97.8% |
| **Our proposed work** | **CNN – MobileNet V1** | **Kaggle (4. 072 Images)** | **99.5%** |

Source : (Research Results, 2023)

## CONCLUSION

Considering the importance of agriculture and crops in Indonesia, and given the various plant diseases currently, especially in potato plants, a reliable approach to detecting and classifying diseases in potato leaves can produce accurate results. This research contributes to the field of agricultural technology by developing a highly accurate and efficient model for the detection and classification of potato leaf diseases using the Convolutional Neural Network (CNN) with the MobileNetV1 architecture. With the aid of computer technology and deep learning, this research, utilizing the CNN algorithm with a simple MobileNetV1 architecture, managed to achieve a satisfactory accuracy of 99.5%. This value demonstrates a highly accurate and rapid detection capability against types of diseases in potato leaves. Various trial levels and optimization methods can also be considered for further research with the proposed system. Furthermore, there's hope to develop the proposed model by building an expert system for identification and classification of diseases in potato leaves.

## REFERENCE

Akther, J., Harun-Or-Roshid, M., Nayan, A. A., & Kibria, M. G. (2021). Transfer learning on VGG16 for the Classification of Potato Leaves Infected by Blight Diseases. *2021 Emerging Technology in Computing, Communication and Electronics, ETCCE 2021*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ETCCE54784.2021.9689792

Anim-Ayeko, A. O., Schillaci, C., & Lipani, A. (2023). Automatic blight disease detection in potato (Solanum tuberosum L.) and tomato (Solanum lycopersicum, L. 1753) plants using deep learning. *Smart Agricultural Technology*, *4*.

https://doi.org/10.1016/j.atech.2023.100178

Arshad, F., Mateen, M., Hayat, S., Wardah, M., Al-Huda, Z., Gu, Y. H., & Al-antari, M. A. (2023). PLDPNet: End-to-end hybrid deep learning framework for potato leaf disease prediction. *Alexandria Engineering Journal*, *78*, 406–418. https://doi.org/10.1016/j.aej.2023.07.076

Dasgupta, S. R., Rakshit, S., Mondal, D., & Kole, D. K. (2020). Detection of Diseases in Potato Leaves Using Transfer Learning. *Advances in Intelligent Systems and Computing*, *999*, 675–684. Springer. https://doi.org/10.1007/978-981-13-9042-5_58

Gaikwad, V. P., & Musande, V. (2023). Potato Plant leaf disease detection using CNN Model. *Eur. Chem. Bull*, *12*(1), 516–527.

Gao, J., Westergaard, J. C., Sundmark, E. H. R., Bagge, M., Liljeroth, E., & Alexandersson, E. (2021). Automatic late blight lesion recognition and severity quantification based on field imagery of diverse potato genotypes by deep learning. *Knowledge-Based Systems*, *214*. https://doi.org/10.1016/j.knosys.2020.106723

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv*. Retrieved from http://arxiv.org/abs/1704.04861

Islam, F., Hoq, M. N., & Rahman, C. M. (2019). Application of Transfer Learning to Detect Potato Disease from Leaf Image. *2019 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2019*, 127–130. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/RAAICON48939.2019.53

Jaya, T. S., & Sahlinal, D. (2022). *Potato Leaf Disease Detection with Convolutional Neural Network Method*. https://doi.org/10.25181/icoaas.v3i3.2867

Khan, M. A., Armghan, A., Alenezi, F., Abbas, S., Alhaisoni, M., Tariq, U., … Thinnukool, O. (2023). Crops Leaf Diseases Recognition: A Framework of Optimum Deep Learning Features Performance evaluation of deep e-CNN with integrated spatial-spectral features in hyperspectral image classification View project Crops Leaf Diseases Recognition: A Framework of Optimum Deep Learning Features. *Computers, Materials & Continua*, *74*(1), 1139–1159. https://doi.org/10.32604/cmc.2022.028824

Lesmana, A. M., Fadhillah, R. P., & Rozikin, C. (2022). Identifikasi Penyakit pada Citra Daun Kentang Menggunakan Convolutional Neural Network (CNN). *Jurnal Sains Dan Informatika*, *8*(1), 21–30. https://doi.org/10.34128/jsi.v8i1.377

M, T. A., & Kristiyanti, D. A. (2023). Mobilenetv2 Architecture for Identification of Potato Leaf Disease. Journal of Theoretical and Applied Information Technology, 101(16), 6273–6285.Retrieved from https://www.researchgate.net/publication/373656978

Nishad, M. A. R., Mitu, M. A., & Jahan, N. (2022). Predicting and Classifying Potato Leaf Disease using K-means Segmentation Techniques and Deep Learning Networks. *Procedia Computer Science*, *212*(C), 220–229. Elsevier B.V. https://doi.org/10.1016/j.procs.2022.11.006

Rashid, J., Khan, I., Ali, G., Almotiri, S. H., Alghamdi, M. A., & Masood, K. (2021). Multi-level deep learning model for potato leaf disease recognition. *Electronics (Switzerland)*, *10*(17). https://doi.org/10.3390/electronics10172064

Rozaqi, A. J., Sunyoto, A., & Arief, M. R. (2021a). Implementasi Transfer Learning pada Algoritma Convolutional Neural Network Untuk Identifikasi Penyakit Daun Kentang. *Procedia of Engineering and Life Science*, *1*(1).

Rozaqi, A. J., Sunyoto, A., & Arief, R. (2021b). *Deteksi Penyakit pada Daun Kentang Menggunakan Pengolahan Citra dengan Metode Convolutional Neural Network Detection of Potato Leaves Disease Using Image Processing with Convolutional Neural Network Methods*.

Sagar, A., & Jacob, D. (2020). *On Using Transfer Learning For Plant Disease Detection*. https://doi.org/10.13140/RG.2.2.12224.15360/1

Saputra, R. A., Wasyianti, S., Supriyatna, A., & Saefudin, D. F. (2021). Penerapan Algoritma Convolutional Neural Network Dan Arsitektur MobileNet Pada Aplikasi Deteksi Penyakit Daun Padi. *JURNAL SWABUMI*, *9*(2). Retrieved from https://archive.ics.uci.edu/ml/datasets/Rice

Sharma, A. K., Kang, B., & Kim, K. K. (2021). LightNet: A Lightweight Neural Network for Image Classification. *In 2021 18th International SoC Design Conference (ISOCC)* (pp. 419-420). IEEE. https://doi.org/10.1109/ISOCC53507.2021.9613865

Sharma, P., Berwal, Y. P. S., & Ghai, W. (2020). Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. *Information Processing*

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.4718

*in Agriculture*, 7(4), 566-574.. https://doi.org/10.1016/j.inpa.2019.11.001

Suganthi, M., & Sathiaseelan, J. (2020). An Exploratory of Hybrid Techniques on Deep Learning for Image Classification. *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 1–4. https://doi.org/10.1109/ICCCSP49186.2020.9315270

Tiwari, D., Ashish, M., Gangwar, N., Sharma, A., Patel, S., & Bhardwaj, S. (2020). Potato Leaf Diseases Detection Using Deep Learning. *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)*, 461–466.

Wani, J. A., Sharma, S., Muzamil, M., Ahmed, S., Sharma, S., & Singh, S. (2022, January 1). Machine Learning and Deep Learning Based Computational Techniques in Automatic Agricultural Diseases Detection: Methodologies, Applications, and Challenges. *Archives of Computational Methods in Engineering*, Vol. 29, pp. 641–677. Springer Science and Business Media B.V. https://doi.org/10.1007/s11831-021-09588-5

# DEVELOPMENT OF CINEVERSE FILM WEBSITE UTILIZING THEMOVIEDB'S API FOR DYNAMIC CONTENT MANAGEMENT

**Siti Nur Fadhilah[1*]; Fandy Setyo Utomo[2]**

Informatika[1,2]
AMIKOM Purwokerto University, Purwokerto, Indonesia[1,2]
https://www.amikompurwokerto.ac.id/[1,2]
fdhilah1202@gmail.com[1*], fandy_setyo_utomo@amikompurwokerto.ac.id[2]
(*) Corresponding Author



***Abstract***—*The development of websites in this digital era is crucial to creating captivating and relevant online experiences. The combination of server-side programming and client-side technologies along with MySQL database management forms the foundation for a dynamic user interface emphasizes the significance of integrating various technologies to achieve this goal. This project involves the use of PHP, HTML, CSS, JavaScript, and MySQL, with the integration of The Movie Database (TMDB) API, showcasing the intricate fusion of creativity, technical prowess, and data integration. The resulting website offers a comprehensive list of films with detailed information and posters, enhancing the user experience and making it an essential read for those interested in crafting immersive online experiences. The abstract of this research aims to explore the process of website development using diverse technologies and data integration and to analyze its impact on user experience. By examining aspects such as security, performance, and routine maintenance, this study aims to provide in-depth insights into producing captivating and relevant online experiences in the context of modern web development.*

***Keywords:*** *movies, PHP, the movie database API, websites.*

**Abstrak**—*Pengembangan situs web di era digital ini sangat penting untuk menciptakan pengalaman online yang menarik dan relevan. Gabungan pemrograman sisi server dan teknologi sisi klien bersama dengan manajemen database MySQL membentuk dasar antarmuka pengguna yang dinamis menekankan pentingnya mengintegrasikan berbagai teknologi untuk mencapai tujuan ini. Proyek ini melibatkan penggunaan PHP, HTML, CSS, JavaScript, dan MySQL, dengan integrasi API The Movie Database (TMDB), yang menampilkan perpaduan rumit antara kreativitas, kecakapan teknis, dan integrasi data. Situs web yang dihasilkan menawarkan daftar film yang komprehensif dengan informasi dan poster yang detail, meningkatkan pengalaman pengguna dan membuatnya menjadi bacaan penting bagi mereka yang tertarik dalam menciptakan pengalaman online yang mendalam. Abstrak penelitian ini bertujuan untuk mengeksplorasi proses pengembangan situs web menggunakan teknologi dan integrasi data yang beragam serta menganalisis dampaknya terhadap pengalaman pengguna. Dengan meneliti aspek-aspek seperti keamanan, performa, dan pemeliharaan rutin, penelitian ini bertujuan untuk memberikan wawasan yang mendalam tentang bagaimana menghasilkan pengalaman online yang menarik dan relevan dalam konteks pengembangan web modern.*

***Kata Kunci****:* film, *PHP, the movie database API,* website.

## INTRODUCTION

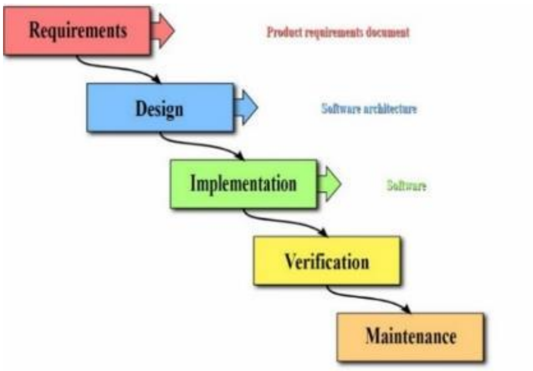
Film industry is experiencing a significant surge, creating new enthusiasm for filmmakers across the country. This growth has also resulted in a growing variety of film genres , including comedy, politics, drama, musicals, and works that raise national themes (Karolina et al., 2020)**.** The term "film" typically refers to a motion picture or movie, which consists of a sequence of static images that, when projected onto a screen, create the illusion of continuous motion, thanks to the phenomenon known as persistence of vision (Hjort, 2019). As a mass communication medium, film is considered effective because of its ability to present messages audio-visually, allowing complex stories to be conveyed in a relatively short time. Through the experience of watching a film, viewers feel the ability to cross the boundaries of time and space, allowing them to connect with the lives depicted in the film and perhaps be influenced by the message conveyed to the audience (Prima, 2022). Films have various themes that function as a means to entertain and convey messages to the audience. The advantage of the audio-visual format in films is that it is able to access and influence the emotions and morality of the audience. Many filmmakers use this medium as a way to convey implicit moral messages to their intended audience. Certain messages in a film are communicated for the viewer to read, or decode, and subsequently influence the viewer's individual understanding. A website is a compilation of web pages and associated content distinguished by a shared domain name and hosted on one or more web servers. Websites are commonly reachable through the Internet or a restricted local area network (Vargas et al., 2020). As time goes by, film websites are becoming more and more interested. However, the problem is that almost all film websites require users to pay before wanting to watch a film.

Writer do research that can help overcome the problem above with make a free movie website. Study it also uses an API that can accessed for free. API, abbreviation from Application Programming Interface, Historically, API, which stands for Application Programming Interface, has been used since the early days of personal computers for exchanging data between two or more programs. APIs have become increasingly vital in the modern software ecosystem for building large-scale software solutions on top of common technology platforms (Ofoeda et al., 2019). APIs facilitate swift and inventive app development by enabling applications to engage with external systems. They are pivotal in crafting diverse application platforms like IoT, mobile apps, and web applications (Idris et al., 2022). An API is an interface that connects various application systems, allowing simultaneous access to some or all of the functions of these systems (Paramitha et al., 2022).

The previous research cited in the passage involves two relevant studies pertaining to the topic of creating a free movie streaming website. Firstly, the study conducted by Adam Adhitama in 2022, focused on the successful creation of a user interface (UI) using Figma for streaming websites. The results of this research indicated that the UI created was effective in providing a good user experience (Adhitama et al., 2022). Secondly, the study by Estu Prayoga in 2023, discussed the use of the Waterfall method in booking cinema tickets, with results showing that this method could shorten the time required for ticket purchasing and booking. Both of these studies provide crucial groundwork for the current research in developing the Cineverse website. The author utilized the Waterfall method supported by system design using UML, with the primary goal of providing assistance to users who wish to watch films with easy access and without requiring prior payment (Prayoga et al., 2023).

## MATERIALS AND METHODS

In this research, system development was carried out using the waterfall system development model with the application of the Unified Modeling Language (UML). Model Waterfall has a series of stages consisting of requirements analysis, system design, system implementation, testing, and maintenance (Muni & Ihwan 2021) . Look at Figure 1 below:




Source: (Ridoh & Putra, 2021)

Figure 1. Waterfall Method

Figure 1 is a brief explanation of the stages of the Waterfall Model. The following is an explanation of the waterfall method (Badrul, 2021).

### a. Needs Analysis

This stage is the requirements gathering stage including documents and interfaces for analyze / specifying software requirements so that user needs can be understood in order to determine the software solution that will be used in the system computerization process. Necessary requirements _ in study This is:

a. Hardware
   1. Computer or laptop with at least 2GB RAM
   2. Computer or laptop minimum 128GB Hard Disk
   3. Mouse
   4. Keyboards
b. Software
   1. Windows Operating System
   2. XAMPP
   3. MySQL
   4. Web Browser

c. User
   1. Admin is someone who has the right and obtains several policies to manage the website. Admins can only control several parts, namely, likes, comments, user registration, and managing admin data.
   2. Users is someone who uses this website. The skills you have must also be able to use every device and conditions required to operate the website.

**b. Design**

During software program creation, the author designs data structures, software architecture, interface representation, and coding procedures. Unified Modeling Language (UML) is utilized to visually illustrate the program's design, including Activity Diagrams, Use Case Diagrams is to identify the various functions in the system and who has the right to use these function (Musthofa & Adiguna 2022), and Sequence Diagrams (Pecoraro and Luzi, 2022). UML serves as a tool or model for designing object-oriented software development (Sonata, 2019). Sequence Diagrams is a sequence of dynamic models that describes the instances of classes participating in a use case and the messages that pass between them over time (Fowler, 2021). For database design, Logical Record Structure (LRS) is employed, outlining record arrangement within tables derived from various entities (Gumelar, 2023). LRS also a description of the structure of records in tables formed from the results between sets of entities to determine the number of tables and foreign keys ( FK ) (Syafi'i & Fajarita 2019).

1. Program Code (Implementation)
   After do analysis and design device soft, step next is implement in form a given movie website Name Cineverse. It's involving use a number of Language programming for operate desired functions _ on the website, namely:
   a. PHP
      PHP, known as "Hypertext Preprocessor," is widely utilized as a server-side scripting language in web development, facilitating the creation of dynamic web pages. One notable framework built upon PHP, known as "Hypertext Preprocessor," is widely utilized as a server-side scripting language in web development, facilitating the creation of dynamic web pages (Vidal-Silva et al., 2020).
   b. MySQL
      MySQL is software or software that is open or can be accessed by many people. Its function is to create a database. SQL can be called an abbreviation of Structured Query Language (Bintang et al., 2023).
   c. TMDB API
      The Movie Database API is an API service intended for programmers who are interested in using images or data from films, TV shows, or actors in the applications they want to create. TMDB API is a system provided for programmers to programmatically retrieve and use data and/or images in the API (https://developer.themoviedb.org/docs/faq).

**c. Testing**

Testing focuses on the software from a logical and functional perspective and ensures that all parts have been tested so that the output produced is as desired. At this stage the test was carried out by the author using a black box testing. Blackbox Testing is a software testing method that tests the functionality of an application without peeking into its internal structure or how it works (Dwi & Wardah, 2021). This testing method can be applied to virtually any level of software testing: units, integration, system, and acceptance.

**d. Support or Maintenance (Maintenance)**

Defining effort - development efforts for the system that is being created to deal with to anticipate developments and changes in the system concerned related to hardware and software. The hardware used is the Windows 11 Home Single Language operating system specifications 64-bit, Intel Core i5-1132H CPU 3.20 G H z, Memory RAM 16 GB.

## RESULTS AND DISCUSSION

Research results include all scientific activities and methods used during the research process. In this context, research results are realized in the form of an online web application.

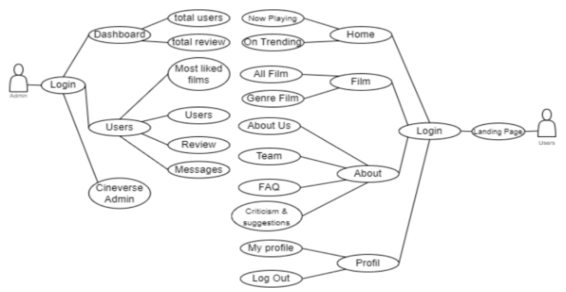
**1. UML (Unified Modeling Diagram) Design**

Below describe and discuss the results of the design process for creating the Cineverse film website.

**2. Use Cases**

Figure 2 is a use case diagram from the Cineverse film website. In figure 2 is a brief description of the use case used in this research. It can be seen that every user or admin who wants to access this website is required to log in first. If not, then you will not be able to access all the pages on this website.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5210

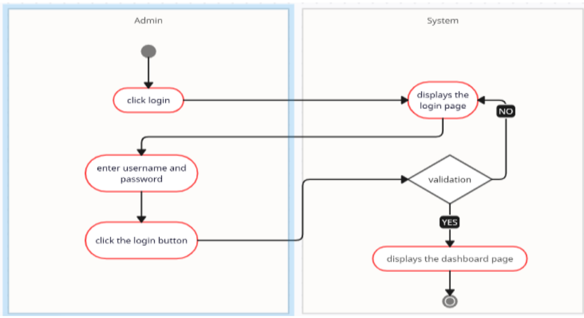



Source: (Research Results, 2024)
Figure 2. Use case diagram

## 3. Activity Diagrams

The following is an activity diagram from the website proposed Cineverse films.

a. Activity Diagram Admin Login

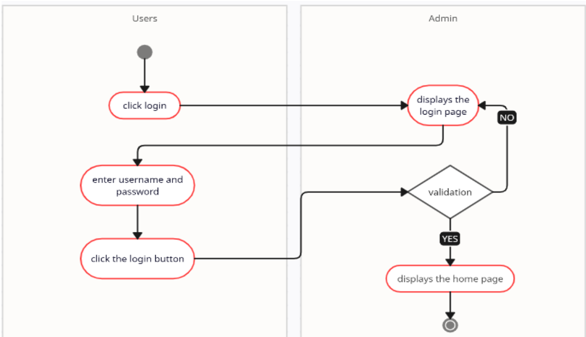



Source: (Research Results, 2024)
Figure 3. Admin login activity diagram

Figure 3 is the admin's way of logging into the website. When the admin enters the admin page, he is required to log in by entering the username and password that was entered previously. Admin does not have a registration page and must go through testing and approval.

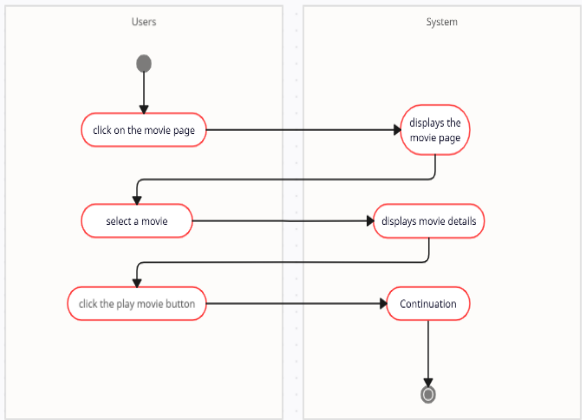
b. Activity Diagram User Login

Figure 4 is the login mechanism in the user section. When users enter a website page, they are required to log in first for authentication. If they don't have an account, they can create an account in the registration section.




Source: (Research Results, 2024)
Figure 4. User Login Activity Diagram

c. Activity Diagram Watching Movies User



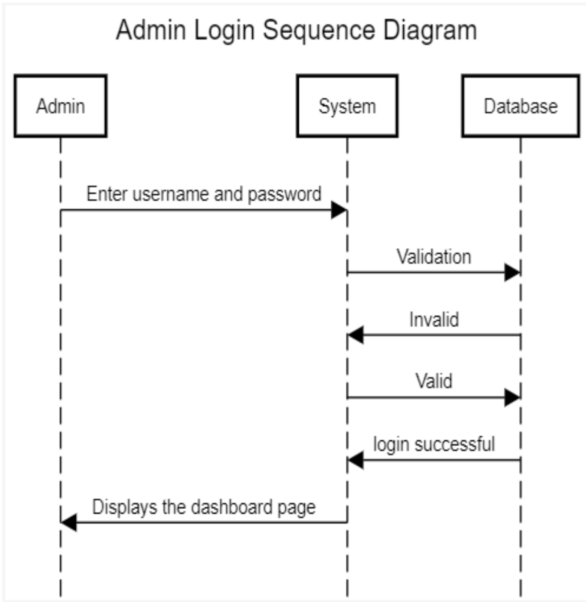

Source: (Research Results, 2024)
Figure 5. User Movie Watching Activity Diagram

Figure 5 is the mechanism when a user selects a film and wants to watch the film. When the user selects a film, the details of the film will be displayed and when pressing the play button, the system will play the film.

## 4. Sequence Diagrams

Following is the sequence diagram from the website proposed Cineverse films.

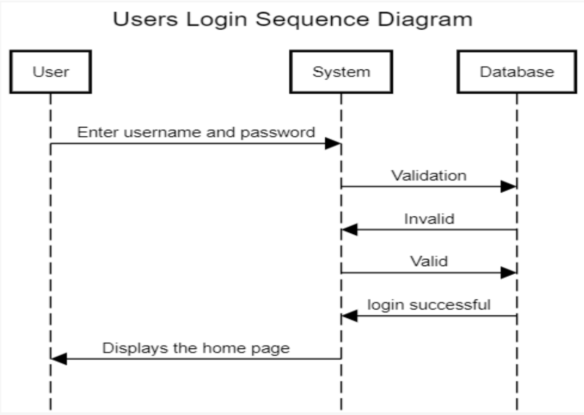
a. Admin Login Sequence Diagram




Source: (Research Results, 2024)
Figure 6. Admin Login Sequence Diagram

Figure 6 is how the system and database collaborate to verify and ensure that the username and password entered by the admin are correct with those in the database or not.

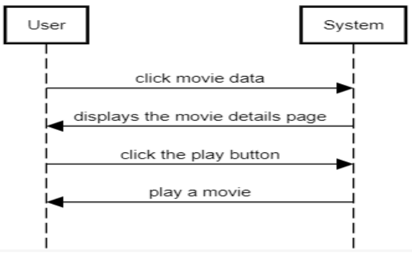
b. User Login Sequence Diagram




Source: (Research Results, 2024)
Figure 7. User Login Sequence Diagram

Figure 7 is how the system and database collaborate to verify and ensure that the username and password used by the user are correct with those in the database or not.

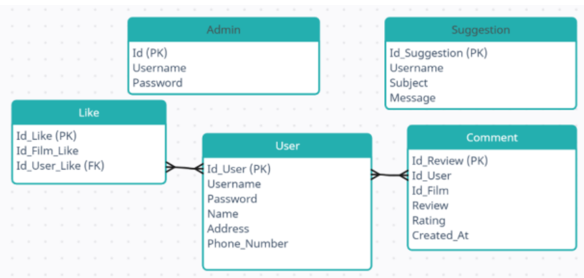
c. User Movie Playback Sequence Diagram




Source: (Research Results, 2024)
Figure 8. Sequence Diagram for Film Screening Users

Figure 8 depicts the system flow that occurs when a user selects a movie and plans to play it. In the initial stage, users can explore the list of films presented with detailed information and attractive posters. After the user selects the film of interest, the next step is to clicking the play button to start movie playback, ensuring a smooth and enjoyable viewing experience.

## 5. LRS (Logical Record Structure) Design



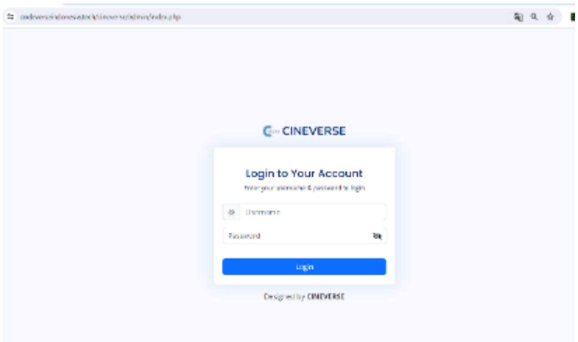

Source: (Research Results, 2024)
Figure 9. Design of LRS (Logical Record Structure)

Figure 9 shows that the user table is related to the likes and comments table, while the admin and suggestions table is not related to any table.

## 6. Implementation

After analyzing and designing the system, the next step is to implement it as a test of the program that has been created, which becomes a benchmark for further development, so that this implementation becomes a representation of all activities and scientific methods used in this research. At stage it also displays results from implementation of The Movie Database API. In this research, the author used several services provided by The Movie Database API, including:

1. Admin Login Page




Source : Source: (Research Results, 2024)
Figure 10. Admin Login Page

Figure 10 is an implementation of the login display for admin. This page contains 2 text fields and you are required to enter the correct username and password. If it is wrong, it will return to this page until the username and password provided are correct.

2. Admin Dashboard Page

Source : Source: (Research Results, 2024)
Figure 11. Admin Dashboard Page

Figure 11 is the page when the admin has entered the correct username and password. This page displays the number of visitor reviews of films

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5210

shown and watched, the number of likes or preferences for films, and the total number of users.

3. Cineverse Admin User Data Page

Source : Source: (Research Results, 2024)
Figure 12. Cineverse Admin User Data Page

Figures 12 are pages for managing users or website visitors. On this page, the admin can only delete the user and cannot create or edit users.

4. Admin Review Data Page

Source : Source: (Research Results, 2024)
Figure 13. Admin Review Data Page

Figure 1 3 is a page where the admin views or manages reviews from viewers. Admin can only delete the review if necessary.

5. Admin Like Data Page

Source : Source: (Research Results, 2024)
Figure 14. Admin Like Data Page

Figure 14 is a page where the admin manages likes given by users to a film that he thinks is good.

6. Criticism and Suggestions Data page

Source: (Research Results, 2024)
Figure 15. Criticism and Suggestions Data Page

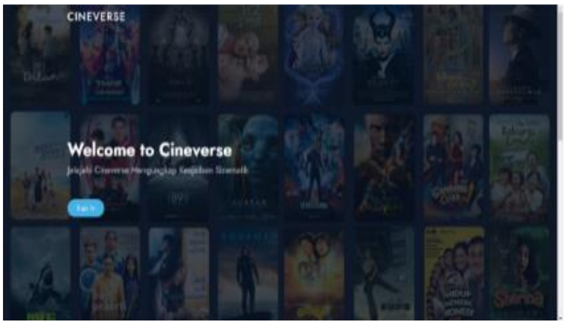
Figure 15 shows the criticism and suggestions page. This page manages all criticism and suggestions given by users via the form provided.

7. Admin Data Page

Source: (Research Results, 2024)
Figure 16. Admin Data Page

Figure 16 shows the admin data page. Here, admins can manage admin data themselves, such as deleting, editing, and even adding new admins.

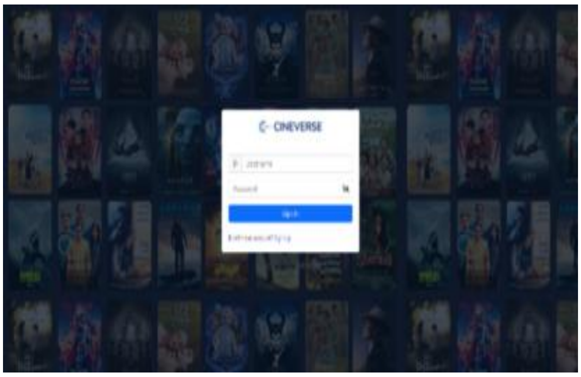
8. Landing Pages




Source: (Research Results, 2024)
Figure 17. User Landing Page

Figure 17 is the first page a user accesses the website. This is a mandatory page that you must go through if you are accessing this website for the first time.

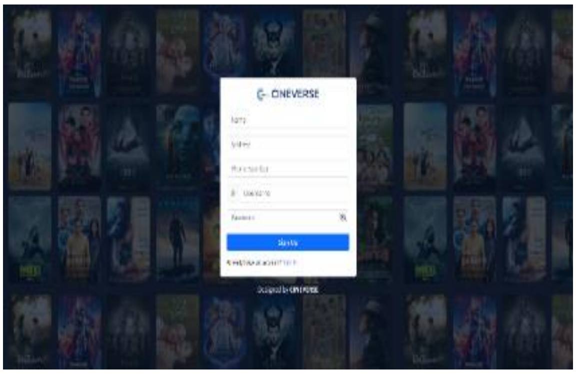
9. User Login Page




Source: (Research Results, 2024)
Figure 18. User Login Page

Figure 18 is the user login page , the page used to access access this website with contains 2 textfields that must be filled in correctly. If you don't have an account, you can register first .

10. Register Page Users

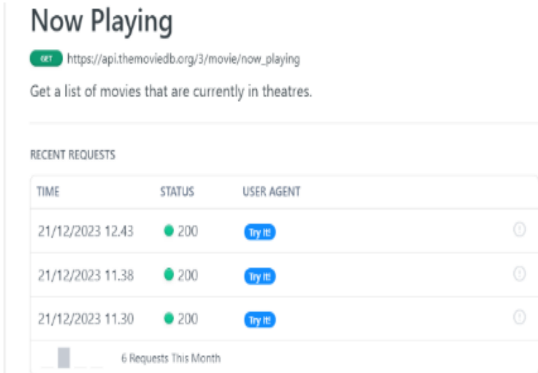



Source: (Research Results, 2024)
Figure 19 . User Registration Page

Figure 19 is the page that the user will go through if they don't have an account to log in. On this page you are required to fill in all the forms provided so that registration can be successful.

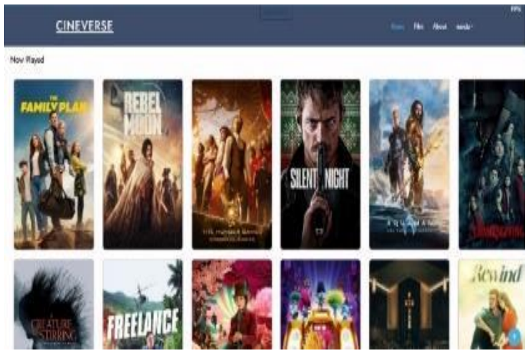
11. User Home Page

On the page Here , there are two parts main screen , namely "Now Playing" which displays the current film played moment this , and "On Trending" which displays current films trend , with both of them uses the TMDB API for serve information in a way dynamic .



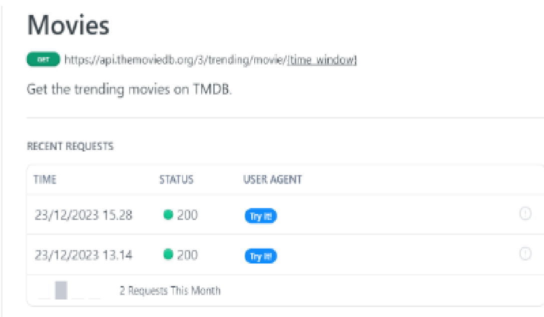

Source: (Research Results, 2024)
Figure 20 . API Now Playing

Figure 20 is an API used to call films that are currently playing in theaters.




Source: (Research Results, 2024)
Figure 21 . Home menu display Now Playing section

Figure 21 is results of API implementation on the website and can be seen by the user after logging in .




Source: (Research Results, 2024)
Figure 22 . Trending Movies API

The function of this API service is to display films that are currently trending or that have been requested a lot by other users . Figure 22 is the API used to call films that are trending at that time.

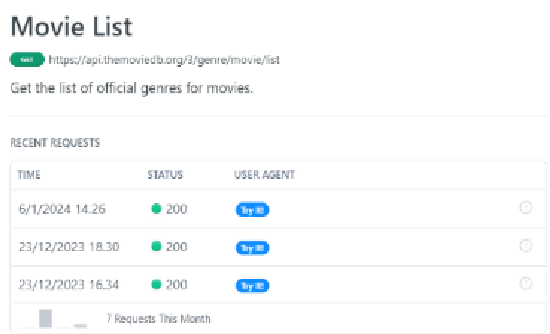
PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5210

Source: (Research Results, 2024)
Figure 23 . Home menu display On Trending section

Figure 23 is the API implementation on the existing website under "now playing".

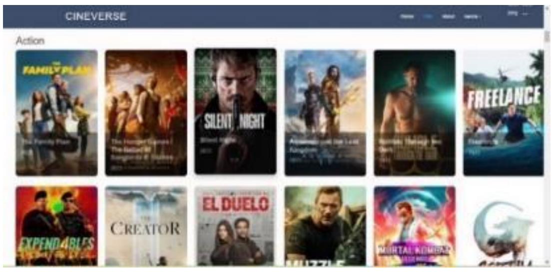
12. User Movie Page

On page here, there is appearance various films grouped based on each genre, where genre information _ obtained in a way dynamic from the TMDB API.



Source: (Research Results, 2024)
Figure 24 . Movie List API

The function of this API service is to display films based on previously entered genres. Figure 24 is the API used to call films based on their genre.

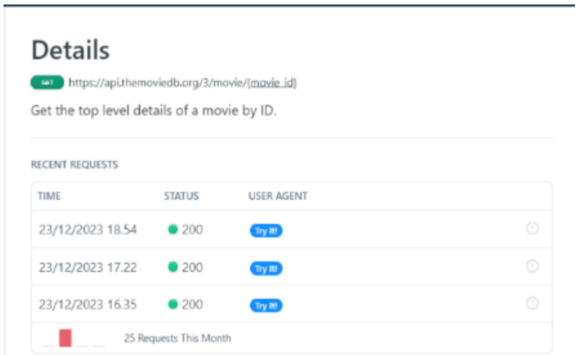


Source: (Research Results, 2024)
Figure 25 . Film Menu Website Display

Figure 25 shows the results of the API implementation derived from Figure 10, which now includes the addition of style elements to increase the attractiveness and aesthetics of the appearance.

13. User Movie Details Page

This page displays film details including banner, poster, title , year release , rating , video, synopsis , cast, film recommendations and comments accessed by users. A number of feature implemented with utilizing the API, except for ratings and comments .

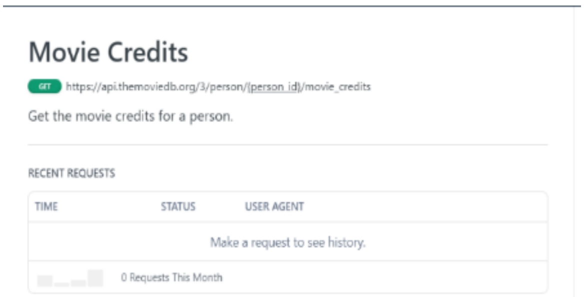


Source: (Research Results, 2024)
Figure 2 6. API Movies Details

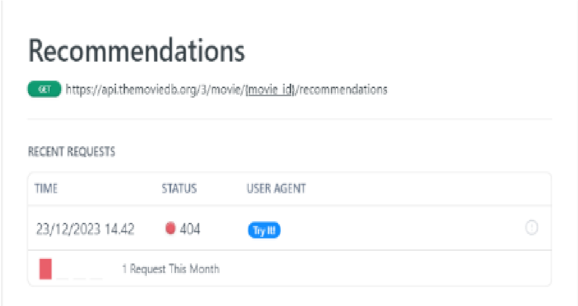
The function of this API service is to display whatever information is needed in a film. Figure 2 6 is an API that is used to display details of a film starting from the title, duration, genre, and so on.

Source: (Research Results, 2024)
Figure 2 7. View of the film details

Figure 2 7 is API implementation results on the website that will display when the user selects a movie to watch.

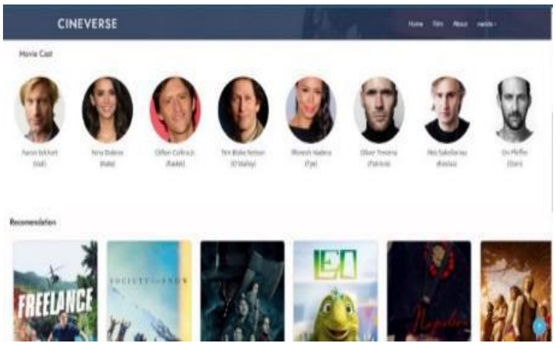


Source: (Research Results, 2024)
Figure 28. Movie Credits




Source: (Research Results, 2024)
Figure 29. Recommendations Movies

The function of this API service is to display information about Who just actor in a film. Figure 28 is the API used to display the player list a film. Figure 29 is the API used for bring up movie recommendations.




Source: (Research Results, 2024)
Figure 30. Detailed view of the film

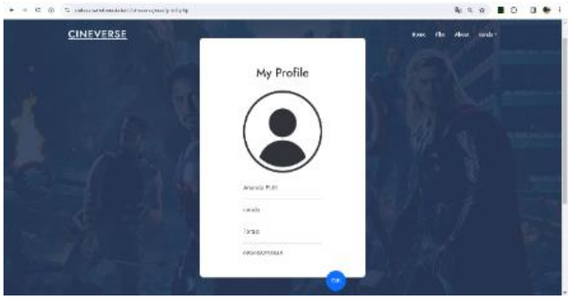
Figure 30 is API implementation results on the website that will display movie cast and recommendations.

14. About User page


Source: (Research Results, 2024)
Figure 31. About User page

Figure 31 is the about page in the about menu in the top navbar which provides at a glance information about the website.

15. User Profile Page




Source: (Research Results, 2024)
Figure 32 . User Profile Page

Figure 32 shows the edit page in the profile menu in the top navbar which can be accessed by clicking on the user's username section.

## 7. Testing Results with Black Box Testing

The results obtained in experiments using Black Box Testing can be seen in the table 1.

Table 1. Black Box Testing Results

| No | Scenario | Expected results | Conclusion |
|---|---|---|---|
| 1 | Enter the Login page by entering the wrong username and password | A warning or alert appears about an incorrect password and username | Valid |
| 2 | Enter the home page | There are 12 films on trending and 12 films now playing | Valid |
| 3 | Click on one of the films | Details of the film emerge | Valid |
| 4 | Write a comment and click send | Successfully sent to the database and appears above the comments column | Valid |
| 5 | Press the love icon below the film poster | The page will reload and the love icon will turn red | Valid |
| 6 | Select the movie menu in the navbar | Features 12 films based on genre | Valid |
| 7 | Select the About menu in the navbar | The about menu appears | Valid |
| 8 | Select a profile in the navbar | A new page appears containing the user profile | Valid |

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5210

| No | Scenario | Expected results | Conclusion |
|---|---|---|---|
| 9 | Pressing the change button | Raises capital to change user information | Valid |
| 10 | Change the information and press the change button | The user's personal information will be changed on the page and in the database | Valid |

Source: (Research Results, 2024)

## CONCLUSION

Creating a native website by utilizing a combination of PHP, HTML, CSS, and JavaScript for the client side, as well as MySQL as a database management system, forms a solid foundation for presenting dynamic and responsive content. PHP as a server-side language provides data processing power on the server, while HTML and CSS are responsible for page structure and layout. JavaScript enables enhanced interactivity, creating a more dynamic user experience. One of the important elements of this project is the API integration of The Movie Database (TMDb), which enriches the website content with up-to-date information about films. The use of APIs allows websites to automatically update movie listings, cast information, and reviews, providing a more relevant and dynamic user experience. The main functions of the website include the ability to display movie listings, provide detailed information, and display poster images, thereby providing users with a complete experience. Search and categories help users find films easily, while interactive features, such as ratings and reviews, increase user engagement. However, in implementing this project, attention to security is crucial. It requires measures such as input validation and use of bound parameters to prevent SQL injection attacks and other security measures. Routine maintenance also needs to be carried out, including performance monitoring, bug fixes, and security updates to maintain smooth operations and website security. Thus, the project of creating this website is not just about presenting film information, but also involves aspects of development, security, and maintenance to ensure its long-term success and meet user expectations.

## REFERENCE

Adhitama, A., Widiyaningsih, W., & Lailasari, M. (2022). Perancangan User Interface Pada Website Streaming Anime Menggunakan Aplikasi Figma. Jurnal Nasional Teknologi Komputer, 2(4), 159-176.

Badrul, M. (2021). Penerapan Metode Waterfall Untuk Perancangan Sistem Informasi Inventory Pada Toko Keramik Bintang Terang. PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer, 8(2), 57-52. doi: 10.30656/prosisko.v8i2.3852.

Fowler, M. (2018). UML distilled: a brief guide to the standard object modeling language. Addison-Wesley Professional.

Hjort, M. (2019). The public value of film: Moving images, health and well-being. Journal of Scandinavian Cinema, 9(1), 7-23.

Idris, M., Syarif, I., & Winarno, I. (2022). Web application security education platform based on OWASP API security project. EMITTER international journal of engineering technology, 246-261. doi: 10.24003/emitter.v10i2.705.

Karolina, C. M., Maryani, E., & Sjuchro, D. W. (2020). Implikasi genre film dan pemahaman penonton film tuna netra di “Bioskop Harewos”. ProTVF, 4(1), 123-142. doi: 10.24198/ptvf.v4i1.25035.

Muni, A., & Ihwan, K. (2021). Perangcangan Sistem Informasi Film Berbasis Web. JUTI UNISI, 5(2), 28-33.

Musthofa, N., & Adiguna, M. A. (2022). Perancangan Aplikasi E-Commerce Spare-Part Komputer Berbasis Web Menggunakan CodeIgniter Pada Dhamar Putra Computer Kota Tangerang. OKTAL: Jurnal Ilmu Komputer Dan Sains, 1(03), 199-207.

Ofoeda, J., Boateng, R., & Effah, J. (2019). Application programming interface (API) research: A review of the past to inform the future. International Journal of Enterprise Information Systems (IJEIS), 15(3), 76-95. doi: 10.4018/IJEIS.2019070105.

Paramitha, I. A. K. P., Wiharta, D. M., & Arsa, I. M. (2022). Perancangan Dan Implementasi Restful Api Pada Sistem Informasi Manajemen Dosen Universitas Udayana. Jurnal SPEKTRUM Vol, 9(3). doi: 10.24843/spektrum.2022.v09.i03.p3.

Prayoga, E., Anandita, C. A. M. R., Putri, S. A., & Sumantri, R. B. B. (2023). Rancang Bangun Aplikasi Pemesanan Tiket Bioskop Xxi Cibaduyut Berbasis Website Dengan Metode Waterfall. Jurnal Sistem Informasi Kaputama (JSIK), 7(1), 29-35. doi: 10.59697/jsik.v7i1.70.

Prima, D. A. M. (2022). Analisis Isi Film" The Platform". Journal of Digital Communication and Design (JDCODE), 1(2), 127-136.

Ridoh, A., & Putra, Y. I. (2021). Perancangan dan Implementasi Sistem Informasi Dokumen Layanan Publik Berbasis Web Untuk

Mempermudah Masyarakat Memperoleh Informasi Pada Pemerintah Kabupaten Bungo. Jurnal Basicedu, 5(5), 4227-4235. doi: 10.31004/basicedu.v5i5.1525.

Sonata, F. (2019). Pemanfaatan UML (Unified Modeling Language) dalam perancangan sistem informasi e-commerce jenis customer-to-customer. Jurnal Komunika: Jurnal Komunikasi, Media Dan Informatika, 8(1), 22-31. doi: 10.31504/komunika.v8i1.1832.

Syafi'i, M. A., & Fajarita, L. (2019). Pemodelan Sistem Informasi Persediaan Barang pada SDIT Lentera Ilmu dengan Metodologi Berorientasi Obyek. IDEALIS: InDonEsiA journaL Information System, 2(1), 166-175.

Vidal-Silva, C., Jiménez, C., Madariaga, E., & Urzúa, L. (2020). Applying PHP codeigniter for easy web development.

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5243

# CLASSIFICATION OF CUSTOMERS' REPEAT ORDER PROBABILITY USING DECISION TREE, NAÏVE BAYES AND RANDOM FOREST

**Amelia Citra Dewi[1*]; Arief Hermawan; Donny Avianto[3]**

Master of Information Technology[1,2,3]
Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia[1,2,3]
https://uty.ac.id/[1,2,3]
ameliacitradewi@gmail.com[1*]; ariefdb@uty.ac.id[2]; donny@uty.ac.id[3]
(*) Corresponding Author



***Abstract***—*Limited customer information in sales data on e-commerce in Indonesia hinders companies in determining targeted marketing strategies, especially in targeting groups of potential customers to make repeat purchases. Sales data in the form of customers' names and cellphone numbers has been hidden by e-commerce, and only data is available in the form of products purchased, number of purchases, and customer addresses. So far, the methods used to determine potential customers mostly use more complete data features. Research that uses limited e-commerce data to determine potential customers is scarce. Several algorithms for predicting repeat purchases in e-commerce also have been widely used. However, the comparison of the performance of these methods in the context of e-commerce in Indonesia with limited data has yet to be discovered. In this research, the Decision Tree, Naive Bayes, and Random Forest methods were compared to classify potential customers using Maschere brand sales data from two e-commerce sites, namely Tokopedia and Shopee. The research results show that the Decision Tree algorithm achieved an accuracy of 90.91%, Naive Bayes achieved an accuracy of 37.50%, and Random Forest achieved the best level of accuracy, namely 93.94%. These results show that the Random Forest method is the best method for classifying customers' probability of repeat purchases. In the future, the results of this research can be developed again as a decision-making system to determine potential customers.*

***Keywords:*** *customer classification, decision tree, e-commerce, naive bayes, random forest.*

**Abstrak**—Keterbatasan informasi pelanggan dalam data penjualan pada e-commerce di Indonesia menghambat perusahaan dalam menentukan strategi marketing yang terarah, terutama dalam menargetkan kelompok calon pelanggan potensial untuk melakukan pembelian berulang. Data penjualan berupa nama dan nomor handphone pelanggan sudah disembunyikan pihak e-commerce, dan hanya tersedia data berupa produk yang dibeli, jumlah pembelian, beserta alamat pelanggan. Selama ini metode yang digunakan untuk menentukan pelanggan potensial sebagian besar menggunakan fitur data yang lebih lengkap. Adapun penelitian yang menggunakan keterbatasan data e-commerce untuk menentukan pelanggan potensial sangat jarang dijumpai. Beberapa algoritma untuk memprediksi pembelian berulang di e-commerce sudah banyak digunakan, namun perbandingan performa metode tersebut dalam konteks e-commerce di Indonesia dengan keterbatasan data belum diketahui. Pada penelitian ini, dibandingkan metode Decision Tree, Naive Bayes, dan Random Forest untuk mengklasifikasikan calon pelanggan potensial dengan menggunakan data penjualan merk Maschere dari dua e-commerce, yaitu Tokopedia dan Shopee. Hasil penelitian menunjukkan algoritma Decision Tree mencapai akurasi 90.91%, Naive Bayes memiliki mencapai akurasi 37.50%, dan Random Forest mencapai tingkat akurasi yang terbaik yaitu sebesar 93.94%. Dari hasil tersebut diketahui bahwa metode Random Forest menjadi metode terbaik dalam mengklasifikasikan probabilitas pelanggan untuk pembelian berulang. Di masa mendatang, hasil penelitian ini dapat dikembangkan kembali sebagai sistem penentu keputusan untuk menentukan calon pelanggan potensial.

***Kata Kunci****: klasifikasi pelanggan, decision tree, e-commerce, naive bayes, random forest.*

## INTRODUCTION

E-commerce helps a company increase the reach and number of potential customers it can target for product sales. This is because the transaction process of exchanging goods in e-commerce involves internet and computer networks (Farras et al., 2022). E-commerce groupings are divided based on the parties involved in the transaction. Business to consumer (B2C) e-commerce occurs when transactions occur in retail, where the business consumers are individuals. Business to Business (B2B) e-commerce occurs when both parties in transactions are organizations, and Consumer to Consumer (C2C) occurs when both parties are individuals (Man, 2020).

Maschere is one of the brands that carries out B2C sales through e-commerce in Indonesia, such as Tokopedia and Shopee. In an effort to increase sales, Maschere carries out marketing strategies in the form of discounts on specific dates or promotional events run by e-commerce parties. Apart from discount prices, Maschere is also active in e-commerce marketing by implementing a cost-per-click advertising system. However, this method is considered less efficient. It does not have a long-term impact on sales because many buyers only hunt for discounts or do window shopping: browsing online stores and looking at the products on offer without making a purchase (Ma et al., 2020).

In business practice, companies use sales data to monitor sales and forecast trends. Historical sales data supports analysis to optimize prices, predict demand, manage stock, and plan marketing strategies (Chee et al., 2022; Farras et al., 2022). However, sales data from the Tokopedia e-commerce platform is currently limited. With the implementation of the buyer's data protection policy on Tokopedia and Shopee, some information on sales data is disguised (Pusat Edukasi Penjual Tokopedia, 2023; Pusat Edukasi Penjual Shopee Indonesia, 2023). This can influence the company in determining business strategy, considering that marketing strategy is critical for increasing sales (Djami et al., 2023).

Finding patterns, trends, and useful information from databases, in this case sales data, can be done using the Knowledge Discovery in Database (KDD) process. In KDD, data mining methods are used to identify patterns in data (Kotawadekar, 2022).

Classification is a technique in data mining that is used to divide data into categories, classes, or groups based on specific characteristics (Kotawadekar, 2022). In the process, classification techniques analyze data sets that already have labels to determine rules for classifying new data into existing class labels (Djami et al., 2023).

Decision Tree, Random Forest, and Naive Bayes are three algorithms that can be used in classification techniques. Decision Tree is an algorithm that uses a decision tree structure to classify objects based on a series of questions asked of the object, where each node represents a question and each branch represents the answer to that question (Charbuty & Abdulazeez, 2021; Talekar, 2020). Meanwhile, Random Forest is an algorithm that uses several decision trees to make predictions. Each decision tree is built separately from other decision trees, and the final prediction in this algorithm is made based on combining the predictions from all decision trees (Papakyriakou & Barbounakis, 2022). Naive Bayes stands out as a straightforward yet effective classification algorithm. Naive Bayes classifier assumes that each feature in the data is independent. Naive Bayes can estimate the probability that an object belongs to a particular class by multiplying the probabilities of each feature of the object (Wickramasinghe & Kalutarage, 2021).

These three algorithms were selected to explore their comparative effectiveness in accurately classifying customers' repeat order probabilities, a crucial aspect for targeted marketing strategies in e-commerce, especially in the context of limited customer information availability from e-commerce in Indonesia.

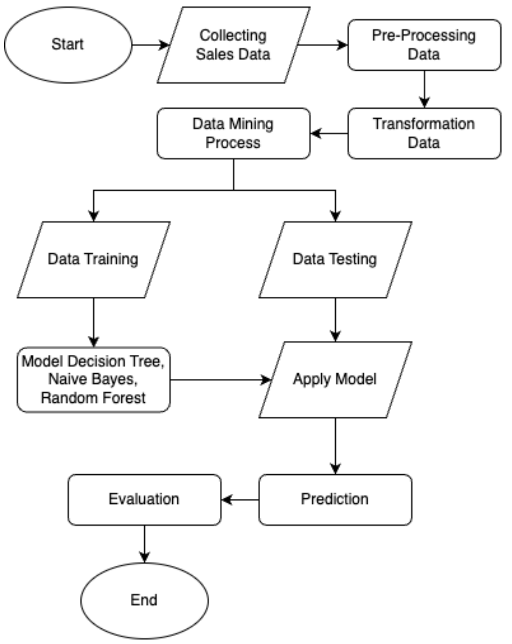
Previous research has shown how machine learning algorithms are used in predicting sales or classifying customer profiles. Naive Bayes has been used to predict the number of sales based on the best-selling to least-selling product categories with an accuracy rate of 54% (Djami et al., 2023). This algorithm is also used to classify the distribution of customer locations with an accuracy level of 92% (Putro et al., 2020). Decision Trees have been used to predict customer purchasing patterns to maintain product stock availability with an accuracy level of 98.86% (Budilaksono et al., 2021). Random Forest has also been used to predict when a customer will make their next purchase with an accuracy rate of up to 89% (M K et al., 2021). However, all of these studies use more complete datasets sourced from sample datasets from Kaggle repository (M K et al., 2021), or internal company records which do not originate from third-party data like e-commerce, and are not used to analyze the possibility of repeat orders by customers (Budilaksono et al., 2021; Djami et al., 2023; Putro et al., 2020).

This research aims to fill in possible differences in results due to differences between the data used in previous research and the current research: internal data that can be adjusted and curated by the company and e-commerce sales data that the company cannot change. In this research, a classification of potential customers in e-commerce is proposed based on the probability of repeat customer purchases. The dataset that will be used is Maschere Indonesia sales data on Tokopedia and Shopee as the third-party data providers. The three classification methods that will be used are Naive Bayes, Random Forest, and Decision Tree.

By identifying which buyers have the potential to become loyal customers and make repeat purchases, companies can target specific promotions only to potential loyal customers so that promotional costs can be more efficient in boosting sales. This research also aims to contribute to the broader field by serving as a valuable reference for similar companies that conduct sales through e-commerce platforms like Tokopedia and Shopee. This research provides insights and methodologies that can be adapted to improve sales strategies and promotional efficiency for other companies operating in the digital marketplace. This approach offers a strategic advantage by enabling more targeted marketing efforts and optimizing promotional expenditures, thereby potentially enhancing customer loyalty and sales outcomes in the competitive e-commerce landscape.

## MATERIALS AND METHODS

In this paper, Knowledge Discovery in Databases (KDD) is used in conjunction with Data Mining to discover specific patterns from data sets (Kotawadekar, 2022). The flow of the research methods used can be seen in Figure 1.




Source : (Research Results, 2024)
Figure 1. Research Method Flow Chart

In Figure 1, Sales data collection is the first step in this research, followed by data pre-processing, which involves cleaning and transforming data, data training, and data testing with data mining models such as: Decision Tree, Naive Bayes, and Random Forest, data testing, then finally evaluating classification results (Maryoosh & Hussein, 2022). The Maschere brand sales data used in this research is sales data for the period August to October 2023.

### A. Collecting Sales Data

Sales data for the Maschere brand was obtained by providing limited access to the Tokopedia and Shopee e-commerce seller dashboards. On the Tokopedia and Shopee seller dashboard pages, sales data for the period August to October 2023 can be downloaded in .csv file format.

In that period, a total of 496 records were found, of which 396 records came from sales via Tokopedia and 100 records came from Shopee sales. The downloaded data contains information in the form of Customer Name, Cell Phone Number, Email, Street Name, Province, City, Country, Date Registered as a customer based on date of first purchase, Total Orders, and Total Nominal Spending.

### B. Pre-processing Sales Data

Before it can be used for this research, sales data is then entered into the data pre-processing stage. Sales data from the two e-commerce sites is first combined into one file. Next, sales data needs to go through a data cleansing process first (Fan et al., 2021).

At this stage, unnecessary data, such as empty, incomplete, incorrect, or irrelevant data for analysis purposes, is cleaned (Fan et al., 2021). This is done to reduce noise or interference with research results due to inaccurate data (Arhami & Nasir, 2020). Empty column data without any information such as: Email and Street Name are removed from the dataset.

After the pre-processing stage, the sales data is selected into 10 features. These features are determined by their contribution to the research objective. It was critical to perform machine learning with these attributes to capture customer purchasing behavior dynamics, which may influence the repeat order probability. As displayed in Table 1, “Name” and “Phone Number” represent the customer’s name and phone number, “Province” is the customer’s originating province, “City” is the customer’s originating city, “Country” is the customer’s originating country, “Date Regist.” is the date registration when the customer made first purchased at Maschere through e-commerce and made another purchased between August to

October 2023, “Total Orders Shopee” and “Total Orders Tokopedia” represent the number of times a customer has placed orders on each e-commerce platform, “Total Amount” is the total amount of the customer’s purchases across all e-commerce platforms, and “Cust. Category” is the label for customer categorization.

Table 1. Maschere’s Sales Data After Pre-Processing

| Name | Phone Number | Province | City | Country | Date Regist. | Total Orders Shopee | Total Orders Tokopedia | Total Amount | Cust. Category |
|---|---|---|---|---|---|---|---|---|---|
| Digda | *********0451 | Jawa Barat | Kota Bekasi | Indonesia | 2023-05-31 16:32:37 | 0 | 2 | 291300 | Regular |
| M******i | ******47 | Banten | Kab. Pandeglang | ID | 2023-06-01 01:39:39 | 9 | 0 | 1919093 | Loyal |
| L******i | ******46 | Jawa Barat | Kota Bekasi | ID | 2023-06-22 19:56:31 | 2 | 0 | 64382 | Non-commited |
| Albertus | *********4987 | D.I. Yogyakarta | Kab. Sleman | Indonesia | 2023-05-31 16:32:26 | 0 | 1 | 139000 | Regular |
| D**n | ******00 | Jambi | Kota Jambi | ID | 2023-06-01 01:40:01 | 20 | 0 | 1191691 | Occasionals |

Source : (Research Results, 2024)

Based on Maschere sales data, Table 2 shows that 8.47% are loyal customers, 14.93% are regular customers, 55.85% are occasional customers, and 20.77% are non-attached customers.

Table 2. Customer Ratio Data by Categories

| Cust. Category | Data Ratio | Number of Customer |
|---|---|---|
| Loyal | 8.47 % | 42 |
| Regular | 14.92 % | 74 |
| Occasional | 55.85 % | 277 |
| Non-Commited | 20.77 % | 103 |

Source : (Research Results, 2024)

## C. Transformation Sales Data

Classification of potential customers will be carried out using RapidMiner Studio Version 10.3 software running on the Linux Debian 12 operating system. The distribution of sales data is 80% for training data and 20% for test data. Customer status in sales data will be converted into 4 class labels: Level 1 for loyal customers, Level 2 for Regular customers, Level 3 for occasional customers, and Level 4 for unattached customers.

## D. Data Mining Methods

In this research, the data mining classification algorithms that will be used are Decision Tree, Naive Bayes, and Random Forest. Although there are many other algorithms in machine learning, limiting the comparison to these three algorithms is also influenced by time and resource constraints, enabling a comprehensive and focused comparison.

Decision Tree is an algorithm that uses a decision tree structure to classify objects based on a series of questions asked of the object, where each node represents a question and each branch represents the answer to that question (Charbuty & Abdulazeez, 2021).

Naive Bayes is a classification algorithm based on the Bayes Theorem, which uses the assumption of independence between features in expressing posterior probability relationships (Papakyriakou & Barbounakis, 2022).

$$P(c|B) = \frac{P(c \cap B)}{p(B)} = \frac{P(c).P(B|c)}{P(B)} \qquad (1)$$

where $P(c|B)$ is the probability that data $B$ is included in class $c$, $P(c)$ is the probability that data is included in class, and $P(B/c)$ is the probability that data $B$ has features that correspond to class $c$.

Meanwhile, in the Random Forest algorithm, object classification is carried out using several decision trees, which are built separately from other decision trees to avoid overfitting. The final classification results in the Random Forest model are made based on combining all decision trees and using the Gini Index indicator (Papakyriakou & Barbounakis, 2022).

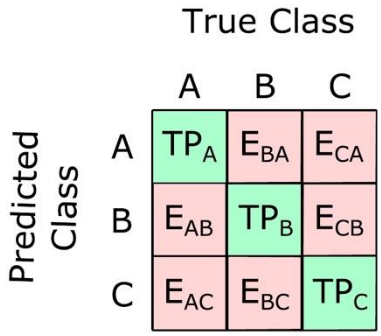
$$G = 1 - \sum_{i=1}^{c} (p(i))^2 \qquad (2)$$

where C represent the total number of classes, and $p(i)$ denote the probability of selecting a data point belonging to class $i$.

The classification results obtained through the Decision Tree, Naive Bayes, and Random Forest algorithms will then be evaluated using the confusion matrix table (Tharwat, 2021). In Figure 2,

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5243

$TP_A$ is the amount of class A data that is correctly predicted as class A. $E_{AB}$ is the amount of class A data that is incorrectly classified as class B, and so on (Tharwat, 2021).




Source : (Tharwat, 2021)

Figure 2. Confusion Matrix for a Multi-Class Classification

The model test results can be derived from the information in the confusion matrix table, including accuracy, precision, recall, and F1-Score values. Accuracy is a performance measure that reflects the percentage of data that the model predicts correctly. Precision assesses the ratio of accurate positive predictions relative to all positive predictions, while recall evaluates the proportion of correctly predicted positive data among all actual positive data. Meanwhile, F1-Score is used to measure classification performance by combining precision and recall values (Varoquaux & Colliot, 2023).

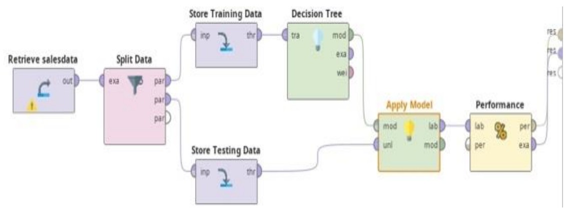
The selection of Accuracy, Precision, Recall, and F1 Score as evaluation indicators in this research is based on their relevance and effectiveness in measuring the performance of classification models. The chosen four indicators represent critical aspects of model performance: overall prediction accuracy (Accuracy), the model's ability to accurately identify positive instances (Precision), the model's ability to find all actual positive cases (Recall), and the harmony between Precision and Recall (F1 Score). These four indicators provide a comprehensive view of the model's quality in the context of this research, allowing researchers and practitioners to understand the model's capability in accurately predicting customer repeat purchases. Although other indicators could add to the understanding of model performance, combining these four indicators is sufficient to provide a robust and representative evaluation of the research objectives.

By carrying out an evaluation, you can find out which algorithm produces the best output so that it can be used to determine potential customers for the Maschere brand.

## RESULTS AND DISCUSSION

In this research, three main experiments will be carried out to determine the classification of potential customers based on repeat order probability using Decision Tree, Naive Bayes, and Random Forest algorithms. Figure 3 displays the operator architecture and functions used in this research. There are several operators used in training and testing data, such as (RapidMiner, 2023):

a. The Retrieve operator is used to import sales data.
b. The Split Data operator splits imported sales data into two data subsets. In this research, the data split is divided into 80:20 for training and test data.
c. Store Data Operators are used to store sales data that has been split into a database.
d. Algorithm models are used to determine algorithms in learning training data for classification and prediction.
e. The Apply Model operator is used to apply the training model to test data to produce classifications or predictions.
f. The Performance Classification operator is used to evaluate model performance results on test data.




Source : (Research Results, 2024)

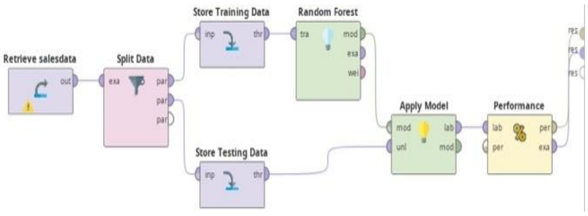
Figure 3. Training and Testing Data Process with Decision Tree Algorithm

In Figure 3, the training data is trained using the Decision Tree algorithm. The training is output as a model, which test data can then use. After the training model is applied to the test data, performance evaluation can then be carried out to check the accuracy of the test data's results.

In the first experiment, the Decision Tree model set the depth value to a maximum of 10, split the data 80:20, and did not use pruning or pre-pruning. In terms of performance results, the Decision Tree algorithm shows an accuracy level of 90.91%, a precision of 94.05%, a recall of 88.33%, and an F1 score of 0.91.

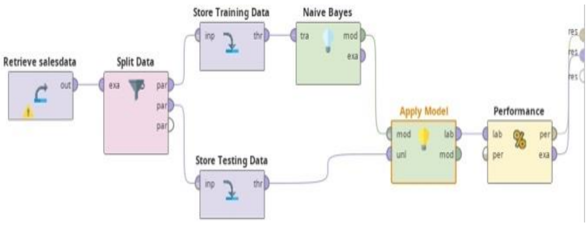
The second experiment, seen in Figure 4, uses a Random Forest model with the depth value set to a maximum of 10, split data 80:20, and does not use pruning. The performance results of the Random Forest algorithm show an accuracy level of

93.94%, precision of 95.14%, recall of 91.90%, and F1 Score of 0.93.




Source : (Research Results, 2024)
Figure 4. Training and Testing Data Process with Random Forest Algorithm

In the third experiment depicted in Figure 5, the Naive Bayes algorithm was applied with an 80:20 data split. The outcomes indicate an accuracy rate of 36.36%, precision reaching 51.70%, recall at 47%, and an F1 Score of 0.49.




Source : (Research Results, 2024)
Figure 5. Training and Testing Data Process with Random Forest Algorithm

Based on these three experiments, classification using Maschere sales data shows that the Random Forest algorithm has the highest level of accuracy compared to other algorithms, namely 93.94%. Then followed by the Decision Tree algorithm, which achieved an accuracy level of 87.88%, and Naive Bayes with the lowest accuracy level of 36.36%. The higher the accuracy level, the higher the correct model prediction value.

Table 3. Data Test Results with Naive Bayes, Decision Tree and Random Forest Algorithms

| Algorithms | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Naive Bayes | 36.36 % | 51.70 % | 47.00 % | 0.49 |
| Decision Tree | 90.91 % | 94.05 % | 88.33 % | 0.91 |
| **Random Forest** | **93.94 %** | **95.14 %** | **91.90 %** | **0.93** |

Source : (Research Results, 2024)

The Random Forest algorithm yielded the highest precision, recall, and F1 scores among the tested models. This shows that the Random Forest algorithm has good performance in predicting classes.

There are many reasons why the Random Forest algorithm can have a better level of accuracy than Decision Tree. Random Forest can produce more accurate models than Decision Trees for various types of data, including complex data. By building several decision trees randomly, Random Forest can also provide more accurate prediction results by combining the prediction results from each decision tree.

Naive Bayes provides results that are much lower in accuracy compared to other algorithms. In the fourth experiment, training and retesting were carried out on the Naive Bayes algorithm with several types of composition of training data and test data. Split data is carried out with a ratio of 70:30, 60:40, 50:50, 40:60, and 30:70 for training data and test data respectively. Table 4 displays the results of the data testing.

Table 4. Naive Bayes Data Test Results with Different Split Data Composition

| Split Data | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 90:10 | 30.61 % | 59.03 % | 43.93 % | 0.50 |
| 70:30 | 26.17 % | 38.95 % | 34.34 % | 0.37 |
| 60:40 | 25.13 % | 40.59 % | 34.21 % | 0.37 |
| **50:50** | **37.50 %** | **47.96 %** | **44.92 %** | **0.46** |
| 40:60 | 27.27 % | 38.45 % | 37.71 % | 0.48 |
| 30:70 | 21.61 % | 39.83 % | 29.23 % | 0.34 |

Source : (Research Results, 2024)

Even though several different data split compositions have been carried out, the Naive Bayes algorithm was only able to achieve the highest accuracy of 37.50% at a 50:50 data split composition. Even when using a 90:10 data split, the accuracy only reached 30.61%. This can be caused by imbalanced data, where the number of comparisons between classes in Maschere sales data is not evenly distributed. Higher precision in the 90:10 split data does not correspond to higher overall accuracy, indicating that Naive Bayes has a known weakness when predicting the majority class in imbalanced datasets. Because the Naive Bayes algorithm assumes independence for each attribute, it tends to predict the majority class and is not suitable for use for all types of data.

In the next stage, testing the Decision Tree and Random Forest training models was carried out by adding pruning settings. In the fifth experiment, Table 5 and Table 6 show the results of testing the Decision Tree and Random Forest algorithms, which are adjusted with additional pruning at Confidence levels ranging from 0.1 to 0.5.

Table 5. Decision Tree Algorithm Data Test Results with Pruning

| Pruning | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 0.1 | 88.89 % | 90.61 % | 87.42 % | 0.89 |
| **0.2** | **90.91 %** | **94.05 %** | **88.33 %** | **0.91** |
| **0.3** | **90.91 %** | **94.05 %** | **88.33 %** | **0.91** |
| 0.4 | 88.89 % | 90.61 % | 87.42 % | 0.89 |
| 0.5 | 88.89 % | 90.61 % | 87.42 % | 0.89 |

Source : (Research Results, 2024)

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5243

Table 6. Random Forest Algorithm Data Test Results with Pruning

| Pruning | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 0.1 | 91.92 | 91.73 | 91.00 | 0.91 |
| **0.2** | **93.94** | **95.14** | **91.90** | **0.93** |
| **0.3** | **93.94** | **95.14** | **91.90** | **0.93** |
| 0.4 | 91.92 | 91.73 | 91.00 | 0.91 |
| 0.5 | 91.92 | 91.73 | 91.00 | 0.91 |

Source : (Research Results, 2024)

Based on the results of the fifth experiment, data testing using the Decision Tree and Random Forest algorithms with pruning did not significantly impact increasing accuracy. The accuracy, recall, precision, and F1-Score values of both Decision Tree and Random Forest algorithms with pruning settings remain the same as for tests without pruning. This condition is caused by Maschere's sales data needing to be more balanced. The model accuracy level also decreased in the Decision Tree algorithm to 88.89% and in the Random Forest to 91.92%, which could be caused by pruning, which removed too many decision tree branches. This causes the model to be too simple and unable to capture data patterns well. In this case, the use of pruning can have a significant impact if used on data with overfitting conditions or more complex data.

## CONCLUSION

In this research, the Random Forest method shows better performance compared to the Decision Tree and Naive Bayes methods in classifying customers based on the probability of repeat orders. The high level of accuracy, namely 93.94%, indicates that the Random Forest method can be used as an effective tool in identifying potential customers. This is particularly important as limited customer information in e-commerce sales data poses an obstacle to determining effective marketing strategies. Although this research was conducted using sales data from the Maschere brand on two e-commerce sites in Indonesia, namely Tokopedia and Shopee, its results can provide general insights useful for other companies selling via e-commerce to optimize their marketing strategies. Sales data stored in a database can be a valuable source of information for predicting sales trends, optimizing prices, managing stock, and planning marketing strategies. This research lays the groundwork for further development in the form of a decision-making system to support companies in better identifying and targeting potential customers in the future. However, the results of this research are contextual and can be influenced by factors that are not included in the analysis model, such as economic factors and changes in market trends. Further research and regular data updates are recommended to ensure the continued success of potential customer classification.

## REFERENCE

Arhami, M., & Nasir, M. (2020). *Data Mining—Algoritma dan Implementasi*. Penerbit Andi.

Budilaksono, S., Jupriyanto, J., Suwarno, M. A. S., Suwartane, I. G. A., Azhari, L., Fauzi, A., … Effendi, M. S. (2021). Customer Profilling for Precision Marketing using RFM Method, K-MEANS algorithm and Decision Tree. *SinkrOn*, *6*(1), 191–200. https://doi.org/10.33395/sinkron.v6i1.11225

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Chee, C. C. F., Leng Chiew, K., Sarbini, I. N. B., & Jing, E. K. H. (2022). Data Analytics Approach for Short-term Sales Forecasts Using Limited Information in E-commerce Marketplace. *Acta Informatica Pragensia*, *11*(3), 309–323. https://doi.org/10.18267/j.aip.196

Djami, A. S. M., Utami, N. W., & Paramitha, A. I. I. (2023). The Prediction Of Product Sales Level Using K-Nearest Neighbor and Naive Bayes Algorithms (Case Study: PT Kotamas Bali). Jurnal Pilar Nusa Mandiri, 19(2), 77-84. https://doi.org/10.33480/pilar.v19i2.4420

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*, 652801. https://doi.org/10.3389/fenrg.2021.652801

Farras, M., Friscily, F., Gabriela, G., Sera, S., Sherinne, S., & Mulyawan, B. (2022). *Implementation of Big Data in E-Commerce to Improve User Experience:* Presented at the 3rd Tarumanagara International Conference on the Applications of Social Sciences and Humanities (TICASH 2021), Jakarta, Indonesia. Jakarta, Indonesia. https://doi.org/10.2991/assehr.k.220404.326

Kotawadekar, M. S. V. (2022). Data Mining and Knowledge Discovery Process. *International Journal of Creative Research Thoughts*, *10*(11).

M K, S., Gupta, R., & Gupta, K. (2021). Predicting Customers' Next Order. *International*

*Journal of Engineering Research and Technology*, *10*(6), 418–421. https://doi.org/10.17577/IJERTV10IS060166

Ma, L., Zhang, X., Ding, X., & Wang, G. (2020). How Social Ties Influence Customers' Involvement and Online Purchase Intentions. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(3), 395–408. https://doi.org/10.3390/jtaer16030025

Man, L. (2020). Comparison between Community Shopping Mode and Traditional Cross-border E-commerce Modes. *International Journal of Scientific Engineering and Science*, *4*(3), 8–10. https://doi.org/10.5281/zenodo.3749669

Maryoosh, A. A., & Hussein, E. M. (2022). A Review: Data Mining Techniques and Its Applications. *International Journal of Computer Science and Mobile Applications*, *10*(3), 1–14. https://doi.org/10.47760/ijcsma.2022.v10i03.001

RapidMiner. (2023, October). Operator Manual—RapidMiner Documentation. Retrieved January 16, 2024, from https://docs.rapidminer.com/latest/studio/operators/

Papakyriakou, D., & Barbounakis, I. S. (2022). Data Mining Methods: A Review. *International Journal of Computer Applications*, *183*(48), 5–19. https://doi.org/10.5120/ijca2022921884

Pusat Edukasi Penjual Tokopedia. (2023, 14 Desember). Penerapan Kebijakan Pelindungan Data Pribadi Pembeli. Retrieved January 16, 2024, from Pusat Seller website: https://seller.tokopedia.com/edu/kebijakan-data-pembeli/kebijakan-data-pembeli

Pusat Edukasi Penjual Shopee Indonesia. (2023, March 20). Perubahan Tampilan Informasi Pembeli. Retrieved January 16, 2024, from https://seller.shopee.co.id/edu/article/14910

Putro, H. F., Vulandari, R. T., & Saptomo, W. L. Y. (2020). Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan. *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, *8*(2). https://doi.org/10.30646/tikomsin.v8i2.500

Talekar, B. (2020). A Detailed Review on Decision Tree and Random Forest. *Bioscience Biotechnology Research Communications*, *13*(14), 245–248. https://doi.org/10.21786/bbrc/13.14/57

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Varoquaux, G., & Colliot, O. (2023). Evaluating Machine Learning Models and Their Diagnostic Value. In O. Colliot (Ed.), *Machine Learning for Brain Disorders* (pp. 601–630). New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-3195-9_20

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5307

# IDENTIFICATION OF POTATO LEAF DISEASES USING ARTIFICIAL NEURAL NETWORKS WITH EXTREME LEARNING MACHINE ALGORITHM

**Moh. Erkamim[1*]; Ri Sabti Septarini[2]; Mursalim Tonggiroh[3]; Siti Nurhayati[4]**

Smart City Information Systems Study Program[1]
Universitas Tunas Pembangunan, Surakarta, Indonesia[1]
https://utp.ac.id/[1]
erkamim@lecture.utp.ac.id[1*]

Informatics Engineering Study Program[2]
Universitas Muhammadiyah Tangerang, Tangerang, Indonesia[2]
https://umt.ac.id/[2]
risabtis@ft-umt.ac.id[2]

Information Systems Study Program[3,4]
Universitas Yapis Papua, Jayapura, Indonesia[3,4]
https://uniyap.ac.id/
mursalim.t@gmail.com[3], nurhayatist.siti21@gmail.com[4]
(*) Corresponding Author



**Abstract**—Potato plants have an important role in providing a source of carbohydrates for society. However, potato production is often threatened by various plant diseases, such as leaf disease, which can cause a decrease in yields. Identification of diseases on potato leaves is currently mostly done by farmers manually, so it is not always efficient and accurate. So the aim of this research is to identify diseases on potato leaves with artificial neural networks using the ELM (Extreme Learning Machine) approach and the GLCM (Gray Level Co-Occurrence Matrix) method for feature extraction. The GLCM approach functions to obtain texture features on objects by measuring how often certain pairs of pixel intensities appear together at various distances and directions in the image. Meanwhile, the ELM algorithm is used for image identification by adopting a one-time training method without iteration, which involves randomly determining weights and biases in hidden layers, thus allowing training to be carried out quickly and efficiently. Evaluation of the model by looking for the level of accuracy produces a value of 84.667%. The results show that the model developed is capable of accurate identification.

**Keywords:** artificial neural networks, ELM, extreme learning machine, GLCM, potato leaf disease.

***Abstrak****—Tanaman kentang mempunyai peranan penting dalam menyediakan sumber karbohidrat bagi masyarakat. Namun produksi kentang seringkali terancam oleh berbagai penyakit tanaman, seperti penyakit daun yang dapat menyebabkan penurunan hasil. Identifikasi penyakit pada daun kentang saat ini banyak dilakukan petani secara manual sehingga tidak selalu efisien dan akurat. Sehingga tujuan dari penelitian ini adalah mengidentifikasi penyakit pada daun kentang dengan jaringan syaraf tiruan menggunakan pendekatan ELM (Extreme Learning Machine) dan metode GLCM (Gray Level Co-Occurrence Matrix) untuk ekstraksi fitur. Pendekatan GLCM berfungsi untuk memperoleh fitur tekstur pada objek dengan mengukur seberapa sering pasangan intensitas piksel tertentu muncul bersamaan pada berbagai jarak dan arah pada gambar. Sedangkan algoritma ELM digunakan untuk identifikasi citra dengan mengadopsi metode pelatihan satu kali tanpa iterasi, yang melibatkan penentuan bobot dan bias pada lapisan tersembunyi secara acak, sehingga memungkinkan pelatihan dilakukan dengan cepat dan efisien. Evaluasi model dengan mencari tingkat akurasi menghasilkan nilai sebesar 84,667%. Hasilnya menunjukkan bahwa model yang dikembangkan mampu melakukan identifikasi secara akurat.*

***Kata Kunci****: jaringan syaraf tiruan, ELM, extreme learning machine, GLCM, penyakit daun kentang.*

## INTRODUCTION

Agriculture has a crucial role in meeting global food needs. One plant that has an important role is the potato plant (*Solanum tuberosum*), which is the main source of carbohydrates for people in various parts of the world. Potato production in Indonesia is increasing gradually from year to year. This increasing market demand for potatoes motivates farmers and producers to increase production to meet market demand. This is proven by potato production in Indonesia in 2022 reaching 1.5 million metric tons, an increase of 10.5% compared to the previous year (Armavillia, 2023). However, potato crop production is often threatened by various diseases, which can cause a reduction in yield and quality. Diseases on potato leaves are one of the main problems faced by farmers. Identification of diseases on potato leaves is currently mostly done manually by farmers, which may not always be efficient and accurate. Usually, identifying diseases on potato leaves is done by looking at the characteristics of the leaves. However, with the large number of potato plants and the large area of land, it is difficult for farmers to identify diseases on potato leaves. Quick and accurate disease identification is a key step in effective crop management. In this context, a digital image processing approach can be applied to solve the problem of identifying diseases on potato leaves.

The image processing process involves a series of techniques to improve quality, extract information, and understand image structure (Rao et al., 2021). Several studies have implemented image processing on plant diseases with leaf images using various algorithms. There is research regarding the identification of Siamese orange plant diseases by applying the K-Nearest Neighbor (KNN) method (Ariesdianto et al., 2021). This research produces an accuracy of 70%, and this approach carries out classification that works based on the distance principle. However, the KNN algorithm is sensitive to outliers in the dataset and can be computationally intensive when using large data. Further research regarding the classification of diseases in rice leaves uses the Support Vector Machine (SVM) approach (Prastyo et al., 2023). The model built produces an accuracy of up to 80% by carrying out classification by searching for optimal hyperplanes that maximize the margin between different classes. However, the SVM approach is sensitive to parameter selection, which can affect performance. The next research on identifying diseases on grape leaves uses the backpropagation artificial neural network approach (Ansah et al., 2022). This approach is able to produce accuracy values of up to 82%. The Backpropagation artificial neural network algorithm is able to identify images through a series of hidden layers; this network produces output that reflects the class or label that corresponds to the image. However, the Backpropagation algorithm experiences convergence problems, causing the training process to potentially slow down or become trapped in a suboptimal local minimum.

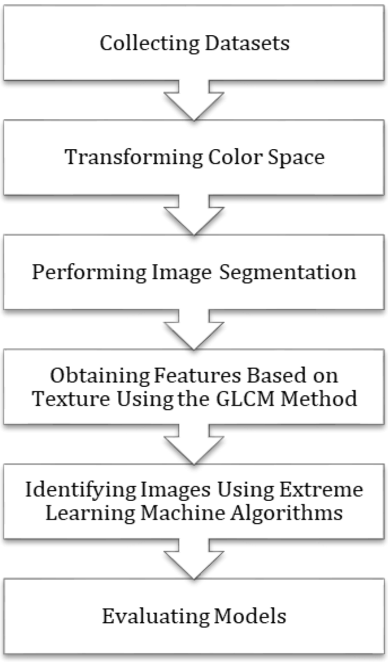
The gap obtained from previous research is that the image identification process requires a fast and effective approach so that it can reduce the computational burden of the process. In the forthcoming study, the identification process will be conducted employing artificial neural networks. This choice is predicated on the recognition of artificial neural networks as an effective method for image identification or classification, owing to their ability to mimic the functionality of human neural networks (Herdiansah et al., 2022). The researchers used artificial neural networks and applied the ELM (Extreme Learning Machine) algorithm to conduct the study. The ELM algorithm is known for its ability to process information quickly and efficiently, so it can be useful in classifying plant diseases with a high level of accuracy (Mayatopani et al., 2021). This fast initialization process allows ELM to skip the iterative and time-consuming training phase typically encountered in traditional neural networks (Wahid et al., 2021). Next, ELM evaluates the model using a closed analytical solution to determine the optimal weights in the output layer without the need for iterative adjustments. For this case study, the leaf characteristics extracted are texture features. The texture feature used is the Gray Level Co-Occurrence Matrix (GLCM), where this approach is used to describe the spatial relationship between image pixel intensities. Extracting features from GLCM provides important information about image structure and texture, enhancing the system's ability to recognize patterns and objects in images (Andrian et al., 2020).

The objective of this study is to discern diseases affecting potato foliage through the utilization of artificial neural networks employing the Extreme Learning Machine (ELM) approach alongside the Gray-Level Co-occurrence Matrix (GLCM) method for feature extraction. It is hoped that the application of this technological framework can lead to the creation of a system that is able to recognize disease quickly and precisely. Therefore, the primary contribution of this investigation lies in the refinement of the ELM algorithm, enabling proficient classification of potato leaf diseases, coupled with furnishing a comprehensive comprehension of these maladies through a meticulous analysis of the traits exhibited by infected foliage.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5307

## MATERIALS AND METHODS

To carry out research, research stages are required, which contain a series of steps or processes carried out in order to carry out research. The stages implemented in this research are illustrated in Figure 1.




Source : (Research Results, 2024)
Figure 1. Research Steps

The stages shown in Figure 1 explain the research steps, which contain the methods used to solve the problem. Further details regarding the research steps are explained as follows:

### 1. Collecting Datasets

A dataset refers to an organized collection of data, whether in the form of tables, files, or other data structures, used for analysis, research, or machine learning. Datasets in digital image processing refer to collections of image data that are used to train and test algorithms or models for various image processing tasks (Muraina, 2022). The goal of this is to provide enough variation in image conditions so that the model can learn and work well on new data that has never been seen before (Ahmad et al., 2022). The dataset used in this research contains images of potato leaves collected directly using a camera with the same light level. The classes used consist of 3 classes, namely healthy leaves, early blight disease, and late blight disease. A total of 600 images were collected to create the dataset. The data obtained is then distributed into the data used for training and testing with a distribution percentage of 70% and 30%. This division is based on the use of data that is not large, so most of the data is used for training to create more relevant learning patterns (Muraina, 2022). So, we obtained training data for 450 images and testing data for 150 images.

### 2. Transforming Color Space

Color space transformation refers to the conversion of color values from one color space system to another. This is done to make it easier to represent images and obtain the required information. The color space transformation carried out is from the Red-Green-Blue (RGB) image into the CIE L* a* b* color format, also known as CIELAB. The CIELAB color system was designed to provide color representation that is more consistent and more closely related to human perception of color (Malounas et al., 2024). The CIELAB color space transformation offers several advantages, primarily because it better reflects the way the human eye sees and perceives color (Baek et al., 2022). This color space consists of three main components: L* (brightness), a* (green and red parts), and b* (blue and yellow parts). Additionally, the CIELAB color space transformation offers color universality, where differences in the same color are considered uniform across the color spectrum. The values of L*, a* and b* can be obtained by equations (1), (2) and (3).

$$L^* = 116\left(\frac{Y}{Y_n}\right)\frac{1}{3}16\ for\ \frac{Y}{Y_n} > 0.008856 \text{......................} (1)$$

$$a^* = 500\left(f\left(\frac{X}{X_n}\right)\right) - f\left(\frac{Y}{Y_n}\right) \text{................................} (2)$$

$$b^* = 200\left(f\left(\frac{Y}{Y_n}\right)\right) - f\left(\frac{Z}{Z_n}\right) \text{................................} (3)$$

### 3. Performing Image Segmentation

Image segmentation is an important step in image processing that deals with dividing or separating an image into different segments based on related properties or attributes (Miao et al., 2023). In this case study, the desired object is separated from the background using the K-Means Clustering approach. This method divides the image into k groups, or clusters, where each pixel is grouped into the cluster that has the average intensity that is closest to the intensity of that pixel (Javidan et al., 2023). To obtain groups through the cluster process in K-Means Clustering, it can be calculated using equation (4).

$$\overline{V_{ij}} = \frac{1}{N}\sum_{k=0}^{n_i} x_{kj} \text{.......................................................} (4)$$

where $\overline{V_{ij}}$ refers to the center of cluster $i$ in attribute $j$, ni refers to the number of groups in cluster $i$, and $x_{kj}$ refers to the value $k$ in attribute $j$.

**4. Obtaining Features Based on Texture Using the GLCM Method**

Feature extraction is a procedure that seeks to extract significant data or distinctive attributes from a picture (Borman et al., 2023). The main information in this case study is the texture characteristics of the object, which are represented by the texture feature. One of the approaches used to obtain the texture characteristics of popular objects is the GLCM (Gray Level Co-Occurrence Matrix). The GLCM approach measures the relative appearance of pairs of gray levels that are close to each other in an image (Wanti et al., 2021). This process involves the formation of a matrix that represents the spatial distribution of gray between pixels (Andrian et al., 2020). In each matrix cell, an entry describes the frequency of occurrence of a particular pair of gray values in a particular direction.

Extraction parameters from GLCM, such as contrast, correlation, energy, and homogeneity, provide very useful information for obtaining information about image texture. The contrast parameter is used to measure the contrast between high and low intensity. High contrast values indicate sharp differences between pixel intensities. To produce a contrast value, it can be obtained through equation (5).

$$Contrast = \sum_i \sum_j (i-j)^2 pd(i,j) \qquad (5)$$

The correlation parameter is used to measure the extent to which pixels that have the same or different intensities are correlated in one particular direction. The correlation value can be searched using equation (6).

$$Correlation = \sum_i \sum_j \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \qquad (6)$$

The next parameter is energy, where this parameter functions to measure the homogeneity of the intensity distribution, and the higher the value, the more homogeneous the image. Equation (7) calculates the energy value of an image.

$$Energy = \sum_i \sum_j p_2^d(i,j) \qquad (7)$$

The homogeneity parameter measures the extent to which the intensity distribution approaches a single intensity distribution. A high homogeneity value indicates that the pixel intensity tends to be uniform. Equation (8) calculates the homogeneity value.

$$Homogenity = \sum_i \sum_j \frac{pd(i,j)}{i+|i-j|} \qquad (8)$$

**5. Identifying Images Using Extreme Learning Machine Algorithms**

ELM is a machine learning algorithm known for its ability to quickly and efficiently train artificial neural networks (Zabala-Blanco et al., 2020). ELM differentiates itself by adopting a one-step training approach where the input weights and hidden weights of hidden layers can be calculated directly and randomly, without the need for iterative adjustments required by some conventional neural network algorithms (Wahid et al., 2021). In the training stage, determining the weighting for the input layer and hidden layers is done with random values and in one step without requiring repeated adjustments.

To prepare the hidden layer in the ELM network through several processes. If $N$ represents the input and the target, which is denoted by $(\boldsymbol{x_i}, \boldsymbol{t_i})$, it refers to $\mathrm{x_i} = [\mathrm{x_{i1}}, \mathrm{x_{i2}}, \cdots \mathrm{x_{in}}]^{\mathrm{T}} \in \mathrm{R^n}$ and $\mathrm{t_i} = [\mathrm{t_{i1}}, \mathrm{t_{i2}}, \cdots \mathrm{t_{in}}]^{\mathrm{T}} \in \mathrm{R^m}$. So, the number of hidden layers is $\tilde{N}$ and the activation function *g(x)* is obtained through equation (9).

$$H \sum_{i=1}^{\tilde{N}} \beta_i g_i(x_{ij}) = \sum_{i=1}^{\tilde{N}} \beta_i g_i(w_i \cdot x_j + b_i) = o_j \qquad (9)$$

where $w_i$ is the notation for the weight vector that connects the input and hidden layers, $\beta_i$ refers to the weight value connecting the hidden layer and the target, $b_i$ refers to the threshold result in the hidden layer, and $w_i \cdot x_j$ refers to the multiplication value.

From equation (9), simplification is carried out to become equation (10).

$$H\beta = T \qquad (10)$$

where $H$ is the notation for the hidden layer input matrix and $T$ is the notation for the destination matrix.

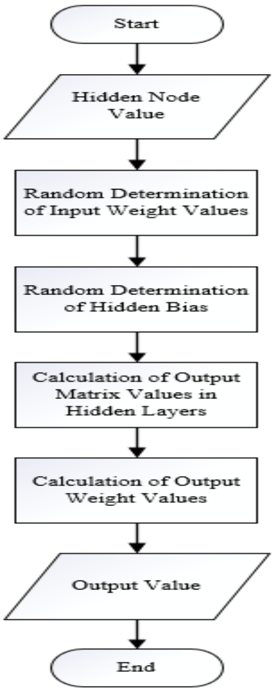
To obtain the input values of hidden weights and biases, random values are given to the network, so that the output weights connected to the hidden layer are found using equation (11).

$$\beta = H \dagger T \qquad (11)$$

where $\beta$ refers to the output weight value, $H$ refers to the input hidden layer matrix, while $T$ refers to the target matrix.

ELM has a neural network architecture consisting of input and hidden layers and is usually used for classification or identification tasks. In the training stage, the weights between the input and hidden layers are generated randomly and performed in a single step without requiring repeated retuning. For more details regarding the ELM algorithm process, it is visualized in the flowchart in Figure 2.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5307




Source : (Research Results, 2024)
Figure 2. Flowchart for the ELM Algorithm Process

The flowchart in Figure 4 illustrates the process of the ELM algorithm, where randomization is used to assign weights to the hidden layers, which contain a greater number of neurons. Next, give weight to the outer layer based on the analysis method with matrix inversion. Below is a description of the stages of the ELM network procedure:

Input : Input pattern $(x_i, t_i)$
Output : Input weight $(w_i)$, output weight $(\beta_i)$ and hidden bias $(b_i)$
The stages :
Stage 1 : Prepare the activation function $(g(x))$, then the hidden node value $(\tilde{N})$
Stage 2 : Get random values for the input weight values $(w_i)$ and hidden bias $(b_i)$
Stage 3 : Calculate the output matrix value (H) in the hidden layer
Stage 4 : Calculate the output weight value $(\beta)$

## 6. Evaluating Models

The evaluation of this stage measures the performance and accuracy of a model being developed. The evaluation approach relies on the confusion matrix. A confusion matrix is a table that compares the model classification results with the actual values of the data (Wu, 2022). The confusion matrix consists of four main cells: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP is the amount of data that is truly correctly classified into the positive class, while TN is the amount of data that is truly correctly classified into the negative class. FP is the amount of data that should be classified as a negative class but is incorrectly classified as a positive class, while FN is the amount of data that should be classified as a positive class but is incorrectly classified as a negative class. Using these values, several model evaluation metrics can be calculated, such as precision, recall, and accuracy. The study utilized accuracy, precision, and recall evaluations because they offer crucial insights into the performance of the classification system in identifying potato leaf diseases. These values can be obtained through equations (12), (13), and (14).

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

## RESULTS AND DISCUSSION

To identify potato diseases based on leaf images, a dataset is first prepared, which is used as training and testing data. In this research, the dataset used is images of potato leaves collected by taking them directly using a camera with the same light level. The classification involves three classes: healthy leaves, early blight disease, and late blight disease. A total of 600 images were collected, with the data split into 70% for training (450 images) and 30% for testing (150 images). Table 1 displays sample images from each class used in the dataset.

Table 1. Examples of Images Used as Datasets

| No | Class Name | Image Sample |
|---|---|---|
| 1 | Healthy Leaves |  |
| 2 | Early Blight |  |
| 3 | Late Blight |  |

Source : (Research Results, 2024)

Table 1 is an example of a dataset used for training and testing data. The developed model is implemented in the MATLAB application. The stages begin with transforming the RGB image into a CIELAB image, to make it easier to represent the image and obtain the necessary information. An example of the results of changing an RGB image to a CIELAB image is visualized in Figure 3.

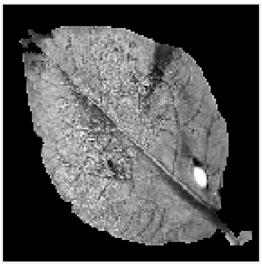
Source : (Research Results, 2024)
Figure 3. (a) RGB image; (b) CIELAB Image

Figure 2 (b) illustrates an instance of the outcomes of image transformation conducted in the CIELAB color space. Subsequently, a segmentation process is executed to distinguish images based on similar characteristics, whereby the foreground and background are segregated. In this particular case study, the foreground is isolated from the background utilizing the K-means clustering technique. This approach partitions the image into k groups or clusters, assigning each pixel to the cluster whose average intensity most closely matches that of the pixel. The outcomes of foreground separation employing K-means clustering are depicted in Figure 4.

Source : (Research Results, 2024)
Figure 4. (a) CIELAB Color Space Image; (b) Segmented Image

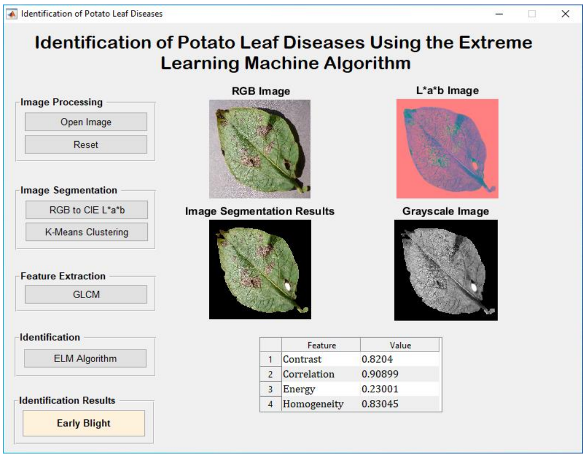
Figure 4 (b) shows the foreground of the segmented image, from which features will be extracted. The main information in this case study is the texture characteristics of the object, which are represented by the texture feature. The texture feature extraction approach applied is the Gray Level Co-Occurrence Matrix (GLCM). In the GLCM approach, it is calculated by measuring how often certain pairs of pixel intensities appear together at various distances and directions in the image. Extraction parameters from GLCM, such as contrast, correlation, energy, and homogeneity, provide very useful information for obtaining information about image texture. Table 2 displays the GLCM parameter values for the segmented images.

Table 2. Results of GLCM Parameter Values

| Segmented Image Converted to Grayscale | GLCM Parameters | Value |
|---|---|---|
|  | Contrast | 0.82040 |
| | Correlation | 0.90899 |
| | Energy | 0.23001 |
| | Homogeneity | 0.83045 |

Source : (Research Results, 2024)

Table 2 shows the input values of all GLCM parameters used for the identification process using the Extreme Learning Machine (ELM) algorithm with artificial neural networks. Next, the model built was tested to measure the performance of the model implemented in MATLAB software. The implementation of the model used for testing in MATLAB software is shown in Figure 5.



Source : (Research Results, 2024)
Figure 5. Model Implementation Interface

To measure the performance of the developed model, an evaluation was carried out using a confusion matrix. The matrix compares the model classification results with the actual data values. Using these values, several model evaluation metrics can be calculated, such as precision, recall, and accuracy. The evaluation process uses 150 test data by comparing model identification findings with actual facts. Figure 6 presents the results of the model evaluation's confusion matrix.

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5307

Truth data

| Classifier results | Healthy Leaves | Early Blight | Late Blight | Classification overall | User's accuracy (Precision) |
|---|---|---|---|---|---|
| Healthy Leaves | 47 | 2 | 2 | 51 | 92.157% |
| Early Blight | 3 | 39 | 6 | 48 | 81.25% |
| Late Blight | 3 | 7 | 41 | 51 | 80.392% |
| Truth overall | 53 | 48 | 49 | 150 | |
| Producer's accuracy (Recall) | 88.679% | 81.25% | 83.673% | | |

Source : (Research Results, 2024)

Figure 6. Obtained Confusion Matrix Values

The evaluation resulting from the confusion matrix visualized in Figure 6 obtained measurement values such as precision, recall, and accuracy. The selection of accuracy, precision, and recall assessments in this investigation was deliberate as these three metrics offer crucial and comprehensive insights into the classification system's effectiveness in detecting potato leaf diseases. These values are displayed in Table 3.

Table 3. Precision, Recall and Accuracy Values

| Class Name | Precision | Recall | Accuracy |
|---|---|---|---|
| Healthy Leaves | 92.157% | 88.679% | |
| Early Blight | 81.250% | 81.250% | 84.667% |
| Late Blight | 80.392% | 83.673% | |

Source : (Research Results, 2024)

Table 3 displays the outcomes of the model assessment, revealing an overall accuracy of 84.667%. After that, we put these results into criteria groups by referring to the following values: “Very Poor” with a score of less than 40%; “Poor” with a score between 40% and 55%; “Fairly Good” with a score between 56% and 75%; and “Good” with a score between 76% and 100% (Harjanti, 2022). The accuracy results obtained in this study are in the "Good" category.

This research identifies potato leaf diseases by applying an artificial neural network approach using the ELM approach based on texture features using the GLCM method. This research uses traditional machines because they can handle data limitations and have model interpretability. If it is related to previous research which used machine learning to identify plant diseases using leaf images, the results of the proposed model have better results. Research on the identification of Siamese orange plant diseases using the K-Nearest Neighbor (KNN) method produces an accuracy of 70% (Ariesdianto et al., 2021); research on disease classification in rice leaves through the application of the Support Vector Machine (SVM) approach produces an accuracy value of 80% (Prastyo et al., 2023); and research on disease identification on grape leaves uses the Backpropagation artificial neural network approach produces an accuracy rate of 82% (Ansah et al., 2022). Even though the plant object used is different from previous research, namely potato leaves, the proposed model obtains a higher accuracy, namely 84.667%. The reason behind this is because ELM utilizes artificial neural networks, which employ a single training process without any iteration. This process entails the random assignment of weights and biases in the hidden layers. Not only that, the ELM training process involves randomly determining weights and biases in hidden layers, which allows training to be carried out quickly and efficiently. As a result, ELM often provides competitive performance with shorter training times compared to some conventional neural network algorithms.

However, from the model evaluation results, the accuracy value was 84,444%, so the error rate reached 15,556%. This inaccuracy is caused by several factors, including: 1) In the ELM algorithm, the weighting is obtained randomly, which affects the various input values; 2) The ELM algorithm cannot tune parameters adaptively during training, so it is less than optimal for understanding complex relationships in images that require iterative adjustments; 3) The features used are based on texture alone, thereby ignoring other information that might represent the object; 4) The classes used have almost similar characteristics, so to get representative data, pre-processing is needed to get optimal data.

## CONCLUSION

Our study created a way to find potato leaf diseases by using artificial neural networks with the ELM (Extreme Learning Machine) algorithm and the GLCM (Gray Level Co-Occurrence Matrix) method to pull out texture features. The GLCM method effectively captures texture features by assessing the frequency of specific pairs of pixel intensities at diverse distances and orientations within the image. These extracted features serve as inputs for the identification process conducted through the ELM algorithm, known for its swift and efficient one-time training approach. Evaluation of the model via the confusion matrix yielded an accuracy of 84.667%, showcasing the model's adeptness in precise disease identification. Nevertheless, enhancements are needed, particularly in refining the weighting process within the ELM algorithm, which currently relies on random assignment. This variability affects input values, suggesting potential integration of complementary algorithms like fuzzy logic to yield more consistent weight values. Additionally, our research underscores the necessity for incorporating diverse feature sets beyond texture alone to enhance the representativeness and robustness of disease identification systems.

## REFERENCE

Ahmad, I., Rahmanto, Y., Borman, R. I., Rossi, F., Jusman, Y., & Alexander, A. D. (2022). Identification of Pineapple Disease Based on Image Using Neural Network Self-Organizing Map (SOM) Model. *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*.

Andrian, R., Maharani, D., Muhammad, M. A., & Junaidi, A. (2020). Butterfly Identification Using Gray Level Co-Occurrence Matrix (GLCM) Extraction Feature and K-Nearest Neighbor (KNN) Classification. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *6*(1), 11–21. https://doi.org/10.26594/register.v6i1.1602

Ansah, M. A., Kasih, P., & Dara, M. A. D. W. (2022). Identifikasi Penyakit Daun Anggur Berdasarkan Fitur Warna Dan Tekstur Dengan Metode Backpropagation Berbasis Android. *Seminar Nasional Inovasi Teknologi*, 265–271.

Ariesdianto, R. H., Fitri, Z. E., Madjid, A., & Imron, A. M. N. (2021). Identifikasi Penyakit Daun Jeruk Siam Menggunakan K-Nearest Neighbor. *Jurnal Ilmu Komputer Dan Informatika (JIKI)*, *1*(2), 133–140.

Armavillia, K. E. (2023). *Produksi Kentang Di Indonesia Naik Secara Bertahap*. GoodStats. https://data.goodstats.id/statistic/elmaarmavillia/produksi-kentang-di-indonesia-naik-secara-bertahap-P6HbQ

Baek, S.-H., Park, K.-H., Jeon, J.-S., & Kwak, T.-Y. (2022). Using the CIELAB Color System for Soil Color Identification Based on Digital Image Processing. *Journal of The Korean Geotechnical Society*, *38*(5), 61–71.

Borman, R. I., Kurniawan, D. E., Styawati, Ahmad, I., & Alita, D. (2023). Classification of maturity levels of palm fresh fruit bunches using the linear discriminant analysis algorithm. *AIP Conference Proceedings*, *2665*(1), 30023. https://doi.org/10.1063/5.0126513

Harjanti, T. W. (2022). Implementation of Inference Engine with Certainty Factor on Potential Diagnosis of Brain Tumor Disease. *PILAR Nusa Mandiri: Journal of Computing and Information System*, *18*(1), 25–30. https://doi.org/10.33480/pilar.v18i1.xxxx

Herdiansah, A., Borman, R. I., Nurnaningsih, D., Sinlae, A. A. J., & Al Hakim, R. R. (2022). Klasifikasi Citra Daun Herbal Dengan Menggunakan Backpropagation Neural Networks Berdasarkan Ekstraksi Ciri Bentuk. *JURIKOM (Jurnal Riset Komputer)*, *9*(2), 388–395. https://doi.org/10.30865/jurikom.v9i1.3846

Javidan, S. M., Banakar, A., Vakilian, K. A., & Ampatzidis, Y. (2023). Diagnosis of grape leaf diseases using automatic K-means clustering and machine learning. *Smart Agricultural Technology*, *3*, 1–14. https://doi.org/10.1016/j.atech.2022.100081

Malounas, I., Lentzou, D., Xanthopoulos, G., & Fountas, S. (2024). Testing the suitability of automated machine learning, hyperspectral imaging and CIELAB color space for proximal in situ fertilization level classification. *Smart Agricultural Technology*, *8*, 100437. https://doi.org/https://doi.org/10.1016/j.atech.2024.100437

Mayatopani, H., Borman, R. I., Atmojo, W. T., & Arisantoso, A. (2021). Classification of Vehicle Types Using Backpropagation Neural Networks with Metric and Ecentricity Parameters. *Jurnal Riset Informatika*, *4*(1), 65–70. https://doi.org/10.34288/jri.v4i1.293

Miao, Y., Li, S., Wang, L., Li, H., Qiu, R., & Zhang, M. (2023). A single plant segmentation method of maize point cloud based on Euclidean clustering and K-means clustering. *Computers and Electronics in Agriculture*, *210*, 107951. https://doi.org/https://doi.org/10.1016/j.compag.2023.107951

Muraina, I. O. (2022). Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts. *7th International Mardin Artuklu Scientific Researches Conference*, 496–505.

Prastyo, B. A., Istiadi, I., & Rahman, A. Y. (2023). Klasifikasi Penyakit Daun Padi Menggunakan Support Vector Machine Melalui Tekstur Dan Warna Daun Dengan HSV Dan Gabor Filter. *The 6th Conference on Innovation and Application of Science and Technology (CIASTECH)*, 567–575.

Rao, V. C. S., Venkratamulu, S., & Sammulal, P. (2021). *Digital Image Processing and Applications*. Horizon Books (A Division of Ignited Minds Edutech P Ltd).

Wahid, R. R., Anggraeni, F. T., & Nugroho, B. (2021). Brain Tumor Classification with Hybrid Algorithm Convolutional Neural Network-Extreme Learning Machine. *International Journal of Computer, Network Security and Information System*, *3*(1), 29–33. https://doi.org/https://doi.org/10.33005/ijconsist.v3i1.53

Wanti, E. P., Pariyandani, A., Zulkarnain, S., & Idrus, S. (2021). Utilization of SVM Method and GLCM Feature Extraction in Classifying Fish

Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5307

Images with Formalin. *Scientific Journal of Informatics*, *8*(1), 168–175. https://doi.org/10.15294/sji.v8i1.26806

Wu, M.-T. (2022). Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom. *Scientific Reports*, *12*(1), 3095. https://doi.org/10.1038/s41598-022-07137-z

Zabala-Blanco, D., Mora, M., Hernandez-Garcıa, R., & Barrientos, R. J. (2020). The Extreme Learning Machine Algorithm for Classifying Fingerprints. *39th International Conference of the Chilean Computer Science Society (SCCC)*. https://doi.org/https://doi.org/10.1109/SCCC51225.2020.9281232

# IMPLEMENTATION OF ARMA MODEL FOR BENGAWAN SOLO RIVER WATER LEVEL AT JURUG MONITORING POST

**Sri Siswanti[1]; Retno Tri Vulandari[2*]; Setiyowati[3]**

Informatics[1], Information technology[2], Information System[3]
STMIK Sinar Nusantara, Surakarta, Indonesia
www.sinus.ac.id
syswanty@sinus.ac.id[1], retnotv@sinus.ac.id[2*], setiyowati@sinus.ac.id[3]
(*) Corresponding Author



**Abstract**—The amount of annual rainfall in the Bengawan Solo watershed causes high water flow (water discharge) in several rivers. In addition, high flow rates significantly increased the water surface level at some observation sites. The Bengawan Solo River burst its banks in November 2016, causing flooding in several areas in eastern Solo. At that time, the river stage at the Jurug monitoring post passed ten. Therefore, a flood early warning system would be useful for predicting water levels in this context. Every day, one post on the Bengawan Solo River measures the water level. The time series data used in this study is the water level. Autoregressive Moving Average (ARMA) is a predictive method for measuring time set data. The assumption of homoscedasticity or constant error variance is used in this model. However, the study will use the ARMA model if the variance changes randomly. The study used 60 pieces of data from January to February 2018. This study can directly use ARMA because the data results are stationary based on ADF value 0.0036. After the first pause, the ACF and PACF are disconnected based on the correlogram pattern. This shows that the water level of the Bengawan Solo River in that period can appear on the Autoregressive Moving Average with orders p = 1 and q = 1 ARMA(1,1). Thus, the total average residue is 0.0668384, so the short error is 6.68384%.

**Keywords:** autoregressive moving average (ARMA), jurug monitoring post, time series analysis, water level.

**Abstrak**—Jumlah curah hujan tahunan di DAS Bengawan Solo menyebabkan aliran air (debit air) yang tinggi di beberapa sungai. Selain itu, laju aliran yang tinggi meningkatkan ketinggian permukaan air secara signifikan di beberapa lokasi pengamatan. Sungai Bengawan Solo meluap pada November 2016, menyebabkan banjir di beberapa daerah di timur Solo. Pada saat itu, tahap sungai di pos pemantauan Jurug melewati sepuluh. Oleh karena itu, sistem peringatan dini banjir akan berguna untuk memprediksi ketinggian air dalam konteks ini. Setiap hari, satu pos di Sungai Bengawan Solo mengukur tinggi airnya. Data time series yang digunakan dalam penelitian ini adalah ketinggian muka air. Autoregressive Moving Average (ARMA) adalah metode prediksi untuk mengukur data set waktu. Asumsi homoskedastisitas atau varians kesalahan konstan digunakan dalam model ini. Namun, penelitian ini akan menggunakan model ARMA jika variansnya berubah secara acak. Studi ini menggunakan 60 data dari Januari hingga Februari 2018. Penelitian ini dapat langsung menggunakan ARMA karena hasil datanya stasioner berdasarkan nilai ADF 0,0036. Setelah jeda pertama, ACF dan PACF terputus berdasarkan pola correlogram. Hal ini menunjukkan bahwa tinggi muka air Sungai Bengawan Solo pada periode tersebut dapat muncul pada Autoregressive Moving Average dengan order p=1 dan q=1 ARMA(1,1). Dengan demikian, total rata-rata residunya adalah 0,0668384, sehingga kesalahan singkatnya adalah 6,68384%.

**Kata Kunci**: autoregressive moving average (ARMA), pos pemantauan jurug, analisa runtun waktu, tinggi muka air.

## INTRODUCTION

River basin is the most important information for water resource management. In addition, the information of its peak flow can be useful to design better flood control buildings (Hidayat et al., 2022). Meanwhile, the data of small river stream is needed for useful water location planning, especially in long dry season. The average annual flow can give better pictures of water resource potential which can offer advantages from a river basin (Saidah & Hanifah, 2020). Water debit is water flow rate that passes a cross-section in the river per unit time. In technical reports, its flow rate usually appears in hydrograph (Biantoro et al., 2021). Hydrograph shows behaviour of water flow rate as a response of the changes of biogeophysical characteristic which happens because of the river basin management activity and seasonal or annual fluctuation, like local climate changes (Park et al., 2023).

The annual rainfall intensity in Bengawan Solo river basin causes its big water debit in some streams. In addition, the flows also increase the height of water level in some dam posts. The measurement of water level in every dam can certainly be useful to prevent flood (Trinugroho et al., 2022). At the end of 2016, Bengawan Solo River was overflowing and the flood covered East Solo areas. According to a report in Jurug monitoring post, the river stage passed over 10 and the flood covered 10 districts of East Solo. At this case, it needs an efficient model to predict the water level, especially flood stage, as flood early warning in order not to bring back the past disaster in Solo. This research uses the rate of water level, which is measured every day, at Jurug monitoring post as time series data.

Arrangement with a stationary model, such as Autoregressive Moving Average (ARMA), is possible with Autocorelation Function (ACF) and Partial Autocorelation Function (PACF). ARMA model has the assumption of homoscedacity or a constant error variance, but it turns into heteroscedacity if time series data from the water surface have a constant error variance every time (Nguyen, 2020).

## MATERIALS AND METHODS

### 1. Data Processing Procedure

Procedures for processing data used in this study are as follows:

a. Data Collection
This study requires time series data and measurement results, including water level data from the Jurug monitoring post from 2009 to June 2018.
b. Input Data Processing
First, the ARMA-GARCH method was used to analyse the data, then the original data plot was used to identify the distribution pattern of the data, and then stationary tests were performed (Berutu et al., 2023).
c. Data Sharing (Load Data)
The data distribution consists of testing data, namely TMA data from January - February 2018 which is taken randomly, training data, namely the water level of the Bengawan Solo river in 2009 - June 2018.

### 2. Data Analysis

Stationary tests, identification of the ACF and PACF models, estimation of the ARMA model parameter, diagnostic tests, and then GARCH error variance models are all part of the process of creating the ARMA stationary model.

a. Stasionecity Test

The unit root test can be used to determine the data's stationarity. The hypothesis of the examination is written as

H0 : $\eta = 1$ (data has unit root)
H1 : $\eta < 1$ (data has no unit root)

Test statistics are the ratio of the estimated coefficient minus 1 compared to its standard deviation (Fauzi & Irviani, 2023). The Augmanted Dickey-Fuller (ADF), also known as the t-ratio, is formulated in the same way as in formula (1).

$$ADF = \frac{\eta-1}{\sigma(\eta)} \text{.........................................................} (1)$$

with $x_0$ = 0, T is the sample size and is the $x_t$ t-th observation data. H0 is rejected when the ratio t > t $\alpha$ ,(T − 1)

b. ACF and PACF Model Identification

The tools for identifying ARMA models are ACF and PACF. The autocorrelation function shows the magnitude of the correlation between observations at t-time and observations at previous times, while the PACF is a function that shows the magnitude of the partial correlation between observations at t-time and observations at previous times (Maulidiyah & Fauzy, 2023).

c. Stationary Model Parameter Estimation

ARMA model contains two components: the AR model and the MA model, where the order of AR is p and the order of MA is q (Gustiansyah, Rizki, & Apriyanti, 2023). Here is a stationary model according to (Safwandi, 2023).

(1) Autoregressive (AR)
A observation at time t that is expressed as a linear function against p at the prior time plus a random error $e_t$ is autoregressive (AR). The general form of the p-order autoregressive model is formulated in formula (2).

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + e_t \text{.......(2)}$$

(2) Moving Average (MA)
A phenomenon that an observation at time t is represented as a linear combination of a number of random errors is called moving average (MA). The general form of the q-order moving average model is formulated in formula (3).

$$Y_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \cdots - \beta_a e_{t-q} \text{........(3)}$$

(3) Autoregressive Moving Average (ARMA)
The main model of Autoregressive Moving Average (ARMA) (p,q) is expressed in formula (4). ARMA is a combination of AR and MA.

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \cdots - \beta_a e_{t-q} \text{......................(4)}$$

d. Diagnostic Model
The diagnostic model is used to find out if the model is suitable for use. Error is an excellent indicator of model compatibility if it is homogeneous, autocorrelation-free, and has a low mean square error (MSE) value (Marheni & Triyanto, 2023). As a result, computations of the MSE value, variance homogeneity tests, and error autocorrelation tests are performed. Re-identifying and estimating is required if the error does not satisfy these three requirements, indicating that the model developed does not match the data (Dewi & Indah, 2022).

e. Model ARCH dan Model GARCH
The ARCH model and GARCH model (Sumiyati & Wilujeng, 2022) are models used for time series data that have high volatility and have inconstant error variances (heteroscedasticity). Engle developed the ARCH model with mean and variance modeled simultaneously. General form of the ARCH(p) model is formulated in formula (5).

$$\sigma_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \alpha_2 e_{t-2}^2 + \cdots + \alpha_p e_{t-p}^2 \text{.............(5)}$$

with $\sigma_t^2$ residual variances and $e_{t-p}^2$ residual squares of past periods. Bollerslev developed the ARCH model by taking into account residual variances of past periods. General form of the GARCH(p,q) model is formulated in formula (6).

$$\sigma_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \alpha_2 e_{t-2}^2 + \cdots + \alpha_p e_{t-p}^2 + \lambda_1 \sigma_{t-1}^2 + \cdots + \lambda_q \sigma_{t-q}^2 \text{ ..........................................(6)}$$

with $\sigma_{t-q}^2$ residual variances of past periods.

f. Research Mindset
The operational steps to accomplish the research goals are listed below.
(1) To determine the distribution pattern of the data, create a data plot;
(2) Use a unit root test to perform a stationary test; if the data is stationary, the direct data can be modeled.
(3) The ln transformation is carried out if the data is not stationary. Next, run the unit through another root test.
(4) ACF and PACF graphs were used to identify the model after the stationary data. After that, it calculates the parameter's magnitude and draws conclusions from the data's stationary model.
(5) Diagnostic tests on errors produced by the model are then conducted following the formation of the stationary model. A mistake is the distinction between prediction data and actual data.
(6) If the assumption of homogeneity of variance is not met, then it means that the data has variable errors.
(7) Perform modeling for error variance correction using GARCH.
(8) Perform these simulation stages using MATLAB.

## RESULTS AND DISCUSSION

### a. Description and Data Pattern

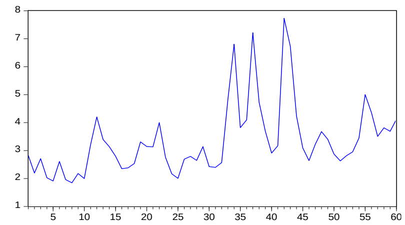
A time series that remains immobile is one whose characteristics remain constant regardless of the observation time of the sequence. Time series that exhibit a trend or seasonality are therefore non-stationary, and the presence of seasonality will impact the time series' value at different points in time. A white noise sequence, on the other hand, is stationary and does not change in appearance over time; it should appear relatively constant. A stagnant time series will typically not exhibit any predictable patterns over an extended period of time. Time plots will reveal that the series has a constant variance and is essentially level.

Table 1. Stationary Data Test

| Null Hypothesis: TMA has a unit root | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -3,906148 | 0,0036 |
| Test critical values: | 1% level | -3,546099 | |
| | 5% level | -2,911730 | |
| | 10% level | -2,593551 | |

(Research Results, 2018)

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5004




(Research Results, 2018)

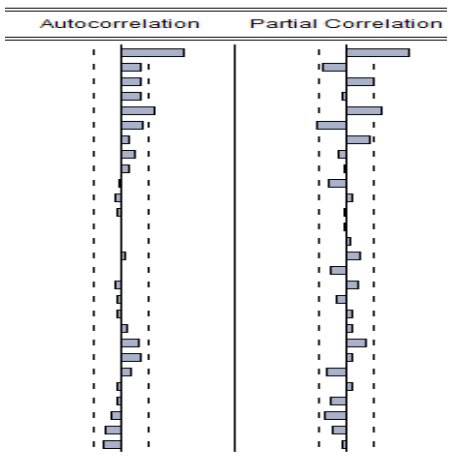
Figure 1. Time Series Plot of Water Level

Based on the research method, the first step to do is data description. Figure 1 shows the average data of water level is stationary in inconstant variance and its stationary and unit root test strengthens this circumstance. For the result of the test, the probability value of its Augmented Dickey-Fuller (ADF) is 0,0036. In this case, that probability value is smaller than significant level $\alpha = 0{,}05$. Moreover, it can be proven by statistic value t that $|t|_{TMA} = 2{,}911730 > t_{(0{,}05;59)} = -1{,}671$, in which $H_0$ is successfully rejected. It means the data have no unit root (Table 1) and it turns to be stationary. As a result of stationary tendency toward its average, this research provides the average model firstly before its variance model.

**b. Formation of Conditional Mean Model in Stationary Process**

**1) Identification Model**

In this research, conditional mean model from stationary data can use ARMA. This research also applies ACF and PACF to identify ARMA comprehensively. In Figure 2, ACF and PACF value cut off after the first lag then ARMA will be fit for the conditional mean model.




(Research Results, 2018)

Figure 2. Water Level of ACF and PACF

**2) Parameter Estimation**

ARMA (1, 1) is a process of autoregressive order 1 and process of moving average order 1. It shows in the formula (7).

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

Then, it obtains auto covariance function, as in formulas (8) and (9).

For k = 0,

$$E(Y_t Y_t) = \gamma_0 = \phi\gamma_1 + \sigma_e^2 - \theta(\phi - \theta)\sigma_e^2 \ldots(8)$$

For $k = 1$,

$$E(Y_t Y_{t-1}) = \gamma_1 = \phi\gamma_0 - \theta\sigma_e^2 \ldots\ldots\ldots\ldots\ldots(9)$$

Applied with substitution of equation (2) to equation (1), it obtains formula as formulated in formula (10) and (11).

$$\gamma_0 = \frac{(1-2\theta\phi+\theta^2)}{(1-\phi^2)}\sigma_e^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

$$\gamma_1 = \frac{(1-\theta\phi)(\phi-\theta)}{(1-\phi^2)}\sigma_e^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots(11)$$

For $k = 2$, auto covariance function shows as in formula (12), (13), (14, (15) and (16).

$$\gamma_2 = \frac{(1-\theta\phi)(\phi-\theta)}{(1-\phi^2)}\phi\sigma_e^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots(12)$$

For $k = k$,

$$\gamma_k = \frac{(1-\theta\phi)(\phi-\theta)}{(1-\phi^2)}\phi^{k-1}\ \sigma_e^2 \ldots\ldots\ldots\ldots\ldots(13)$$

Mean while, auto correlation function for $k = 1$ is shown in the formula (14).

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{(1-\theta\phi)(\phi-\theta)}{1-2\theta\phi+\theta^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(14)$$

For $k = 2$,

$$\rho_2 = \frac{\gamma_2}{\gamma_0} = \frac{(1-\theta\phi)(\phi-\theta)\phi}{1-2\theta\phi+\theta^2} \ldots\ldots\ldots\ldots\ldots\ldots(15)$$

For $k = k$,

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{(1-\theta\phi)(\phi-\theta)\phi^{k-1}}{1-2\theta\phi+\theta^2} \ldots\ldots\ldots\ldots\ldots(16)$$

For the next one, the parameter estimation of ARMA (1,1) model shows a result in table 2.

Table 2. ARMA (1,1) Model

| Variable | Coefficient |
|---|---|
| C | 3.376777 |
| AR(1) | 0.075557 |
| MA(1) | 0.948131 |

(Research Results, 2018)

Table 2 shows $\hat{\phi}_1 = 0{,}075557, \hat{\theta}_1 = 0{,}0{,}948131$ and the intercept value is 3,376777. As a result, ARMA(1,1) model is: $Y_t = 0{,}075557Y_{t-1} + 3{,}376777 + e_t - 0{,}0{,}948131e_{t-1}$ with $Y_t$ is water level at time t and $e_t$ is ARMA errors at time t

**3) Diagnostic Test**

In this auto correlation test, good conditional mean models have no auto correlation in the errors. Auto correlation shows the inter-correlation of its observations. Breusch-Godfrey test is a statistical test to find out the presence of auto correlation in the errors (Chaudhary et al., 2022). Table 3 shows this Breusch-Godfrey test of the research.

Table 3. Auto Correlation Test

| | |
|---|---|
| Uji Breusch-Godfrey | 0,2134 |
| Error lag-1 | 0,7652 |
| Error lag-2 | 0,7213 |
| Error lag-3 | 0,4512 |
| Error lag-4 | 0,2111 |
| Error lag-5 | 0,1202 |
| Error lag-6 | 0,0912 |
| Error lag-7 | 0,2307 |
| Error lag-8 | 0,3087 |

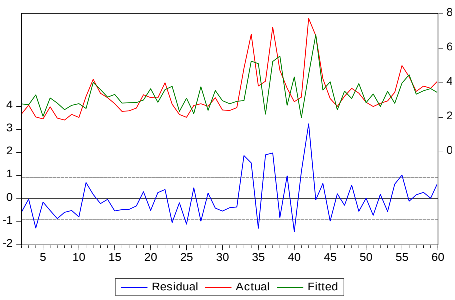
(Research Results, 2018)

Heteroscedacity has inconstant variability in the errors for each observation. Its data values tend to have rapid fluctuation. Meanwhile, volality enables to measure and draw fluctuation of data. Volality can be defined as data variance over the time series. Volality also shows the tendency of data to changes rapidly over the time so that error variance changes each time. In Table 4, White test can present the volality with data probability value less than $\alpha = 0,05$. As a result, it obtains $R^2 = 9,213801 > \chi^2_{0,05;2} = 5,991$ and $H_0$ is rejected which shows the heteroscedacity in the data.

Table 4. White's Heteroscedastic Test

| | |
|---|---|
| Obs. $R^2$ | 9,213801 |
| Probabilitas $\chi^2$ | 0,0100 |

(Research Results, 2018)




(Research Results, 2018)
Figure 3. Residual, Actual, and Fitted

Based on Figure 3, it can be seen that the actual data with the fitted data has the same pattern. The red line indicates the actual data pattern, while the green line indicates the prediction data pattern. Actual data patterns are similar to fitted data pattern. Blue lines indicate residual or data errors, having the same pattern.

## CONCLUSION

The water level of Bengawan Solo River on Januari – February 2018 uses ARMA (1,1) model because its result tends to be stationary toward the average but it has inconstant data variance. Based on research that has been done, the ARMA model (1.1) can be used to predict the water level of the Bengawan Solo river. The model was successfully used as an early warning to prevent flooding, this can be seen from the diagnostic test comparing actual data and fitted data in April and May 2018, having the same data pattern. For the following research, it will be great if there is data variance model using GARCH or TARCH model. Time series data is difficult to show in a certain model because its data fluctuation gets influence from many factors based on its individual characteristics such as TGARCH, MGARCH, and APARCH. Those models are fit for asymmetric time series data.

## REFERENCE

Berutu, S. S., Budiati, H., & Gulo, F. (2023). Data Preprocessing Approach for Machine Learning-based Sentiment Classification. *Jurnal Informatics, Telecommunication, and Electronics Vol 15 No 4*, 34-43.

Biantoro, A. W., Iskendar, & Subekti. (2021). The Effects of Water Debit and Number of Blades on the Power Generated of Prototype Turbines Propeller as Renewable Electricity. *Jurnal Rekayasa Mesin Vol 12 No 1*, 54-64.

Chaudhary, A. A., Nazir, N., Riaz, A., Sadiq, N., & Riaz, N. (2022). Autocorrelation: what happens if the error or disturbance terms are correlated in time-series data. *Competitive Education Research Journal*, *3*(2), 154-163.

Dewi, W. I., & Indah, N. P. (2022). Pengaruh Perputaran Kas dan Modal Kerja terhadap Profitabilitas Perusahaan di Bursa Efek Indonesia. *Jurnal Ekonomi, Manajemen, dan Akuntansi (JEBM) Vol 24 No 3*, 624-629.

Fauzi, & Irviani, R. (2023). Analysis of The Effect of Macropredential and Macroprudential Indicators on The Stock Price Index of Islamic Banks in Indonesia Using The Error Correction Model. *Jurnal Syntax Admiration Vol 4 No 5*, 1-19.

Gustiansyah, M. A., Rizki, A., & Apriyanti, B. (2023). Aplikasi Model ARIMA dalam Peramalan Data Harga Emas Dunia. *Jurnal Statistika dan Aplikasinya Vol 7 No 1*, 84-92.

Hidayat, D., Legowo, S., & Farid, M. (2022). Clarification of Watershed Recharge in Cisadane River Basin throughGround Test. *Journal of the Civil Engineering Forum,*, 217-224.

Marheni, R., & Triyanto, E. (2023). Pengaruh Dana Alokasi Umum, Dana Alokasi Khusus dan Dana Bagi Hasil terhadap Belanja Modal di Kabupaten/Kota Provinsi Jawa Tengah. *Jurnal Cakrawala Ilmiah Vol 2 No 11*, 224-232.

Maulidiyah, W., & Fauzy, A. (2023). Perbandingan Metode Peramalan Double Exponential Smoothing dengan Damped Parameter dan Autoregressive Integrated Moving Average. *Emerging Statistics and Data Science Journal Vol 1 No 3*, 361-377.

Nguyen, X. H. (2020). Combining Statistical Machine Learning Models with ARIMA for Water Level forecasting: The case of the Red river. *Advances in Water Resources Journal Vol 142*, 47-52.

Park, K., Seong, Y., & Jung, Y. (2023). Development of Water Level Prediction Improvement Method Using Multivariate Time Series Data by GRU Model. *MDPI Journal Issu 15*, 1-14.

Safwandi. (2023). Time Series Model Using Autoregressive Integrated Moving Average (ARIMA) Method for Inflation in Indonesia. *Jurnal Investasi Islam Vol 8 No 1*, 13-25.

Saidah, H., & Hanifah, L. (2020). Analysis of Water System Condition for Monitoring of Critical Area in Jangkok Watershed. *Jurnal Tanah dan Sumberdaya Lahan Vol 7 No 2* , 237-248.

Sumiyati, & Wilujeng, P. R. (2022). Metode ARCH/GARCH untuk Memprediksi Hubungan Economic Uncertainty (COVID 19) dan Volatilitas Saham. *Jurnal Bisnis dan Akuntansi Vol 24 No 1*, 117-130.

Trinugroho, M. W., Arif, S. S., & Susanto, S. (2022). Changes in Rainfall Pattern in Bengawan Solo Sub-Watershed. *Journal of Soil Science and Argoclimatology Vol 19 No 2*, 125-136.

# DEEP LEARNING FOR AUTOMATIC CLASSIFICATION OF AVOCADO FRUIT MATURITY

**Wina Widiati[1*]; Toto Haryanto[2]**

Departemen Ilmu Komputer[1]
Fakultas Matematika dan Ilmu Pengetahuan Alam[2]
Institute Pertanian Bogor, Bogor, Indonesia[1,2]
Universitas Bina Sarana Informatika[1]
https://ipb.ac.id
amoywina@apps.ipb.ac.id[1*]; wina.wnw@bsi.ac.id[1]; totoharyanto@apps.ipb.ac.id[2]
(*) Corresponding Author



***Abstract***—*Avocado (Persea Americana), a fleshy fruit with a single seed, has increased in popularity globally, especially in tropical and Mediterranean climates, thanks to its commercial and nutritional value. Rich in bioactive compounds, avocados contribute to the prevention and treatment of various diseases, including cardiovascular problems and cancer. Avocado production in Indonesia, for example, is showing a significant increase, reflecting the growing demand. Avocado ripeness affects shelf life and quality, making the determination of ripeness level a critical aspect of postharvest management. Skin color and pulp firmness change during storage, affecting quality and nutritional value. Proper classification of ripeness is important to reduce post-harvest losses, improve quality and optimize export costs. Recent research shows the use of technologies such as machine learning and YOLO (You Only Look Once) version 9 in real-time detection of avocado ripeness, offering innovative solutions to reduce post-harvest losses and improve distribution efficiency. This approach not only benefits farmers and consumers but also ensures consumer satisfaction and reduces economic losses. This study highlights the importance of real-time detection in monitoring avocado ripeness, where the training process was conducted for 89,280 iterations resulting in a new model for avocado ripeness detection. The final model has a mean Average Precision (mAP) validation value of 84.3%, mAP 84.3% signifies the optimal level of accuracy in object recognition in avocado fruit maturity images using the YOLO model that has undergone an intensive training process.*

***Keywords:*** *avocado maturity, postharvest management, real-time detection, YOLO (You Only Look Once) technology.*

**Abstrak**—Alpukat (*Persea Americana*), buah berdaging dengan biji tunggal, telah meningkat popularitasnya secara global, terutama di iklim tropis dan Mediterania, berkat nilai komersial dan nutrisinya. Kaya akan senyawa bioaktif, alpukat berkontribusi pada pencegahan dan pengobatan berbagai penyakit, termasuk masalah kardiovaskular dan kanker. Produksi alpukat di Indonesia, misalnya, menunjukkan peningkatan signifikan, mencerminkan permintaan yang meningkat. Kematangan alpukat memengaruhi umur simpan dan kualitas, menjadikan penentuan tingkat kematangan sebagai aspek kritis dalam manajemen pascapanen. Warna kulit dan kekencangan daging buah berubah selama penyimpanan, mempengaruhi kualitas dan nilai nutrisi. Klasifikasi kematangan yang tepat penting untuk mengurangi kerugian pasca panen, meningkatkan kualitas, dan mengoptimalkan biaya ekspor. Penelitian terbaru menunjukkan penggunaan teknologi seperti machine learning dan YOLO (*You Only Look Once*) versi 9 dalam deteksi real-time kematangan alpukat, menawarkan solusi inovatif untuk mengurangi kerugian pasca panen dan meningkatkan efisiensi distribusi. Pendekatan ini tidak hanya menguntungkan petani dan konsumen tetapi juga memastikan kepuasan konsumen dan mengurangi kerugian ekonomi. Studi ini menyoroti pentingnya deteksi real-time dalam memantau kematangan alpukat, dimana proses pelatihan dilakukan sebanyak 89.280 iterasi yang menghasilkan model baru untuk deteksi kematangan buah alpukat. Model akhir memiliki nilai validasi *mean Average Precision* (mAP) sebesar 84.3%, mAP 84.3% menandakan tingkat akurasi yang optimal dalam pengenalan objek pada citra kematangan buah alpukat menggunakan model

YOLO V9 yang telah menjalani proses pelatihan secara intensif.

***Kata Kunci**: kematangan alpukat, manajemen pascapanen, deteksi real-time, teknologi YOLO (You Only Look Once.*

## INTRODUCTION

Avocado (Persea Americana) is a commercially valuable fruit cultivated in tropical and Mediterranean climates around the world. In recent decades, avocado fruit has grown in popularity due to increased consumer awareness of its dietary value. Avocado fruit is a fleshy, single-seeded fruit. The fruit has a thick skin and is green or purple-black in colour. Avocado is a climacteric fruit that continues to undergo physiological changes after harvest (Shrestha, 2022). Avocado fruit is very important due to its nutritional and health benefits. It contains bioactive compounds that have been linked to the prevention and treatment of diseases such as macular degeneration, osteoarthritis, cardiovascular problems, and cancer (Adetuyi, et al., 2022). Avocado consumption has been linked to protective effects on human health, including the prevention of cardiovascular disease, diabetes, and certain forms of cancer (Pacheco, et al., 2022). Avocado (Persea Americana) is one of the climacteric fruits that has increased production from year to year. Based on data obtained from BPS, the total avocado production in Indonesia in 2019 reached 40170 quintals, in 2020 it reached 33173 quintals and in 2021 it reached 87377([BPS] Badan Pusat Statistik, 2022). Overall, avocado fruit plays an important role in promoting human nutrition and health.

Determination of the level of maturity is an important factor considering that the level of maturity affects the shelf life, where the higher the level of maturity for avocado fruit (Cho, et al., 2020). Therefore, a postharvest management methodology is needed to determine the ripening stage of avocados to prevent fruit loss due to quality deterioration (Han, et al., 2023). The skin colour and flesh firmness of avocado fruits change during storage. Colour is considered a basic physical property of agrofood products and can be correlated with other quality, other attributes such as nutrition and visual or non-visual defects (Cho, et al., 2020). Classification of avocado fruit ripeness is important to determine optimal consumption ripeness, improve fruit quality, and reduce export costs and losses (Jaramillo-Acevedo, et al., 2020). In addition, classification of avocado fruit maturity is important for many reasons. It helps to reduce post-harvest losses by sorting fruits based on their duration until they are ready to be eaten (Jaramillo-Acevedo, et al., 2020). In addition, it contributes to improved fruit quality and decreased export costs and losses (Al-Dairi et al., 2023). Proper ripeness classification is also beneficial for growers, consumers, vendors (Ronaghi, 2021); (Lin et al. 2020). Overall, avocado ripeness classification is essential to optimise fruit sorting, processing and storage, and to ensure consumer satisfaction and reduce economic losses.

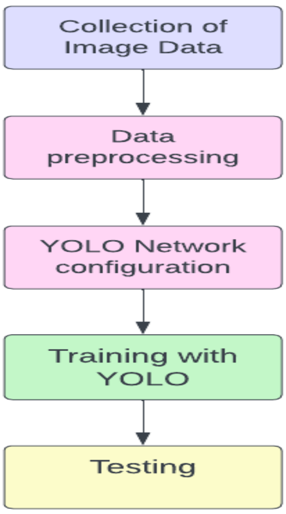
Some research related to avocado fruit maturity has been carried out and obtained quite good results such as research conducted by (Tama, et al., 2022) getting 81% accuracy results, and research using machine learning and deep learning getting quite good results, but for detection in real time has never been done. classifying avocado fruit maturity in real time is important because it can help reduce post-harvest losses by sorting fruit according to its duration until it is ready to eat, ensuring that fruit is not wasted and can be marketed effectively, contributing to improving fruit quality and reducing export costs and losses.

Real-time object detection is an important task in computer vision with various applications. YOLO (You Only Look Once) is a popular algorithm for real-time object detection (Majumder & Wilmot, 2023) which processes the entire image at once and predicts bounding box and class probabilities for identified objects (V, et al., 2022). YoloV5 is an updated version of YOLO, which incorporates a feature pyramid network (FPN) and anchor boxes to improve detection accuracy (Fang, et al., 2023); (Hui, et al., 2024). The YOLO algorithm has shown high performance and fast inference speed, making it widely used and improved in various applications (Diwan, et al., 2023). This research highlights the importance of real-time detection in monitoring avocado ripeness, where the training process was carried out for 89,280 iterations which resulted in a new model for avocado ripeness detection. The final model has a mean Average Precision (mAP) validation value of 84.3%, mAP 84.3% indicates the optimal level of accuracy in object recognition in avocado fruit maturity images using the YOLO V9 model that has undergone an intensive training process.

## MATERIALS AND METHODS

The stages in this research illustrate the steps generally taken in the development of an image-based detection system. In the initial stage, the focus is on image data collection followed by image data pre-processing. This pre-processing involves labelling and adjusting the image size to suit the needs of further processing. After that, the next important step is the configuration of the YOLO neural network that is tailored to the characteristics and distinctive features of the collected image dataset. The training process is then carried out so that the YOLO model can learn from existing image data, forming a new model that is responsive and accurate in detecting certain objects (Diwan, et al., 2023). The final stage of the research focuses on testing the trained model using real-time smartphone camera data, to test the performance and responsiveness of the model in practical situations (Paul et al., 2024).

The steps presented in this research describe the general process of developing an image-based detection system. Starting from data collection, pre-processing, to the formation and training of the YOLO model, each of these important steps is the foundation that ensures responsiveness and accuracy in the detection of the object that is the focus of the research. Testing the model in a real-time environment using smartphone camera data is a crucial final step in evaluating the reliability and accuracy of the developed model. Thus, these steps illustrate the journey from data collection to testing that helps in building a reliable and adaptive detection system to practical situations.




Source : (Research Results, 2023)
Figure 1. Research Scheme

## A. Image Data Collection

Image data collection for avocado ripeness detection research covering three classes, namely unripe, ripe, and overripe, is a process that involves capturing images of avocados from various angles and conditions. These images are taken using a camera device that is specially calibrated to capture the critical details that describe the ripeness level of the avocado. This image capture process can be done in the field at the time of harvest, taking into account variations in colour, texture and size of the avocado fruit within each grade. This image data is important for training the ripeness detection algorithm, as each class has different visual characteristics that need to be accurately identified by the system.

After image data collection, the next step is data processing to prepare a suitable dataset for training the avocado ripeness detection model. This processing involves preprocessing the images to adjust the image quality, segmenting the avocado fruit from the background, and tagging each image with the appropriate class label of unripe, ripe, or overripe. Good dataset quality is key in training the detection model, so each image belonging to the three ripeness classes must be represented precisely and accurately. This image data collection is an important foundation in the development of a reliable avocado ripeness detection system to support the agricultural industry in managing harvest more efficiently. Data taken from : https://data.mendeley.com/research-data/?search=avocado%20fruit

## B. Data Preprocessing

The image data processing process started with 900 data, consisting of 300 images for each avocado ripeness class, namely unripe, ripe, and overripe. After going through the augmentation process, the total data reached 1365 images covering a wider variety within each class. Of the total data, 1095 images were allocated as training data to train the detection model, while 180 images were used as validation data to test the performance of the model during the training process. Furthermore, 90 images were taken as testing data to measure the accuracy and final performance of the trained detection model.

This proportional split is very important in the development of the avocado ripeness detection model. Larger training data allows the model to learn better and recognise more complex patterns in maturity classification. Validation data helps measure the generalisation ability of the model, while test data ensures that the model can accurately recognise avocado ripeness on a dataset that has never been seen before. With this proportional arrangement, the model can produce consistent and reliable results in classifying avocado ripeness under various conditions.

Table 1. Dataset Distribution

| Description | Unripe | Ripe | Over Ripe | Total |
|---|---|---|---|---|
| Initial Amount | 300 | 300 | 300 | 900 |
| After Augmentasi | 155 | 155 | 155 | 1365 |
| Training Data | 630 | | | 630 |
| Validation Data | 180 | | | 180 |
| Testing Data | 90 | | | 90 |

Source : (Research Results, 2023)

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5043

### C. YOLO Network Configuration

In the process of configuring YOLO for avocado ripeness detection, determining the batch size is crucial. The use of a batch size of 64 is recommended in the YOLO system for an optimal training process. This batch size is the number of images processed simultaneously in one iteration during model training. With an appropriate batch size, such as 64, the system can process a sufficiently large number of images for each iteration, improving training speed and efficiency. The use of a customized batch size recommended by the YOLO system helps in optimising the learning process of the detection model, ensuring that the model can efficiently and accurately learn avocado ripeness characteristics from the existing image dataset.

### D. Training with YOLO

The YOLO model applied in this study has a neural network structure consisting of several convolutional layers with a 3 × 3 kernel and a max-pooling layer with a 2 × 2 kernel, as seen in Figure 2. The last convolutional layer uses a 1 × 1 kernel that aims to reduce the data to a grid size of 13 × 13 × 40. The 13 × 13 dimension represents the grid size of the input image processed by the model, while the number 40 is obtained from the sum according to the filter formula used in that layer.

This shows that the YOLO model has been specifically designed with a neural network structure that allows processing images of 13 × 13 grid size with success in extracting information related to avocado ripeness from the given image data.

### E. Testing

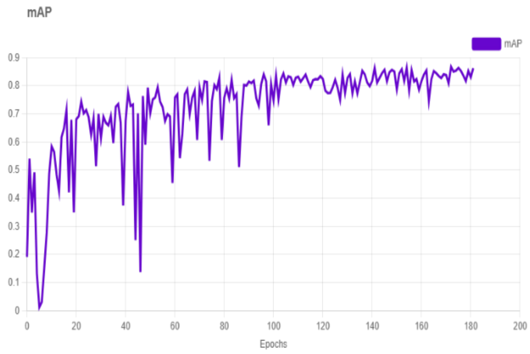
Testing is done in real-time using a smartphone camera as the main tool. The use of a smartphone camera is a test method used to measure accuracy in detecting the ripeness level of avocados by utilising a pre-trained detection model. In this testing process, the smartphone camera acts as an image capture tool that allows direct evaluation of the responsiveness and accuracy of the detection model to avocado fruit objects in real situations, presenting real-time data to evaluate model performance more carefully.

## RESULTS AND DISCUSSION

This study documents avocado fruit ripeness results in real-time using the YOLO (You Only Look Once) approach in image analysis. This method enables precise and efficient identification and classification of avocado fruit ripeness, allowing accurate direct observation of changes in fruit ripeness in an instantaneous and continuous manner tailored to the research scheme.

### a. Training Results

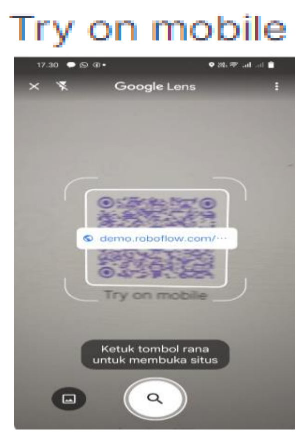
The training process was carried out as many as 89,280 iterations which resulted in a new model for avocado ripeness detection. The final model created has a mean Average Precision (mAP) validation value of 86.4%, as seen in Figure 2. The mAP figure of 86.4% signifies an optimal level of accuracy in object recognition in avocado fruit ripeness images using the YOLO model that has undergone an intensive training process. This confirms that the model is able to very well identify avocado fruit ripeness in the image, showing a maximum level of accuracy in ripeness classification.




Source : (Research Results, 2023)
Figure 2. mean Average Precision (mAP)

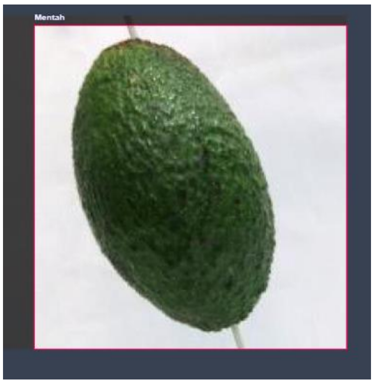
### b. Test Results on smartphone camera

To conduct initial testing, it is done by scanning the barcode that has been built with YOLO for all smartphone cameras using Google Lens as shown in Figure 3.




Source : (Research Results, 2023)
Figure 3. Barcode from YOLO to The Link Demo

One of the test results on an image with an avocado object can be seen in Figure 4. Based on Figure 4, the object recognition accuracy using a smartphone camera is 99%. Object recognition on video can give different accuracy results when the recognised object moves position.

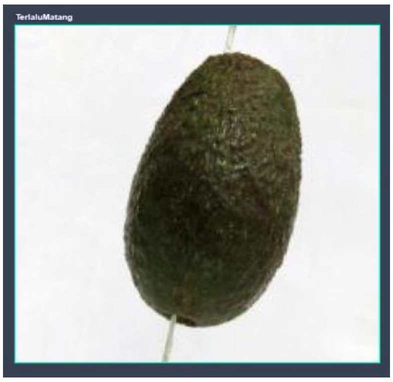



Source : (Research Results, 2023)
Figure 4. Demo of Detection Result Image with Unripe Category

Figure 4 is a demo of a ripeness detection result for an avocado fruit in the "Unripe" category, showing a green avocado fruit placed on a white background. The avocado fruit displays the typical characteristics of ripeness that has not yet been reached; an oval shape with a light green colour, a smooth skin without brown spots or indentations, and still attached to the stalk. The text accompanying the image, "Unripe," explains that the avocado has not reached the ideal level of ripeness for consumption. This image can educate users about the characteristics of a ripe avocado, which usually has a darker skin and brown spots, as well as soft and creamy flesh. By illustrating the visual difference between unripe, ripe and overripe avocados, this image can help in choosing the right avocado for consumption.


Source : (Research Results, 2023)
Figure 5. Demo Image of Detection Result with Ripe Category

Figure 5 shown depicts an avocado fruit with a "Ripe" category that matches its category. The avocado has an elongated shape with a dark green colour. The skin appears smooth with some brown patches visible on the underside. In addition, this avocado fruit has been detached from the stalk. In this experiment, the results obtained are suitable for the classification of avocado ripeness.




Source : (Research Results, 2023)
Figure 6. Demo Image of Detection Result with Over Ripe Category

Figure 6 is a demo of the detection results using the "Overripe" category with an experiment conducted in real-time using the YOLO model. The image shows an avocado fruit that has reached an excessive level of ripeness. This avocado is characterised by an oval shape but with a darker skin colour, looking very dark green and even brown. It appears that the skin has developed some brown patches in various parts, showing signs of overripe. In the context of experimenting with the YOLO model in real-time, this image provides a clear picture of the level of ripeness of an avocado that has passed its ideal ripeness point.

Based on several experiments, it was found that the model built with YOLO for the classification of avocado ripeness in real time obtained significant results and the accuracy results exceeded the research conducted by (Tama, et al., 2022) here the accuracy obtained was only 81%.

## CONCLUSION

Avocado ripeness classification is an important aspect of post-harvest management and distribution. This process affects the fruit's shelf life, skin colour, flesh firmness and quality during storage. Determining the right ripeness helps in effective management and also reduces costs and export losses. This classification also provides benefits to growers, consumers, and vendors, ensuring consumer satisfaction and reducing economic losses. This study emphasises the importance of real-time detection in monitoring avocado ripeness. This study highlights the importance of real-time detection of avocado ripeness, which is an area of research that has not been done before. The method of using YOLO (You Only Look Once) to monitor avocado ripeness directly at the time of ripeness change is an innovative step. This research highlights the importance of real-time detection in monitoring avocado ripeness, where the training process was carried out for 89,280 iterations which resulted in a new model for avocado ripeness detection. The final model has a mean Average Precision (mAP) validation value of 84.3%, mAP 84.3% indicates the optimal level of accuracy in object recognition in avocado fruit maturity images using the YOLO V9 model that has undergone an intensive training process. Further research can improve the accuracy of avocado ripeness detection in more diverse test conditions, such as light and background variations or integrate other sensor or image processing technologies that can assist in detecting avocado ripeness more accurately and efficiently in real situations.

## REFERENCE

[BPS] Badan Pusat Statistik. 2022. Total Produksi Alpukat Tahun 2021. https://sulut.bps.go.id/statictable/2022/06/24/200/produksi-buah-buahan-dan-sayuran-tahunan-menurut-jenis-tanaman-kuintal-2019-2021.html.

Adetuyi, B. O., Odine, G. O., Olajide, P. A., Adetuyi, O. A., Atanda, O. O., & Oloke, J. K. (2022). Nutraceuticals: role in metabolic disease, prevention and treatment. World News of Natural Sciences, 42, 1-27.

Al-Dairi, M., Pathare, P. B., Al-Yahyai, R., Jayasuriya, H., & Al-Attabi, Z. (2023). Postharvest quality, technologies, and strategies to reduce losses along the supply chain of banana: A review. Trends in Food Science & Technology, 134, 177-191.

Cho, B. H., Koyama, K., Olivares Díaz, E., & Koseki, S. (2020). Determination of "Hass" avocado ripeness during storage based on smartphone image and machine learning model. Food and Bioprocess Technology, 13(9), 1579-1587.

Diwan, T., Anirudh, G., & Tembhurne, J. V. (2023). Object detection using YOLO: Challenges, architectural successors, datasets and applications. multimedia Tools and Applications, 82(6), 9243-9275.

Fang, Y., Ma, Y., Zhang, X., & Wang, Y. (2023). Enhanced YOLOv5 algorithm for helmet wearing detection via combining bi-directional feature pyramid, attention mechanism and transfer learning. Multimedia Tools and Applications, 82(18), 28617-28641.

Han, Y., Bai, S. H., Trueman, S. J., Khoshelham, K., & Kämper, W. (2023). Predicting the ripening time of 'Hass' and 'Shepard'avocado fruit by hyperspectral imaging. Precision Agriculture, 24(5), 1889-1905.

Hui, Y., You, S., Hu, X., Yang, P., & Zhao, J. (2024). SEB-YOLO: An Improved YOLOv5 Model for Remote Sensing Small Target Detection. Sensors, 24(7), 2193.

Jaramillo-Acevedo, C. A., Choque-Valderrama, W. E., Guerrero-Álvarez, G. E., & Meneses-Escobar, C. A. (2020). Hass avocado ripeness classification by mobile devices using digital image processing and ANN methods. International Journal of Food Engineering, 16(12), 20190161.

Majumder, M., & Wilmot, C. (2023). Automated vehicle counting from pre-recorded video using you only look once (YOLO) object detection model. Journal of imaging, 9(7), 131.

Pacheco, L. S., Li, Y., Rimm, E. B., Manson, J. E., Sun, Q., Rexrode, K., ... & Guasch-Ferré, M. (2022). Avocado consumption and risk of cardiovascular disease in US adults. Journal of the american heart association, 11(7), e024014.

Paul, A., Machavaram, R., Kumar, D., & Nagar, H. (2024). Smart solutions for capsicum Harvesting: Unleashing the power of YOLO for Detection, Segmentation, growth stage Classification, Counting, and real-time mobile identification. Computers and Electronics in Agriculture, 219, 108832.

Ronaghi, M. H. (2021). A blockchain maturity model in agricultural supply chain. Information Processing in Agriculture, 8(3), 398-408.

Shrestha, A. (2022). Oxidative Rancidity Of Avocado Oil (Hass Variety) Produced By Screw Pressing And Solvent Extraction Method (Doctoral dissertation, Tribhuvan University Institute of Science and Technology Food Technology Instruction Committee Central Campus of Technology, Dharan).

Tama, B. A., Vania, M., Lee, S., & Lim, S. (2023). Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals. Artificial Intelligence Review, 56(5), 4667-4709.

V, V., K, C. R., & C., R. A. (2022). Real Time Object Detection System with YOLO and CNN Models: A Review. Journal of Xi'an University of Architecture & Technology, XIV(7), 144–151. Retrieved from http://arxiv.org/abs/2208.00773

# APPLICATION OF FUZZY LOGIC AND GENETIC ALGORITHM APPROACHES IN EVALUATION OF GAME DEVELOPMENT

**Daniati Uki Eka Saputri[1]; Faruq Aziz[2*]; Nurul Khasanah[3]; Taopik Hidayat[4]; Rendi Septian[5]**

Information Systems[1,2,5], Informatics[3], Data Sains[4]
Universitas Nusa Mandiri, Jakarta, Indonesia[1,2,3,4,5]
https://www.nusamandiri.ac.id[1,2,3,4,5]
daniati.due@nusamandiri.ac.id[1], faruq.fqs@nusamandiri.ac.id[2*], nurul.nuk@nusamandiri.ac.id[3], taopik.toi@nusamandiri.ac.id[4], 14002348@nusamandiri.ac.id[5]
(*) Corresponding Author



***Abstract***— *The gaming industry is undergoing rapid evolution, presenting developers with intricate challenges in selecting compelling and successful game concepts. To tackle these challenges, decision support systems (DSS) play an increasingly crucial role in facilitating accurate decision-making. Despite their growing importance, the adoption of DSS within the gaming sector remains limited. Therefore, scientific research focused on developing DSS to evaluate optimal game concepts is essential to foster innovation in gaming industries. This study aims to construct a decision support system utilizing fuzzy logic and optimized with genetic algorithms to assess and identify game concepts with the highest potential for success in the market. Evaluation results highlight the system's effectiveness in recommending top-quality games like "Clash of Clans," "Honor of Kings," and "Genshin Impact," renowned for delivering exceptional gaming experiences and receiving high ratings. The system evaluation achieved an average Mean Squared Error (MSE) of 0.0246, indicating accurate prediction of game ratings with minimal error. The significance of this research extends beyond advancing decision support systems in gaming, opening avenues for further advancements in optimizing game evaluations and similar technologies across industries grappling with data-driven decision-making challenges.*

***Keywords:*** *Decision Support Systems, Games, Fuzzy Logic, Genetic Algorithms.*

**Abstrak**— Industri game sedang mengalami evolusi cepat, memberikan tantangan kompleks bagi pengembang dalam memilih konsep game yang menarik dan sukses. Untuk mengatasi tantangan ini, sistem pendukung keputusan (DSS) memainkan peran yang semakin penting dalam memfasilitasi pengambilan keputusan yang akurat. Meskipun pentingnya semakin meningkat, adopsi DSS dalam sektor game masih terbatas. Oleh karena itu, penelitian ilmiah yang berfokus pada pengembangan DSS untuk mengevaluasi konsep game optimal sangat penting untuk mendorong inovasi dalam industri game. Studi ini bertujuan untuk membangun sistem pendukung keputusan yang menggunakan logika fuzzy dan dioptimalkan dengan algoritma genetika untuk menilai dan mengidentifikasi konsep game dengan potensi kesuksesan tertinggi di pasar. Hasil evaluasi menunjukkan efektivitas sistem dalam merekomendasikan game berkualitas tinggi seperti "Clash of Clans," "Honor of Kings," dan "Genshin Impact," yang terkenal karena memberikan pengalaman bermain game yang luar biasa dan mendapatkan rating tinggi. Evaluasi sistem ini mencapai Mean Squared Error (MSE) rata-rata sebesar 0.0246, menunjukkan prediksi rating game yang akurat dengan kesalahan minimal. Signifikansi dari penelitian ini tidak hanya berdampak pada kemajuan sistem pendukung keputusan dalam industri game, tetapi juga membuka jalan untuk kemajuan lebih lanjut dalam mengoptimalkan evaluasi game dan teknologi serupa di berbagai industri yang menghadapi tantangan pengambilan keputusan berbasis data.

***Kata Kunci****: Sistem Pendukung Keputusan, Game, Fuzzy Logic, Algoritma Genetika.*

## INTRODUCTION

The modern gaming industry continues to experience rapid growth, making it one of the most dynamic and innovative sectors worldwide (Rahajaan & Yaurwarin, 2022). The accessibility of games across various platforms, such as computers and mobile devices, has provided users with a wide range of entertainment options while creating profitable business opportunities for game developers, e-sports athletes, and sellers of gaming accessories (Dhanandjaya et al., 2022).

As the gaming industry continues to expand rapidly, game developers face a multitude of complex challenges in creating games that not only capture attention but are also favored by gamers. To achieve success, developers must demonstrate ingenuity and originality in their selection of concepts, designs, and features that most appeal to players (Dimitriadou et al., 2021). A game can no longer rely solely on technical aspects such as graphics and gameplay; careful consideration must also be given to elements such as game mechanics, design aesthetics, and complexity levels (Spieler & Kemeny, 2020).

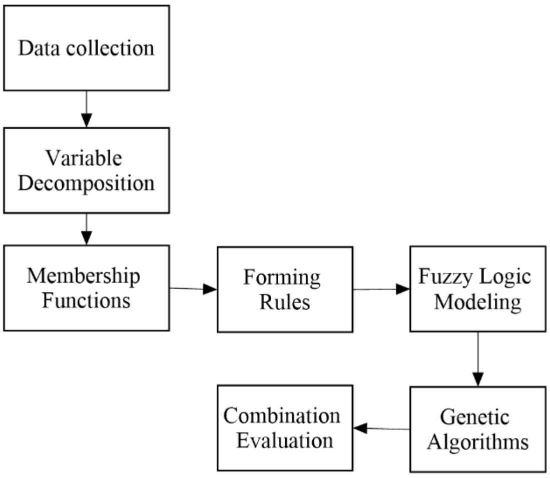
The role of Decision Support Systems (DSS) is becoming increasingly vital in aiding developers in making accurate decisions. A DSS is a system designed to facilitate the decision-making process by providing essential information, models, and tools for evaluating and analyzing data(Mahendra et al., 2023), (Aziz et al., 2024), without replacing the input from decision-makers (Novita et al., 2022). Specifically in the gaming industry, DSS can contribute to assessing the quality of a game based on a set of predetermined criteria, thereby enabling developers to make more precise and efficient decisions (Mu'alimin & Latipah, 2021).

The concept of Decision Support Systems (DSS) has been applied in various commercial and industrial scenarios (Wibowo et al., 2020), (Aziz et al., 2020), (Su et al., 2023), (Kabadurmus et al., 2023), (Purnamawati et al., 2023), (Prasetyo & Prasetyaningrum, 2023). Research on decision support systems in the gaming world has been conducted (Altanny & Johan, 2023), where a website was designed to rank the best characters in the game Genshin Impact based on five combat roles. Using the Simple Additive Weighting (SAW) method, it successfully helped F2P players choose the most suitable character. Previous research by (Mu'alimin, 2021) developed an application to help parents select games for young children according to predetermined criteria using the Topsis method. While decision support systems can serve as a reference for future evaluations, no study has specifically implemented fuzzy methods and genetic algorithms to evaluate the best game concepts from a developer's perspective. Evaluating game concepts often faces high uncertainty and complexity due to the numerous factors that must be considered, making conventional evaluation methods potentially ineffective in addressing these issues.

This research aims to develop a decision support system to assist game developers in evaluating and selecting game concepts that possess the greatest potential for success in the market. This system takes into account various key aspects such as graphics, gameplay, narrative, innovation, sound quality, and developer reputation. By implementing fuzzy methods and optimizing with genetic algorithms, this study aims to create an evaluation model that can enhance accuracy and effectiveness in the selection process of the most likely successful games, while addressing the uncertainties and complexities in evaluation. Therefore, this research is expected to make a positive contribution towards improving the quality and performance of games under development.

## MATERIALS AND METHODS

The following represents the research stages conducted by the researcher, as shown in Figure 1. The dataset used in this study was obtained from the Kaggle website, which contains data on mobile and PC games for the game development process (Fadly, 2023). This dataset includes information about various well-known games on mobile platforms. This information encompasses several assessment factors used to determine the quality of each game, covering aspects such as graphics, gameplay, storyline, innovation, audio quality, and developer. The evaluation for each of these factors employs a scale from 1 to 10, with higher numbers indicating superior quality.




Source: (Research Results, 2024)
Figure 1. Research methodology

In complex decision-making, we often encounter situations where available data is incomplete or uncertain, and the criteria used to evaluate alternatives can be highly diverse and difficult to measure directly. It is in this context that the strengths of two distinct approaches, namely fuzzy logic and genetic algorithms, can be combined to provide a more effective solution.

Fuzzy logic is a mathematical algorithm that utilizes computers to mimic the way humans comprehend and make decisions in the real world (Kharisma et al., 2023). This technique simplifies problem modeling by allowing for uncertainty and ambiguity in human judgment. Through fuzzy logic, we can create criteria that are more flexible and precise for evaluating complex situations, thereby handling uncertain data more efficiently (Junaidi, 2023). For example, in game evaluation, criteria such as "graphics," "gameplay," and "storyline" are often difficult to measure accurately, and fuzzy logic enables us to describe these criteria in finer shades of gray.

On the other hand, genetic algorithms have the advantage of being adaptable and efficient without requiring gradient information (Acampora et al., 2023). Their ability to explore search spaces globally and locally makes them highly effective in solving various optimization problems. Genetic algorithms are also easy to implement, parallelizable, and flexible (Junaidi, 2023), (Acampora et al., 2023). With genetic algorithms, we can tackle various optimization challenges and search for solutions (Aziz et al., 2024).

The combination of fuzzy logic and genetic algorithms can enhance system performance by addressing uncertainty, optimizing parameters, and identifying optimal input combinations. This enables stronger solutions that are more adaptable to complex problems. Therefore, their combination can offer superior decision-making solutions, particularly when dealing with uncertain data and evaluations involving numerous complex criteria.

## RESULTS AND DISCUSSION

### A. Datasets

Table 1 presents several datasets of mobile and PC games containing various criteria, including game title, publisher, rating, graphics, gameplay, storyline, innovation, sound, and developer (Fadly, 2023).

Table 1. Mobile and PC Game Data

| Game Title | Publisher | Rating | Graphics | Gameplay | Storyline | Innovation | Sound | Developer |
|---|---|---|---|---|---|---|---|---|
| Clash of Clans | Supercell | 4,7 | 6 | 7 | 5 | 6 | 6 | 7 |
| Honor of Kings | Tencent | 4,7 | 8 | 8 | 5 | 7 | 8 | 8 |
| Genshin Impact | miHoYo | 4,6 | 9 | 9 | 9 | 8 | 9 | 9 |
| Minecraft | Mojang | 4,6 | 8 | 9 | 6 | 8 | 8 | 6 |
| Candy Crush Saga | King | 4,6 | 6 | 7 | 5 | 6 | 6 | 6 |

Source: (Fadly, 2023)

### B. Variable Decomposition

In this stage, fuzzy variables are initialized for each assessed aspect of the game, with antecedents for inputs and consequents for outputs within the control system. The input variables include graphics, gameplay, storyline, innovation, sound, and developer, while the output variable is the rating. These input variables are considered to have a significant influence on the overall quality of the game. Each variable is defined within a range of values from 0 to 10, with an interval of 1, as depicted in Figure 2.

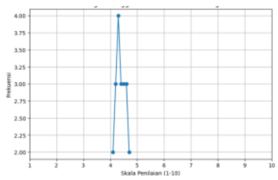


Figure 2(a) Rating Scale

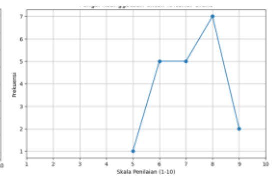


Figure 2(b) Graphic Scale

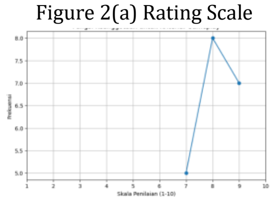


Figure 2(c) Gameplay Scale

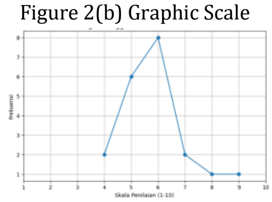


Figure 2(d) Story Scale

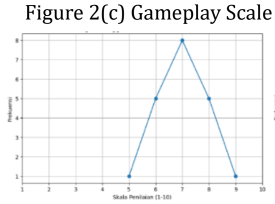


Figure 2(e) Innovation Scale

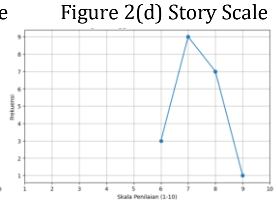


Figure 2(f) Sound Scale

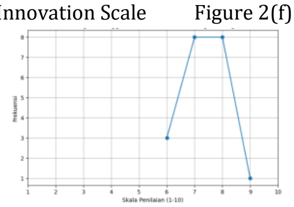


Figure 2(g) Developer Scale

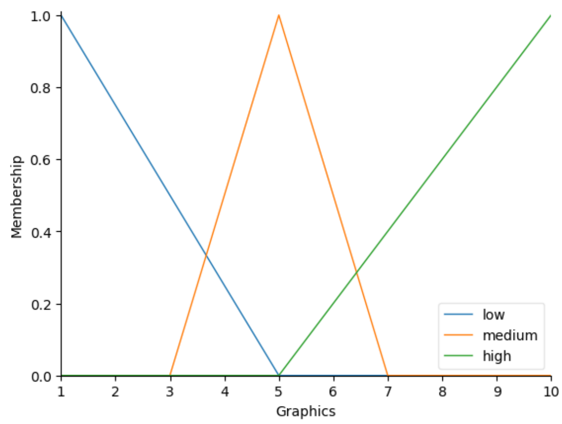
Source: (Research Results, 2024)

Figure 2. Scale of Each Variable

In Figure 2, the horizontal axis of each image represents the rating scale from 1 to 10, which is used to assess the games based on the established criteria. The vertical axis indicates the frequency or the number of games receiving a particular rating on this scale. Each point on the plot signifies how many games have received a specific rating. The higher the position of the point, the greater the number of games that have obtained that rating.

## C. Defining Membership Functions

The researcher then employed the automf(3) function, also known as the simple triangular membership function, as it is straightforward to implement and sufficiently effective for the system. The automated process generates membership functions divided into three categories: low, medium, and high. These membership functions determine how input values are categorized into each linguistic label. An illustration of the membership function for one of the variables defined in Figure 3 below is provided, where the low criterion resides in trimf [1, 1, 5], the medium criterion in trimf [3, 5, 7], and the high criterion in trimf [5, 5, 10].




Source: (Research Results, 2024)
Figure 3. One of the Membership Functions in Graphic Variables

## D. Forming Fuzzy Rules

In this stage, fuzzy rules are formulated based on fuzzy logic as elucidated in Table 2. These rules dictate how the combination of fuzzy input values will yield specific fuzzy output values.

Table 2. Fuzzy Rules

| Rules | Description |
|---|---|
| Rules 1 | If any of the graphics, gameplay, story, innovation, sound, or developer is 'low', then the rating is 'low' |
| Rules 2 | If all graphics, gameplay, story, innovation, sound, and developer are 'medium', then the rating is 'medium' |
| Rules 3 | If any of the graphics, gameplay, story, innovation, sound, or developer is 'high', then the rating is 'high' |

Source: (Research Results, 2024)

## E. Fuzzy Logic Modeling and Genetic Algorithms

In this stage, a fuzzy logic model is constructed to evaluate games based on predetermined criteria. In this model, both input and output variables are utilized to determine the quality rating of the games using fuzzy rules. The outcomes triggered by the rules are then aggregated to produce the final fuzzy output, which is subsequently defuzzified (converted from fuzzy values to crisp/single values) using the centroid method to generate the final rating score.

The model is constructed by inputting values for each variable, namely graphics, gameplay, storyline, innovation, sound, and developer, with scores of 8, 9, 7, 8, 9, and 8 respectively. The fuzzy control system generates a rating of 8.1, involving the evaluation of various predefined fuzzy rules that integrate diverse assessment aspects to produce a comprehensive output. The system yields a relatively high rating because many rules supporting high outputs are triggered, particularly emphasizing graphics and gameplay, which are crucial factors in this evaluation, both obtaining high scores.

Table 3. Game Rating Scores Using the Fuzzy Logic Model

| Game Title | Publisher | Rating | Graphics | Gameplay | Storyline | Innovation | Sound | Developer | Rating score |
|---|---|---|---|---|---|---|---|---|---|
| Fortnite | Epic Games | 4.2 | 9 | 9 | 6 | 7 | 7 | 8 | 8.2777778 |
| Genshin Impact | miHoYo | 4.6 | 9 | 9 | 9 | 8 | 9 | 9 | 8.2777778 |
| Roblox | Roblox Corporation | 4.4 | 8 | 8 | 7 | 8 | 8 | 8 | 8.1428571 |

Source: (Research Results, 2024)

In Table 3, games such as "Fortnite" and "Genshin Impact" obtained high scores, reflecting their high quality in various aspects such as graphics, gameplay, and developers. The next step involves the application of genetic algorithms to select combinations of games that provide the highest total rating. This process is conducted by representing each game as an individual in the population, where each individual has genes to indicate whether the game is selected or not.

Genetic algorithms execute an evolutionary process using the described genetic operators and run for 50 generations with a crossover probability of 0.5 and a mutation probability of 0.2. In each generation, individuals with the best fitness values are selected, and this process repeats to create a new generation. Through this evolutionary process, the population of individuals is analyzed and modified to search for optimal solutions.

Table 4. Best Games Based on Genetic Algorithm Rating Scores

| Game Title | Publisher | Rating | Graphics | Gameplay | Storyline | Innovation | Sound | Developer | Rating score |
|---|---|---|---|---|---|---|---|---|---|
| Clash of Clans | Supercell | 4.7 | 6 | 7 | 5 | 6 | 6 | 7 | 7.7407741 |
| Honor of Kings | Tencent | 4.7 | 8 | 8 | 5 | 7 | 8 | 8 | 8.1428557 |
| Genshin Impact | miHoYo | 4.6 | 9 | 9 | 9 | 8 | 9 | 9 | 8.2777778 |
| Candy Crush Saga | King | 4.6 | 6 | 7 | 5 | 6 | 6 | 6 | 7.7407741 |
| Minecraft | Mojang | 4.6 | 8 | 9 | 6 | 8 | 8 | 6 | 8.1428557 |

Source: (Research Results, 2024)

After the genetic algorithm explored through various combinations of populations, the top 5 games with the highest ratings from the best-found population are Clash of Clans, Honor of Kings, Genshin Impact, Candy Crush Saga, and Minecraft as seen in Table 4 above.

### F. Combination Evaluation

After the model optimization process is completed, an evaluation is conducted on the best population to select the recommended games. The selected games are then assessed based on ratings and other criteria. From the five selected games, we can observe a good diversification both in terms of publishers (Supercell, Tencent, miHoYo, King, Mojang) and the types of gameplay offered, ranging from strategy and adventure to puzzle, indicating that the algorithm successfully selects games that are not only high in ratings but also offer a variety of experiences for players. All selected games have relatively high ratings, indicating that the algorithm successfully selects games that are not only popular but also of high quality in various aspects such as graphics, gameplay, and storyline.

The system evaluation was conducted using cross-validation by calculating the Mean Squared Error (MSE) to measure the accuracy of the predicted ratings compared to the actual values. The average MSE obtained was 0.0246, indicating that the fuzzy system is quite accurate in predicting game ratings with a low prediction error.

## CONCLUSION

The evaluation results demonstrate that this system is capable of providing high-quality game recommendations based on aggregate values generated through the defuzzification process. The recommended games, including "Clash of Clans," "Honor of Kings," "Genshin Impact," "Candy Crush Saga," and "Minecraft," indicate that the model can select games that are not only popular but also have high ratings and offer a quality gaming experience. Furthermore, this study opens up opportunities for further development and research in optimizing game evaluation systems and similar technologies in other sectors facing similar challenges in data-driven decision making. This includes expanding the dataset to encompass more types of games and user data, as well as integrating other machine learning techniques such as deep learning and reinforcement learning. A more comprehensive evaluation with cross-validation techniques and real-world testing is necessary for further validation.

## REFERENCE


Acampora, G., Chiatto, A., & Vitiello, A. (2023). Genetic algorithms as classical optimizer for the Quantum Approximate Optimization Algorithm. *Applied Soft Computing*, *142*, 110296. https://doi.org/10.1016/j.asoc.2023.110296

Altanny, P., & Johan, M. E. (2023). Web-based Decision Support System for Characters Selection in Game Genshin Impact with SAW Method. *Ultima InfoSys : Jurnal Ilmu Sistem Informasi*, *14*(1), 40–51. https://doi.org/10.31937/si.v14i1.3207

Aziz, F., Said, F., & Sudrajat, A. (2020). Penerapan Konsep Finite State Automata Dalam Proses Pendaftaran Kelas Kursus Bahasa

Inggris Pada Tempat Kursus. *MATICS: Jurnal Ilmu Komputer Dan Teknologi Informasi (Journal of Computer Science and Information Technology)*, *12*(2), Article 2. https://doi.org/10.18860/mat.v12i2.9330

Aziz, F., Yanto, Y., & Juningsih, E. H. (2024). Rancang Bangun Sistem Penunjang Keputusan Penentuan Beasiswa Menggunakan Metode Fuzzy Tsukamoto Dengan Optimasi Genetic Algorithm. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), Article 1. https://doi.org/10.36040/jati.v8i1.9338

Dhanandjaya, P. B. G., Budiartha, I. N. P., & Arini, D. G. D. (2022). Penyalahgunaan Benda Virtual dalam Permainan Game Online di Indonesia. *Jurnal Konstruksi Hukum*, *3*(3), Article 3. https://doi.org/10.55637/jkh.3.3.5349.569-575

Dimitriadou, A., Djafarova, N., Turetken, O., Verkuyl, M., & Ferworn, A. (2021). *Challenges in Serious Game Design and Development: Educators' Experiences*. *52*(2). https://doi.org/10.1177/1046878120944197

Fadly, D. (2023). *Data Game Mobile and PC*. https://www.kaggle.com/datasets/gamerengineering/data-game-mobile-and-pc

Junaidi. (2023). Implementasi Fuzzy Logic Dengan Metode Mamdani Untuk Sistem Pendukung Keputusan Kinerja Dosen. *Jurnal Information System*, *3*(1), 17–27. https://doi.org/10.61488/jis.v3i1.256

Kabadurmus, O., Kayikci, Y., Demir, S., & Koc, B. (2023). A data-driven decision support system with smart packaging in grocery store supply chains during outbreaks. *Socio-Economic Planning Sciences*, *85*, 101417. https://doi.org/10.1016/j.seps.2022.101417

Kharisma, L. P. I., Yahya, S. R., Sepriano, Handayanto, R. T., Herlawati, Gunawan, I. M. A. O., Handika, I. P. S., Hatta, H. R., & Syamil, A. (2023). *Metode Spk Favorit Di Masa Depan: Teori dan Contoh*. PT. Sonpedia Publishing Indonesia.

Mahendra, G. S., Tampubolon, L. P. D., Herlinah, Arni, S., Kharisma, L. P. I., Resmi, M. G., Sudipa, I. G. I., Khairunnisa, Ariana, A. A. G. B., Syam, S., & Edi. (2023). *Sistem Pendukung Keputusan (Teori dan Penerapannya dalam berbagai Metode)*. PT. Sonpedia Publishing Indonesia.

Mu'alimin, M. (2021). Sistem Pendukung Keputusan Aplikasi Pemilihan Game Android Untuk Anak Usia Dini. *JSiI (Jurnal Sistem Informasi)*, *8*(1). https://doi.org/10.30656/jsii.v8i1.3027

Mu'alimin, M., & Latipah, L. (2021). Sistem Pendukung Keputusan Aplikasi Pemilihan Game Android Untuk Anak Usia Dini. *JSiI (J-Urnal Si-Stem I-Nformasi)*, *8*(1), Article 1.

Novita, L., Khasanah, S. N., & Saputri, D. U. E. (2022). Sistem Penunjang Keputusan Pemilihan Salesman Terbaik Menggunakan Metode Analytical Hierarki Process Pada PT.Cahaya Esa Karunia Optima. *Indonesian Journal on Software Engineering (IJSE)*, *8*(1), Article 1. https://doi.org/10.31294/ijse.v8i1.11483

Prasetyo, H. A., & Prasetyaningrum, P. T. (2023). Sistem Pendukung Keputusan Pemilihan Supplier Bahan Baku Furniture Terbaik Menggunakan Metode Multi-Objective Optimization By Ratio Analysis (MOORA). *Technologia : Jurnal Ilmiah*, *14*(2), Article 2. https://doi.org/10.31602/tji.v14i2.7838

Purnamawati, A., Winarto, M. N., & Saputri, D. U. E. (2023). Sistem Pendukung Keputusan Penentuan Produk Terbaik Menggunakan Metode Preference Selection Index. *CHAIN: Journal of Computer Technology, Computer Engineering, and Informatics*, *1*(2), Article 2. https://doi.org/10.58602/chain.v1i2.28

Rahajaan, J. A., & Yaurwarin, W. (2022). Bisnis Startup Dalam Kompleksitas Hukum di Indonesia. *Journal of Business Application*, *1*(1), Article 1. https://doi.org/10.51135/jba.v1.i1.p63-72

Spieler, B., & Kemeny, F. (2020). *Design, Complexity, and Coding: A Framework to Evaluate Games*. https://bernadette-spieler.com/wp-content/uploads/2020/09/ECGBL-Proceedings-581-589.pdf

Su, D., Zhang, L., Peng, H., Saeidi, P., & Tirkolaee, E. B. (2023). Technical challenges of blockchain technology for sustainable manufacturing paradigm in Industry 4.0 era using a fuzzy decision support system. *Technological Forecasting and Social Change*, *188*, 122275. https://doi.org/10.1016/j.techfore.2022.122275

Wibowo, A., Maduretno, V. S. A. T. W., Masduki, L. R., & Supiyandi. (2020). Gaming Mobile Selection with Decision Support System using Simple Additive Weighting. *International Journal of Advanced Science and Technology*, *29*(04), Article 04.