JoTI

Journal of Technology and Informatics

Exploring the Application of Machine Learning for Automatic Inbound Email Classification in CRM System at XYZ Company
Lukman Arif Sanjani, Raden Bimo Mandala Putra, Umi Laili Yuhana

Penerapan Algoritma Apriori untuk Rekomendasi Produk Spare Part Mobil
Reza Aditya Permana, Primandani Arsi, Pungkas Subarkah

Pengaruh Big Data dalam Akuntansi Syariah: Analisis Prediktif dan Pengambilan Keputusan Strategis
Adhi Riza Aulia, Sulis Saputra, Arga Dwi Titandy, Mohammad Zacky, Agus Arwani

Prediksi Harga Mobil Audi Bekas Menggunakan Model Regresi Linear dengan Framework Streamlit
Putri Aulia Azhar, Muhammad Arya Pratama, Risma Fitriani

Perancangan E-Commerce Distributor Clothing Company Berdasarkan Product Knowledge
Nisa Eridiar, Tan Amelia, Ayouvi Poerna Wardhanie

Arrhythmia Classification with ECG Signal using Extreme Gradient Boosting (XGBoost) Algorithm
Diah Asmawati, Lukman Arif Sanjani, Christiant Dimas Renggana, Chastine Fatichah, Tanzilal Mustaqim

Membangun Sistem Katalog Digital untuk Perpustakaan SMP: Solusi Tepat Mempermudah Pencarian Buku
Mahendra Bagaskara, Erwin Sutomo, Ayuningtyas

Identifikasi Pengenalan Wajah Berdasarkan Jenis Kelamin Menggunakan Metode Convolutional Neural Network (CNN)
Natasya Chalista Imanuela Natun, Maria Angelica Santhia, Yampi R. Kaesmetan

Operating Systems (OS): An Insight Investigative Research Analysis and Future Directions
Zarif Bin Akhtar

Machine Learning Approach for Classification of Cyber Threats Actors in Web Region
Anthony Edet, Saviour Inyang, Imeh Umoren, Ubong E. Etuk

Jurnal JoTI | Volume : 6 | No : 1 | Halaman 1- 77 | E-ISSN : 2686-6102 | P-ISSN : 2721-4842

Jurnal of Technology Informatics (JoTI)
Volume 6 Nomor 1, Oktober 2024
E-ISSN : 2686-6102
P-ISSN: 2721-4842

Jurnal of Technology Informatics (JoTI) merupakan media penyampaian hasil penelitian untuk semua bidang keilmuan Teknik Informatika dan Teknik Elektro yang terbit dua kali dalamsetahun yaitu April dan Oktober, dengan E-ISSN 2686-6102 dan P-ISSN 2721-4842, yang diterbitkan oleh Universitas Dinamika pertama kali tahun 2019.

TEAM EDITORIAL

Publisher:
- Universitas Dinamika

Website:
- http:// e-journals.dinamika.ac.id/joti

Email:
- joti@dinamika.ac.id

Editor's Address:
- Raya Kedung Baruk No. 98 Surabaya

Jurnal of Technology Informatics (JoTI)
Volume 6 Nomor 1, Oktober 2024
E-ISSN : 2686-6102
P-ISSN: 2721-4842

TABLE CONTENT

Exploring the Application of Machine Learning for Automatic Inbound Email Classification in CRM System at XYZ Company 1-7
Lukman Arif Sanjani, Raden Bimo Mandala Putra, Umi Laili Yuhana

Penerapan Algoritma Apriori untuk Rekomendasi Produk Spare Part Mobil 8-16
Reza Aditya Permana, Primandani Arsi, Pungkas Subarkah

Pengaruh Big Data dalam Akuntansi Syariah: Analisis Prediktif dan Pengambilan Keputusan Strategis 17-21
Adhi Riza Aulia, Sulis Saputra, Arga Dwi Titandy, Mohammad Zacky, Agus Arwani

Prediksi Harga Mobil Audi Bekas Menggunakan Model Regresi Linear dengan Framework Streamlit 22-28
Putri Aulia Azhar, Muhammad Arya Pratama, Risma Fitriani

Perancangan E-Commerce Distributor Clothing Company Berdasarkan Product Knowledge 29-35
Nisa Eridiar, Tan Amelia, Ayouvi Poerna Wardhanie

Arrhythmia Classification with ECG Signal using Extreme Gradient Boosting (XGBoost) Algorithm 36-42
Diah Asmawati, Lukman Arif Sanjani, Christiant Dimas Renggana, Chastine Fatichah, Tanzilal Mustaqim

Membangun Sistem Katalog Digital untuk Perpustakaan SMP: Solusi Tepat Mempermudah Pencarian Buku 43-49
Mahendra Bagaskara, Erwin Sutomo, Ayuningtyas

Identifikasi Pengenalan Wajah Berdasarkan Jenis Kelamin Menggunakan Metode Convolutional Neural Network (CNN) 50-57
Natasya Chalista Imanuela Natun, Maria Angelica Santhia, Yampi R. Kaesmetan

Operating Systems (OS): An Insight Investigative Research Analysis and Future Directions 58-69
Zarif Bin Akhtar

Machine Learning Approach for Classification of Cyber Threats Actors in Web Region 70-77
Anthony Edet, Saviour Inyang, Imeh Umoren, Ubong E. Etuk

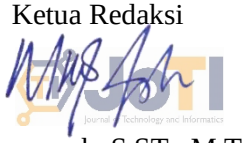
Kata Pengantar

Puji Syukur kehadirat Tuhan Yang Maha Esa atas berkat dan karunia-Nya makalah ilmiah *Jurnal of Techology Informatics* dapat terbit sebagaimana yang telah direncanakan.

Sebagai Tenaga Profesional Dosen, memiliki kewajiban mengajar, meneliti, dan melakukan pengabdian masyarakat. Setiap hasil penelitian sebaiknya dipublikasikan untuk membagi apa yang telah diteliti dan memberitahu kepada masyarakat luas mengenai hasil penelitian. JoTI diharapkan, menjadi wadah dan sarana untuk penyebaran ilmu pengetahuan dan hasil penelitian di bidang Teknik Informatika dan Teknik Elektro secara berkelanjutan. JoTI juga diharapkan menjadi wadah pertemuan para penelitian dan dunia industri yang tertarik pada hasil penelitian. Terbitan JoTI dilakukan dua kali (April dan Oktober) dalam satu tahun melalui proses *review* yang berpengalaman dan sudah memiliki makalah yang diterbitkan di jurnal Internasional.

Kami mengucapkan terimakasih kepada peneliti yang telah mengirimkan hasil penelitiannya lewat JoTI, kepada Mitra Bestari yang sudah meluangkan waktu guna me-*review* malakah yang kami ajukan, serta kepada Universitas Dinamika yang mendukung penuh atas pengelolaan jurnal ini, dan kami mengucapkan terimakasih kepada semua pihak, baik yang terlibat secara langsung ataupun tidak langsung.

Ketua Redaksi



Musayyanah, S.ST., M.T.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, October 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.715

Exploring the Application of Machine Learning for Automatic Inbound Email Classification in CRM System at XYZ Company

Lukman Arif Sanjani[1], **Raden Bimo Mandala Putra**[2], **Umi Laili Yuhana**[3]
[1,2,3]Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail: 6025232027@student.its.ac.id[1], 6025232022@student.its.ac.id[2], yuhana@if.its.ac.id[3]
* Corresponding author: E-mail: 6025232027@student.its.ac.id

Abstract: *Customer service has become increasingly crucial in today's business landscape, necessitating companies to provide fast, responsive, and personalized assistance to their clientele. However, amidst the challenges posed by surges in email volume, manual categorization and response strategies often lead to performance declines. To address this, we propose a system leveraging Machine Learning techniques for automated email classification. Our evaluation reveals promising results, with SVM achieving the highest accuracy of 96.59%, followed by XGB (96.02%) and RF (95.27%). These models exhibit commendable precision, recall, F1 scores, and Matthews Correlation Coefficient (MCC), showcasing their effectiveness in improving customer service efficiency and responsiveness. This integration of technology not only enhances operational efficiency but also fosters harmonious customer relationships, ultimately leading to increased loyalty and profitability for companies.*

Keywords*: Classification, Customer Service, Machine Learning, Random Forest, SVM, XGBoost*

INTRODUCTION

Customer service has become an increasingly important aspect in the current era, especially in the context of intensifying business competition. It is no longer sufficient to compete solely on the basis of product quality or service; companies must also be able to provide fast, responsive, and comfortable service to their customers. Customer satisfaction fosters harmonious relationships between product / service providers and consumers, ultimately leading to consumer loyalty and profitability for the company. Amidst rapid technological advancements and rising customer expectations, the integration of technology to enhance customer service has become imperative. Automation is a crucial aspect of the current Digital Era. With advancements in Artificial Intelligence (AI) and Machine Learning (ML), many companies are adopting these technologies across various operational aspects. Automation not only allows companies to improve efficiency but also enables them to deliver more personalized and responsive service to customers.

PT. XYZ is a company operating in the field of Customer Relationship Management (CRM). Being part of an industry heavily reliant on customer service, there are often significant spikes in email volume at certain times, causing difficulties for available customer service personnel to respond promptly to the influx of emails. This leads to a decline in the performance rating of the Customer Service team. To address this issue, the Customer Service team manually reads each incoming message, categorizes them, and then responds using pre-existing templates tailored to each category. While this method is somewhat helpful, it is still deemed ineffective in improving Customer Service performance during email spikes, as the team must individually read and categorize each email. To address this challenge, a system utilizing Machine Learning approaches is needed, where the system will automatically classify incoming email categories based on patterns in the text content. In this study, three machine learning algorithms were employed, as follows:

1. Support Vector Machines (SVM): Machine learning is widely used in classification, such as using SVM in email classification systems. For instance, Drucker created a spam classification system using SVM that incorporated a thousand features and spanned seven thousand dimensions, resulting in shorter training times and enhanced accuracy and speed [1]. Paper [2]proposed a spam filtering algorithm with SVM, achieving higher accuracy by reducing false positive rates. The result produced by this classifier is a hyperplane that aims to maximize the distinction between feature vectors belonging to different classes. When presented with a new instance, SVM assigns a label based on the subspace to which its feature vector is aligned.
2. Random Forest: A well-known approach in Supervised Learning is the Random Forest technique. It's proficient in addressing both classification and regression challenges within machine learning. Random Forest has demonstrated its efficacy in implementing ensemble learning concepts, where numerous classifiers are merged to tackle intricate problems and enhance model accuracy[3]. This method operates on two fundamental principles: firstly, it generates numerous decision trees and amalgamates them into a unified model, and secondly, it makes its final decision based on the majority consensus among the trees under consideration.
3. XGBoost: XGBoost Classification stands out as a robust machine learning technique tailored for classification purposes. It represents a specialized adaptation of the XGBoost algorithm crafted explicitly to address classification challenges. XGBoost, which stands for "Extreme Gradient

Boosting," has become popular due to its ability to provide highly accurate predictions in various classification tasks [4]. XGBoost has become popular for its high performance in machine learning tasks. It builds a series of decision trees sequentially, with each model correcting the errors of the previous ones. The final prediction is made by weighted averaging of all models' forecasts.

Previous research about this topic has examined email classification, primarily focusing on spam detection and grouping emails into various folders based on different categories [5][6]. Additionally, previous studies have also explored hierarchical folder classification methods [7]. While previous research focused on classifying emails into spam and non-spam categories, our research has developed a new API system that can automatically classify email categories based on the text content. The category names can be dynamically set by users of a CRM application. Furthermore, the system developed in this research has the potential to be implemented in CRM systems of different companies.

METHOD

In this research, a separate system will be developed, which will later be connected to the CRM system through an API. This is done to facilitate easier application maintenance and further development in the future. For the development of this classifier application, the Waterfall methodology was employed. The waterfall model was chosen due to its structured, sequential approach. This method covers gathering requirements, design, implementation, verification and maintenance stages, ensuring clear direction and control, and minimizing errors [8]. The stages of the Waterfall method are illustrated in Figure 1.

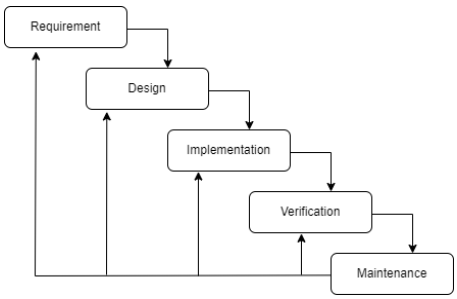


Figure 1. Waterfall SDLC Method

Requirements Analysis

In this stage, the developer initiates dialogue to grasp users' requirements and software limitations. This insight is commonly acquired through interviews, discussions, or in-person surveys. Following this, the collected data undergoes analysis to extract essential user-centric information.

System Design

During this phase, the requirements specification acquired from the preceding stage is reviewed, and the system design is formulated. The purpose of system design is to identify hardware and system prerequisites, as well as to outline the overall architecture of the system. In this study, a Use Case Diagram was also created. It was used to illustrate the interaction between actors and activities within the use case. Additionally, this diagram provides a comprehensive model of what is being done, who is involved internally, and which individuals or actors play a role externally in order to explain the various functions that users / actors can perform within the Email Classification Application [9].

Implementation

At this point, the system undergoes its initial development phase, where it is broken down into small programs termed as units, which will later be integrated in the following stage. Every unit will individually developed and then will be tested for functionality, a procedure commonly named by unit testing.

Integration Testing And Verification

Once each unit crafted during the implementation stage undergoes thorough individual testing, they are seamlessly amalgamated into the system. Following integration, rigorous testing of the entire system ensues, aiming to unearth any latent flaws or discrepancies.

Operation And Maintenance

In the concluding stage of the waterfall model, the completed software is launched and sustained through maintenance operations. This maintenance phase includes addressing any overlooked errors and optimizing system components, as well as enhancing system functionality to accommodate emerging needs.

RESULT AND DISCUSSION

Requirement

The required elements in the development of this application are the functional requirements list, as well as data from previous emails which will be needed to create the machine learning model. The main features that must be possessed by the application are Dataset Storage, Model Training, & Data Prediction.

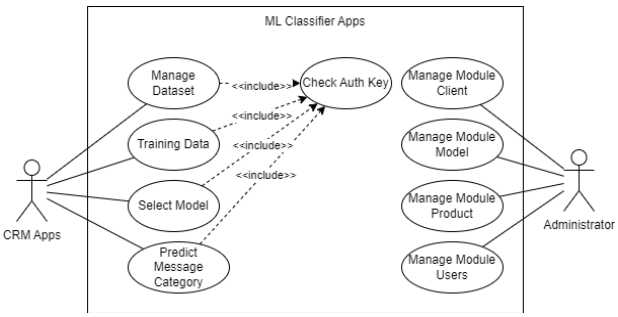


Figure 2. Usecase Diagram Email Classification Apps

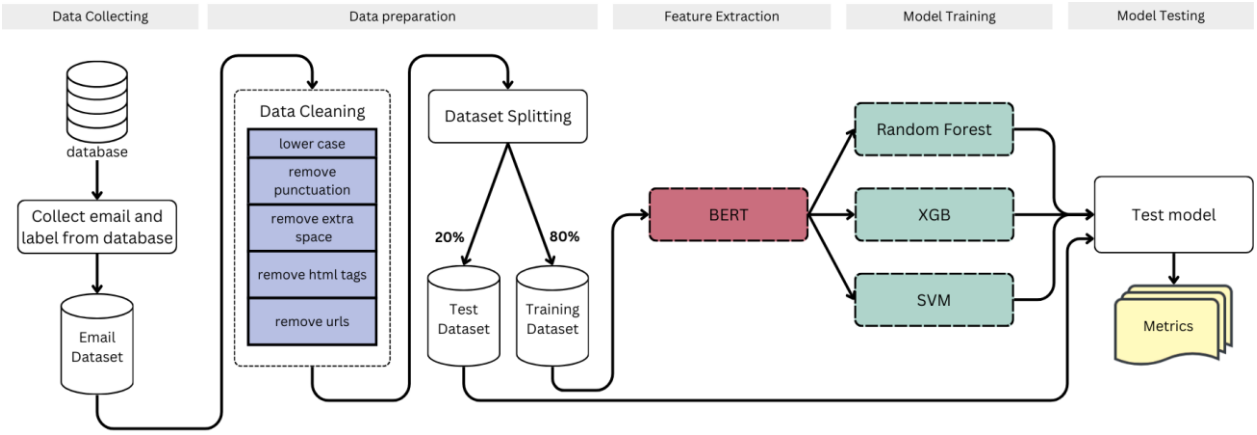


Figure 3. Machine Learning Workflow and Architecture

System Design

1. API Architecture

Initial design is required to ensure that the resulting application meets the needs and business processes of XYZ Company. This system is built with the aim that users / customer service leaders can send datasets for training, perform data training, and choose the machine learning model to be used independently. This can make the company's business processes more efficient. Of course, for the first few weeks of implementation, Customer Service must manually classify procedures first. When the data has been collected, it can be sent to the email classifier application through the existing CRM application for data training. Below is the Use Case Diagram explaining the processes that can be performed by the Customer Service Leader on the Email Classifier Application in Figure 2.

The data in this application will be stored in a Relational Database Management System (RDBMS) MySQL. All datasets received from the CRM System through the API will be stored in the training_datas table. Previously, the Administrator of this Application will configure data related to Client, available Machine Learning Models, Name of Product / Campaign / Division of the client, and user data from this application. For the model data, it will be stored in a File, then the path along with the results will be stored in the model_product table. Figure 4 shown the complete overview of the Entity Relational Diagram in this application. Here is an explanation of the functions of each table in Figure 4:

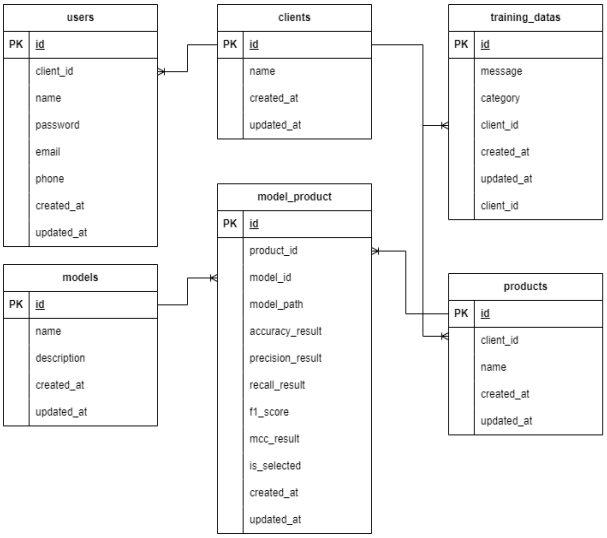


Figure 4. Entity Relational Diagram Database Email Classifier Apps

- users: This table is used to store data on the application's users in this research.
- clients: This table indicates the company from which the users engaging with the application originate. This is provided to facilitate potential integration with other companies' CRM applications.
- training_datas: This table is used to store various training data input by the users.
- models: This table stores the names of the machine learning algorithms used in this research, namely SVM, Random Forest, and XGBoost.
- products: Each client may have several service products, each with different training data. This table is needed to distinguish between these products.
- model_products: This table is used to store the results of the machine learning modeling. The model files are stored, and their paths are recorded in the model_path column. Additionally, the evaluation results of each mode are stored. This table also keeps records of which models are used by users for predictions when new email data is received into their CRM application.

In this application, communication between Customer Service and the application cannot be done directly, but must go through the CRM application commonly used by Customer Service. Eventually, the two applications will be connected through API communication. The main API used in this application is listed in Table 1. When accessing each endpoint, validation of the Authorization Key will be performed first, where this Key will be created by the Administrator of the email classifier application when adding a new Client. This is done to ensure that every HTTP Request

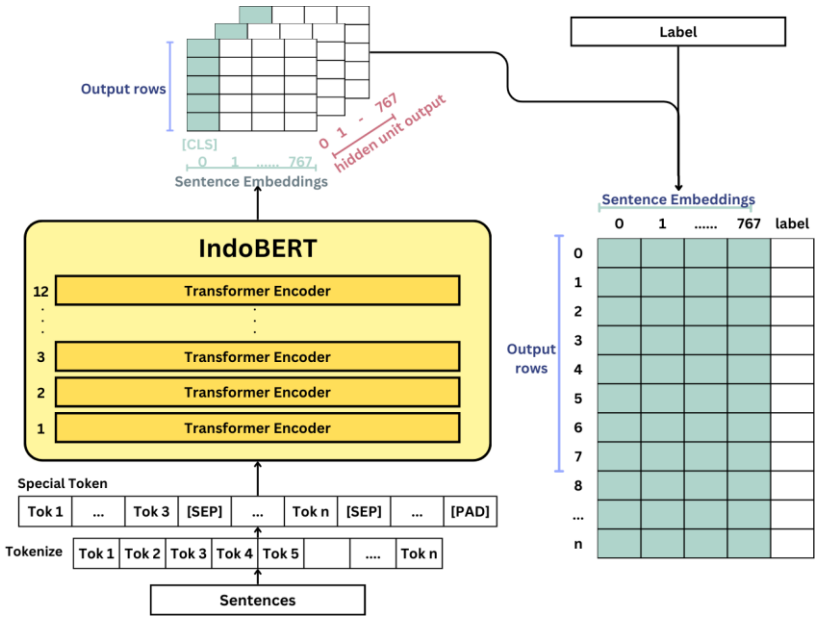


Figure 5. Feature Extraction Workflow

transaction is secure and not infiltrated by requests from other entities. The first API is used to send datasets to the system, which includes functions to delete old datasets and input new datasets for the client. Then, the second API is used to perform the Train Model of Machine Learning action. The third API is used to select the model that will be used for auto prediction/auto classification. Finally, the last API is used to perform prediction/classification if there are any mixed emails entering the CRM system.

Table 1. List of Main API Endpoint

Method	URL API	Parameter
POST	https://{base_url}/api/training-data/create	client_id, product_id, data (contains an array of Message & Category)
POST	https://{base_url}/api/model/train	client_id
POST	https://{base_url}/api/model/choose	client_id, model_id
POST	https://{base_url}/api/model/predict	client_id, message

2. Machine Learning Workflow and Architecture

Figure 3 describes the approach employed in our experiments. The study's workflow is segmented into four main sections: data collection, data preparation, feature extraction, and training of machine learning.

2.1. Data Collecting

The data is sourced from the API database, where clients have previously uploaded data to the training data table via the API. To evaluate the system's performance, we employ the XYZ Company dataset as a reference point for evaluating the architecture's capabilities. This dataset is loaded into the training_datas table before retrieving the data from the database for analysis. The evaluation dataset utilized consists of emails sent by clients to the XYZ Company helpdesk via email. These emails have been annotated by workers affiliated with XYZ Company. The dataset comprises a total of 1112 emails, with labels classified into three distinct categories: 'Question', 'Incident', and 'Request'. The distribution of emails across these categories is as follows: 'Question': 739, 'Incident': 115, 'Request': 258

2.2. Data Preparation

In order to evaluate the model's performance, we divided the evaluation dataset into 80% for training data and 20% for testing data. Subsequently, in pursuit of enhanced model accuracy, we implemented cleaning procedures on both the training and testing datasets to eliminate noise. We employed various text cleaning methods, including:

a) Lower casing
b) Remove Html tags.
c) Remove URL.
d) Remove certain punctuation
e) Remove extra space

2.3. Feature Extraction

In our research, we employed pre-trained IndoBERT, a transformer-based model similar to BERT, trained on an Indonesian corpus. It functioned as an extractor of feature, harnessing contextual information to generate word embeddings from textual data. [10]. The BERT model was selected because previous studies have already conducted a comparison of the performance of various Indonesian BERT models, with Indo BERT emerging as the top performer [11]. In this study, IndoBERT serves primarily as a feature extraction tool with two main functions tokenization and embedding. Main process flow of IndoBERT can be shown on Figure 5.

BERT, a sophisticated embedding model, possesses the remarkable ability to grasp contextual nuances, generating fluid embeddings for words within

Table 2. Training Result

Model	Precision	Accuracy	Recall	MCC	F1 Score
Random Forest	0.952657	0.952652	0.952652	0.905307	0.952651
XGBoost	0.960390	0.960227	0.960227	0.920613	0.960222
SVM	0.966334	0.965909	0.965909	0.932238	0.965899

their contextual framework. It perceives sentences as sequences of tokens and can adeptly handle both single sentences and sentence pairs. Token embeddings are meticulously crafted from a vocabulary pool assembled from Word Piece embeddings, encompassing an extensive repertoire of 30,000 tokens.[12]. In our research, we leverage the AutoTokenizer library to facilitate the tokenization process, harnessing the power of pre-trained IndoBERT sourced from Hugging Face. BERT's functionality extends beyond mere token embeddings, incorporating positional embeddings and segment embeddings for each token, enriching the model's contextual understanding. [13]. Positional embeddings hold information about the sequence of tokens, while segment embeddings indicate sentence boundaries: tokens from the first sentence receive an embedding of 0, and tokens from the second sentence receive an embedding of 1 [14]. Prior to encoding text with BERT, the tokenizer inserts the [CLS] token at the beginning of the initial sentence for classification tasks, and the [SEP] token at the conclusion of each sentence . This process ensures that BERT can effectively capture the semantic meaning of the input text by considering the entire context, while also providing clear markers for sentence boundaries. These tokens play a crucial role in the architecture of BERT, facilitating tasks such as sentence classification, semantic similarity assessment, and named entity recognition [13]. After tokenizing the input text using the pre-trained IndoBERT tokenizer, the next step is to extract the vector embeddings associated with each token. These embeddings capture the contextual information of each token within the given text. Specifically, the vector embedding for each token is obtained from the hidden state that the IndoBERT model outputs for that token.

In our architecture, we focus on utilizing the hidden state corresponding to the special token [CLS] for classification tasks. The [CLS] token represents the aggregated representation of the entire input sequence and is often used as a summarization of the semantic content of the text. By extracting the hidden state of the [CLS] token, we effectively capture the representative semantic meaning of the entire sequence [14]. The extracted vector embeddings are then fed into downstream classification tasks These tasks leverage the contextual embeddings provided by IndoBERT to make predictions or perform further analysis on the input text [14].

2.4. Model Training

We selected the 3 Machine Learning classifiers algorithm, SVM, Random Forest and XGB to choose. They were implemented using Scikit-learn, a Library of Python Machine Learning to build the predictive models. The models will be trained on client private data, then save the model to the model library.

2.5. Test Model

Users of our API can compare the performance of models using five metrics: Precision, Accuracy, F1 Score, Recall, and Matthews Correlation Coefficient (MCC), customizing their evaluation according to their specific needs and preferences.

- Overall Accuracy

This metric computes the ratio of properly predicted occurrences to total instances in the dataset. It offers a comprehensive evaluation of the model's effectiveness However, accuracy alone can be misleading in imbalanced datasets, hence the need for additional metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- Precision

Refers to the accuracy of confident forecasts. It is the proportion of accurately predicted positive observations to total projected positive observations. Precision is crucial when the cost of false positives is high.

.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- Recall (Sensitivity)

Recall also known as sensitivity or true positive rate, represents the proportion of actual positive cases that were accurately recognized by the model. Recall is important in scenarios where missing a positive case is costly.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

- F1 Score

The F1 score represents the balance of precision and recall. It maintains a balance between precision and recall and is effective in classes that are uneven. The F1 score is particularly useful when both false positives and false negatives are important

Table 3. Testing Result

Method	URL API	Parameter	Result
POST	https://{base_main_url}/api/training-data/create	client_id, product_id, data (contains an array of Message & Category)	Success
POST	https://{base_main_url}/api/model/train	client_id	Success
POST	https://{base_main_url}/api/model/choose	client_id, model_id	Success
POST	https://{base_main_url}/api/model/predict	client_id, message	Success

$$F1\ Score\ = 2x\ \frac{Precision\ x\ Recall}{Precision\ +\ Recall} \quad (4)$$

- Matthews Correlation Coefficient (MCC)

MCC represents the correlation coefficient between observed and predicted binary classifications. It considers genuine and false positives and negatives and is widely recognized as a balanced metric. MCC is especially useful for imbalanced datasets because it provides a more informative and truthful score in such situations compared to accuracy.

$$MCC = \frac{(TP\ x\ TN\ -\ FP\ x\ FN)}{\sqrt{(TP\ +\ FP)(TP\ +\ FN)(TN\ +\ FP)(TN\ +\ FN)}} \quad (5)$$

In these equations:

- *TP* means True Positives
- *TN* means True Negatives
- *FP* means False Positives
- *FN* means False Negatives

By using these metrics, users can get a well-rounded understanding of their model's performance, ensuring that they can make informed decisions based on the specific requirements and potential costs associated with false positives or false negatives in their applications.

Implementation

After the analysis and design stages have been completed, the next stage is the implementation stage. The application is built using the Python language with the Flask Framework. The structure used is MVC (Model, View, Controller). This is done considering that the application can be easily further developed in terms of functionality in the future [15]. Flask Framework itself is chosen because this application is in the form of an API service, so it is more suitable to use a Microframework for faster performance. Additionally, Flask is chosen because of its maturity, having been released for 14 years, making this framework stable in functionality and performance. Third-party libraries used for modeling include Scikit-learn and Xgboost. Using the API for training models shows quite satisfactory results, with the highest model accuracy achieved by the SVM model with an accuracy of 0.97. The complete experimental results can be found in Table 2.

Integration Testing And Verification

We have conducted comprehensive testing on all primary APIs, confirming their performance aligns with our expectations. Each API has successfully produced the intended results, meeting all specified requirements and functionality criteria. The complete testing results can be found in Table 3.

Operation And Maintenance

Based on the successful testing outcomes and the application's operational functionality, it can be concluded that the application is ready for use. Any deviations or errors encountered will be addressed promptly through further refinement processes

CONCLUSION AND SUGGESTION

This application has successfully performed classification effectively with high accuracy, thus potentially facilitating future customer service operations.

Further research is needed to investigate how the system performs when dealing with a larger number of categories beyond three. If there is a significant decrease in performance, further development may be necessary, possibly incorporating deep learning techniques

REFERENCES

[1] H. Drucker, D. Wu, and Y. N. Yapnik, “Support Vector Machines for spam categorization,” *IEEE Trans Neural Netw*, vol. 10, no. xx, 1999.

[2] Md. R. Islam, M. U. Chowdhury, and Wanlei Zhou, “An Innovative Spam Filtering Model Based on Support Vector Machine,” in *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06)*, IEEE, pp. 348–353. doi: 10.1109/CIMCA.2005.1631493.

[3] C. Anilkumar, A. Karrothu, N. S. Mouli, and C. B. Tej, “Recognition and Processing of phishing Emails Using NLP: A Survey,” in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, Jan. 2023.

[4] R. Suprayoga, S. Zega, Muhathir, and S. Mardiana, "Classification of Mango Leaf Diseases Using XGBoost Method and HoG Feature Extraction," in *2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA)*, 2023.

[5] K. Schneider, "A comparison of event models for Naive Bayes antispam e-mail filtering," in *Proceedings of the II th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, 2003.

[6] K. Iqbal and M. S. Khan, "Email classification analysis using machine learning techniques," *Applied Computing and Informatics*, May 2022, doi: 10.1108/ACI-01-2022-0012.

[7] J. D. Brutiag and C. Meek, "Challenges of the email domain for text classification," in *ICML*, 2000, pp. 103–110.

[8] A. Putu Candra *et al.*, "Sistem Informasi Penjualan Online Thrift Shop Berbasis Web," *Journal of Technology and Informatics (JoTI)*, vol. 5, no. 2, pp. 116–124, Apr. 2024, doi: 10.37802/joti.v5i2.586.

[9] A. Hidayatur, A. Yuana, A. Wafi, R. Harjo, T. Maulana, and A. Terza, "Pengembangan Proses Bisnis Pelayanan Statistik Terpadu Badan Pusat Statistik Kota Surabaya Menggunakan Metode Prototyping," *Journal of Technology and Informatics (JoTI)*, vol. 5, no. 2, pp. 70–79, Apr. 2024, doi: 10.37802/joti.v5i2.548.

[10] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[11] H. Lucky, Roslynlia, and D. Suhartono, "Towards Classification of Personality Prediction Model: A Combination of BERT Word Embedding and MLSMOTE," in *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, 2021, pp. 346–350. doi: 10.1109/ICCSAI53272.2021.9609750.

[12] Pavithra and Savitha, "Topic Modeling for Evolving Textual Data Using LDA, HDP, NMF, BERTOPIC, and DTM With a Focus on Research Papers," *Journal of Technology and Informatics (JoTI)*, vol. 5, no. 2, pp. 53–63, Apr. 2024, doi: 10.37802/joti.v5i2.618.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".

[14] S. Ravichandiran, *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT*. Packt Publishing Ltd, 2021.

[15] L. A. Sanjani, S. J. Hartati, and P. Sudarmaningtyas, "Rancang Bangun Sistem Informasi Penggajian Pegawai dan Remunerasi Jasa Medis Pada Rumah Sakit Bedah Surabaya," *Jurnal Sistem Informasi dan Komputer Akuntansi (JSIKA)*, vol. 3, no. 1, pp. 87–93, 2014.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, Bulan Oktober Tahun 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.597

Penerapan Algoritma Apriori untuk Rekomendasi Produk *Spare Part* Mobil

Reza Aditya Permana[1], **Primandani Arsi**[2], **Pungkas Subarkah**[3]
[1,2,3]Informatika, Universitas Amikom Purwokerto, Banyumas, Jawa Tengah, Indonesia
e-mail: rezapermana127@gmail.com[1], ukhti.prima@amikompurwokerto.ac.id[2],
subarkah@amikompurwokerto.ac.id[3]
* Penulis Korespondensi: E-mail: subarkah@amikompurwokerto.ac.id

Abstrak: Guna memberikan layanan optimal kepada pelanggan, bisnis *spare part* mobil perlu menerapkan strategi bisnis terbaik. Namun, terkadang beberapa faktor menghambat penentuan strategi tersebut. Salah satu penyebabnya adalah kesulitan dalam melakukan analisis terkait data penjualan pelanggan yang sudah ada. Berdasarkan data pelanggan yang disimpan dalam *database*, perusahaan mengidentifikasi bahwa sistem penjualan saat ini tidak efektif. Kemudian apabila produk banyak yang tidak terjual, maka keterlambatan dalam pengembalian modal bagi penjual juga akan menjadi masalah. Dalam mengatasi hambatan ini, perusahaan perlu mengadopsi strategi promosi penjualan dan menganalisis pola penjualan *spare part* untuk menentukan pola penjualan dan memberikan gambaran keterkaitan antar barang dengan menganalisis data transaksi penjualan. Metode yang akan digunakan dalam menganalisis pola penjualan pada penelitian ini adalah dengan menggunakan *market basket analysis.* Metode ini dihitung menggunakan Algoritma Apriori dengan menggunakan bantuan Google Collab dengan menentukan nilai *minimum support* sebesar 8 % dan *minimum confidence* sebesar 60 %. Perhitungan dalam Google Collab akan menghasilkan kumpulan *item* yang sering dibeli dengan kombinasi 1 *itemset* sampai dengan 5 *itemset* yang dibeli secara bersamaan. Hasil *assosiation rule* tersebut dapat dijadikan acuan kepada pengambil keputusan dalam upaya meningkatkan strategi pemasaran dan promosi produk suku cadang mobil bagi PT Milenia Mega Mandiri yang lebih baik.

Kata Kunci: Algoritma Apriori; Google Collab*; Market Basket Analysis*; *Spare part*

Abstract: *To provide optimal service to customers, a car spare part business needs to implement the best business strategy. However, sometimes several factors hinder the determination of these strategies. One of the reasons is the difficulty in analyzing existing customer sales data. Based on the customer data stored in the database, the company identified that the current sales system was ineffective. Then, if many products are not sold, the delay in the return of capital for the seller will also be a problem. To overcome these obstacles, the company needs to adopt a sales promotion strategy and analyze the sales pattern of spare parts to determine the sales pattern and provide an overview of the interrelationship between goods by analyzing sales transaction data. The method that will be used in analyzing sales patterns in this study is to use market basket analysis. This method is calculated using the Apriori Algorithm using the help of Google Collab by determining the minimum support value of 8% and minimum confidence of 60%. Calculations in Google Collab will produce a collection of items that are often purchased with a combination of 1 itemset to 5 itemsets that are purchased simultaneously. The results of the association rule can be used as a reference for decision makers in an effort to improve the marketing strategy and promotion of auto parts products for PT Milenia Mega Mandiri better.*

Keywords: *Algoritma Apriori; Google Collab; Market Basket Analysis; Spare part*

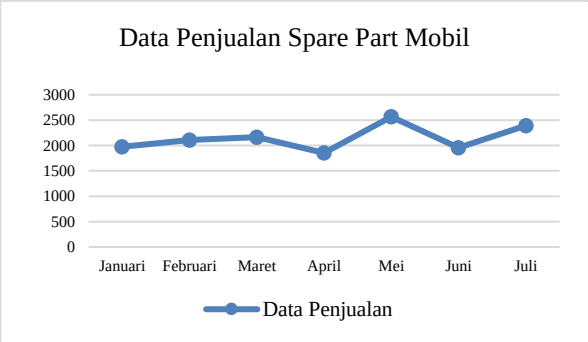
PENDAHULUAN

Semakin pesatnya pertumbuhan teknologi dewasa ini membuat persaingan dunia bisnis khususnya pada penjualan di bidang otomotif menjadi sangat ketat. Disebabkan oleh kompetisi yang tinggi di dunia bisnis, terutama dalam industri penjualan, Beberapa perusahaan yang bergerak di sektor penjualan suku cadang menghadapi penurunan dan mengalami kesulitan dalam merencanakan strategi promosi yang cocok untuk pembeli. *Spare part* berurusan dengan stok barang yang digunakan dalam kegiatan pemeliharaan untuk menjaga peralatan atau produk tetap dalam kondisi operasi [1].

Sebagai strategi penting dalam menghadapi persaingan yang semakin ketat dalam industri penjualan *spare part* mobil, para pelaku bisnis menyadari bahwa keberhasilan usaha mereka tidak hanya bergantung pada kualitas produk saja, melainkan juga pada kemampuan untuk menarik perhatian pelanggan dan meningkatkan jumlah penjualan. Berfokus pada distribusi aksesoris dan audio mobil, PT Milenia Mega Mandiri tidak hanya menyediakan *spare part* berkualitas tinggi, tetapi juga memberikan solusi menyeluruh untuk meningkatkan pengalaman berkendara pelanggan.

PT Milenia Mega Mandiri adalah sebuah perusahaan yang bergerak di bidang distribusi aksesoris dan *audio* mobil. PT Milenia Mega Mandiri berkantor di Jl. Pembangunan I No.1, RT.3/RW.1, Petojo Utara, Kecamatan Gambir, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10130. Bisnis yang berdiri sejak 1998 ini telah berkembang ke dalam kelompok bidang usaha yang berbasis *Original Equipment Manufacturing* (OEM) dan *Automotive Product*. Unit dari PT ini antara lain yaitu

aksesoris mobil, *audio* mobil, perawatan mobil, dan *home audio*.



Gambar 1. Data Penjualan *Spare Part* Mobil

Berdasarkan data penjualan pada Gambar 1, merupakan hasil wawancara dengan Bapak Rizki Ryan selaku karyawan bagian *Finance* dari perusahaan PT. Milenia Mega Mandiri, terdapat 1974 transaksi pada bulan Januari. Sedangkan pada bulan Februari terdapat 2107 transaksi. Selanjutnya pada bulan Maret terdapat transaksi sebanyak 2162. Tetapi pada bulan April terdapat penurunan transaksi yaitu menjadi 1853. Kemudian pada bulan Mei terdapat kenaikan transaksi menjadi 2.565. Selanjutnya pada bulan Juni mengalami penurunan kembali yaitu menjadi 1.956 transaksi. Data transaksi yang terakhir yaitu pada bulan Juli, pada bulan ini transaksi mengalami kenaikan kembali menjadi 2392. Demikian terlihat jelas bahwa data transaksi pada PT Milenia Mega Mandiri tidak stabil yaitu mengalami penurunan dan kenaikan yang tidak menentu. Maka dari itu dibutuhkan strategi bisnis agar tetap memberikan pelayanan terbaik.

Untuk memberikan layanan optimal kepada pelanggan, bisnis *spare part* mobil perlu menerapkan strategi bisnis terbaik. Namun, terkadang beberapa faktor menghambat penentuan strategi tersebut. Salah satu penyebabnya adalah kesulitan dalam melakukan analisis terkait data penjualan pelanggan yang sudah ada. Berdasarkan data pelanggan yang disimpan dalam *database*, perusahaan mengidentifikasi bahwa sistem penjualan saat ini tidak efektif. Kemudian apabila produk banyak yang tidak terjual, maka keterlambatan dalam pengembalian modal bagi penjual juga akan menjadi masalah. Dalam mengatasi hambatan ini, perusahaan perlu mengadopsi strategi promosi penjualan dan menganalisis pola penjualan *spare part*. Maka dari itu, penelitian ini berencana untuk mendorong para pengembang agar Menerapkan strategi yang dapat meningkatkan penjualan dan pemasaran produk yang mereka sediakan. Tujuan ini didasarkan pada hasil penelitian terdahulu yang bertujuan Disediakan sebagai saran kepada para pengambil keputusan untuk meningkatkan strategi pemasaran dan promosi produk suku cadang mobil dengan lebih efektif [2]. Langkah ini diperlukan pendekatan yang tepat dalam menganalisis pola penjualan, salah satunya penerapan teknik data *mining*. Menurut Widodo dan Nabawi data *mining* melibatkan penggunaan teknik analisis data untuk mengidentifikasi dan menemukan pola-pola yang terdapat dalam suatu kumpulan data. [3].

Dengan simpelnya, data *mining* atau penggalian data dapat diartikan sebagai serangkaian proses untuk menemukan nilai tambah dari kumpulan data berupa pengetahuan yang sebelumnya tidak diketahui secara manual [4]. Untuk memahami kebiasaan belanja konsumen, dapat digunakan metode asosiasi yang umumnya dikenal sebagai Analisis Keranjang Belanja (*Market Basket Analysis*). Analisis Keranjang Belanja merupakan suatu pendekatan untuk menganalisis kebiasaan belanja konsumen dengan mengidentifikasi hubungan antara beberapa *item* yang berbeda, yang ditempatkan oleh konsumen dalam keranjang belanja saat suatu transaksi tertentu[5]. Terdapat berbagai algoritma dalam bidang *data mining* yang dapat dipakai untuk memprediksi hasil dari pemrosesan kumpulan data, dan salah satunya adalah dengan menggunakan Algoritma Apriori.

Menurut Toivonen Algoritma Apriori adalah algoritma di bidang *data mining* untuk mengekstraksi aturan asosiasi yang disebut juga dengan *Association Rule Mining* (ARM) [6][7]. Ini memungkinkan perusahaan untuk mencari dan mengidentifikasi pola asosiasi di antara produk yang dijual, sehingga membantu meningkatkan keterkaitan antar barang-barang yang ditawarkan[8].

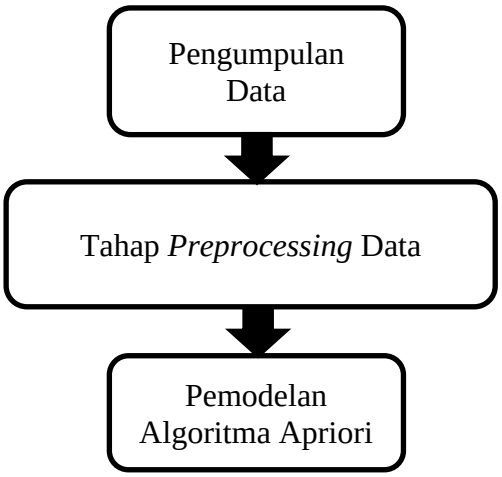
Algoritma Apriori ini telah digunakan dalam beberapa penelitian terdahulu, diantaranya pada penelitian yang berjudul "Prediksi Penjualan *Spare Part* Mobil Daihatsu Menggunakan Algoritma Apriori" algoritma apriori digunakan sebagai panduan bagi para pengambil keputusan dengan upaya meningkatkan efektivitas strategi pemasaran dan promosi produk suku cadang mobil [2]. Kemudian pada penelitian yang berjudul "Penerapan Algoritma Apriori untuk Market Basket Analysis" hasil perhitungan Algoritma Apriori ini digunakan sebagai masukan untuk membantu pemilik bisnis menentukan strategi penjualan untuk bisnisnya [9]. Selanjutnya pada penelitian yang berjudul "Penerapan Algoritma Apriori Pada Transaksi Penjualan Untuk Rekomendasi Menu Makanan Dan Minuman" menghasilkan rekomendasi menu makanan dan minuman yang digunakan sebagai patokan pemilik usaha Warung Tenda untuk menambahkan opsi menu paket [5]. Berikutnya penelitian mengenai rekomendasi *restock* pada toko pojok UMKM. Hasil yang diperoleh dalam penelitian ini yaitu adanya produk yang signifikan masuk ke dalam pengadaan barang[10].

Berdasarkan permasalahan yang telah diungkapkan diatas, penulis yakin bahwa usulan penerapan Algoritma Apriori dapat memberikan kontribusi signifikan dalam memahami pola penjualan. Melalui analisis yang lebih mendalam menggunakan algoritma ini, diharapkan PT Milenia Mega Mandiri dapat meningkatkan penjualan *spare part* mobil, kontribusi penelitian lainnya yaitu dapat merancang strategi promosi yang lebih tepat sesuai dengan preferensi dan kebutuhan pelanggan. Dengan diusulkannya penerapan algoritma Apriori, diharapkan

hal ini mampu memberikan wawasan berharga bagi perusahaan untuk mengoptimalkan kinerja penjualan mereka dan meraih keberhasilan dalam industri yang kompetitif ini.

METODE PENELITIAN

Penelitian ini dilakukan dengan menggunakan algoritma Apriori. Algoritma Apriori digunakan untuk mengetahui pola penjualan guna meningkatkan penjualan dan dengan tepat menentukan strategi promosi yang diberikan pelanggan dalam penjualan *spare part* mobil pada PT Milenia Mega Mandiri. Berikut merupakan tahapan penelitian, yang disajikan pada Gambar 2.



Gambar 2. Tahapan Penelitian

Dibawah ini merupakan penjelasan dari Gambar 2. tahapan penelitian, sebagai berikut:

Pengumpulan Data

Metode pengumpulan data adalah cara yang dapat digunakan untuk mengumpulkan data penelitian. Metode pengumpulan data dibagi menjadi beberapa proses yaitu tahap wawancara, tahap observasi, dan menemukan studi pustaka. Data yang diperoleh melalui proses pengumpulan data adalah data penjualan *spare part* mobil.

Preprocessing Data

Tahap *Preprocessing Data* adalah proses transformasi data mentah atau raw data yang diperoleh dari berbagai sumber yang bertujuan untuk melakukan pembersihan data. Hal ini dimaksudkan untuk membersihkan data yang tidak lengkap, salah memasukkan, atau perlu disesuaikan dengan kebutuhan [11].

Pemodelan Algoritma Apriori

Algoritma Apriori digunakan untuk menemukan kumpulan item yang sering digunakan yang membantu membentuk aturan asosiasi[12][13]. Pada pembentukan *Assosciation Rules* digunakan *tools Google Colab. Google Colab* sebagai *tools* yang Google sediakan guna membantu pengguna dalam proses pemrograman dan pengolahan data[14][15]. Berikut langkah yang dilakukan dalam perhitungan Algoritma Apriori.

1. Pemrosesan Data, *dataset* yang didapat pada tahap preprocessing kemudian diproses menggunakan Algoritma Apriori untuk menghasilkan *frequent itemset*.
2. Menentukan *minimum support* sebagai batas dalam pembentukan *frequent itemset* dan *minimum confidence* dalam pembentukan *association rules*. Penentuan *minimum support* dan *minimum confidence* disesuaikan dengan kebutuhan karena tidak ada ketentuannya. Dalam penelitian ini menggunakan *minimum support* sebesar 8% dan *minimum confidence* sebesar 60%, dengan merujuk penelitian terkait [16]. Peneliti memakai ini dikarenakan sudah ada penelitian yang menghasilkan dengan nilai *minimum support* dan *minimum confidence*.
3. Data yang digunakan untuk perhitungan sebelum dilakukan *preprocessing* adalah 2.392 data dan setelah dilakukan *preprocessing* maka data yang digunakan menjadi 504 data.

HASIL DAN PEMBAHASAN

Pengambilan Data

Dalam penelitian ini, peneliti melakukan pengumpulan data dari PT. Milenia Mega Mandiri dan memastikan bahwa informasi pada variabel yang menjadi fokus dalam data dapat digunakan secara sistematis. Tahap ini adalah tahap awal dalam penerapan Algoritma Apriori yang bertujuan untuk mendapatkan *dataset* guna sebagai subjek yang akan dilakukan uji dalam penerapan Algoritma Apriori.

	Refference	Tanggal	Tempo	Customer	Customer	Alamat	Contact	Telp	Wilayah	Divisi	Brand	Class
2	INV00124	10/07/2023	C	MUV01JB	MUKTI VG	RUKO DA/	BPK MUK	0215434	JAKARTA I	MEGUIAR	MEGUIAR	CAR MAIN
3	INV00124	10/07/2023	C	MUV01JB	MUKTI VG	RUKO DA/	BPK MUK	0215434	JAKARTA I	CAR CARE	MEGUIAR	CAR MAIN
4	INV00124	10/07/2023	C	MUV01JB	MUKTI VG	RUKO DA/	BPK MUK	0215434	JAKARTA I	CAR CARE	MEGUIAR	CAR MAIN
5	INV00124	10/07/2023	C	MUV01JB	MUKTI VG	RUKO DA/	BPK MUK	0215434	JAKARTA I	CAR CARE	MEGUIAR	CAR MAIN
6	INV00124	10/07/2023	C	MUV01JB	MUKTI VG	RUKO DA/	BPK MUK	0215434	JAKARTA I	CAR CARE	MEGUIAR	CAR MAIN
7	INV00124	10/07/2023	C	GS01JSW	Graha Sak	ITC Duta M	Handoko	021-723	JAKARTA S	CAR CARE	RED LINE	MOTOR O
8	INV00124	10/07/2023	C	GS01JSW	Graha Sak	ITC Duta M	Handoko	021-723	JAKARTA S	CAR CARE	RED LINE	MOTOR O
9	INV00124	10/07/2023	C	A9801JSW	AUTO 98	JL. RM HA	ADAM NU	0811890	JAKARTA S	CAR CARE	MEGUIAR	CAR MAIN
10	INV00124	10/07/2023	C	A9801JSW	AUTO 98	JL. RM HA	ADAM NU	0811890	JAKARTA S	MEGUIAR	MEGUIAR	CAR MAIN
11	INV00124	17/07/2023	1B	MA01JUW	PT. MEGA	JL. PECEN	(-CYNTHIA	021-568(	JAKARTA I	BATTERY	NOCO	GENERAL
12	INV00124	03/07/2023	C	JOU01PLM	JOU AUTO	SUKABAN	DESTIART	0812853	PALEMBA	MEGUIAR	MEGUIAR	CAR MAIN
13	INV00124	03/07/2023	C	JOU01PLM	JOU AUTO	SUKABAN	DESTIART	0812853	PALEMBA	MEGUIAR	MEGUIAR	CAR MAIN
14	INV00124	03/07/2023	C	JOU01PLM	JOU AUTO	SUKABAN	DESTIART	0812853	PALEMBA	MEGUIAR	MEGUIAR	CAR MAIN
15	INV00124	03/07/2023	C	JOU01PLM	JOU AUTO	SUKABAN	DESTIART	0812853	PALEMBA	MEGUIAR	MEGUIAR	CAR MAIN
16	...	...	...	...	...	...	...	...	...	...	...	...
17	INV00124	28/07/2023	1B	TA01JUPR	PT. TOYOT	JL. GAYA M	INDRA	6515551	JAKARTA I	PROJECT	ORIGINAL	ALARM

Gambar 3. Data Penjualan *Spare Part* Mobil Bulan Juli

Type	ItemID	Item	Qty	Price	Amount	Return An	Retur No	Sales	Bulan	Tahun
CLEANER (	CMMG01(	MEGUIAR	20	575675,7	7483784	0		AGUS SETI	7	2023
CLEANER (	CMMG01(	MEGUIAR	48	200000	6240000	0		AGUS SETI	7	2023
CLEANER (	CMMG01(	MEGUIAR	30	297297,3	5797297	0		AGUS SETI	7	2023
CLEANER (	CMMG01(	MEGUIAR	30	201801,8	3935135	0		AGUS SETI	7	2023
CLEANER (	CMMG01(	MEGUIAR	30	186486,5	3636487	0		AGUS SETI	7	2023
ADDITIVE	MORL038	REDLINE 8	12	162162,2	1556757	0		WINSTON	7	2023
ADDITIVE	MORL03S	REDLINE S	24	157657,7	3027027	0		WINSTON	7	2023
CLEANER (	CMMG01(	MEGUIAR	6	253153,2	1215135	0		WINSTON	7	2023
CLEANER (	CMMG01(	MEGUIAR	12	54054,05	518918,9	0		WINSTON	7	2023
CHARGER	GALL02NC	NOCO GE	10	1147500	11475000	0		COMPANY	7	2023
CLEANER (	CMMG01(	MEGUIAR	12	54054,05	454054	0		SAIFUL AZ	7	2023
CLEANER (	CMMG01(	MEGUIAR	1	141441,4	127297,3	0		SAIFUL AZ	7	2023
CLEANER (	CMMG01(	MEGUIAR	1	141441,4	127297,3	0		SAIFUL AZ	7	2023
CLEANER (	CMMG01(	MEGUIAR	1	144144,1	129729,7	0		SAIFUL AZ	7	2023
...	...	...	...	...	...	...	...	...	...	...
ALARM	ALOE013	ALARM OE	1	256059	256059	0		IRVANTO	7	2023

Gambar 4. Data Penjualan *Spare Part* Mobil Bulan Juli

Berdasarkan Gambar 3. dan 4. maka data yang diambil melalui proses pengambilan data merupakan data penjualan dalam format *file* Excel. Data penjualan meliputi 23 atribut antara lain *refference*, tanggal, tempo, *costumerId*, *costumer*, alamat, *contact*, telepon, wilayah,

divisi, *brand*, *class*, *type*, *itemId*, *item*, *qty*, *price*, *amount*, *return amount*, *return no*, sales, bulan, tahun. Serta terdapat total transaksi sebanyak 2.392.

Preprocessing Data

Tahap *Preprocessing Data* adalah proses transformasi data mentah atau *raw data* yang diperoleh dari berbagai sumber menjadi informasi yang lebih bersih dan bisa digunakan untuk pengolahan selanjutnya.

Tabel 1. Hasil *Preprocessing Data*

Tanggal	ID Transaksi	Items
01/07/2023	INV00124293	Shock Sensor D03b,Shock Sensor Set
03/07/2023	INV00124117	Meguairs G17748 (247681) Ultimate Wash & Wax,Meguiars A1216ea (61291) Cleaner Wax Liquid,Meguiars G13616 (165672) Quick Interior Detailer,Meguiars G15415 (274429) Endurance Tire Spray (Aerosol)
03/07/2023	INV00124169	Meguairs G17748 (247681) Ultimate Wash & Wax,Meguiars A1624b Quick Wax 24oz,Meguiars G12310 (134781) Plast X,Meguiars G12664 (158044) Nxt Car Wash,Meguiars G13616 (165672) Quick Interior Detailer,Meguiars G14512 (249808) Ultimate Protectant,Meguiars G180124 Ultimate Wheel Cleaner,Meguiars G18616 (274434) Gold Class Leather Conditioner,Meguiars G201024 (10461791) Ultimate Quick Detailer,Meguiars G210516 Ultimate Liquid Wax 16 Oz,Meguiars
03/07/2023	INV00124179	Meguairs G17748 (247681) Ultimate Wash & Wax,Meguiars A1624b Quick Wax 24oz,Meguiars G12310 (134781) Plast X,Meguiars G12664 (158044) Nxt Car Wash,Meguiars G13616 (165672) Quick Interior Detailer,Meguiars G14512 (249808) Ultimate Protectant,Meguiars G180124 Ultimate Wheel Cleaner,Meguiars G18616 (274434) Gold Class Leather Conditioner,Meguiars G201024 (10461791) Ultimate Quick Detailer,Meguiars G210516 Ultimate Liquid Wax 16 Oz,Meguiars
03/07/2023	INV00124180	Meguiars G190315 Ultimate Insane Tire Spray,Meguiars G191564eu Ultimate Snow Foam,Meguiars G220216 Ultimate Inshine Protectan 16 Oz,Meguiars M10501 Mirror Glaze Ultra Cut Compound
.....	.....	.....
31/07/2023	INV00124811	Auto Folding Mirror Set,Shock Sensor Set (G)
31/07/2023	INV00124812	Auto Folding Mirror Set,Security System Set,Shock Sensor Set (G)

Berdasarkan Tabel 1. proses *preprocessing data* ini dilakukan secara *manual* dengan diambil atribut yang diperlukan seperti *reference*, tanggal, dan item, serta mengerucutkan data agar sesuai dengan *reference* sehingga menjadi 504 data. Proses ini dilakukan agar file data penjualan dapat dilakukan perhitungan Algoritma Apriori dengan menggunakan *Google Collab.*

Pemodelan Algoritma Apriori

Pada tahap ini, peneliti melakukan proses data mining atau proses mencari pola informasi menarik dalam data penjualan yang telah dilakukan *preprocessing* dengan menggunakan *Google Collab*. Dalam proses ini terdapat beberapa proses antara lain sebagai berikut:

1. *Pre-Coding*

Pada proses ini dilakukan penginstalan *library* dan gunakan perintah *import* dalam *python* untuk memasukkan beberapa pustaka yang diperlukan, seperti *pandas*, *numpy*, *dan apyori.*

```
[ ]  # Menginstall Library
     !pip install pandas
     !pip install numpy
     !pip install apyori
```

Gambar 5. *Menginstall library*

```
[ ]  # Memanggil library yang dibutuhkan
     import pandas as pd
     import numpy as np
     from apyori import apriori
```

Gambar 6. Memanggil *library*

a) *Library pandas*, *library pandas* adalah pustaka *Python* yang kuat dan fleksibel untuk analisis data. Ini menyediakan struktur data dan fungsionalitas berbagai alat yang digunakan untuk manipulasi, pembersihan, analisis, dan visualisasi data dengan mudah.
b) *Library* numpy, *library numpy* sanget berguna untuk melakukan perhitungan numerik dengan cepat
c) *Library apyori*, *library apyori* digunakan untuk menerapkan algoritma apriori yang digunakan untuk asosiasi data.

2. *Data Collection*

Pada tahap *data collection* menggunakan pandas untuk membaca *file* Excel dengan nama Data Penjualan.xlsx. Selanjutnya menampilkan beberapa baris pertama *DataFrame* yang dimuat dari *file Excel* dengan menggunakan *df.head()*. Secara *default*, ini menampilkan lima baris pertama saja. Berikut pada Gambar 7. menunjukkan *input* dan *output* dari proses dalam pengumpulan data dari data penjualan.

```
# Load data Google Colab  ( File Excel pada Google Drive)
df= pd.read_excel('/content/drive/MyDrive/dataset/Data Penjualan.xlsx')
df.head()
```

	Tanggal	ID Transaksi	Items
0	2023-07-01	INV00124293	SHOCK SENSOR D03B,SHOCK SENSOR SET
1	2023-07-03	INV00124117	MEGUAIRS G17748 (247681) ULTIMATE WASH & WAX,M...
2	2023-07-03	INV00124169	MEGUAIRS G10307 (118743) SCRATCH X,MEGUIARS G1...
3	2023-07-03	INV00124179	MEGUAIRS G17748 (247681) ULTIMATE WASH & WAX,M...
4	2023-07-03	INV00124180	MEGUIARS G190315 ULTIMATE INSANE TIRE SPRAY,ME...

Gambar 7. *Output Data Collection*

Variabel – variabel penyusun:
Tanggal : Tanggal pembelian barang
ID transaksi: Id dari transaksi
Items : Barang yang dibeli pada sebuah transaksi

3. *Data Cleansing*

Dalam tahap pembersihan data atau *cleansing*, beberapa tindakan dilakukan seperti mengeliminasi atau menghapus variabel tanggal dan id transaksi dengan menggunakan *data= df.drop(['Tanggal','ID Transaksi'], axis=1). Penerapan axis=1* dilakukan agar parameter sumbu diatur ke 1, yang berarti operasi harus diterapkan sepanjang kolom. Langkah ini dilakukan untuk menghapus kolom tertentu, bukan baris. *Input* dan *output* dari proses yang terjadi pada data *cleansing* dari data penjualan.

```
[4]  # Membuang kolom Tanggal, ID Transaksi
     data=df.drop(['Tanggal','ID Transaksi'],axis=1)
```

Gambar 8. *Output Data Cleansing*

4. *Data Transformation*

Pada tahap *transformation* dilakukan untuk mengonvernsi data transaksi pembelian barang yang ada pada data penjualan menjadi sebuah *list* yang ditunjukkan pada Gambar 9.

```
# Membuat list dalam list dari transaksi pembelian barang
records = []
for i in range(data.shape[0]):
    records.append([str(data.values[i,j]).split(',') for j in range(data.shape[1])])

trx = [[] for trx in range(len(records))]
for i in range(len(records)):
    for j in records[i][0]:
        trx[i].append(j)
trx
```

Gambar 9. Input *Data Transformation*

```
[['SHOCK SENSOR D03B', 'SHOCK SENSOR SET'],
 ['MEGUAIRS G17748 (247681) ULTIMATE WASH & WAX',
  'MEGUIARS A1216EA (61291) CLEANER WAX LIQUID',
  'MEGUIARS G13616 (165672) QUICK INTERIOR DETAILER',
  'MEGUIARS G15415 (274429) ENDURANCE TIRE SPRAY (AEROSOL)'],
 ['MEGUIARS G10307 (118743) SCRATCH X',
  'MEGUIARS G19216 (274435) ULTIMATE POLISH',
  'MEGUIARS G210516 ULTIMATE LIQUID WAX 16 OZ',
  'MEGUIARS G210608 ULTIMATE PASTE WAX 80 OZ'],
 ['MEGUAIRS G17748 (247681) ULTIMATE WASH & WAX',
  'MEGUIARS A1624B QUICK WAX 24OZ',
  'MEGUIARS G12310 (134781) PLAST X',
  'MEGUIARS G12664 (158044) NXT CAR WASH',
  'MEGUIARS G13616 (165672) QUICK INTERIOR DETAILER',
  'MEGUIARS G14512 (249808) ULTIMATE PROTECTANT',
  'MEGUIARS G180124 ULTIMATE WHEEL CLEANER',
  'MEGUIARS G18616 (274434) GOLD CLASS LEATHER CONDITIONER',
  'MEGUIARS G201024 (10461791) ULTIMATE QUICK DETAILER',
  'MEGUIARS G210516 ULTIMATE LIQUID WAX 16 OZ',
  'MEGUIARS G7164 (124738) GOLD CLASS CAR WASH & CONDITIONER',
  'MEGUIARS G7624 GOLD CLASS QUICK DETAILER',
  'MEGUIARS G8408 PERFECT CLARITY GLASS COMPOUND/POLISH-8OZ',
  'MEGUIARS G8504 PERFECT CLARITY GLASS SEALANT',
  'MEGUIARS M3401 FINAL INSPECTION'],
```

Gambar 10. *Output Data Transformation*

Berdasarkan Gambar 10, *input* untuk membuat *list* dari transaksi pembelian barang dijelaskan sebagai berikut.

a) *Records* merupakan *list* kosong yang akan digunakan untuk menyimpan data transaksi yang telah dipisahkan. *Loop* pertama (*for* i *in range*(data.shape[0])) digunakan untuk mengiterasi melalui setiap baris dalam *DataFrame 'data'*.
b) Pada *loop* kedua *(for j in range(data.shape[1]))* untuk mengiterasi melalui setiap kolom dalam setiap baris, dan Anda menggunakan metode .split(',') untuk memisahkan data dalam setiap sel yang mungkin mengandung beberapa *item* yang dipisahkan oleh tanda koma. Hasil pemisahan ini kemudian disimpan dalam bentuk list dan dimasukkan ke dalam *list records*. Dengan demikian, setiap elemen dalam *records* akan berisi *list item-item* transaksi dari satu baris *DataFrame 'data'*.
c) Membuat *list* kosong trx yang akan digunakan untuk menyimpan hasil akhir.

d) Pada *loop* ketiga *(for i in range(len(records)))* untuk mengiterasi melalui setiap elemen dalam *records*. Di dalam *loop* ini, Anda mengakses elemen pertama dari setiap elemen *records* (elemen pertama adalah *list item-item* transaksi) dan menambahkannya ke dalam *list* 'trx'.

5. *Data Exploration*

Pada *data exploration* digunakan untuk Membuat variabel yang mencakup beberapa barang yang sering dibeli atau muncul dalam transaksi menggunakan perintah Apriori.

```
[ ]  # Menggunakan fungsi apriori untuk membuat asosiasi
     association_rules = apriori(trx, min_support=0.08, min_confidence=0.60,min_lift=1)
     # Membuat list hasil dari algoritma apriori
     association_results = association_rules
```

Gambar 11. Perhitungan menggunakan Algoritma Apriori dan membuat List

Berdasarkan Gambar 10, dilakukan perhitungan data yang berasal dari *dataframe* trx dengan *minimum support* 0.08 / 8%, *minimum confidence* 0.6 / 6% dan dengan *min lift ratio* sebesar 1. Selanjutnya membuat list dari hasil perhitungan Algoritma Apriori.

Tahap selanjutnya yaitu menampilkan hasil dari perhitungan algoritma apriori dengan variabel diantaranya *rule*, *support*, dan *confidence.*

```
# Menampilkan hasil asosiasi dari item
pd.set_option('max_colwidth', 1000)
Result=pd.DataFrame(columns=['Rule','Support','Confidence'])
for item in association_results:
    pair = item[2]
    for i in pair:
        items = str([x for x in i[0]])
        if i[3]!=1:
            Result=Result.append({
                'Rule':str([x for x in i[0]])+ " -> " +str([x for x in i[1]]),
                'Support':str(round(item[1]*100,2))+'%',
                'Confidence':str(round(i[2] *100,2))+'%'
                },ignore_index=True)
Result
```

Gambar 12. Menampilkan Hasil dari Perhitungan Algoritma Apriori

Berdasarkan Gambar 12, dilakukan untuk menampilkan hasil dari perhitungan Algoritma Apriori dengan penjelasan sebagai berikut:

a) *pd.set_options('max_colwidth', 1000)* : mengatur lebar kolom maksimum menjadi 1000 karakter.
b) *Result = pd.DataFrame(columns=['Rule', 'Support', 'Confidence'])* : membuat *DataFrame* kosong yang disebut 'result' dengan tiga kolom *'Rule','Support','Confidence'*.
c) Kode kemudian melakukan iterasi melalui *association_results.* Di dalam *loop*, kode mengekstrak informasi tentang setiap aturan asosiasi dan menambahkannya ke dalam *DataFrame 'Result'*.
d) *pair = item[2]* : mengekstrak informasi pasangan dari *association_results.*
e) *for i in pair*: melakukan iterasi melalui pasangan item pada aturan asosiasi.
f) *items = str([x for x in i[0]])* mengonversi sisi kiri dari aturan asosiasi menjadi *string*, yang mewakili *item-item* yang terlibat dalam bagian premis dari aturan.
g) Kode memeriksa apakah tingkat kepercayaan (*confidence*) dari aturan tidak sama dengan 1 (artinya, aturan yang memiliki kepercayaan sempurna diabaikan). Jika tingkat kepercayaan tidak sama dengan 1, maka kode menambahkan baris baru ke dalam *DataFrame Result* dengan informasi berikut:

a. *'Rule'*: Representasi *string* dari aturan asosiasi, misalnya, "A -> B."
b. *'Support'*: Dukungan *(support)* aturan (dibulatkan hingga 2 desimal) yang diwakili dalam bentuk persentase.
c. *'Confidence'*: Tingkat kepercayaan aturan (dibulatkan hingga 2 desimal) yang diwakili dalam bentuk persentase.

h) *DataFrame Result* kemudian diisi dengan aturan asosiasi dan nilai dukungan *(support)* serta tingkat kepercayaan *(confidence)* yang terkait kemudian ditampilkan.

Gambar 13. Tampilan Hasil dari Perhitungan menggunakan Algoritma Apriori

Dari Gambar 13. menghasilkan aturan asaosiasi dari hasil perhitungan menggunakan Algoritma Apriori dengan rincian sebagai berikut:

1. Aturan pertama mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* maka akan membeli *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 93.18% nilai *confidence.*
2. Aturan kedua mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* maka akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 100% nilai *confidence*.
3. Aturan ketiga mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* maka akan membeli *Security System Set Type G HV* menghasilkan nilai *support* sebesar 8.55% dari transaksi dan menghasilkan 97.73% nilai *confidence*.

4. Aturan keempat mengatakan bahwa ketika membeli *Security System Set Type G HV* maka akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* menghasilkan nilai *support* sebesar 8.55% dari transaksi dan menghasilkan 97.73% nilai *confidence*.
5. Aturan kelima mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* maka akan membeli *Shock Sensor Set* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 93.18% nilai *confidence*.
6. Aturam keenam mengatakan bahwa ketika membeli *Shock Sensor Set* maka akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 95.35% nilai *confidence*.
7. Aturan ke tujuh mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* maka akan membeli *Surround Monitor* menghasilkan nilai *support* sebesar 8.55% dari transaksi dan menghasilkan 97.73% nilai *confidence*.
8. Aturan ke delapan mengatakan bahwa ketika membeli *Surround Monitor* maka akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* menghasilkan nilai *support* sebesar 8.55% dari transaksi dan menghasilkan 70.49% nilai *confidence*.
9. Aturan ke sembilan mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* maka akan membeli *Surround Monitor* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 100% nilai *confidence*.
10. Aturan ke sepuluh mengatakan bahwa ketika membeli *Surround Monitor* maka akan membeli *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 67.21% nilai *confidence*.
11. Aturan ke sebelas mengatakan bahwa ketika membeli *Security System Set Type G GAS* maka akan membeli *Surround Monitor* menghasilkan nilai *support* sebesar 8.35% dari transaksi dan menghasilkan 100% nilai *confidence*.
12. Aturan ke dua belas mengatakan bahwa ketika membeli *Surround Monitor* maka akan membeli *Security System Set Type G GAS* menghasilkan nilai *support* sebesar 8.35% dari transaksi dan menghasilkan 68.85% nilai *confidence*.
13. Aturan ke tiga belas mengatakan bahwa ketika membeli *Security System Set Type G HV* maka akan membeli *Surround Monitor* menghasilkan nilai *support* sebesar 8.55% dari transaksi dan menghasilkan 97.73% nilai *confidence*.
14. Aturan ke empat belas mengatakan bahwa ketika membeli *Surround Monitor* maka akan membeli *Security System Set Type G HV* menghasilkan nilai *support* sebesar 8.55% dari transaksi dan menghasilkan 70.49% nilai *confidence*.
15. Aturan ke lima belas mengatakan bahwa ketika membeli *Shock Sensor Set* maka akan membeli *Surround Monitor* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 95.35% nilai *confidence*.
16. Aturan ke enam belas mengatakan bahwa ketika membeli *Surround Monitor* maka akan membeli *Shock Sensor Set* menghasilkan nilai *support* sebesar 8.15% dari ransaksi dan menghasilkan 67.21% nilai *confidence*.
17. Aturan ke tujuh belas mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)*, kemungkinan besar juga akan membeli *Surround Monitor* dan *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 93.18% nilai *confidence*.
18. Aturan ke delapan belas mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* Dan *Surround Monitor* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 100% nilai *confidence*.
19. Aturan ke sembilan belas mengatakan bahwa ketika membeli *Surround Monitor*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 67.21% nilai *confidence*.
20. Aturan ke dua puluh mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* kemungkinan besar juga akan membeli *Surround Monitor* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 100% nilai *confidence*.
21. Aturan ke dua puluh satu mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Surround Monitor*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 95.35% nilai *confidence*.
22. Aturan ke dua puluh dua mengatakan bahwa ketika membeli *Surround Monitor* dan *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* menghasilkan nilai *support* sebesar 8.15% dari transaksi dan menghasilkan 100% nilai *confidence*.
23. Aturan ke dua puluh tiga mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV*

(Attitude Black MC), kemungkinan besar juga akan membeli *Surround Monitor* dan *Security System Set Type G HV* menghasilkan nilai *support* sebesar 8.35% dari transaksi dan menghasilkan 95.45% nilai *confidence*.

24. Aturan ke dua puluh empat mengatakan bahwa ketika membeli *Security System Set Type G HV*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Surround Monitor* menghasilkan nilai *support* sebesar 8.35% dari transaksi dan menghasilkan 95.45% nilai *confidence*.
25. Aturan ke dua puluh lima mengatakan bahwa ketika membeli *Surround Monitor*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Security System Set Type G HV* menghasilkan nilai *support* sebesar 8.35% dari transaksi dan menghasilkan 68.85% nilai *confidence*.
26. Aturan ke dua puluh enam mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Security System Set Type G HV*, kemungkinan besar juga akan membeli *Surround Monitor* menghasilkan nilai *support* sebesar 8.35% dari transaksi dan menghasilkan 97.67% nilai *confidence*.
27. Aturan ke dua puluh tujuh mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Surround Monitor*, kemungkinan besar juga akan membeli *Security System Set Type G HV* menghasilkan nilai *support* sebesar 8.35% dari transaksi dan menghasilkan 97.67% nilai *confidence*.
28. Aturan ke dua puluh tujuh mengatakan bahwa ketika membeli *Surround Monitor* dan *Security System Set Type G HV*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* menghasilkan nilai *support* sebesar 8.35% dari transaksi dan menghasilkan 97.67% nilai *confidence*.

Dari perhitungan menggunakan Algoritma Apriori diatas menghasilkan enam pola penjualan dengan nilai *confidence* sebesar 100% antara lain:

1. Aturan kedua dengan *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* maka akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dengan nilai *support* sebesar 8.15%.
2. Aturan ke sembilan mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* maka akan membeli *Surround Monitor* dengan nilai *support* sebesar 8.15%.
3. Aturan ke sebelas mengatakan bahwa ketika membeli *Security System Set Type G GAS* maka akan membeli *Surround Monitor* dengan nilai *support* sebesar 8.35%.
4. Aturan ke delapan belas mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Surround Monitor* dengan nilai *support* sebesar 8.15%.
5. Aturan ke dua puluh mengatakan bahwa ketika membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dan *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* kemungkinan besar juga akan membeli *Surround Monitor* dengan nilai *support* sebesar 8.15%.
6. Aturan ke dua puluh dua mengatakan bahwa ketika membeli *Surround Monitor* dan *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)*, kemungkinan besar juga akan membeli *Clearance Sensor Set Q HV & V HV (Attitude Black MC)* dengan nilai *support* sebesar 8.15%.

KESIMPULAN DAN SARAN

Kesimpulan yang didapatkan pada penelitian ini adalah peneliti berhasil menganalisis data transaksi PT. Milenia Mega Mandiri dari Juli 2023 dengan jumlah transaksi sebanyak 2.392 data transaksi penjualan. Algoritma Apriori dapat diterapkan untuk mengidentifikasi suku cadang yang sering dibeli oleh konsumen dengan memperhatikan kebiasaan transaksi konsumen. Hasil analisis dengan menetapkan batas *minimum support* 8% dan *minimum confidence* 60% menghasilkan dua puluh delapan aturan asosiasi. Contohnya, jika seorang konsumen melakukan pembelian *Clearance Sensor Set Q HV & V HV (Platinum White Pearl MC)* maka 100% akan membeli *Surround Monitor*. *D*ari hasil aturan yang didapatkan dari data, dapat disimpulkan suku cadang apa yang sering dibeli bersamaan oleh setiap konsumen. Informasi ini dapat menjadi nilai tambah untuk meningkatkan penjualan dengan mengetahui kebiasaan pembelian suku cadang konsumen. Oleh karena itu, PT. Milenia Mega Mandiri dapat membuat keputusan bisnis dengan membuat rekomendasi *spare part* mobil. Saran yang dapat diberikan untuk penelitian ini adalah data yang dianalisis masih menggunakan nilai *minimum support* sebesar 8% dan *minimum confidence* sebesar 60%. Untuk penelitian selanjutnya, disarankan agar peneliti dapat mengeksplorasi variasi nilai *minimum support*, baik lebih dari 8% maupun kurang dari 8%, serta variasi nilai *minimum confidence*, baik lebih dari 60% maupun kurang dari 60%. Hal ini dapat dilakukan guna mendapatkan hasil aturan asosiasi yang lebih optimal dan menyeluruh, sehingga dapat memberikan pemahaman yang lebih baik terhadap pola hubungan antar data dalam konteks penjualan *spare part*.

DAFTAR PUSTAKA

[1] S. Zhang, K. Huang, and Y. Yuan, “Spare parts inventory management: A literature review,” *Sustain.*, vol. 13, no. 5, pp. 1–23, 2021, doi: 10.3390/su13052460.

[2] M. Ramadhan, J. Hutagalung, M. Dahria, I. Zulkarnain, and H. Jaya, “Prediksi Penjualan Spare Part Mobil Daihatsu Menggunakan

Algoritma Apriori," *Techno.Com*, vol. 22, no. 1, pp. 156–166, 2023, doi: 10.33633/tc.v22i1.7192.

[3] E. Widodo and Nabawi, "IMPLEMENTASI DATA MINING DENGAN METODE ALGORITMA APRIORI UNTUK MENENTUKAN POLA PEMBELIAN DI PT DONG SUNG TOOLS," *J. Teknol. Pelita Bangsa*, vol. 10, p. 64, 2019, doi: 10.1134/s0320972519100129.

[4] K. Annisa, B. S. Ginting, and M. A. Syari, "Penerapan Data Mining Pengelompokan Data Pengguna Air Bersih Berdasarkan Keluhannya Menggunakan Metode Clustering Pada PDAM Langkat," vol. 6341, no. April, 2022.

[5] N. N. Merliani, N. I. Khoerida, N. T. Widiawati, L. A. Triana, and P. Subarkah, "Penerapan Algoritma Apriori Pada Transaksi Penjualan Untuk Rekomendasi Menu Makanan Dan Minuman," *J. Nas. Teknol. dan Sist. Inf.*, vol. 8, no. 1, pp. 9–16, 2022, doi: 10.25077/teknosi.v8i1.2022.9-16.

[6] S. Styawati, A. Nurkholis, and K. N. Anjumi, "Analisis Pola Transaksi Pelanggan Menggunakan Algoritme Apriori," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 5, no. 2, pp. 619–626, 2021.

[7] I. Tahyudin *et al.*, *Big Data dan Analytics*. Purwokerto: Zahira Media Publiser, 2023.

[8] H. Bin Wang and Y. J. Gao, "Research on parallelization of Apriori algorithm in association rule mining," *Procedia Comput. Sci.*, vol. 183, pp. 641–647, 2021, doi: 10.1016/j.procs.2021.02.109.

[9] N. F. FAHRUDIN, "Penerapan Algoritma Apriori untuk Market Basket Analysis," *MIND J.*, vol. 1, no. 2, pp. 13–23, 2019, doi: 10.26760/mindjournal.v4i1.13-23.

[10] A. Salsabiela, A. P. Kuncoro, P. Subarkah, and P. Arsi, "Rekomendasi Restock Barang di Toko Pojok UMKM Menggunakan Algoritma K-Means Clustering," *J. Technol. Informatics*, vol. 5, no. 2, pp. 87–92, 2024, doi: 10.37802/joti.v5i2.554.

[11] I. Nurdiyanto, O. Nurdiawan, A. Irma Purnamasari, and D. Ade Kurnia, "Penentuan Keputusan Pemberian Pinjaman Kredit Menggunakan Algoritma C.45," *J. Dadta Sci. dan Inform.*, vol. 2, no. 1, pp. 1–5, 2022.

[12] A. G. Andersen *et al.*, "Evaluation of an a priori scatter correction algorithm for cone-beam computed tomography based range and dose calculations in proton therapy," *Phys. Imaging Radiat. Oncol.*, vol. 16, no. May, pp. 89–94, 2020, doi: 10.1016/j.phro.2020.09.014.

[13] F. Navidi, I. L. Gørtz, and V. Nagarajan, "Approximation algorithms for the a priori traveling repairman," *Oper. Res. Lett.*, vol. 48, no. 5, pp. 599–606, 2020, doi: 10.1016/j.orl.2020.07.009.

[14] H. Yu, "Apriori algorithm optimization based on Spark platform under big data," *Microprocess. Microsyst.*, vol. 80, no. November 2020, p. 103528, 2021, doi: 10.1016/j.micpro.2020.103528.

[15] M. Giacomini, L. Borchini, R. Sevilla, and A. Huerta, "Separated response surfaces for flows in parametrised domains: Comparison of a priori and a posteriori PGD algorithms," *Finite Elem. Anal. Des.*, vol. 196, no. November 2020, p. 103530, 2021, doi: 10.1016/j.finel.2021.103530.

[16] P. Haryandi, Y. Widiastiwi, and N. Chamidah, "Penerapan Algoritma Apriori untuk Mencari Pola Penjualan Produk Herbal (Studi Kasus: Toko Hanawan Gemilang)," *Inform. J. Ilmu Komput.*, vol. 17, no. 3, p. 218, 2021, doi: 10.52958/iftk.v17i3.3655.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, Bulan Oktober Tahun 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.704

Pengaruh Big Data dalam Akuntansi Syariah: Analisis Prediktif dan Pengambilan Keputusan Strategis

Adhi Riza Aulia[1], Sulis Saputra[2], Arga Dwi Titandy [3], Mohammad Zacky [4], Agus Arwani[5]
[1,2,3,4]Akuntansi Syariah, UIN K.H. Abdurrahman Wahid, Pekalongan, Indonesia
e-mail: adhi.riza.aulia@mhs.uingusdur.ac.id[1], sulis.saputra@mhs.uingusdur.ac.id[2],
arga.dwi.titandy@mhs.uingusdur.ac.id[3], mohammad.zacky@mhs.uingusdur.ac.id[4]
* Penulis Korespondensi: E-mail: adhi.riza.aulia@mhs.uingusdur.ac.id

Abstrak: Penelitian ini membahas pengaruh Big Data dalam akuntansi syariah, dengan fokus pada analisis prediktif dan pengambilan keputusan strategis. Permasalahan yang dihadapi adalah kompleksitas dan volume data yang sulit diproses dengan metode tradisional. Tujuan dari penelitian ini adalah untuk mengevaluasi dampak Big Data pada akuntansi syariah dan mengidentifikasi manfaat serta konsekuensi strategisnya. Metode yang digunakan dalam penelitian ini adalah studi literatur. Pencarian dilakukan menggunakan basis data akademik seperti Google Scholar, ResearchGate, ScienceDirect, dan Wiley Online Library dengan menggunakan kata kunci yang relevan. Literatur yang dipilih memenuhi kriteria inklusi berdasarkan relevansi dengan Big Data dalam konteks akuntansi syariah, kualitas metodologi penelitian, dan kebaruan. Hasil penelitian menunjukkan bahwa penggunaan Big Data dalam akuntansi syariah dapat meningkatkan efisiensi proses akuntansi melalui otomatisasi pemrosesan transaksi. Selain itu, analisis prediktif berbasis big data dapat membantu meramalkan hasil keuangan di masa depan. Sedangkan analisis strategis dilakukan terhadap penelitian yang terpilih untuk mengidentifikasi temuan-temuan utama, metodologi yang digunakan, data yang dikumpulkan, serta hasil dan kesimpulan penelitian tersebut. Pengambilan keputusan strategis berdasarkan informasi dari analisis Big Data memungkinkan pengambilan keputusan yang lebih tepat waktu dan informatif. Peningkatan efisiensi dapat dibandingkan dengan metode manual melalui indikator seperti waktu pemrosesan transaksi, akurasi laporan keuangan, dan pengurangan kesalahan manusia. Studi empiris yang melibatkan data historis transaksi keuangan syariah menunjukkan bahwa waktu pemrosesan transaksi dapat berkurang hingga 50% dengan penggunaan Big Data dibandingkan dengan metode manual.

Kata Kunci: Akuntansi Syariah; Big Data; Pengambilan Keputusan

Abstract: *This research discusses the influence of Big Data in Islamic accounting, focusing on predictive analysis and strategic decision making. The problem faced is the complexity and volume of data that is difficult to process with traditional methods. The purpose of this study is to evaluate the impact of Big Data on Islamic accounting and identify its strategic benefits and consequences. The method used in this research is a literature study. The search was conducted using academic databases such as Google Scholar, ResearchGate, ScienceDirect, and Wiley Online Library using relevant keywords. The selected literature met the inclusion criteria based on relevance to Big Data in the context of Islamic accounting, quality of research methodology, and novelty. The results showed that the use of Big Data in Islamic accounting can improve the efficiency of accounting processes through the automation of transaction processing. This research discusses the influence of Big Data in Islamic accounting, focusing on predictive analysis and strategic decision making. The problem faced is the complexity and volume of data that is difficult to process with traditional methods. The purpose of this study is to evaluate the impact of Big Data on Islamic accounting and identify its strategic benefits and consequences. The method used in this research is a literature study. The search was conducted using academic databases such as Google Scholar, ResearchGate, ScienceDirect, and Wiley Online Library using relevant keywords. The selected literature met the inclusion criteria based on relevance to Big Data in the context of Islamic accounting, quality of research methodology, and novelty. The results showed that the use of Big Data in Islamic accounting can improve the efficiency of accounting processes through the automation of transaction processing.*

Keywords: *Big Data; Decision Making; Sharia Accounting*

PENDAHULUAN

Sebagai cabang akuntansi yang berbasis syariah, akuntansi syariah telah menjadi bidang yang semakin penting dalam dunia keuangan. Perkembangan teknologi informasi di era modern telah membuka peluang baru yang signifikan untuk meningkatkan efektivitas dan efisiensi proses akuntansi syariah [1]. Big Data adalah salah satu teknologi yang sedang menjadi perhatian besar saat ini. Potensi besar yang dimiliki teknologi ini untuk mengubah cara orang mengumpulkan, menganalisis, dan menggunakan data dalam banyak bidang, seperti akuntansi [2].

Dalam akuntansi syariah, "Big Data" merupakan istilah yang digunakan untuk

menggambarkan kumpulan data yang sangat besar, kompleks, dan beragam yang sulit untuk diproses dengan metode tradisional [3]. Kumpulan data ini dapat berasal dari transaksi keuangan, interaksi pelanggan, dan berbagai sumber lainnya. Oleh karena itu, penggunaan Big Data dapat menawarkan manfaat yang signifikan dalam mengatasi masalah yang terkait dengan volume dan kompleksitas data yang ada.

Adapun tujuan pada penelitian ini adalah untuk mengevaluasi dampak Big Data pada akuntansi syariah, dengan penekanan pada analisis prediktif dan pengambilan keputusan strategis. Dalam konteks ini, analisis prediktif mengacu pada teknik dan algoritma analisis data yang digunakan untuk menemukan pola, tren, dan perilaku yang dapat membantu meramalkan hasil keuangan di masa depan. Sementara pengambilan keputusan strategis mengacu pada penggunaan informasi yang diperoleh dari analisis Big Data untuk meramalkan hasil keuangan di masa depan.

Dalam menghadapi tantangan yang dihadapi oleh organisasi yang bergerak pada bidang akuntansi syariah, penelitian ini sangat penting. Perkembangan teknologi informasi telah menciptakan peluang baru dan tantangan dalam manajemen data akuntansi syariah di dunia digital saat ini. Memahami dampak Big Data dapat membantu organisasi mengoptimalkan penggunaan sumber daya, meningkatkan efisiensi operasional, dan membuat keputusan yang lebih baik.

Studi telah menunjukkan bahwa menggunakan big data dalam akuntansi syariah dapat menguntungkan. Misalnya, penelitian yang dilakukan oleh [4] menyelidiki penggunaan analisis big data dalam audit syariah dan menemukan bahwa penggunaan alat analisis big data dapat meningkatkan efisiensi dan efektivitas proses audit syariah. Mengevaluasi dampak Big Data pada akuntansi syariah melalui studi literatur dan analisis empiris. Mengidentifikasi manfaat strategis penggunaan Big Data dalam meningkatkan efisiensi dan pengambilan keputusan dalam akuntansi syariah. Selain itu, penelitian yang dilakukan oleh [5] menemukan bahwa analisis risiko dalam perbankan syariah dapat diperkuat, membantu dalam pengambilan keputusan yang lebih baik, dan meningkatkan manajemen risiko. Penelitian ini dapat memberikan wawasan berharga tentang bagaimana Big Data dapat digunakan untuk mengatasi tantangan dalam akuntansi syariah. Dan Menawarkan dasar untuk pengembangan metode dan praktik yang lebih baik untuk mengintegrasikan Big Data ke dalam proses akuntansi syariah. Dalam penelitian ini juga diharapkan dapat membantu organisasi akuntansi syariah untuk mengoptimalkan penggunaan sumber daya dan meningkatkan efisiensi operasional. Dan mampu memberikan dasar untuk pengambilan keputusan yang lebih baik berdasarkan analisis data yang mendalam.

Penggunaan Big Data dalam akuntansi syariah memiliki konsekuensi strategis selain keuntungan operasional [6] menunjukkan bahwa analisis prediktif berbasis big data' sangat penting untuk memahami perilaku dan preferensi pelanggan di industri keuangan syariah. Penelitian ini membantu Perusahaan dalam membuat strategi pemasaran yang lebih baik dan memenuhi kebutuhan pasar.

Beberapa batasan penelitian ini harus dipertimbangkan. Pertama, penelitian ini terbatas pada dampak Big Data pada akuntansi syariah dan tidak mencakup aspek lain dari teknologi informasi yang relevan. Kedua, penelitian ini didasarkan pada studi literatur yang relevan dan analisis kritis, dan tidak melibatkan penelitian empiris. Akibatnya, hasil penelitian ini bersifat konseptual dan tidak dapat digunakan secara langsung dalam dunia nyata.

Dengan mempertimbangkan latar belakang, tujuan, relevansi, dan keterbatasan penelitian ini, penelitian ini diharapkan dapat memberikan dorongan wawasan yang berharga tentang dampak besar data pada akuntansi syariah. Selain itu, penelitian ini juga dapat memberikan dasar untuk pengembangan metode dan praktik yang lebih baik untuk mengintegrasikan big data ke dalam proses akuntansi syariah. Selain itu, penelitian ini juga dapat mengidentifikasi area penelitian yang menarik untuk penelitian masa depan.

METODE

Pengaruh Big Data dalam akuntansi syariah diteliti melalui metode studi literatur. Berikut ini adalah tahapan metodologi yang digunakan. Pertama, penelitian literatur dilakukan dengan menggunakan kata kunci yang relevan, seperti "Big Data", "akuntansi syariah", "teknologi informasi", dan "pengaruh menggunakan basis data akademik seperti Google Scholar, ResearchGate, ScienceDirect, dan Wiley Online Library adalah beberapa platform di mana pencarian dilakukan.

Pencarian Literatur: Gunakan basis data akademik seperti Google Scholar, ResearchGate, dan database jurnal terkemuka seperti ScienceDirect dan Wiley Online Library untuk melakukan pencarian literatur. Kata kunci seperti "Big Data", "akuntansi syariah", "teknologi informasi", dan "pengaruh" digunakan dalam pencarian.

Kedua, literatur yang relevan dengan fokus penelitian ini dipilih berdasarkan kriteria inklusi yang meliputi relevansi dengan Big Data dalam konteks akuntansi syariah, kualitas metodologi penelitian, dan kebaruan [7]. Penelitian yang menggunakan pendekatan kualitatif maupun kuantitatif, baik penelitian penuh maupun ringkasan, dipertimbangkan.

Ketiga, Analisis kritis dilakukan terhadap penelitian yang terpilih untuk mengidentifikasi temuan-temuan utama, metodologi yang digunakan, data yang dikumpulkan, serta hasil dan kesimpulan penelitian tersebut [8]. Dalam analisis kritis, kekuatan dan kelemahan setiap penelitian dievaluasi untuk membangun pemahaman yang komprehensif tentang pengaruh Big Data dalam akuntansi syariah.

HASIL DAN PEMBAHASAN

1. Pengaruh Big Data terhadap Efisiensi Proses Akuntansi Syariah

Akuntansi syariah dapat menggunakan Big Data untuk meningkatkan efisiensi prosesnya [9]. Ini karena jumlah data yang sangat besar ini memiliki kemampuan untuk mengumpulkan, menyimpan, dan menganalisis [10]. Ini memungkinkan akuntan syariah untuk mengotomatiskan tugas-tugas biasa seperti memproses transaksi dan laporan keuangan. Hal ini dapat mengurangi waktu dan upaya yang dibutuhkan untuk proses akuntansi, yang memungkinkan penggunaan sumber daya yang lebih efisien dan pengalokasian waktu yang lebih baik untuk analisis yang lebih mendalam.

Di mana salah satu faktor penting dalam meningkatkan efisiensi dalam proses akuntansi syariah adalah otomatisasi pemrosesan transaksi [11]. Dengan memanfaatkan Big Data, perusahaan dapat menggabungkan sistem akuntansi mereka dengan teknologi canggih, seperti sistem basis data terdistribusi dan teknik pemrosesan data secara *real-time* [12]. Ini memungkinkan perusahaan untuk mendapatkan informasi keuangan yang lebih cepat dan akurat, mengurangi keterlambatan dalam pemrosesan transaksi, dan memungkinkan dalam pengambilan keputusan yang lebih tepat waktu [13].

Big Data memungkinkan akuntan syariah untuk mengoptimalkan pelaporan keuangan mereka [14]. Perusahaan dapat menghasilkan laporan keuangan yang lebih lengkap dan relevan dengan menganalisis semua data, termasuk data internal perusahaan dan data eksternal seperti tren industri dan informasi pasar [15]. Selain itu, informasi yang diperoleh dari Big Data juga dapat digunakan untuk mengidentifikasi potensi risiko dan peluang yang dapat mempengaruhi kinerja keuangan perusahaan.

Dalam akuntansi syariah, penggunaan Big Data juga dapat meningkatkan pengawasan dan penegakan kepatuhan syariah. Bisnis dapat menemukan kemungkinan ketidaksesuaian dengan prinsip-prinsip syariah, seperti transaksi yang melanggar larangan riba atau *gharar*, dengan menganalisis data secara menyeluruh [16]. Ini memungkinkan perusahaan untuk mengambil tindakan yang diperlukan secara proaktif dan meminimalkan risiko pelanggaran prinsip syariah.

Namun, penting untuk diingat bahwa penggunaan Big Data dalam akuntansi syariah menghadirkan sejumlah masalah. Salah satu tantangan utama adalah kebutuhan akan infrastruktur teknologi yang memadai dan keahlian dalam menganalisis data yang kompleks [17]. Perusahaan perlu menginvestasikan sumber daya dalam pengembangan dan pemeliharaan sistem teknologi yang dapat mengelola dan menganalisis data Big Data dengan efisien dan akurat.

Big Data dapat meningkatkan efisiensi akuntansi syariah dengan mengotomatisasi tugas-tugas rutin seperti pemrosesan transaksi dan pelaporan keuangan. Penelitian sebelumnya juga disebutkan dalam penggunaan analisis Big Data dalam audit syariah dapat meningkatkan efisiensi dan efektivitas proses audit hingga 30%. Analisis risiko dalam perbankan syariah diperkuat dengan Big Data, yang membantu dalam pengambilan keputusan yang lebih baik dan manajemen risiko yang lebih efektif. Dalam penelitian ini diperlukan kajian yang lebih komprehensif dengan menggabungkan berbagai studi yang telah dilakukan terkait penggunaan Big Data dalam akuntansi syariah.

Analisis kritis terhadap berbagai metodologi yang digunakan dalam studi-studi tersebut untuk mengidentifikasi keunggulan dan keterbatasannya.

2. Analisis Prediktif Menggunakan Big Data dan Pengambilan Keputusan Strategis

Salah satu keunggulan utama analisis prediktif menggunakan Big Data adalah kemampuannya untuk mengolah *volume* Big data dan beragam dengan cepat [18]. Data yang dikumpulkan dari berbagai sumber, seperti transaksi keuangan, data pasar, dan data operasional, dapat digabungkan dan dianalisis secara menyeluruh. Hal ini memungkinkan perusahaan untuk mendapatkan wawasan yang lebih mendalam tentang pola dan hubungan yang ada dalam data tersebut.

Pengambilan keputusan strategis dengan menggunakan Big Data melibatkan proses analisis data yang mendalam untuk mendukung pengambilan keputusan tingkat atas dalam organisasi [19]. Dengan mengumpulkan, mengolah, dan menganalisis big data, organisasi dapat memahami tren pasar, mengidentifikasi peluang dan risiko, serta merencanakan strategi bisnis yang tepat. Pengambilan keputusan yang didasarkan pada data memungkinkan organisasi untuk mengoptimalkan kinerja operasional, meningkatkan kepuasan pelanggan, dan mencapai tujuan bisnis mereka dengan lebih efisien [20].

3. Tantangan Penggunaan Big Data dalam Akuntansi Syariah

Penggunaan Big Data dalam akuntansi Syariah menghadapi sejumlah tantangan yang harus diatasi untuk memaksimalkan manfaatnya. Salah satu tantangan utamanya adalah kebutuhan infrastruktur teknologi yang memadai [21]. Mengelola dan menganalisis Big Data memerlukan sistem yang mampu mengumpulkan, menyimpan, dan memproses data dalam jumlah besar dengan cepat dan efisien [22]. Sumber daya untuk mengembangkan dan memelihara infrastruktur teknologi yang memenuhi kebutuhan mereka.

Tantangan lainnya adalah adopsi budaya dan perubahan dalam kebiasaan kerja. Penggunaan Big Data membutuhkan pendekatan yang berbeda dalam pengumpulan, pengolahan, dan analisis data. Perusahaan perlu mengubah budaya kerja dan memastikan perusahaan memiliki pemahaman dan keterampilan yang sesuai. Pelatihan dan pengembangan karyawan menjadi penting untuk memastikan bahwa mereka dapat memanfaatkan

teknologi Big Data dengan efektif [23]. Perubahan dalam sistem kerja dan proses bisnis juga mungkin diperlukan untuk mengintegrasikan penggunaan Big Data secara efisien dalam akuntansi syariah.

Perusahaan perlu waspada dalam menghadapi tantangan-tantangan yang terjadi, perusahaan perlu memiliki komitmen yang kuat untuk mengadopsi Big Data dan mengatasi hambatan yang mungkin muncul [24]. Dengan menghadapi tantangan ini, perusahaan dapat memanfaatkan potensi Big Data dalam akuntansi syariah untuk meningkatkan efisiensi, akurasi, dan pengambilan keputusan yang lebih baik.

KESIMPULAN DAN SARAN

Penelitian pengaruh Big Data dalam Akuntansi Syariah: Analisis Prediktif dan Pengambilan Keputusan Strategis dalam konteks akuntansi syariah dengan fokus pada analisis prediktif dan pengambilan keputusan strategis. Penggunaan Big Data dalam akuntansi syariah memberikan manfaat signifikan dalam meningkatkan efisiensi proses akuntansi, mempercepat pemrosesan transaksi, dan mengoptimalkan pelaporan keuangan. Studi empiris menunjukkan bahwa penggunaan Big Data dapat meningkatkan efisiensi hingga 50% dibandingkan dengan metode manual. Analisis risiko yang diperkuat dengan Big Data membantu dalam pengambilan keputusan yang lebih baik dan manajemen risiko yang lebih efektif. Hasil penelitian menunjukkan bahwa penggunaan Big Data dapat memberikan manfaat signifikan dalam meningkatkan efisiensi proses akuntansi syariah. Dengan kemampuan untuk mengumpulkan, menyimpan, dan menganalisis jumlah data yang besar, akuntan syariah dapat mengotomatiskan tugas-tugas rutin dan mempercepat pemrosesan transaksi serta laporan keuangan. Hal ini menghasilkan penggunaan sumber daya yang lebih efisien dan memberikan waktu yang lebih banyak untuk analisis mendalam. Selain itu, penggunaan Big Data juga dapat memperkuat analisis risiko, membantu pengambilan keputusan yang lebih baik, dan meningkatkan manajemen risiko dalam perbankan syariah. Oleh karena itu, pemahaman yang lebih baik tentang dampak Big Data dalam akuntansi syariah dapat membantu organisasi mengoptimalkan penggunaan sumber daya dan membuat keputusan yang lebih baik.

Terdapat beberapa saran yang dapat diberikan untuk pengembangan lebih lanjut dalam penggunaan Big Data dalam akuntansi syariah. Organisasi yang bergerak dalam bidang akuntansi syariah sebaiknya mempertimbangkan investasi dalam infrastruktur teknologi yang mendukung pengolahan Big Data. Hal ini akan memungkinkan mereka untuk mengumpulkan, menyimpan, dan menganalisis data dengan efisien dan efektif, Diperlukan pelatihan dan peningkatan kompetensi bagi para akuntan syariah dalam hal penggunaan alat dan teknik analisis data yang berkaitan dengan Big Data. Ini akan memastikan bahwa mereka dapat memanfaatkan potensi penuh teknologi ini dan menghasilkan wawasan yang bernilai dalam pengambilan keputusan, dan Adopsi Big Data dalam akuntansi syariah harus dilakukan dengan memperhatikan aspek keamanan dan privasi data. Organisasi perlu mengimplementasikan langkah-langkah yang tepat untuk melindungi kerahasiaan dan integritas data yang dikumpulkan.

Ucapan Terimakasih

Penulis hanya menyampaikan ucapan terima kasih yang tulus kepada semua pihak yang telah berkontribusi dalam penelitian ini. Tanpa dukungan dan bantuan mereka, penelitian ini tidak akan terwujud. Penulis mengungkapkan rasa terima kasih kepada Institusi yaitu UIN K.H. Abdurrahman Wahid, Pekalongan, Jawa Tengah atas dukungan dan fasilitas yang diberikan dalam menjalankan penelitian ini. Institusi ini telah menyediakan lingkungan yang kondusif bagi penulis untuk mengembangkan pengetahuan dan melaksanakan penelitian. Rekan Peneliti yang telah bekerja sama dalam penelitian ini. Kolaborasi penulis dalam mengumpulkan data, menganalisis informasi, dan menyusun penelitian ini telah memberikan kontribusi yang berharga terhadap kesuksesan penelitian. Akhir kata, penulis menyampaikan apresiasi kepada semua pihak yang telah berperan dalam penelitian ini. Kontribusi dan dukungan mereka telah berperan penting dalam kesuksesan penelitian ini. Semoga hasil penelitian ini dapat memberikan manfaat dan memberikan kontribusi positif dalam pengembangan akuntansi syariah.

DAFTAR PUSTAKA

[1] Ahmad Taufiq Harahap, "Perkembangan Akuntansi Syariah Di Indonesia," *J. War. Ed.*, vol. 53, no. 1829–7463, p. Jurnal Warta, 2020.

[2] D. Heryana, L. Setiawati, and B. Suhendar, "Sistem Informasi Dan Potensi Manfaat Big Data Untuk Pendidikan," *Gunahumas*, vol. 2, no. 2, pp. 350–357, 2021, doi: 10.17509/ghm.v2i2.23023.

[3] M. Favaretto, E. de Clercq, C. O. Schneble, and B. S. Elger, "What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade," *PLoS One*, vol. 15, no. 2, pp. 1–20, 2022, doi: 10.1371/journal.pone.0228987.

[4] M. S. Al-Tamimi, H. A., & Al-Mazrooei, "Big Data Analytics in Islamic Finance: A Review and Directions for Future Research," *J. Islam. Account. Bus. Res.*, vol. 9, no. 2, pp. 203–221, 2022.

[5] Y. Hassan, A., Aziz, S. A. A., & Zainuddin, "Big Data Analytics and Risk Management in Islamic Banking: An Exploratory Study," *J. Islam. Account. Bus. Res.*, vol. 10, no. 5, pp. 1028–1045, 2023.

[6] K. A. Al-Harrasi, R. S., Al-Busaidi, K. A., Al-Musawi, A. S., & Al-Ashoori, "Predictive Analytics and Big Data in Islamic Banking: A Systematic Literature Review and Future Research Directions," *J. Financ. Serv. Mark.*, vol. 25, no. 1, pp. 35–48, 2021.

[7] S. V. Siregar, "Big Data: Tantangan dan Peluang

dalam Akuntansi Syariah," *urnal Akunt. dan Audit. Indones.*, vol. 24, no. 1, pp. 53-66., 2022.

[8] M. Hidayat, R., & Sulaiman, "Analisis Big Data dalam Mendukung Efisiensi Reksa Dana Syariah di Indonesia," *J. Ris. Akunt. dan Keuang.*, vol. 6, no. 2, pp. 229–238, 2024.

[9] Y. Rahmawati, "Akuntansi Syariah di Indonesia dalam Era Digital," *Indones. J. Islam. Econ. Financ.*, vol. 2, no. 1, pp. 1–12, 2022, doi: 10.37680/ijief.v2i1.1366.

[10] V. N. A. Z. A. Santari, "'Pengaruh Teknologi Big Data Terhadap Pengambilan Keputusan Akuntansi,'" no. september 2016, pp. 1–6, 2022.

[11] S. Oktaviani, A. F. Hanifah, R. Achmad, A. Shiroth, and M. Wf, "Peranan Teknologi dalam Meningkatkan Efisiensi Manajemen," no. November, 2023.

[12] A. A. Bakri, Y. Yusni, and N. Botutihe, "Analisis Efektivitas Penggunaan Teknologi Big Data dalam Proses Audit: Studi Kasus pada Kantor Akuntan Publik di Indonesia," *J. Akunt. Dan Keuang. West Sci.*, vol. 2, no. 03, pp. 179–186, 2023, doi: 10.58812/jakws.v2i03.641.

[13] Endang Jumali, "Peranan_Akuntansi_Syariah_Dalam_Perkembangan_Keuan," *J. Akunt. DAN Keuang. Islam*, vol. 02, no. 01, pp. 1–14, 2022.

[14] E. Manif, A. Jarboui, M. K. Hassan, and K. Mouakhar, "Big data tools for Islamic financial analysis," *Intell. Syst. Accounting, Financ. Manag.*, vol. 27, no. 1, pp. 10–21, 2020, doi: 10.1002/isaf.1463.

[15] M. M. Hasan, J. Popp, and J. Oláh, "Current landscape and influence of big data on finance," *J. Big Data*, vol. 7, no. 1, 2022, doi: 10.1186/s40537-020-00291-z.

[16] S. Mardian, "Tingkat Kepatuhan Syariah di Lembaga Keuangan Syariah," *J. Akunt. Dan Keuang. Islam*, vol. 3, no. 1, pp. 57–68, 2024, doi: 10.35836/jakis.v3i1.41.

[17] Rahmawati R., "Pengaruh Big Data Terhadap Efisiensi Proses Akuntansi Syariah," *J. Int. Ekon. Perdagang. dan Manaj.*, vol. 10, no. (2), pp. 92-103.2021.

[18] Santari PAK, "Analisis Big Data dalam Profesi Akuntansi: Suatu Tinjauan.," *J. Stud. Akuntansi, Keuangan, dan Audit*, vol. ,6, no. (2), pp. 1-13.2021.

[19] dkk. Bakri, A., "Analisis Big Data dan Sistem Informasi Akuntansi Real-Time untuk Pengambilan Keputusan.," *J. Penelit. Akunt. dan Bisnis Islam.*, vol. 14, no. (2), pp. 252-267.2021.

[20] Y. Mardian, "Big Data dan Pemantauan Kepatuhan dalam Keuangan Islam.," *J. Ekon. Islam. Perbank. dan Keuang.*, vol. 15, no. (1), pp. 134-145.2024.

[21] dkk. Oktaviani, D., "Peran Otomatisasi Data dalam Meningkatkan Efisiensi Proses Akuntansi.," *J. Akunt. dan Manaj. Keuang.*, vol. 1, no. (1), pp. 45-62.2023.

[22] dkk. Hasan, M., "Peran Big Data dalam Meningkatkan Analisis Kinerja Keuangan.," *J. Penelit. Akunt. dan Bisnis Islam.*, vol. 11, no. (5), pp. 863-880.2021.

[23] D. Manif, F., "Analisis Big Data dan Kualitas Pelaporan Keuangan.," *J. Int. Ekon. Perdagang. dan Manajemen*, vol. 8, no. (1), pp. 98-117.2023.

[24] Endang Jumali., "Dampak Big Data Terhadap Ketepatan Waktu Pelaporan Keuangan.," *J. Ekon. dan Keuang. Islam.*, vol. 8, no. (1), Tahun 2023, pp. 87-102.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, Bulan Oktober Tahun 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.763

Prediksi Harga Mobil Audi Bekas Menggunakan Model Regresi Linear dengan *Framework* Streamlit

Putri Aulia Azhar[1], **Muhammad Arya Pratama**[2], **Risma Fitriani**[3]
[1,2,3]Akuntansi, Universitas Singaperbangsa Karawang, Karawang, Jawa Barat, Indonesia
e-mail: 2210631030128@student.unsika.ac.id[1], 2210631030115@student.unsika.ac.id[2],
risma.fitriani@ft.unsika.ac.id[3]
* Penulis Korespondensi: E-mail: 2210631030128@student.unsika.ac.id

Abstrak: Prediksi harga mobil bekas menjadi aspek krusial dalam pasar otomotif yang terus berkembang, hal ini akan memengaruhi keputusan penjual dan pembeli serta dinamika pasar secara keseluruhan. Meskipun kompleksitas dan variasi kondisi mobil bekas sering menjadi hambatan, adopsi teknologi dan metode yang tepat diharapkan mampu mengatasi tantangan tersebut. Melalui program prediksi harga, peneliti bermaksud untuk memberikan solusi praktis dan efisien bagi penjual dan pembeli mobil bekas Audi, dengan tujuan meningkatkan keputusan yang terinformasi dan tepat. Penelitian ini meggunakan algoritma *Linear Regression* untuk menemukan hubungan linear antara variabel-variabel yang digunakan. Tahapan metode yang digunakan meliputi *problem recognition, research plan, dataset collection, data pre-processing, error calculation,* dan *web deployment*. Hasilnya menunjukkan nilai akurasi sebesar 94% dan nilai *Mean Absolute Error* (MAE) sebesar 0,08 atau 8%. Hal ini memberikan indikasi yang kuat bahwa model *linear regression* yang digunakan telah berhasil dalam menghasilkan prediksi yang akurat untuk harga mobil bekas Audi.

Kata Kunci: *Machine Learning*; Prediksi; Regresi Linear; *Streamlit*

Abstract: *Used car price prediction is a crucial aspect of the ever-evolving automotive market, affecting the decisions of sellers and buyers as well as the overall market dynamics. While the complexity and varied conditions of used cars are often an obstacle, the adoption of appropriate technologies and methods is expected to overcome these challenges. Through a price prediction program, the researcher intends to provide a practical and efficient solution for sellers and buyers of Audi used cars, with the aim of improving informed and precise decisions. This research utilizes the Linear Regression algorithm to find a linear relationship between the variables used. The stages of the method used include problem recognition, research plan, dataset collection, data pre-processing, error calculation, and web deployment. The results show an accuracy value of 94% and a Mean Absolute Error (MAE) value of 0.08 or 8%. This gives a strong indication that the linear regression model used has been successful in producing accurate predictions for Audi used car prices.*

Keywords: *Machine Learning; Prediction; Linear Regression; Streamlit*

PENDAHULUAN

Prediksi merupakan analisis data dan informasi untuk meramalkan peristiwa di masa depan. Dalam bidang ilmiah, prediksi berguna untuk memproyeksikan hasil penelitian, perubahan iklim, atau dampak dari aktivitas tertentu. Proyeksi ini menggunakan informasi dari data historis, model matematika, dan analisis statistik. Dalam praktiknya, prediksi harus dilakukan secara terstruktur dan logis, serta mempertimbangkan berbagai faktor yang dapat mempengaruhi hasil akhirnya. Prediksi dapat digunakan sebagai landasan untuk mengambil keputusan strategis yang tepat di berbagai bidang, termasuk riset, kebijakan, dan bisnis [1].

Prediksi harga mobil bekas merupakan suatu aspek penting dalam pasar otomotif yang terus berkembang. Dalam konteks ini, prediksi menjadi sebuah instrumen vital bagi para penjual dan pembeli untuk mengambil keputusan yang tepat. Prediksi harga mobil bekas tidak hanya mempengaruhi keputusan individu, tetapi juga dinamika pasar secara keseluruhan. Kemampuan untuk meramalkan harga dengan akurat memungkinkan penjual untuk menentukan harga yang kompetitif dan mengoptimalkan keuntungan, sementara pembeli dapat membuat keputusan yang lebih informasional dan tepat waktu. Namun, kendala-kendala yang dihadapi dalam menentukan harga jual maupun beli mobil bekas sering kali menjadi hambatan yang signifikan. Meskipun demikian, kompleksitas pasar dan variasi yang luas dalam kondisi mobil bekas sering kali membuat proses prediksi menjadi rumit dan membutuhkan pendekatan yang cermat. Dengan adopsi teknologi dan metode yang tepat, diharapkan bahwa program prediksi harga ini dapat menjadi solusi yang efektif dalam mengatasi tantangan-tantangan tersebut, menyediakan informasi yang berharga bagi para pemangku kepentingan dalam industri otomotif [2].

Mobil bekas Audi telah menjadi populer di pasar, terutama di kalangan penggemar mobil yang mencari kualitas, teknologi, dan desain yang tinggi namun dengan biaya yang lebih terjangkau. Salah satu alasan popularitas mobil bekas Audi adalah karena kualitas yang sangat baik dan teknologi yang diterapkan pada mobil-mobil ini. Merek ini juga dikenal dengan inovasi teknologi yang terus-menerus dilakukan, seperti pengembangan transmisi *multitronic* yang

memungkinkan pengemudi untuk mengendarai mobil dengan lebih mudah dan efisien. Dalam beberapa tahun terakhir, Audi juga telah meningkatkan penjualan mobil bekas dengan meluncurkan berbagai model yang lebih murah dan lebih efisien. Hal ini membuat mobil bekas Audi menjadi pilihan yang lebih terjangkau bagi banyak penggemar, tanpa mengorbankan kualitas dan teknologi yang tinggi [3].

Dengan menghadapi kompleksitas dan ketidakpastian dalam menentukan harga mobil bekas, kebutuhan akan alat yang dapat memberikan panduan yang jelas dan efektif semakin mendesak. Program prediksi harga ini dikembangkan tidak hanya menghadirkan kemudahan, tetapi juga akurasi dalam proses penentuan harga mobil bekas. Program ini dirancang dengan tujuan utama untuk memberikan solusi yang praktis dan efisien bagi penjual dan pembeli, sehingga mereka dapat mengambil keputusan yang terinformasi dan tepat. Melalui layanan *website* yang *user-friendly*, pengguna dapat dengan mudah mengakses prediksi harga yang disesuaikan dengan kriteria mereka, meningkatkan efisiensi dan keselarasan dalam transaksi jual-beli mobil bekas [4].

Machine Learning (ML) adalah disiplin ilmu yang mempelajari cara mesin belajar dari data untuk meningkatkan kemampuan prediksi tanpa bergantung pada aturan yang telah ditetapkan secara spesifik. Dalam ML, mesin mampu membuat model dari data yang disediakan dan memperbaiki model tersebut melalui iterasi pelatihan. Pendekatan ini menggunakan teknik *Supervised Learning (SL)*, yang terbukti efektif dalam analisis data historis untuk menghasilkan prediksi yang akurat. Dengan menerapkan SL, sistem dapat belajar dari data masa lalu dan mengembangkan model untuk memprediksi hasil yang konsisten dengan data *input*, sehingga meningkatkan akurasi prediksi yang diberikan [5].

Streamlit merupakan sebuah *framework open-source* yang bermanfaat untuk menciptakan aplikasi web yang dapat diakses menggunakan bahasa pemrograman Python. Dengan *framework* ini, pengembang dapat membangun aplikasi dengan mudah menggunakan skrip Python yang sederhana, tanpa perlu memiliki pengetahuan mendalam tentang pengembangan web di *frontend*. Penggunaan Streamlit memungkinkan pembuatan antarmuka pengguna (UI) responsif dengan cepat untuk menampilkan visualisasi data, tabel, grafik, dan elemen interaktif lainnya. Selain itu, Streamlit juga menyediakan fitur-fitur yang mempermudah penyebaran aplikasi sehingga dapat diakses secara daring dengan lancar oleh pengguna lain. Karena alasan tersebut, Streamlit telah menjadi pilihan populer di kalangan pengembang *data science* dan *machine learning* untuk mempercepat proses pembuatan dan distribusi aplikasi berbasis web [6].

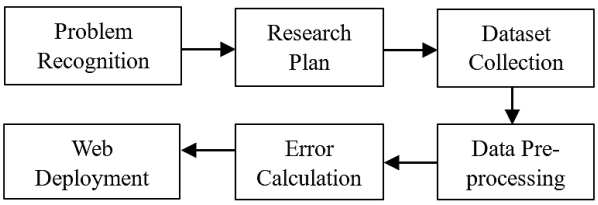
Sebuah studi sebelumnya dilakukan oleh Ernianti Hasibuan, dkk (2022), yang menggunakan regresi linear untuk memprediksi harga mobil dengan variabel numerik, mencapai akurasi model sebesar 76% [7]. Selain itu, Amalia (2021) juga mengadakan penelitian serupa untuk prediksi harga mobil dengan menggunakan model *Gradient Boost Regression* serta penyetelan *hyper-parameter* untuk mengubah data variabel teks menjadi nilai numerik, dan mendapatkan akurasi prediksi sebesar 67% [8]. Selanjutnya penelitian oleh Dea Miftahul Huda, dkk (2024) menghasilkan prediksi harga mobil bekas menggunakan regresi linear berganda dengan hasil *relative error* sebesar 11,89% yang menunjukkan hasil akurat [9].

Studi terdahulu lain oleh Sulaiman dan Edi Surya Negara (2023) menyatakan tentang prediksi harga mobil bekas menggunakan perbandingan algoritma *K-Nearest* dan *Random Forest* yang menunjukkan hasil dari *Random Forest* lebih baik yaitu 96,38% dibanding *K-Nearest* sebesar 59,17% [10]. Selanjutnya terdapat penelitian oleh Reynaldi, dkk (2021) yang menggunakan metode *fuzzy tsukamoto* dan *sugeno* dalam memprediksi harga mobil. Hasil yang didapat yaitu metode *fuzzy tsukamoto* mempunyai hasil yang lebih akurat yaitu tingkat error 8% [11].

Berdasarkan penelitian relevan sebelumnya, terdapat perbedaan akurasi untuk prediksi dimana tipe variabel yang digunakan hanya tipe data numerik, sedangkan untuk tipe data obyek tidak digunakan serta terdapat perubahan menjadi angka. Selain itu terdapat pula perbedaan algoritma yang digunakan. Penelitian ini akan mengikutsertakan tipe data obyek untuk dijadikan acuan prediksi harga mobil bekas. Oleh karena itu, penulis memutuskan untuk membuat program analisis prediksi harga mobil Audi bekas menggunakan *machine learning* dengan algoritma *linear regression* melalui *framework* streamlit.

METODE

Metode penelitian yang tepat merupakan pondasi utama dalam membangun keandalan dan validitas hasil suatu penelitian ilmiah. Penelitian ini menggunakan metode yang ditunjukkan oleh langkah-langkah pada Gambar 1.



Gambar 1. Langkah-langkah Penelitian

Problem Recognition

Pada langkah pertama ini ditemukan bahwa dalam industri mobil bekas, para penjual dan pembeli sering mengalami kendala dalam menentukan harga yang sesuai. Hal ini disebabkan oleh keterbatasan informasi yang jelas dan terintegrasi mengenai faktor-faktor yang mempengaruhi harga mobil bekas. Saat ini, meskipun informasi tersedia secara luas di *internet*, kebanyakan data yang ditemukan tersebar di berbagai sumber yang terpisah. Hal ini mengakibatkan proses pengambilan keputusan terasa rumit dan memakan banyak waktu.

Kendala utama yang dihadapi adalah ketidakpastian dalam penilaian harga mobil bekas karena informasi yang tidak terstruktur dan seringkali tidak akurat. Para penjual dan pembeli harus menyusun informasi dari berbagai sumber yang berbeda, seperti situs web jual-beli, forum diskusi otomotif, dan ulasan pengguna untuk mencoba memahami faktor-faktor yang mempengaruhi harga mobil bekas. Namun, karena ketidaktersediaan data yang lengkap dan terintegrasi, keputusan yang diambil tidak selalu optimal atau didasarkan pada informasi yang cukup.

Selain itu, proses pengumpulan informasi yang tidak terstruktur juga mengakibatkan pemborosan waktu yang signifikan. Penjual dan pembeli harus menghabiskan waktu mereka untuk mencari dan memverifikasi informasi dari berbagai sumber, yang pada akhirnya bisa saja tidak memberikan gambaran yang jelas atau akurat mengenai harga yang sebenarnya.

Dengan mengidentifikasi permasalahan ini, dapat dipahami bahwa diperlukan solusi yang menyediakan akses terpadu dan akurat terhadap informasi mengenai harga mobil bekas. Hal ini akan membantu meningkatkan kepercayaan dan efisiensi dalam proses jual-beli mobil bekas, serta mengurangi ketidakpastian yang mungkin dialami oleh para pelaku pasar.

Research Plan

Langkah selanjutnya yaitu melakukan rencana penelitian dengan menentukan metode yang tepat terkait prediksi yang dilakukan. Peneliti menggunakan metode *linear regression* atau regresi linier yang digunakan untuk menemukan korelasi linier antara variabel dan untuk memprediksi harga mobil bekas berdasarkan data historis yang tersedia [12]. Tujuan utama metode ini adalah untuk menemukan garis lurus terbaik yang dapat menjelaskan hubungan antara variabel *input* dan *output* dalam data.

Prosesnya dimulai dengan menghitung hubungan antara variabel *input* dan *output* menggunakan teknik seperti metode kuadrat terkecil. Ini melibatkan penghitungan koefisien yang tepat untuk variabel *input* yang dapat meminimalkan kesalahan antara nilai prediksi dan nilai aktual dari data yang diamati. Hasil dari *linear regression* dapat digunakan untuk memprediksi nilai *output* berdasarkan nilai *input* yang diberikan dengan rumus pada persamaan 1 [13].

$$Y = a + b1X1 + b2X2 + \cdots + bnX \quad (1)$$

Ket:

Y = variabel terikat

a = konstanta

$b \ldots n$ = koefisien regresi

$X \ldots n$ = variabel bebas

Dataset Collection

Langkah ini mencakup proses pengumpulan data atau himpunan data yang terstruktur dari berbagai sumber untuk keperluan pengembangan model. Sumber data yang akan digunakan berasal dari dataset publik yang tersedia di Kaggle dengan judul "100,000 UK Used Car Data set" oleh Aditya. Dataset ini disimpan dalam format *csv* yang terdiri dari 9 kolom dan 10.668 baris yang berisi informasi mengenai model mobil, tahun mobil, harga, transmisi, jarak tempuh, jenis bahan bakar, besaran pajak, konsumsi bahan bakar, dan ukuran mesin [14]. Adanya fitur-fitur kategorial dan kombinasi data akan menambah kompleksitas analisis, memungkinkan model untuk mempertimbangkan berbagai faktor yang berpengaruh terhadap harga mobil bekas secara menyeluruh. Berikut tampilan awal dari dataset yang ditampilkan pada Gambar 2.

	A	B	C	D	E	F	G	H	I
1	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
2	A1	2017	12500	Manual	15735	Petrol	150	55.4	1.4
3	A6	2016	16500	Automatic	36203	Diesel	20	64.2	2
4	A1	2016	11000	Manual	29946	Petrol	30	55.4	1.4
5	A4	2017	16800	Automatic	25952	Diesel	145	67.3	2
6	A3	2019	17300	Manual	1998	Petrol	145	49.6	1
7	A1	2016	13900	Automatic	32260	Petrol	30	58.9	1.4
8	A6	2016	13250	Automatic	76788	Diesel	30	61.4	2
9	A4	2016	11750	Manual	75185	Diesel	20	70.6	2
10	A3	2015	10200	Manual	46112	Petrol	20	60.1	1.4
11	A1	2016	12000	Manual	22451	Petrol	30	55.4	1.4
12	A3	2017	16100	Manual	28955	Petrol	145	58.9	1.4
13	A6	2016	16500	Automatic	52198	Diesel	125	57.6	2
14	Q3	2016	17000	Manual	44915	Diesel	145	52.3	2
15	A3	2017	16400	Manual	21695	Petrol	30	58.9	1.4

DatasetAudi

Gambar 2. Tampilan Awal Dataset

Data Pre-processing

Langkah ini merupakan langkah penting dalam *pipeline machine learning* yang melibatkan persiapan data untuk analisis lebih lanjut. Pertama, kita memisahkan fitur (X) dan target (y) dari dataset. Dalam kasus ini, target (y) adalah harga (price), sementara fitur (X) mencakup variabel *engine size*, *mpg*, *tax*, *fuel type*, *mileage*, *transmission*, *year*, dan *model*. Selanjutnya membagi data menjadi dua bagian yaitu data pelatihan dan data pengujian. Ini dapat dilakukan dengan menggunakan fungsi *train_test_split*, yang memungkinkan kita untuk memiliki dua set data terpisah yang digunakan untuk melatih model dan menguji performanya secara terpisah. Data pengujian sebesar 15% dan 85% merupakan data pelatihan yang terlampir pada Gambar 3.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.15, random_state=2)
```

Gambar 3. Pemisahan Set Data

Setelah pemisahan, fitur kategoris memerlukan transformasi menggunakan ColumnTransformer dan OneHotEncoder agar dapat dipahami oleh model yang dibuat pada step 1. Hal ini dengan tujuan menggabungkan prediksi tipe data numerik dengan objek agar dapat diprediksi bersamaan. Setelah itu dibentuk pelatihan model dengan algoritma *linear regression* menggunakan data pelatihan pada step 2. Proses ini dapat dilihat pada Gambar 4.

```
step1 = ColumnTransformer(transformers=[
    ('col_tnf',OneHotEncoder(sparse=False,drop='first',handle_unknown='ignore'),[0,2,4])
], remainder='passthrough')

step2 = LinearRegression()

pipe = Pipeline([
    ('step1', step1),
    ('step2', step2)
])

pipe.fit(X_train,y_train)

y_pred = pipe.predict(X_test)
```

Gambar 4. Transformasi Fitur dan Pelatihan Model

Hasil akhir didapatkan 9.067 data pelatihan dengan 8 kolom berisi variabel *model*, *year*, *transmission*, *mileage*, *fuelType*, *tax*, *mpg*, dan, *engineSize* yang akan digunakan dalam pembuatan model seperti yang dapat dilihat pada Gambar 5.

	model	year	transmission	mileage	fuelType	tax	mpg	engineSize
5900	Q5	2019	SemiAuto	5937	Diesel	145	38.2	2.0
2964	A4	2015	SemiAuto	20385	Diesel	125	56.5	2.0
5388	A4	2018	SemiAuto	31850	Petrol	145	36.7	3.0
5649	A4	2017	Automatic	38221	Petrol	145	38.7	3.0
4359	A1	2019	Manual	1308	Petrol	145	47.9	1.0
...	...	...	...	...	...	...	...	...
1099	A3	2019	Manual	4567	Petrol	145	44.8	1.5
2514	A4	2014	Manual	73206	Diesel	125	60.1	2.0
6637	A4	2018	Manual	16217	Petrol	145	51.4	1.4
2575	Q3	2016	Manual	21421	Petrol	145	49.6	1.4
7336	A1	2016	Manual	33500	Petrol	30	58.9	1.4

9067 rows × 8 columns

Gambar 5. Tampilan Akhir Data *Pre-processing*

Error Calculation

Penerapan algoritma *linear regression* untuk menghitung *error* atau kesalahan dengan membandingkan harga mobil yang diprediksi dengan harga sebenarnya. Untuk melihat hal tersebut digunakan metrik *R2 score* dan MAE (*Mean Absolute Error*) pada Gambar 6.

```
print('R2 score', r2_score(y_test,y_pred))
print('MAE', mean_absolute_error(y_test,y_pred))
```

Gambar 6. Metrik Kesalahan Prediksi

R2 Score atau dikenal sebagai koefisien determinasi, adalah sebuah metrik statistik yang umum digunakan dalam analisis regresi. Rentang nilai *R2 Score* adalah dari 0 hingga 1, yang menggambarkan seberapa baik model regresi sesuai dengan data yang dianalisis. Semakin tinggi nilai *R2 Score*, semakin baik kemampuan model regresi dalam menjelaskan variasi dari variabel dependen dengan menggunakan variabel independen yang tersedia. Oleh karena itu, *R2 Score* memiliki peran penting dalam mengevaluasi presisi prediksi dari model regresi, serta memberikan gambaran tentang seberapa besar variasi data yang dapat dijelaskan oleh model tersebut [15]. Rumus untuk *R2 Score* dapat dilihat pada persamaan 2.

$$R^2 = 1 - \frac{SS\,res}{SS\,tot} \quad (2)$$

Ket:
SS_{res} = jumlah kuadrat residu
SS_{tot} = jumlah kuadrat total

Sedangkan *Mean Absolute Error* (MAE) adalah teknik yang digunakan untuk menilai presisi model prediksi dengan menghitung rata-rata kesalahan absolut antara nilai prediksi dan nilai sebenarnya. Dengan menggunakan MAE, kita dapat menilai seberapa besar kesalahan yang terjadi dalam prediksi. Pendekatan ini bermanfaat untuk mengevaluasi akurasi model prediksi dalam unit yang sama dengan data aslinya. Dalam analisis data, MAE digunakan untuk mengukur kinerja model dalam memprediksi nilai aktual, serta membantu dalam memahami variasi nilai dalam kumpulan data untuk meningkatkan kualitas prediksi [16]. Rumus untuk *MAE* dapat dilihat pada persamaan 3.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| \quad (3)$$

Ket:
n = jumlah total data
yi = nilai aktual
$\hat{y}i$ = nilai yang diprediksi oleh model
$|yi - \hat{y}i|$= nilai absolut

Hasil yang didapat adalah nilai *R2 score* sebesar 0,94 atau 94% dan nilai MAE sebesar 0,08 atau 8%. Nilai *R2 score* 94% menunjukkan bahwa model *linear regression* yang digunakan sangat baik dalam menjelaskan variasi harga mobil bekas Audi berdasarkan variabel-variabel yang digunakan dalam analisis. Untuk nilai MAE, semakin rendah nilainya, maka semakin kecil kesalahan prediksi model. Untuk nilai MAE 8% menunjukkan bahwa kesalahan prediksi sangat kecil. Hasil perolehan terlampir pada gambar 7.

```
R2 score 0.9411715017762609
MAE 0.08811376199236715
```

Gambar 7. Hasil Perolehan *Error Calculation*

Web Deployment

Langkah selanjutnya adalah menyimpan model yang telah dilatih serta dataset yang telah dipersiapkan. Hal ini dapat dilakukan dengan menggunakan fungsi *pickle.dump*, yang memungkinkan untuk menyimpan objek Python ke dalam *file*. Dengan menyimpan model, kita dapat menggunakan kembali model tersebut di masa mendatang tanpa perlu melatih ulang. Sedangkan, menyimpan dataset dapat berguna untuk melacak versi data yang digunakan dalam pembangunan model, memungkinkan untuk replikasi hasil atau memperbarui model dengan data baru yang tersedia. Model yang digunakan terlampir pada Gambar 8.

```
import pickle

pickle.dump(df,open('df.pkl','wb'))
pickle.dump(pipe,open('pipe.pkl','wb'))
```

Gambar 8. Penyimpanan Model

Dalam proses *web deployment* untuk aplikasi prediksi harga mobil Audi bekas ini menggunakan Streamlit yaitu sebuah *framework* Python yang memudahkan pembuatan antarmuka pengguna web

interaktif. Aplikasi ini dimulai dengan memuat model prediksi dan data yang telah disimpan menggunakan *pickle*. Setelah itu, *interface* pengguna ditampilkan dengan judul sambutan dan gambar mobil Audi untuk memberikan konteks visual. Sebuah paragraf deskriptif mengenai merek Audi diintegrasikan dengan teks yang dibungkus secara rapi agar mudah dibaca. Bagian awal untuk model prediksi ditampilkan pada Gambar 9.

```
import streamlit as st
import pickle
import numpy as np
import pandas as pd
import textwrap

pipe = pickle.load(open('pipe.pkl','rb'))
df = pickle.load(open('df.pkl','rb'))

st.title('Selamat Datang!')
st.image("Audiimage.jpg", width=400)

text = """
Audi adalah merek mobil yang berasal dari Jerman, dikenal karena kombinasi antara desain elegan, teknologi inovatif, dan performa yang canggih.
"""

wrapped_text = textwrap.fill(text, width=70)

st.markdown(f"<div style='text-align: justify'>{wrapped_text}</div>", unsafe_allow_html=True)

st.title("Prediksi Harga Mobil Audi Bekas")
```

Gambar 9. Model Prediksi

Bagian inti dari aplikasi ini adalah fitur input di mana pengguna dapat memilih berbagai parameter mobil, seperti model, tahun, transmisi, jenis bahan bakar, jarak tempuh, pajak, konsumsi bahan bakar, dan ukuran mesin. Pengguna memasukkan informasi ini melalui *select box* dan *number input*. Setelah data dimasukkan, aplikasi menggunakan model prediksi untuk menghitung dan menampilkan estimasi harga mobil bekas Audi dalam bentuk yang mudah dibaca, dengan hasil akhir dikonversi ke dalam mata uang Rupiah. Model akhir ditampilkan pada Gambar 10.

```
model = st.selectbox('Input Model Mobil', df['model'].unique())
year = st.selectbox('Input Tahun Mobil', df['year'].unique())
transmission = st.selectbox('Input Transmisi Mobil', df['transmission'].unique())
fuelType = st.selectbox('Input Jenis Bahan Bakar Mobil', df['fuelType'].unique())
mileage = st.number_input('Input Jarak Tempuh Mobil')
tax = st.number_input('Input Pajak Mobil')
mpg = st.number_input('Input Konsumsi Bahan Bakar Mobil')
engineSize = st.number_input('Input Ukuran Mesin Mobil')

if st.button('Prediksi Harga Mobil'):
    query = pd.DataFrame({
        'model': [model],
        'year': [year],
        'transmission': [transmission],
        'fuelType': [fuelType],
        'mileage': [mileage],
        'tax': [tax],
        'mpg': [mpg],
        'engineSize': [engineSize]
    })

    st.write(f"<p style='font-size: 30px; font-weight: bold;'>Prediksi Harga Mobil Audi Bekas Dalam Rupiah yaitu Rp
             {int(np.exp(pipe.predict(query)[0]) * 20000)}</p>", unsafe_allow_html=True)
```

Gambar 10. Fitur Model Prediksi

Untuk *deploy* lanjutan aplikasi prediksi harga mobil Audi bekas ke Streamlit Share, langkah pertama adalah memindahkan semua file yang diperlukan ke *repository* GitHub. *Repository* ini akan berisi *file* Python yang mendefinisikan logika aplikasi, model yang disimpan dalam format *pickle*, serta file data yang digunakan untuk pelatihan dan prediksi. Setelah *repository* GitHub siap dan semua *file* terunggah dengan benar, langkah berikutnya adalah menyambungkan *repository* tersebut ke Streamlit Share.

Dalam Streamlit Share, pengguna dapat melakukan integrasi dengan GitHub dengan memasukkan URL *repository* dan melakukan konfigurasi yang diperlukan untuk memastikan aplikasi dapat berjalan dengan baik di *platform* tersebut. Setelah proses ini selesai, aplikasi akan tersedia secara publik, memungkinkan pengguna untuk mengakses dan menggunakan sistem prediksi harga mobil Audi bekas secara *online* melalui antarmuka *web* yang interaktif. Tampilan GitHub yang digunakan dapat dilihat pada Gambar 11.

Gambar 11. Tampilan GitHub

HASIL DAN PEMBAHASAN

Pengujian Korelasi

Uji korelasi dilakukan untuk mengukur dan memvisualisasikan hubungan linier antara variabel-variabel dalam *dataset*. Nilai korelasi berkisar dari -1 hingga 1, di mana -1 menunjukkan korelasi negatif sempurna, 1 menunjukkan korelasi positif sempurna, dan 0 berarti tidak ada korelasi. Dengan menggunakan *sns.heatmap()* dari Seaborn, matriks korelasi ini divisualisasikan dalam bentuk *heatmap* yang mudah dibaca. Setiap sel pada *heatmap* menunjukkan nilai korelasi antara dua variabel, dengan warna yang mewakili kekuatan korelasi tersebut. Warna yang lebih gelap atau lebih cerah menunjukkan nilai korelasi yang lebih ekstrem, sedangkan warna yang lebih netral menunjukkan korelasi yang lebih lemah. *Heatmap* korelasi antara variabel ditunjukkan pada Gambar 12.

Gambar 12. *Heatmap* Uji Korelasi

Dapat dilihat bahwa variabel *year* (tahun) yaitu 0.59 menunjukkan korelasi positif yang cukup signifikan antara tahun mobil dan harga. Ini berarti bahwa mobil yang lebih baru cenderung memiliki harga yang lebih tinggi. Korelasi ini dikategorikan sebagai korelasi positif sedang hingga kuat.

Untuk variabel *mileage* (jarak tempuh) yaitu -0.54 menunjukkan korelasi negatif yang moderat antara jarak tempuh dan harga. Ini berarti bahwa mobil dengan jarak tempuh yang lebih tinggi biasanya memiliki harga yang lebih rendah. Korelasi ini dianggap sebagai korelasi negatif sedang.

Selanjutnya variabel *tax* (pajak) yaitu 0.36 menunjukkan korelasi positif yang relatif lemah antara pajak mobil dan harga. Artinya, ada hubungan bahwa pajak yang lebih tinggi akan sedikit berkaitan dengan harga yang lebih tinggi, tetapi pengaruhnya tidak terlalu besar.

Berikutnya variabel *mpg* (konsumsi bahan bakar) yaitu -0.60 menunjukkan korelasi negatif yang kuat antara konsumsi bahan bakar dan harga. Ini berarti bahwa mobil dengan konsumsi bahan bakar yang lebih rendah (lebih efisien) cenderung memiliki harga yang lebih tinggi.

Terakhir untuk variabel *engine size* (ukuran mesin) yaitu 0.59 menunjukkan korelasi positif yang signifikan antara ukuran mesin dan harga. Artinya, mobil dengan ukuran mesin yang lebih besar cenderung memiliki harga yang lebih tinggi. Korelasi ini juga dikategorikan sebagai korelasi positif sedang hingga kuat.

Pengujian *Web Deployment* dan Hasil Prediksi

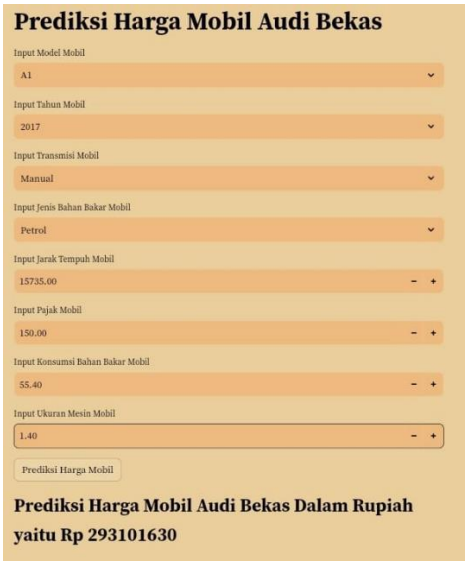
Pengujian *web deployment* untuk aplikasi prediksi mobil menunjukkan bahwa aplikasi telah berhasil di-*deploy* pada Streamlit Share dan berfungsi sesuai dengan tujuan yang diharapkan. Tujuannya adalah supaya tidak ditemukan *error* kritis selama pengujian, dan aplikasi siap untuk digunakan oleh pengguna akhir. Digunakan metode *functional testing* untuk melakukan uji tahap ini. Metode *functional testing* yaitu metode yang memastikan bahwa semua fitur dan fungsi aplikasi bekerja sesuai dengan spesifikasi dan kebutuhan pengguna.

Pengujian ini dimulai dengan identifikasi dan verifikasi seluruh fitur utama aplikasi, seperti kolom *input* data dengan *number input*, memilih data dengan *select box*, dan tampilan hasil prediksi. Pengujian fungsional dimulai dengan menguji kemampuan aplikasi dalam menerima dan memproses berbagai jenis *input* dari pengguna, seperti memasukkan angka tahun jarak tempuh, pajak, konsumsi bahan bakar, dan ukuran mesin. Serta dalam memilih data, seperti kolom model mobil, tahun, transmisi, dan jenis bahan bakar. Setiap kombinasi *input* diuji untuk memastikan bahwa aplikasi menghasilkan *output* yang akurat dan relevan sesuai dengan data yang dimasukkan.

Selain itu, pengujian juga meliputi verifikasi alur kerja aplikasi, mulai dari halaman awal hingga tampilan hasil prediksi, untuk memastikan bahwa tidak ada kesalahan navigasi atau fungsi yang terganggu. Pengujian ini bertujuan untuk memastikan bahwa semua fungsi aplikasi berjalan sesuai dengan desain dan spesifikasi yang diharapkan, serta memberikan pengalaman pengguna yang lancar dan bebas dari kesalahan. Pada Gambar 13 dilampirkan tampilan awal aplikasi yang sudah berhasil di-*running*.

Gambar 13. Tampilan Awal Aplikasi

Hasil yang didapat adalah tidak ditemukan *error* pada seluruh rangkaian pada tampilan *web* aplikasi prediksi. Pengguna dapat menggunakan fitur-fitur *input* dengan baik. Tombol "Prediksi Harga Mobil" juga responsif dan sesuai dengan hasil yang sudah dirancang. Model memberikan hasil prediksi yang cocok dengan nilai *input* yang dimasukkan oleh pengguna. Keberhasilan pengujian ini dapat dilihat pada Gambar 14. Aplikasi prediksi harga mobil Audi ini dapat diakses oleh pengguna pada *link* berikut https://app-projectabcd-2sqk9zqhdahaah26a7ymyz.streamlit.app/.



Gambar 14. Tampilan Aplikasi Prediksi

KESIMPULAN DAN SARAN

Program ini berhasil mengintegrasikan berbagai tahapan dalam pengembangan aplikasi prediksi harga mobil bekas Audi, mulai dari pengumpulan data, analisis, pemrosesan data, pembangunan model, evaluasi, hingga *deployment* aplikasi *web*. *Web* aplikasi juga dapat diakses oleh pengguna umum untuk memudahkan dalam mencari prediksi harga mobil Audi bekas.

Hasil penelitian menunjukkan nilai *R2 score* sebesar 0,94 atau 94% dan nilai MAE sebesar 0,08 atau 8%. Hal ini menunjukkan bahwa model *linear regression*

yang digunakan dalam program memiliki tingkat keakuratan yang tinggi. *R2 score* mendekati 1 artinya berhasil menjelaskan variasi dalam data dengan baik, sementara nilai MAE yang rendah menunjukkan bahwa prediksi harga mobil bekas Audi cenderung memiliki kesalahan yang minim dalam rata-rata nilainya.

Saran penelitian adalah ide atau rekomendasi tentang topik dan arah penelitian yang dapat dieksplorasi lebih lanjut oleh peneliti. Rekomendasi ini muncul setelah meninjau literatur yang ada dan mengidentifikasi celah pengetahuan atau aspek yang belum tercakup dengan baik. Penelitian ini hanya menggunakan MAE dan *R2 Score* untuk menghitung prediksi, oleh sebab itu penelitian selanjutnya diharapkan menambah komponen prediksi lain untuk memaksimalkan hasil dari prediksi agar lebih akurat.

DAFTAR PUSTAKA

[1] T. Muhayat, Jayanta, and N. Chamidah, "Prediksi Harga Smartphone Menggunakan Algoritma Multiple Linear Regression," *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, Vol. 296, pp. 506-525, 2022.

[2] M. A. Saputra, Martanto, and U. Hayati, "Estimasi Harga Mobil Bekas Toyota Yaris Menggunakan Algoritma Regresi Linier," *JATI (Jurnal Mahasiswa Teknik Informatika)*, Vol. 8, No. 2, pp. 1696-1701, 2024.

[3] Tempus, "Why Audi is and always has been a top-quality manufacturer", *Vantage Media Group*, 2021, [Online]. Available: https://tempusmagazine.co.uk/news/why-audi-is-and-always-has-been-a-top-quality-manufacturer/. [Diakses: 24 Juni 2024].

[4] I. N. Simbolon, H. D. S. J. Siburian, and W. A. Manik, "Prediksi Kualitas Air Sungai di Jakarta Menggunakan KNN yang Dioptimalisasi Dengan PSO". *Jurnal Informatika dan Teknik Elektro Terapan*, Vol.12, No. 2, pp. 1193-1203, 2024.

[5] R. G. Wardhana, G. Wang, and F. Sibuea, "Penerapan Machine Learning dalam Prediksi Tingkat Kasus Penyakit di Indonesia," *Journal of Information System Management (JOISM)*, Vol. 5, No. 1, pp. 40-4, 2023.

[6] G. A. Syafarina and Zaenuddin, "Implementasi Framework Streamlit Sebagai Prediksi Harga Jual Rumah Dengan Linear Regresi," *Metik Jurnal*, Vol. 7, No. 2, pp. 121-125, 2023.

[7] E. Hasibuan and A. Karim, "Implementasi Machine Learning untuk Prediksi Harga Mobil Bekas dengan Algoritma Regresi Linear berbasis Web," *Jurnal Ilmiah Komputasi*, Vol. 21, No. 4, pp. 595-602, 2022.

[8] A. Amalia, M. Radhi, S. H. Sinurat, D. R. H. Sitompul, and E. Indra, "Prediksi Harga Mobil Menggunakan Algoritma Regressi dengan Hyper-Parameter Tuning," *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, Vol. 4, No. 2, pp. 28-32, 2021.

[9] D. M. Huda, G. Dwilestari, A. R. Rinaldi, and I. Solihin, "Prediksi Harga Mobil Bekas Menggunakan Algoritma Regresi Linear Berganda", *Jurnal Informatika dan Rekayasa Perangkat Lunak*, Vol. 6, No. 1, pp. 150-157, 2024.

[10] Sulaiman and E. S. Negara, "Komparasi Algoritma K-Nearest Neighbors dan Random Forest Pada Prediksi Harga Mobil Bekas", *Jurnal Penelitian Ilmu dan Teknologi Komputer (JUPITER)*, Vol. 15, No. 1, pp. 337-346, 2023.

[11] Reynaldi, W. Syafrizal, and M. F. A. Hakim, "Analisis Perbandingan Akurasi Metode Fuzzy Tsukamoto dan Fuzzy Sugeno Dalam Prediksi Penentuan Harga Mobil Bekas", *Indonesian Journal of Mathematics and Natural Sciences*, Vol. 44, No. 2, pp. 73-80, 2021.

[12] N. Narwanto, Kusrini, and H. A. Fatta, "Prediksi Peserta Matakuliah Menggunakan Artificial Neural Network –Fuzzy Inferented System(Ann-Fis) Studi Kasus: Universitas Muhammadiyah Surakarta," *Journal of Technology and Informatic (JoTI)*, Vol. 1, No. 2, pp. 84-88, 2020.

[13] A. D. Siburian, D. R. H. Sitompul, S. H. Sinurat, A. Situmorang, Ruben, D. J. Ziegel, and E. Indra, "Laptop Price Prediction with Machine Learning Using Regression Algorithm," *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, Vol. 6, No. 1, pp. 87-91, 2022.

[14] Aditya, "100,000 UK Used Car Data set", *Kaggle*, 2020, [Online]. Available: https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes. [Diakses: 21 Maret 2024].

[15] D. A. Rhamadhani and E. E. D. Saputri, "Analisa Model Machine Learning dalam Memprediksi Laju Produksi Sumur Migas 15/9-F-14H," *Journal of Sustainable Energy Development*, Vol. 1, No. 1, pp. 48-55, 2023.

[16] N. L Azizah, N. Suarna, W. Prihartono, "Prediksi Tingkat Kemiskinan di Provinsi Jawa Barat Menggunakan Algoritma Regresi Linear," *JATI (Jurnal Mahasiswa Teknik Informatika)*, Vol. 7, No. 6, pp. 3377-3381, 2023.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, Bulan Oktober Tahun 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.654

Perancangan E-*Commerce* Distributor *Clothing Company* Berdasarkan *Product Knowledge*

Nisa Eridiar[1], Tan Amelia[2], Ayouvi Poerna Wardhanie[3]
[1,2,3]Sistem Informasi, Universitas Dinamika, Surabaya, Indonesia
e-mail: 18410100191@dinamika.ac.id[1], meli@dinamika.ac.id[2], ayouvi@dinamika.ac.id[3]
* Penulis Korespondensi: E-mail: ayouvi@dinamika.ac.id

Abstrak: E-commerce merupakan salah satu inovasi pemanfaatan internet yang dapat membantu dalam proses jual beli suatu bisnis. Pentingnya pemanfaatan teknologi informasi dalam dunia yang serba digital saat ini, harus didukung pula oleh ketersediaan informasi yang akurat dan sesuai dengan kebutuhan konsumennya. Sebelum melakukan pembelian, biasanya konsumen akan mencari informasi sebanyak-banyaknya terkait produk atau jasa yang akan dibeli sehingga adanya product knowlegde dalam sebuah aplikasi khususnya e-commerce adalah hal yang sangat penting. Oleh sebab itu, penelitian ini bertujuan untuk merancang sebuah e-commerce yang menekankan pada ketersediaan secara lengkap product knowledge yang dibutuhkan konsumen. Adapun metode yang digunakan untuk merancang e-commerce ini menggunakan System Development Life Cycle (SDLC) model Waterfall. Proses SDLC model air terjun ini dipilih karena memiliki proses berurutan dan membuat pengerjaan terjadwal dengan baik. Website diuji menggunakan User Acceptance Testing (UAT) yang dilakukan dengan pemilik dan sales selaku pengelola Toko Bismillah Distributor, serta data pendukung seperti wawancara dan kuesioner kepada konsumen yang masing-masing membutuhkan 7 orang dan 30 sampel. Hasil implementasi variabel product knowledge yang terdapat pada fitur-fitur di dalam e-commerce Toko Bismillah Distributor antara lain variabel mutu produk pada halaman profil berupa deskripsi keunggulan produk, variabel kandungan produk pada detail produk (ukuran, warna, dan bahan produk) dan variabel penyampaian informasi pada halaman kontak, halaman cara belanja, halaman profil, ulasan dan rating. Pengujian UAT menunjukkan website e-commerce dapat digunakan dengan baik.

Kata Kunci: E-*Commerce*; *Product Knowledge*; SDLC Model *Waterfall*; UAT; *Website*

Abstract: *E-commerce is an innovation in using the internet that can help in the buying and selling process of a business. The importance of utilizing information technology in today's digital world must also be supported by the availability of information that is accurate and in line with consumer needs. Before making a purchase, consumers will usually look for as much information as possible regarding the product or service to be purchased, so having product knowledge in an application, especially e-commerce, is very important. Therefore, this research aims to designing an e-commerce platform that emphasizes the complete availability of product knowledge that consumers need. The chosen method for designing the e-commerce website is the System Development Life Cycle (SDLC) model, specifically the waterfall model. The waterfall model allows for well-scheduled development. The website is tested using User Acceptance Testing (UAT), conducted with the owner and sales representatives of Bismillah Distributor, along with supporting data gathered through interviews and questionnaires from the company's consumers. The implementation results in product knowledge variables integrated into features within the e-commerce website for Bismillah Distributor Baby Store. These variables include product quality on the profile page, featuring a description of product excellence, product content variables on the product details page (size, color, and material), and information delivery variables on the contact page, shopping guide page, profile page, reviews, and ratings. UAT results demonstrate that the e-commerce website can be effectively utilized, meeting the needs and expectations of Bismillah Distributor's owner and sales team, as well as providing essential information to consumers.*

Keywords: *E-Commerce; Product Knowledge; SDLC Model Waterfalls; UAT; Website*

PENDAHULUAN

Toko Bismillah Distributor adalah usaha mikro, kecil, dan menengah yang didirikan oleh Bapak Rossi Follan pada tahun 2020 di Depok, Jawa Barat. Usaha ini fokus pada produksi perlengkapan bayi, dengan sebagian besar produknya diproduksi melalui kerja sama dengan perusahaan konveksi. Model penjualan Toko Bismillah Distributor sebelumnya melibatkan interaksi langsung dengan pelanggan, di mana *sales* datang langsung ke konsumen untuk memperlihatkan dan membandingkan kualitas produk dengan produk sejenis. Namun, dampak pandemi Covid-19 telah memberikan tantangan besar bagi Toko Bismillah Distributor. Pembatasan aktivitas manusia selama pandemi menyebabkan *sales* tidak dapat melakukan penjualan *door to door* seperti sebelumnya. Akibatnya, penjualan Toko Bismillah Distributor mengalami penurunan yang signifikan. Penyebab lainnya karena tidak mengenal produk dari Toko Bismillah Distributor, konsumen baru yang berada di luar lokasi Toko Bismillah Distributor menjadi ragu untuk memesan barang. Apabila masih ada perusahaan yang melakukan sistem penjualannya secara tertulis dan

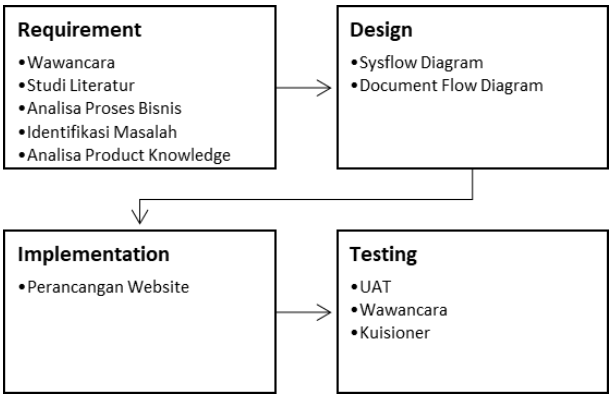
manual, maka tidak jarang akan rentan terjadi banyaknya kesalahan. Oleh karena itu, diharapkan dengan adanya layanan jasa berupa *e-commerce*, produk-produk dari suatu perusahaan dapat dinikmati secara luas oleh pelanggan secara cepat[1]

Pada permasalahan yang telah diidentifikasi sebelumnya, penelitian ini memberikan solusi yaitu membuat *website e-commerce* khusus untuk Toko Bismillah Distributor, yang berfokus pada *product knowledge* konsumen agar sesuai dengan kebutuhan konsumennya. *Website e-commerce* ini memiliki tujuan utama untuk mempermudah proses pemesanan produk secara *online* bagi masyarakat [2]. Fitur yang ditambahkan pada *website e-commerce* akan disesuaikan dengan *product knowledge*, antara lain, foto produk, ulasan produk, serta rincian atau informasi produk yang terkait dengan bahan, ukuran, berat, dan warna [3]. Informasi produk yang disajikan secara rinci dan komprehensif bertujuan untuk memberikan pemahaman yang lebih baik kepada konsumen mengenai kualitas, kegunaan, dan manfaat dari produk yang ditawarkan [4].

Dalam merancang sebuah *website e-commerce* untuk Toko Bismillah Distributor, pemanfaatan *product knowledge* menjadi fokus utama. *Product knowledge* penting dalam merancang *website* karena memahami produk dari perspektif konsumen dan penjual meningkatkan kualitas pengalaman pengguna [5]. Dengan informasi produk yang komprehensif, *website* dapat memberikan pemahaman yang baik kepada konsumen tentang keunggulan dan manfaat produk [6].

METODE

Dalam melakukan rancang bangun *website* Toko Bismillah Distributor digunakan metode SDLC *(System Development Life Cycle)* model *waterfall*. Metode *waterfall* sangat tepat digunakan pada pengembangan sistem jangka pendek, sehingga banyak perusahaan yang menggunakan metode ini dalam mengembangkan sistem yang perusahaan punya, ataupun memperbaiki sistem yang sudah ada sesuai dengan kebutuhan perusahaan masing-masing [7]. Selain itu, keuntungan dari model *waterfall* adalah dokumentasi yang lengkap dan detail di setiap tahapannya sehingga semua kebutuhan dapat dipenuhi secara menyeluruh [8].



Gambar 1. Model *Waterfall*

Requirement

Proses dalam analisis kebutuhan merupakan serangkaian langkah untuk mengumpulkan informasi terkait perusahaan. Langkah-langkah tersebut melibatkan wawancara, studi literatur, analisis proses bisnis, identifikasi masalah, paham mengenai produk atau layanan, serta analisis kebutuhan, termasuk dalamnya konsep *Input Process Output* (IPO).

Design

Langkah desain adalah fase perencanaan yang dijalankan sebelum pembuatan situs *e-commerce* untuk Toko Bismillah Distributor, yang mencakup perancangan sistem. Perancangan sistem dalam penelitian ini melibatkan penggunaan diagram alur sistem, model data konseptual, dan model data fisik. Diagram alur sistem digunakan untuk mengidentifikasi aliran data di dalam situs web yang sedang dibangun. Sementara itu, model data konseptual dan model data fisik digunakan untuk merancang struktur basis data yang akan digunakan.

Implementation

Implementasi adalah fase di mana *website e-commerce* dibangun untuk mewujudkan desain yang telah direncanakan sebelumnya. Setelah selesai membangun *website*, langkah selanjutnya adalah melakukan pengujian UAT (*User Acceptance Test*) yang ditujukan kepada pemilik dan tim penjualan. Tujuan dari pengujian ini adalah untuk memastikan bahwa fungsi dan tugas *website* sesuai dengan kebutuhan pengguna. UAT memberikan kesempatan kepada pemilik dan tim penjualan untuk menguji secara langsung apakah *website* dapat memenuhi ekspektasi dan kebutuhan mereka sebelum diluncurkan secara resmi.

Testing

Testing merupakan tahap akhir dalam penelitian ini, yang melibatkan pengumpulan data dari *website* melalui *User Acceptance Testing* (UAT) kepada pemilik dan tim penjualan Toko Bismillah Distributor dengan menggunakan wawancara. Pengujian UAT dilakukan setelah sistem selesai dibangun. Pengujian ini berguna untuk melihat sejauh mana sistem yang dibangun dapat berjalan dengan baik serta sesuai dengan kebutuhan *user* [9].

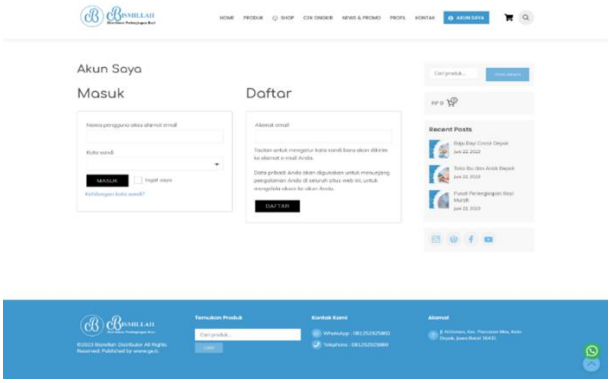
Penelitian ini juga melibatkan wawancara dengan 7 konsumen Toko Bismillah Distributor. Selain itu, data pendukung juga diperoleh melalui penyebaran kuesioner kepada 30 responden sebagai bagian dari tahapan pengumpulan informasi. Keseluruhan proses ini dirancang untuk mendapatkan pandangan dan umpan balik yang komprehensif dari pemilik, tim penjualan, dan konsumen terkait performa dan keefektifan *website e-commerce* yang telah dikembangkan.

HASIL DAN PEMBAHASAN

Variabel Informasi

Analisis variabel informasi pada *product knowledge* pada *website* Toko Bismillah Distributor adalah suatu proses evaluatif yang bertujuan untuk memahami, meningkatkan, dan mengoptimalkan informasi yang disajikan kepada pengguna tentang produk atau layanan tertentu di platform *e-commerce*.

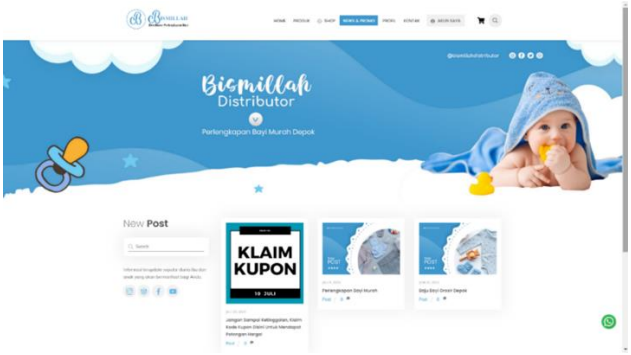
Halaman *Login* dan Registrasi



Gambar 2. Tampilan Halaman *Login* dan Registrasi

Halaman *login* berfungsi sebagai mekanisme pengenalan pengguna yang telah terdaftar, meminta mereka untuk memasukkan nama pengguna dan kata sandi yang telah didaftarkan sebelumnya. Sementara itu, halaman registrasi menyediakan cara bagi pengguna baru untuk mendaftar, dimulai dengan memasukkan alamat email mereka. Setelah itu, pengguna dapat masuk dan mengatur kata sandi serta informasi lain yang diperlukan.

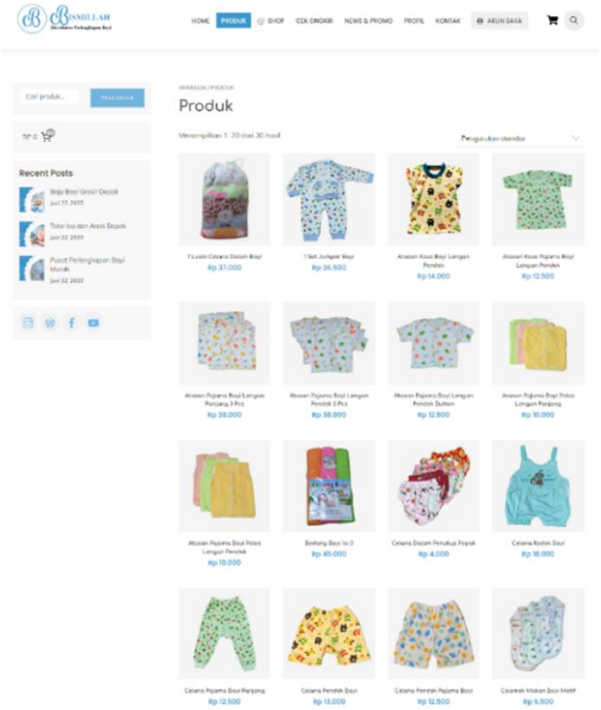
Halaman *News* & Promo



Gambar 3. *News* & Promo

Fitur yang ada pada tampilan tersebut mencerminkan penerapan variabel *Product knowledge*, yang melibatkan penyampaian informasi mengenai pemahaman mengenai promo dan berita terbaru terkait *website*[10]. Hal ini terlihat pada halaman *news* & promo, di mana konsumen dapat mengakses informasi terkait kode promo yang ditawarkan oleh Toko Bismillah Distributor untuk di klaim.

Katalog Produk



Gambar 4. Tampilan Halaman Katalog Produk

Halaman katalog adalah tempat di mana seluruh produk yang ditawarkan oleh Toko Bismillah Distributor dipresentasikan secara menyeluruh. Pada halaman ini, terdapat kategori-kategori produk yang berfungsi sebagai filter untuk mempermudah pencarian, seperti atasan baju, kaos kaki, dan sebagainya. Desain ini mencerminkan implementasi dari variabel *product knowledge*, yang melibatkan penyampaian informasi berupa nama produk dan harga produk [11]. Informasi yang disediakan, seperti nama produk dan harga bertujuan untuk menyajikan informasi yang akurat dan terkini, termasuk potensi diskon atau promosi yang dapat memengaruhi keputusan pembelian.

Variabel Mutu

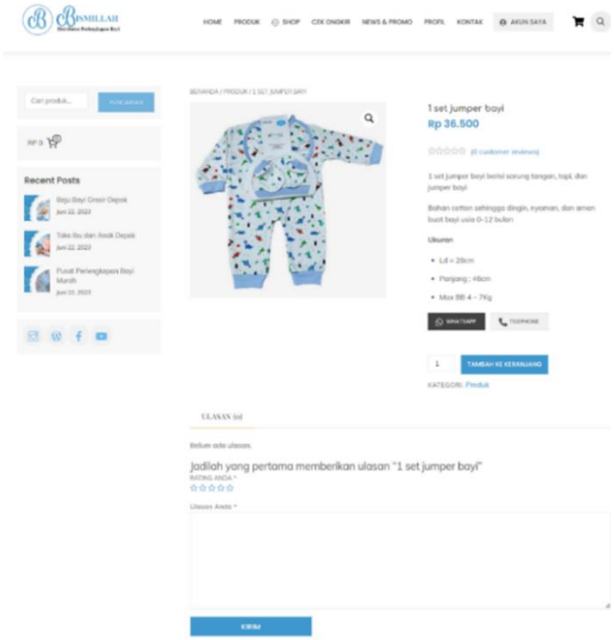
Variabel mutu merupakan variabel yang menjelaskan terkait kualitas produk suatu usaha bisnis, di mana faktor-faktor yang mencakup variabel ini adalah kandungan produk. Menurut Kotler & Kaller dalam Tatang et al (2017) menyatakan para pelanggan menghendaki sebuah laman *website* yang mempunyai mutu tinggi apabila para pelanggan ingin berbelanja *online* [12]. Adapun bentuk implementasi dari variabel mutu ke dalam *website e-commerce* Toko Bismillah Distributor, antara lain:

Gambar 5. Tampilan Halaman Profil

Tampilan ini berfungsi sebagai tempat untuk menyampaikan semua informasi terkait Toko Bismillah Distributor. Implementasi ini mencerminkan *product knowledge* yaitu variabel mutu, di mana halaman tersebut menjelaskan produk yang ditawarkan dan keunggulan-keunggulannya [13].

Variabel Kandungan

Variabel kandungan merupakan variabel yang menjelaskan terkait bahan terbuatnya produk, di mana faktor-faktor yang mencakup variabel ini adalah deskripsi yang tertera pada produk. Adapun bentuk implementasi dari variabel mutu ke dalam *website e-commerce* Toko Bismillah Distributor adalah:



Gambar 6. Tampilan Halaman Detail Produk

Halaman ini muncul ketika pengguna menginginkan informasi rinci tentang suatu produk. Pengguna dapat memilih salah satu produk dan mengakses halaman detail produk untuk menampilkan deskripsi yang mencakup aspek-aspek seperti ukuran, bahan, dan warna pada produk tersebut. Implementasi ini adalah hasil dari penerapan variabel *product knowledge*, yang mengarah pada kandungan produk yang berupa komposisi bahan [14].

Testing

Tahap *testing* digunakan untuk menguji sejauh mana *website e-commerce* yang telah dibuat dapat memenuhi kebutuhan dan berfungsi sesuai harapan pengguna. Metode pengujian yang diterapkan dalam penelitian ini adalah *User Acceptance Testing* (UAT). Proses UAT pada *website e-commerce* dilibatkan pengguna utama, termasuk pemilik dan tim penjualan dari Toko Bismillah Distributor. Pengujian ini bertujuan untuk menilai apakah *website* berjalan sesuai kebutuhan dan apakah kinerjanya memuaskan. Hasil pengujian dengan pemilik dan tim penjualan yang bertanggung jawab atas halaman admin menunjukkan bahwa fungsi pengelolaan kupon diskon, manajemen berita, dan manajemen laporan penjualan dapat diterima dengan baik. Selanjutnya, hasil pengujian dengan pemilik dan tim penjualan untuk mengevaluasi tampilan utama *website*, termasuk halaman beranda, produk, detail produk, berita & promo, keranjang, proses *checkout*, panduan belanja, profil, halaman kontak, ulasan, dan peringkat menunjukkan bahwa *website* memenuhi harapan dan dapat diterima dengan baik oleh pengguna.

Tabel 1. Hasil Uji UAT

Deskripsi & Prosedur Pengujian	***Input***	***Output* yang Diharapkan**	**Diterima**
Login			
Pengguna dapat registrasi di *website*	*Input email*	Pengguna mendapatkan hak autentikasi	✔
Pengguna berhasil *login* di *website*	*Input username* dan *password*	Pengguna mendapatkan hak akses berbelanja	✔
Halaman Produk			
Pengguna dapat klik produk		Memunculkan tampilan detail produk	✔
Pengguna dapat memasukkan kuantitas produk	*Input* kuantitas	Kuantitas produk yang masuk ke dalam *cart* belanja sesuai	✔
Pengguna dapat memasukkan produk ke		Produk masuk ke dalam *cart* belanja	✔

Deskripsi & Prosedur Pengujian	*Input*	*Output* yang Diharapkan	Diterima
dalam *cart* belanja			
Pengguna dapat mencari produk menggunakan fitur pencarian	*Input* nama produk	Memunculkan produk yang dicari	✔
Pengguna dapat klik *icon cart* belanja		Berganti ke halaman *cart* belanja	✔
Halaman *News* & Promo			
Pengguna klik *news* dan salin kode promo		Berganti ke halaman *news* yang di klik	✔
Halaman *Check Out*			
Pengguna dapat memasukkan kode kupon	*Input* kode kupon	Ongkos kirim otomatis terpotong	✔
Pengguna dapat klik tombol *checkout*		Berganti ke halaman pengisian alamat	✔
Pengguna mengisi alamat	*Form* alamat	Muncul ongkos kirim sesuai daerah yang di *input*	✔
Pengguna memilih jenis pengiriman			✔
Halaman Ulasan & *Rating*			
Pengguna mengisi *input*an berupa deskripsi	*Form* ulasan & *rating*		✔
Pengguna klik tombol kirim		Muncul ulasan yang telah diisi	✔
Halaman Cara Belanja			
Pengguna klik menu cara belanja		Berganti ke halaman cara belanja	✔
Halaman Profil			
Pengguna klik menu profil		Berganti ke halaman profil	✔
Halaman Kontak			
Pengguna mengisi *input*an berupa deskripsi	*Form* pesan		✔
Pengguna klik tombol kirim		Pesan akan dikirim ke halaman admin	✔
Halaman Produk – Admin			
Pengguna klik tambah baru		Berganti ke halaman tambah produk	✔
Pengguna mengisi *form* produk	*Form* produk		✔
Pengguna klik tombol simpan		Produk yang disimpan tampil di daftar produk	✔
Pengguna klik tombol sunting		Berganti ke halaman sunting produk	✔
Pengguna mengubah isi *form* produk	*Form* produk	Detail produk berubah	✔
Halaman Laporan Penjualan – Admin			
Pengguna menentukan waktu transaksi	*Input* waktu	Muncul transaksi yang sesuai dengan waktu yang ditentukan	✔

Hasil wawancara dengan pemilik dan tim penjualan sebagai pengelola menunjukkan bahwa *website e-commerce* mendapat tanggapan positif, meskipun mereka menyatakan bahwa belum siap untuk mengimplementasikannya dalam operasional perusahaan. Wawancara dengan konsumen menunjukkan bahwa sebagian besar merespons positif terhadap *website e-commerce* tersebut, namun beberapa konsumen tidak sepakat dengan keberadaannya.

Dalam rangka mendukung data dari wawancara, penelitian ini juga menggunakan kuesioner. Jumlah responden yang dibutuhkan sebanyak 30 orang, yang merupakan konsumen yang pernah melakukan pembelian produk dari Toko Bismillah Distributor. Kuesioner digunakan untuk mendapatkan gambaran yang lebih komprehensif mengenai persepsi konsumen terhadap *website e-commerce* tersebut.

Tabel 2. Hasil Kuesioner Seputar *Website E-commerce* Toko Bismillah Distributor

#	Karakteristik Responden	Keterangan			
		STS	**TS**	**S**	**SS**
1	Karakteristik Responden Berdasarkan Kemudahan *Website E-commerce*	0%	3,4%	43,3 %	53,3 %
	Jumlah	0	1	13	16
2	Karakteristik Responden Berdasarkan Keseluruhan Informasi *Website*	0%	0%	60%	40 %

#	Karakteristik Responden	Keterangan			
		STS	TS	S	SS
	E-commerce Toko Bismillah Distributor Dapat Ditangkap Dengan Mudah dan Jelas				
	Jumlah	0	0	18	12
3	Karakteristik Responden Berdasarkan Tampilan Menarik dari *Website E-commerce* Toko Bismillah Distributor	0%	0%	46,7 %	53,3 %
	Jumlah	0	0	14	16
4	Karakteristik Responden Berdasarkan Fitur - Fitur yang Membantu Dalam Mengakses *Website E-commerce* Toko Bismillah Distributor	0%	0%	50%	50 %
	Jumlah	0	0	15	15
5	Karakteristik Responden Berdasarkan Ketertarikan Informasi Kupon Potongan Harga Pada *Website E-commerce* Toko Bismillah Distributor	0%	0%	53,3 %	46,7 %
	Jumlah	0	0	16	14
6	Karakteristik Responden Berdasarkan Kemudahan dan Kejelasan Informasi Produk Pada *Website E-commerce* Toko Bismillah Distributor	0%	0%	56,7 %	43,3 %
	Jumlah	0	0	17	13
7	Karakteristik Responden Berdasarkan Ketertarikan Untuk Mengunjungi *Website E-commerce* Toko Bismillah Distributor di Masa Mendatang	0%	6,6%	56,7 %	36,7 %
	Jumlah	0	2	17	11
8	Karakteristik Responden Berdasarkan Ketertarikan Membeli Produk Toko Bismillah Distributor Melalui *Website*	0%	13,3 %	46,7 %	40 %
	Jumlah	0	4	14	12
9	Karakteristik Responden Berdasarkan Rekomendasi Produk Pada *Website E-commerce* Toko Bismillah Distributor Kepada Rekan, Teman, atau Saudara	0%	0%	66,7 %	33,3 %
	Jumlah	0	0	20	10

Hasil pendukung berupa kuesioner yang telah diberikan kepada 30 responden konsumen Toko Bismillah Distributor, terdiri dari 9 pria dan 21 wanita, menunjukkan hasil yang cukup memuaskan. Rata-rata pengguna menyatakan setuju dengan *website e-commerce* Toko Bismillah Distributor. Beberapa variabel *product knowledge* yang dicakup dalam kuesioner melibatkan pertanyaan mengenai mutu (nomor 1, 3, dan 4), kandungan (nomor 6), dan informasi (nomor 2 dan 5). Selain variabel *product knowledge*, terdapat pertanyaan tambahan pada nomor 7, 8, dan 9 yang bertujuan untuk menilai tingkat ketertarikan konsumen terhadap penggunaan, pembelian, dan rekomendasi terhadap *website e-commerce* Toko Bismillah Distributor.

KESIMPULAN DAN SARAN

Hasil akhir dalam analisis kebutuhan dan desain sistem melibatkan pemahaman mendalam tentang produk, termasuk informasi detail seperti ukuran, warna, bahan, dan deskripsi keunggulan produk. Fitur-fitur penting seperti detail produk yang jelas, ulasan produk, panduan berbelanja, profil perusahaan, dan kolom kontak serta saran telah diterapkan secara efektif. Pengujian oleh 7 konsumen melalui wawancara dan 30 sampel kuesioner menunjukkan hasil yang dapat diterima, sementara uji *user acceptance testing* mengonfirmasi keseluruhan kesesuaian implementasi *website* dengan kebutuhan pengguna. Dengan demikian, rancangan *website* ini berhasil menciptakan platform *e-commerce* yang memenuhi standar kualitas dan

memberikan pengalaman yang memuaskan bagi pengguna Toko Bismillah Distributor.

Berdasarkan hasil analisis, terdapat beberapa saran yang dapat diimplementasikan untuk meningkatkan fungsionalitas *website*, diantaranya adalah penambahan fitur akun media sosial dan email guna memperluas jangkauan konsumen. Selain itu, disarankan untuk memasukkan fitur notifikasi pada *website* dengan tujuan mempercepat penyampaian informasi terbaru seperti diskon, produk baru, dan *event* khusus kepada pengguna.

Ucapan Terimakasih

Puji syukur penulis kepada Tuhan Yang Maha Esa karena dengan Rahmat-Nya jurnal ini dapat terselesaikan dengan baik. Ucapan terima kasih oleh penulis yang ditujukan kepada keluarga dan rekan-rekan yang selalu mendukung. Penulis juga mengucapkan terima kasih kepada pada dosen Universitas Dinamika yaitu Tan Amelia, S.Kom., M.MT. dan Ayouvi Poerna Wardhanie, S.M.B., M.M. yang telah membantu saya dalam menyelesaikan jurnal ini. Penulis meminta maaf jika terdapat kesalahan dalam penulisan pada jurnal dan mohon dimaklumi. Semoga isi dari jurnal ini dapat bermanfaat bagi kita semua.

DAFTAR PUSTAKA

[1] A. Wirapraja and H. Aribowo, "Pemanfaatan E-Commerce Sebagai Solusi Inovasi Dalam Menjaga Sustainability Bisnis," *Teknika*, vol. 7, no. 1, pp. 66–72, 2018.

[2] S. Sulistiowati, E. Sutomo, and F. Aditya, "Rancang Bangun Aplikasi Penjualan Online Pada UMKM Riot Safe Menggunakan Gamification," *J. Technol. Informatics*, vol. 5, no. 1, pp. 15–24, 2023.

[3] A. Basyir, "Pengaruh fashion lifestyle dan pengetahuan produk terhadap minat beli (studi terhadap konsumen batik tulis madura al-fath KKG Bangkalan)," *J. Pendidik. Tata Niaga(JPTN)*, vol. 7, no. 3, pp. 564–570, 2019.

[4] N. Chasanah, "Penerapan Customer Knowledge Management pada e-commerce Butik Gendis Fashion Muslim," *J. Pendidik. dan Teknol. Indones.*, vol. 1, no. 3, pp. 131–140, 2021.

[5] A. Nirwana, A. Supriyanto, and A. Wardhanie, "Evaluasi Usability, Pedagogical dan User Experience My Brilian Menggunakan Metode Tuxel," *J. Technol. Informatics*, vol. 4, no. 1, pp. 7–12, 2022.

[6] A. P. Wardhanie, A. Z. Naufal, and S. H. Eko Wulandari, "Perancangan Strategi Digital Marketings Dengan Metode Race Pada Layanan Online Food Delivery Berdasarkan Perilaku Pelanggan Generasi Z," *J. Technol. Informatics*, vol. 3, no. 1, pp. 1–11, 2021.

[7] A. Putu Candra *et al.*, "Sistem Informasi Penjualan Online Thrift Shop Berbasis Web," *J. Technol. Informatics*, vol. 5, no. 2, pp. 116–124, 2024.

[8] E. Murni, D. Diniati, M. Mustakim, I. Kusumanto, and A. Anwardi, "Perancangan Dan Implementasi Sistem Website E-commerce Dalam Bisnis Bakery Upaya Meningkatkan Penjualan Dan Pemasaran Menggunakan Metode Waterfall Pada Pabrik Prima Sari Bakery," *J. Rekayasa Sist. Ind.*, vol. 7, p. 122, 2020.

[9] Ayuningtyas, M. Calvin Gunawan, S. Rafi Zulfitra, I. Bintang Pratama, and A. Arfiansyah Juniawan, "Rancang Bangun Website untuk Administrasi Warga Gubeng Kertajaya Surabaya untuk mendukung Penerapan Smart City," *J. Technol. Informatics*, vol. 4, no. 1, pp. 1–6, 2022.

[10] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, and K. Achan, "Product knowledge graph embedding for E-commerce," *WSDM 2020 - Proc. 13th Int. Conf. Web Search Data Min.*, pp. 672–680, 2020.

[11] Julianti and Y. Aini, "Pengaruh Online Costumer Review Dan Online Costumer Rating Terhadap Keputusan Pembelian Online Marketplace (Studi Mahasiswa Universitas Pasir Pengaraian)," *J. Ilm. Cano Ekon.*, vol. 11, no. 2, pp. 51–58, 2022.

[12] S. W. Diah I. P and N. I. K. Wardhani, "Pengaruh Kualitas Website Dan Hedonic Shopping Terhadap Impulse Buying E-Commerce Blibli Di Surabaya," *Sci. J. Reflect. Econ. Accounting, Manag. Bus.*, vol. 6, no. 1, pp. 215–227, 2023.

[13] Agustino; and Syaifullah, "Pengaruh Kualitas Produk Dan Product Knowledge Terhadap Keputusan Pembelian Konsumen Pada Pt Long Time.," *J. EMBA*, vol. 8, no. 1, pp. 627–636, 2020.

[14] P. Elsya and R. Indriyani, "The Impact of Product Knowledge and Product Involvement to Repurchase Intention for Tupperware Products among Housewives in Surabaya, Indonesia," *SHS Web Conf.*, vol. 76, p. 01037, 2020.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, October 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.792

Arrhythmia Classification with ECG Signal using Extreme Gradient Boosting (XGBoost) Algorithm

Diah Asmawati[1], Lukman Arif Sanjani[2], Christiant Dimas Renggana[3], Chastine Fatichah[4], Tanzilal Mustaqim[5]
[1,2,3,4,5]Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail: 6025232014@student.its.ac.id[1], 6025232027@student.its.ac.id[2], 6025232016@student.its.ac.id[3], chastine@if.its.ac.id[4], 7025222005@student.its.ac.id
*Corresponding author: E-mail: 6025232027@student.its.ac.id

Abstract: *Heart disease is one of the most dangerous illnesses because it has the potential to take people's lives. One of the causes of heart disease is arrhythmia, an abnormal condition of the heartbeat. To diagnose arrhythmia, analysis of electrocardiographic (ECG) signals can be performed. However, this analysis is very difficult to do conventionally and has the potential for errors, so there is a need for automatic ECG classification to detect arrhythmia. This study aims to fill the research gap by creating an ECG classification model to detect arrhythmia using the XGBoost algorithm. The results are quite good for each class, with accuracies for class N at 98.87%, class SVEB at 99.37%, class VEB at 99.4%, class F at 99.75%, and class Q at 99.99%. However, compared to existing methods in previous research, these results are still considered not better than those models.*

Keywords*: Arrhythmia; Classification; ECG; XGBoost*

INTRODUCTION

Cardiovascular disease is one of the most common diseases affecting individuals of all ages, from infants to the elderly [1]. It is also the leading cause of death worldwide, accounting for 31% of global deaths [2]. Heart disease has significant economic impacts globally. It is recorded that heart disease causes an annual economic burden of €210 billion in Europe and $555 billion in the United States [3]. One of the causes of heart disease is arrhythmia, a condition where the heart's rhythm is abnormal [4]. A common type of arrhythmia is atrial fibrillation, characterized by a rapid and irregular heartbeat, with a global prevalence of 46.3 million cases [5]. Arrhythmia is responsible for 80% of deaths caused by heart disease [6]. This highlights the significant impact of arrhythmia in addressing heart disease issues.

One method to diagnose arrhythmia is through the use of electrocardiographic (ECG) signals [7]. However, this method presents a significant challenge due to the difficulty in detecting and categorizing different waveforms and morphologies within a signal, making the analysis only possible by specialists in the field [8]. This results in ECG heartbeat classification analysis being very time-consuming and prone to errors due to fatigue [9]. Therefore, there is a need to replace conventional methods involving human labor with an automated heartbeat classification system.

This need can be addressed by leveraging the advancements in technology, which open opportunities to create a heartbeat classification model. Such a model can be developed thanks to the presence of artificial intelligence (AI) today. AI can create a classification model (supervised learning) that requires labeled data as a reference for predictions [10]. This is also supported by AI's ability to develop and improve its performance over time with minimal or no human intervention [11]. The utilization of AI in the form of machine learning for early diagnosis of various diseases has already had a substantial impact previously [12].

Several previous studies have conducted various experiments to create models for classifying types of arrhythmias based on ECG. One study utilized variable length heartbeats for classification with a combination of CNN and LSTM techniques, achieving a final accuracy of 98.10% [7]. Another study using a similar classification approach, CNN-LSTM, achieved a higher accuracy of 99.27% [13]. The use of CNN for feature extraction was also applied to another classifier, XGBoost, which achieved an accuracy of 99.69% [14]. Ensemble learning was also implemented using Random Forest and Support Vector Machine algorithms, resulting in an overall accuracy of 98.21% [15]. Another ensemble learning approach that included the kNN algorithm resulted in a lower accuracy of 97.8% [16].

This study aims to create an arrhythmia classification model based on ECG from a dataset using the XGBoost algorithm. The limitation of this study is that the model will only be tested on one dataset, and the proposed methodology will be implemented using only one algorithm. The expectation is that this research can contribute by filling the research gap where there is still no study that process the MIT-BIH Arrhythmia dataset using the XGBoost classifier with the same approach as this study.

METHOD

The proposed method consists of four major parts: pre-processing, feature extraction using TSFEL, model training, and model testing. The workflow of this method can be seen in Fig 1.

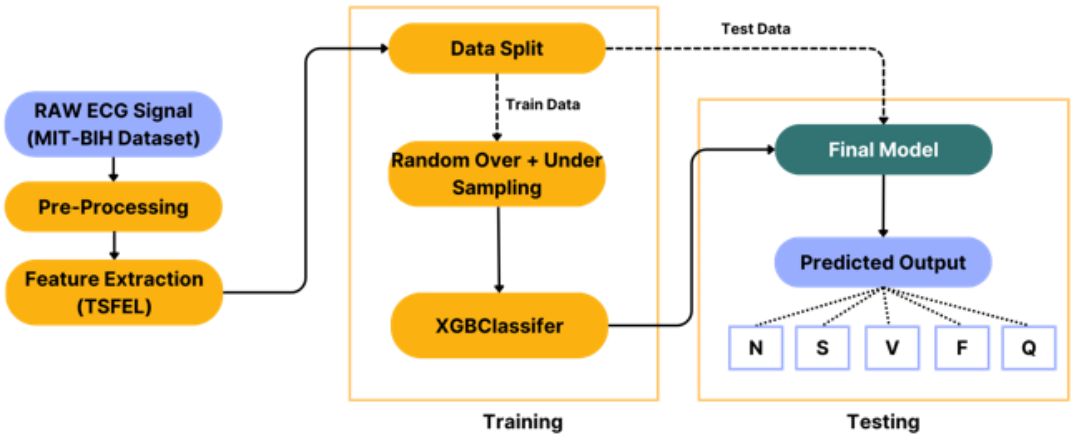


Figure 1. Proposed Methodology Flow

Dataset Description

The dataset used in this study is the public MIT-BIH Arrhythmia Database, which consists of 48 ECG recordings, each spanning 30 minutes. These samples were taken at a frequency of 360Hz per channel with an ADC resolution of 11 bits at a 10mV interval [17], [18]. This dataset began distribution in 1980 and has undergone various updates, with the latest revision edited in 2018. Heartbeat labels are provided in the database based on annotations independently made by one or more cardiologists. Classification labels are then categorized based on the AAMI (Association for the Advancement of Medical Instrumentation) standards, where classes are divided into normal (N), supraventricular (SVEB or S), ventricular (VEB or V), fusion (F), and unclassified (Q) [19]. Each annotation code given in the MIT-BIH dataset is grouped according to the AAMI standards as shown in Table 1.

Table 1. Mapping Of Heartbeat Class Categories in the MIT-BIH Dataset Based on AAMI Standards

ID	Heartbeat Types Based on AAMI Standards	Heartbeat Types Based on MIT-BIH
0	N (Normal or other than S, V, F, Q)	*normal beat (N), left bundle branch block beat (L), right bundle branch block beat (R), atrial escape beat(e), nodal (junctional) escape beat (j)*
1	SVEB (*Supraventricular Ectopic Beat*)	*atrial premature beat (A), aberrated atrial premature beat (a), nodal (junctional) premature beat (J), supraventricular premature beat (S)*
2	VEB (*Ventricular Ectopic Beat*)	*premature ventricular contraction (V), ventricular escape beat (E)*
3	F (*Fusion Beat*)	*fusion of ventricular and normal beat (F)*
4	Q (*Unknown Beat*)	*paced beat (P), fusion of paced and normal beat (f), unclassified beat (U)*

In each recording for each patient, there are 2 readable ECG channels. However, in this case, only one channel, the MDII channel, is used for the classification process.

Pre-process and Segmentation

Pre-processing of the ECG recording data aims to filter and reduce noise in the signal, making it optimal for recognizing classes in the learning approach in the subsequent process. The filtering is performed by applying a median filter with signal widths of 200ms and 600ms, then subtracting the original signal and baseline values to produce a corrected signal. Next, the denoising process is carried out on this signal using the db4 Discrete Wavelet Transform (DWT), applying a high-pass filter to the DWT coefficients, and then inverting it to obtain a cleaner signal.

After pre-processing, the signal is segmented based on the R-Peak annotations provided in the MIT-BIH dataset. As a result, the data generated totals 101,147 samples with the number of each class as

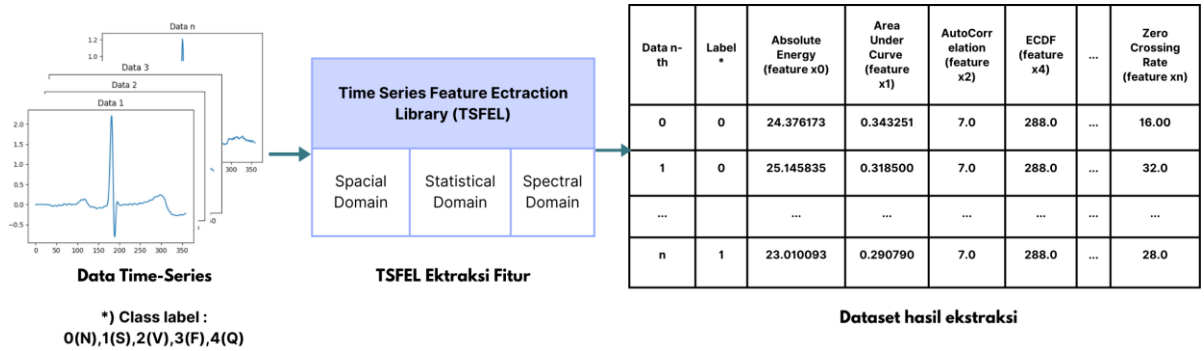


Figure 3. Feature Extraction Using TSFEL Process

follows: 90,320 (N), 7,229 (VEB), 2,781 (SVEB), 802 (F), and 15 (Q). Sample data of each class signal can be illustrated in Fig. 2.

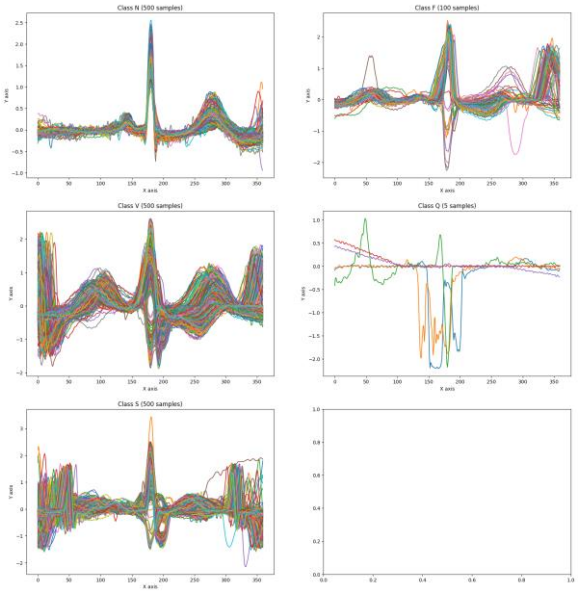


Figure 2. Example of Signal for Each Class

Feature Extraction

The segmented signal is a single data point that represents a specific class. This data inherently contains information that can be extracted and turned into features to be trained in a learning model approach. In this study, the Time Series Feature Extraction Library (TSFEL) is used to extract features from the signal data. TSFEL can handle multidimensional time series data, with available features divided into three domains: statistical, temporal, and spectral [22]. The total categories cover more than 60 features and based on its default parameters used for ECG signals in this study, it can extract 314 features from each data point. The illustration of the feature extraction process using TSFEL can be seen in Fig. 3. The list of features obtained after extraction using TSFEL can be seen in Table 2.

Table 2. Available Features in TSFEL

Domain	Feature Types
Spectral	FFT Mean coefficient Fundamental Frequency Human range energy LPCC (Linear Prediction Cepstral Coefficient) MFCC (Mel-Frequency Cepstral Coefficient) Max power spectrum Maximum frequency Median frequency Power bandwith Spectral centroid Spectral decrease Spectral distance Spectral entropy Spectral kurtosis Spectral positive turning points Spectral roll-off Spectral roll-on Spectral skewness
Statistical	Absolute energy Average power ECDF (Empirical Cumulative Distribution Function) ECDF Percentile ECDF Percentile Count Entropy Histogram Interquartile range Kurtosis Max Mean Mean absolute deviation Median Median absolute deviation Min Peak to peak distance Root mean square Skewness Standard deviation Variance

Domain	Feature Types
Temporal	Area under the curve Autocorrelation Centroid Mean absolute diff Mean diff Median absolute diff Median diff Negative turning points Neighbourhood peaks Positive turning points Signal distance Slope Sum absolute diff Zero crossing rate

Data Balancing

In the previously mentioned dataset distribution, it is evident that the data is imbalanced. Most of the data belongs to the normal class, whereas a good classification requires balanced data across the minority classes. Therefore, data balancing performed to make the dataset balanced [13]. The data balance technique used is a combined Random Over Sampling and Random Under Sampling processes. The result of this phase can be seen in Table 3.

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an algorithm used for classification and regression. XGBoost is an enhancement of the boosting algorithm, which combines several weak models to form a strong model.

Table 3. Data Distribution Before Balanced, After Random Over Sampling, and After Random Under Sampling

Class	Before Data Balancing	After Random Over Sampling	After Random Under Sampling
N	72255	144510	36127
SVEB	5783	11566	5783
VEB	2225	4450	2225
F	642	1284	642
Q	12	24	12

This algorithm develops an ensemble sequential number of trees [14].

In this case, the dateaset will be divided into training and testing data with a ratio of 3:1. The training data will be trained using XGBoost Classifier with the parameter gamma set to 0.1 as a way to control the complexity threshold for decision tree splits. The process is shown in Fig. 4.

Evaluation Matrix

Evaluation of the performance of a classifier is necessary to ensure the accuracy of its predictions. For this evaluation, a recommendation matrix from AAMI is used, which includes TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) [19]. Such a matrix is needed because the classification performed is not binary but multiclass, so the confusion matrix cannot only encompass positive and negative classes. TP (True Positive) represents the condition when a positive instance is correctly predicted, while FP (False Positive) is the incorrect prediction of a negative instance as positive. On the other hand, a result is considered TN (True Negative) when a negative instance is correctly

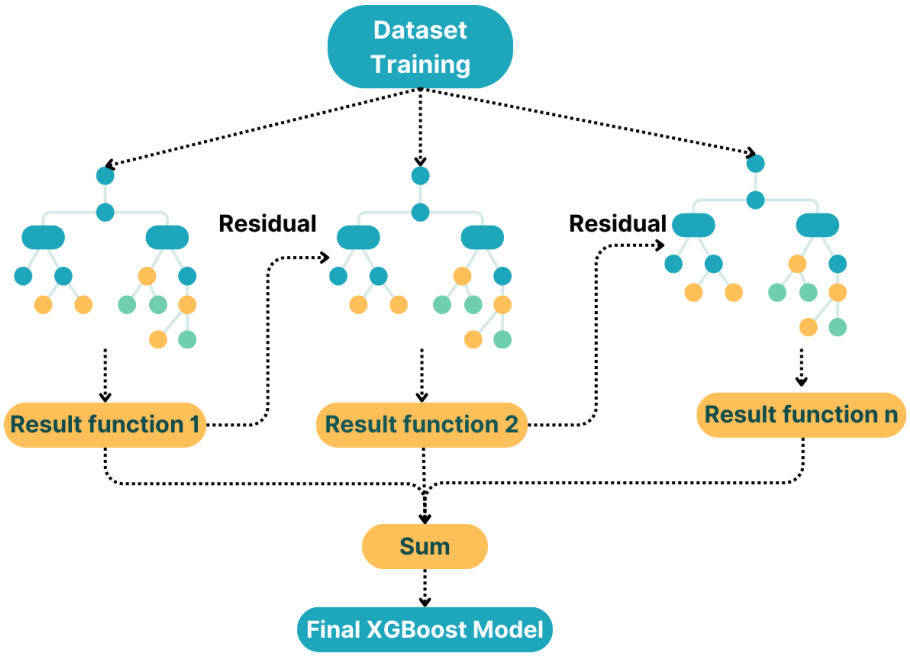


Figure 4. XGBoost Classifier Flow

predicted, whereas FN (False Negative) is the result when a positive instance is incorrectly predicted as negative.

These results are then used to calculate several performance metrics, namely:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}\ x\ 100\% \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN+FP}\ x\ 100\% \quad (3)$$

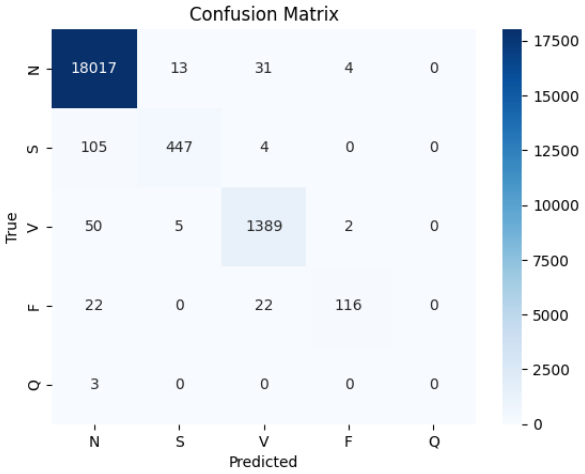
$$\text{F1-Score} = 2\ x\ \frac{Precision*Recall}{Precision+Recall}\ x\ 100\% \quad (4)$$

Formula 1 shows the calculation method for accuracy, which is the ratio of correct predictions to the total number of predictions. Accuracy itself gives a straightforward overview of the prediction results. However, it becomes problematic when dealing with an imbalanced dataset, as accuracy can yield misleading results in such cases. To address this issue, sensitivity is calculated to determine the true positive rate through Formula 2, and specificity is measured to gauge the true negative rate through Formula 3. Finally, to get an understanding of the balance between the true positive rate and the true negative rate, the F1-Score is calculated as shown in Formula 4. The results of these calculations will be used as a reference for comparison with previous research studies.

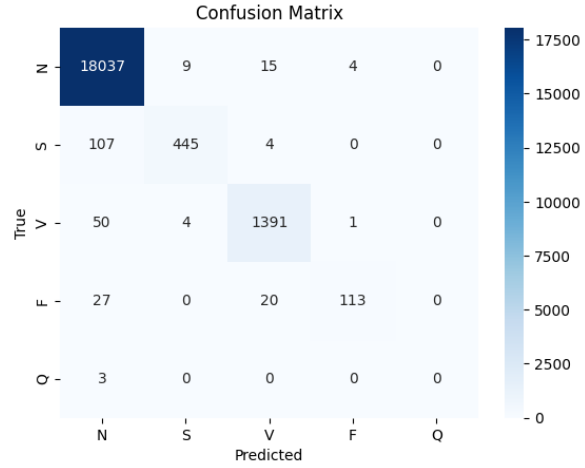
RESULT AND DISCUSSION

The model training was carried out using a dataset that had already undergone pre-processing, segmentation, feature extraction, and data balancing. The data used for training contains 314 features, and the distribution of each class can be seen in Table 2. In this study, experiments were also conducted to predict using a model trained on data that did not undergo the resampling process first.

After successfully training the model and testing it on the test dataset, the prediction results were summarized in a confusion matrix. The confusion matrix helps map the results that can be categorized as TP, TN, FP, or FN. The confusion matrix results for the prediction of 5 classes according to the AAMI recommendations can be seen in Fig. 5. Based on the obtained results, there were a total of 261 prediction errors out of 20,230 test data points in the prediction model using resampled data. In contrast, the prediction model using unresampled data achieved fewer prediction errors, with only 244. However, this cannot be the main reference in determining the performance of each model. Therefore, it is necessary to pay attention to the specific prediction results of each class as shown in Table 4 and Table 5.



(a)



(b)

Figure 5. Confusion Matrix for Resampled Data (a) and Unresampled Data (b)

Table 4. Evaluation Result for Resampled Data

Class	Acc%	Spe%	Sen%	F1%
N	98.87	91.69	99.73	99.37
SVEB	99.37	99.91	80.40	87.56
VEB	99.44	99.70	96.06	96.06
F	99.75	99.97	72.50	82.27
Q	99.99	100	0	0

Table 5. Evaluation Result for Unresampled Data

Class	Acc%	Spe%	Sen%	F1%
N	98.94	91.36	99.85	99.41
SVEB	99.39	99.93	80.04	87.87
VEB	99.54	99.79	96.20	96.73
F	99.74	99.98	70.63	81.36
Q	99.99	100	0	0

The results indicate that almost all evaluation outcomes show that the model with unresampled data is better compared to the model with resampled data. However, it is important to note that the classes that need to be considered are the SVEB, VEB, and F classes. In

the medical field, sensitivity (recall) is an evaluation metric that requires more attention because it can affect patient treatment. The model with resampled data demonstrated better sensitivity values for the VEB and F classes. There was a significant increase, especially in the sensitivity of the F class, which increased by approximately 2%. This shows that the performance of the model with resampled data is better compared to the model with unresampled data in a medical context.

Table 6. Result Comparison

Model	Metric	N	S	V	F	Q
CNN + LSTM [13]	Acc%	97.95	99.16	**99.80**	99.25	99.48
	Sen%	-	-	-	-	-
	Spe%	-	-	-	-	-
AdaBoost + Random Forest [20]	Acc%	**99.24**	**99.58**	99.67	99.79	99.94
	Sen%	**99.95**	**82.61**	**97.45**	70.88	**99.24**
	Spe%	95.86	**99.99**	99.84	**100**	99.99
Ensemble RF & SVM [15]	Acc%	98.48	98.74	99.37	**99.84**	**99.99**
	Sen%	99.50	74.20	94.22	73.21	0
	Spe%	89.82	99.69	99.72	99.88	**100**
CNN + Focal loss [21]	Acc%	98.71	99.16	99.36	99.75	99.84
	Sen%	99.49	77.88	94.54	**82.10**	98.51
	Spe%	**97.60**	99.88	**99.85**	99.92	99.97
Proposed method	Acc%	98.87	99.37	99.44	99.75	**99.99**
	Sen%	99.73	80.40	96.06	72.50	0
	Spe%	91.69	99.91	99.70	99.97	**100**

From the comparison of results, it was found that the proposed method has a lower accuracy compared to existing methods. In Table 6, the bold numbers indicate the best accuracy for the corresponding matrix and class. The method using XGBoost as the classifier only excels in class Q, which cannot be considered a reliable reference due to the severe imbalance in the data, making the results for class Q inherently unreliable. This indicates the subpar performance of XGBoost as a classifier in achieving better accuracy compared to the methods used in previous studies. However, XGBoost does outperform certain methods in specific matrices and classes, but it is not the best when compared to all four other methods simultaneously. For example, the method using the XGBoost classifier outperforms the accuracy of each class when compared to CNN + Focal loss [21]. Another example is that XGBoost can also outperform most of the specificity values produced by the ensemble RF & SVM method, except for class F [15].

This research also has certain limitations that may have prevented the XGBoost algorithm from performing at its best. These limitations include the highly imbalanced dataset, which required sampling that may have compromised data integrity, the scope for further exploration of feature extraction techniques, and the use of hyperparameters to better optimize the XGBoost algorithm. To ensure the performance of XGBoost, further exploration is needed, especially regarding pre-processing, as it also plays a crucial role in determining the final evaluation results.

CONCLUSION AND SUGGESTION

The evaluation results of arrhythmia classification using the XGBoost classifier show that its performance is not quite satisfactory. This is evidenced by the comparison of performance with other models applied in previous research. Although the comparison shows that XGBoost can outperform certain models in specific matrices for certain classes, there are still other models that perform better when compared to XGBoost. Therefore, reconsideration is needed regarding the use of XGBoost for arrhythmia classification. There are certainly several areas for future work, such as trying this classifier on a different and larger dataset, as the current dataset is quite imbalanced. Additionally, ensemble learning could also be applied to improve the accuracy of the created classifier.

REFERENCES

[1] A. K. Sangaiah, M. Arumugam, and G. Bin Bian, "An intelligent learning approach for improving ECG signal classification and arrhythmia analysis," *Artif Intell Med*, vol. 103, Mar. 2020, doi: 10.1016/j.artmed.2019.101788.

[2] Shanthi. Mendis, P. Puska, Bo. Norrving, World Health Organization., World Heart Federation., and World Stroke Organization., *Global atlas on cardiovascular disease prevention and control*. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization, 2011.

[3] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, "A review on deep learning methods for ECG arrhythmia classification," 2020, doi: 10.1016/j.eswax.2020.10.

[4] C. Chen, Z. Hua, R. Zhang, G. Liu, and W. Wen, "Automated arrhythmia classification based on a combination network of CNN and LSTM," *Biomed Signal Process Control*, vol. 57, Mar. 2020, doi: 10.1016/j.bspc.2019.101819.

[5] M. Halomoan, "Epidemiologi Artimia," https://www.alomedika.com/penyakit/kardiologi/aritmia/epidemiologi.

[6] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network," *Information Fusion*, vol. 53, pp. 174–182, Jan. 2020, doi: 10.1016/j.inffus.2019.06.024.

[7] S. L. Oh, E. Y. K. Ng, R. S. Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats," *Comput Biol Med*, vol. 102, pp. 278–287, Nov. 2018, doi: 10.1016/j.compbiomed.2018.06.002.

[8] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation," Apr. 2018, doi: 10.1109/ICHI.2018.00092.

[9] D. K. Atal and M. Singh, "Arrhythmia Classification with ECG signals based on the

Optimization-Enabled Deep Convolutional Neural Network," *Comput Methods Programs Biomed*, vol. 196, Nov. 2020, doi: 10.1016/j.cmpb.2020.105607.

[10] J. Lemantara, "Penerapan Algoritma Naïve Bayes dan ID3 untuk Memprediksi Segmentasi Pelanggan pada Penjualan Mobil," *Journal of Technology and Informatics (JoTI)*, vol. 4, no. 1, pp. 31–40, Oct. 2022, doi: 10.37802/joti.v4i1.265.

[11] N. A. B. Ibrahim, "The Existence of Artificial Intelligence in the Future," *Journal of Technology and Informatics (JoTI)*, vol. 5, no. 1, pp. 25–33, Oct. 2023, doi: 10.37802/joti.v5i1.349.

[12] M. Guhdar, A. Ismail Melhum, and A. Luqman Ibrahim, "Optimizing Accuracy of Stroke Prediction Using Logistic Regression," *Journal of Technology and Informatics (JoTI)*, vol. 4, no. 2, pp. 41–47, Jan. 2023, doi: 10.37802/joti.v4i2.278.

[13] G. R. Patra, M. K. Naik, and M. N. Mohanty, "ECG Signal Classification Using a CNN-LSTM Hybrid Network," in *2023 2nd International Conference on Ambient Intelligence in Health Care, ICAIHC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICAIHC59020.2023.10431449.

[14] R. Mogili and G. Narsimha, "Detection of Cardiac Arrhythmia from ECG Using CNN and XGBoost," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 2, pp. 414–425, Apr. 2022, doi: 10.22266/ijies2022.0430.38.

[15] S. Bhattacharyya, S. Majumder, P. Debnath, and M. Chanda, "Arrhythmic Heartbeat Classification Using Ensemble of Random Forest and Support Vector Machine Algorithm," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 3, pp. 260–268, Jun. 2021, doi: 10.1109/tai.2021.3083689.

[16] R. Kumar and Jyoti, "A Hybrid Approach Using SVM, kNN and Random Forest for ECG Classification," in *ISED 2023 - International Conference on Intelligent Systems and Embedded Design*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ISED59382.2023.10444570.

[17] "MIT-BIH Arrhythmia Database," http://circ.ahajournals.org/content/101/23/e215.full.

[18] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, Jun. 2000, doi: 10.1161/01.CIR.101.23.e215.

[19] "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms," in *ANSI/AAMI EC57:2012/(R)2020; Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms*, AAMI, 2013. doi: 10.2345/9781570204784.ch1.

[20] R. Li *et al.*, "An Intelligent Heartbeat Classification System Based on Attributable Features with AdaBoost+Random Forest Algorithm," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/9913127.

[21] T. F. Romdhane, H. Alhichri, R. Ouni, and M. Atri, "Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss," *Comput Biol Med*, vol. 123, Aug. 2020, doi: 10.1016/j.compbiomed.2020.103866.

[22] M. Barandas et al., "TSFEL: Time series feature extraction library," SoftwareX, vol. 11, 2020, Art. no. 100456.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, Bulan Oktober Tahun 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.778

Membangun Sistem Katalog Digital untuk Perpustakaan SMP: Solusi Tepat Mempermudah Pencarian Buku

Mahendra Bagaskara[1], Erwin Sutomo[2], Ayuningtyas[3*]
[1,2,3]Program Studi Sistem Informasi, Universitas Dinamika, Surabaya, Indonesia
* Penulis Korespondensi: 17410100040@dinamika.ac.id[1], sutomo@dinamika.ac.id[2], tyas@dinamika.ac.id[3*]

Abstrak: Perpustakaan merupakan salah satu unsur penting dalam proses belajar mengajar. Mendapatkan sumber belajar yang tepat dapat membantu proses tersebut. Apabila siswa atau guru kesulitan mendapatkan buku sebagai sumber belajar, proses tersebut dapat terganggu. Penelitian ini bertujuan merancang dan membangun aplikasi katalog buku perpustakaan berbasis web di salah satu SMP Sidoarjo. Permasalahan utama perpustakaan terletak pada sistem katalog manual yang rentan kesalahan, memudarnya isi buku, dan proses pencarian yang lama. Aplikasi ini dikembangkan menggunakan framework CodeIgniter dan mengikuti metodologi SDLC untuk memastikan kelancaran proses pengembangan. Hasil pengujian menunjukkan aplikasi berhasil memvalidasi data dan diharapkan dapat meningkatkan efisiensi dan efektivitas pengelolaan katalog, serta memberikan kemudahan akses bagi pengguna.

Kata Kunci: Aplikasi; Katalog Buku; Perpustakaan Sekolah; *Website*

Abstract: *The library is one of the essential elements in the teaching and learning process. Getting the right learning resources can help the process. If students or teachers have difficulty getting books as learning resources, the process can be disrupted. This study aims to design and build a web-based library book catalog application in one of Sidoarjo's junior high schools. The main problem of the library lies in the manual catalog system, which is prone to errors, the fading of book contents, and the long search process. This application was developed using the CodeIgniter framework and followed the SDLC methodology to ensure a smooth development process. The test results showed that the application successfully validated data and is expected to improve the efficiency and effectiveness of catalog management, as well as provide easy access for users.*

Keywords: *Book Catalog; Application; Library; Website*

PENDAHULUAN

Perpustakaan merupakan sebuah pranata mengelola kumpulan karya tulis, karya cetak, dan/atau karya rekam yang profesional serta sistem yang dapat memenuhi kebutuhan pendidikan, kebutuhan penelitian pelestarian, tempat terciptanya gudang yang penuh akan informasi serta tempat rekreasi untuk menunjang kecerdasan suatu bangsa. Perpustakaan memiliki peran penting upaya dalam melestarikan dan meningkatkan efektiftas belajar-mengajar [1]. Selain itu, perpustakaan juga menyediakan layanan bertujuan untuk mendukung kegiatan para pengunjung agar menjadi lebih nyaman berkegiatan diperpustakaan,

Perpustakaan terbagi menjadi dua perpustakaan umum dan perpustakaan sekolah. Perpustakaan umum merupakan perpustakaan yang digunakan masyarakat umum [2], sedangkan perpustakaan sekolah merupakan perpustakaan yang dimanfaatkan warga sekolah yang terdiri dari peserta didik, guru, dan pegawai. Perpustakaan sekolah sendiri menjadi salah satu fasilitas penunjang dalam mendapatkan informasi untuk warga sekolah, yang terkait materi sekolah, music, cerpen, dan materi lainnya [3].

Studi literatur yang dilakukan menunjukkan adanya perkembangan dalam pengembangan sistem informasi perpustakaan berbasis web. Permasalahan yang umum dihadapi oleh staff perpustakaan seperti pencatatan yang masih manual, kesulitan proses sirkulasi, baik ketika peminjaman ataupun pengembalian juga dihadapi oleh [4], [5], [6]. Penelitian sebelumnya oleh [6] telah berhasil membangun sistem yang mampu mengelola data pengunjung hingga transaksi peminjaman. Namun, penelitian lanjutan oleh [4] berhasil menambahkan fitur klasifikasi buku, sehingga memungkinkan pengelolaan koleksi buku yang lebih terstruktur. Perbedaan lain yang signifikan adalah penambahan fitur denda pada penelitian terbaru, yang memungkinkan perpustakaan untuk melakukan pengawasan terhadap keterlambatan pengembalian buku.

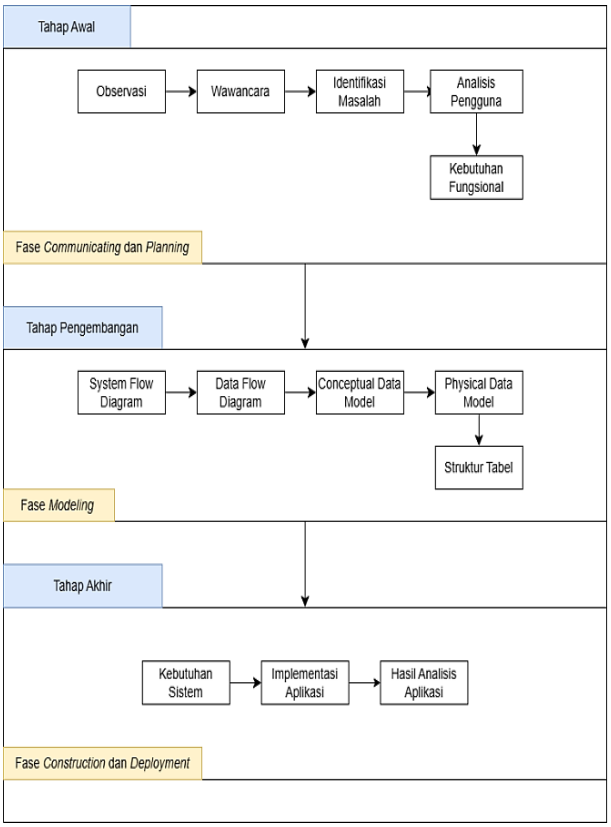
Penelitian dilaksanakan perpustakaan Sekolah Menengah Pertama yang ada di Sidoarjo, yang saat ini perpustakaan tersebut memiliki jumlah katalog buku yang cukup banyak. Dalam mengelola pencatatan katalog buku sekarang masih dilakukan secara manual. Hal tersebut sebenarnya tidak ada salahnya, tetapi memungkinkan memiliki risiko yang dapat dipertimbangkan, seperti pencarian buku menjadi lama, terjadinya kesalahan pada pencatatan buku, dan buku besar yang menyimpan pencatatan buku berisiko rusak ataupun kartu katalog juga hilang. Permasalahan tersebut dapat mempengaruhi proses kualitas pelayanan yang kurang baik pada perpustakaan.

Dengan permasalahan yang terjadi, maka diperlukan sebuah aplikasi katalog buku perpustakaan berbasis *website* yang mengatasi permasalahan pada

proses pencatatan katalog buku menjadi lebih teratur, mempermudah anggota perpustakaan dalam mencari buku, dan meningkatkan kinerja petugas perpustakaan [4]. Dengan adanya aplikasi ini, diharapkan mampu mempermudah dalam pencatatan koleksi buku, dan pencarian buku yang ada di perpustakaan menjadi lebih cepat. Proses penelitian menggunakan metode *Waterfall* dikarenakan mampu mempersingkat waktu dalam mengembangkan aplikasi katalog buku.

METODE

Pada metode penelitian yang diterapkan pada pengerjaan aplikasi katalog buku perpustakaan ini peneliti menggunakan Siklus Hidup Pengembangan Perangkat Lunak (SDLC) melalui Metode *Waterfall* yang tertera pada Gambar 1. [5].Metode *Waterfall* memiliki tahapan dasar, yaitu penentuan spesifikasi awal *(Requirements analysis and definition)*, desain sistem *(System and software design)*, pembuatan atau pengembangan sistem *(Implementation and unit testing)*, dan tahap pemeliharaan *(Operation and maintenance)*. Beberapa tahapan tersebut dibagi menjadi tiga tahapan besar dalam proses pembuatan aplikasi katalog perpustakaan ini [6].



Gambar 1. Metode Penelitian

Pada Gambar 1, penerapan dukungan dalam menciptakan aplikasi katalog buku perpustakaan berbasis *website* melalui tiga tahap [7], antara lain tahap awal (fase *communicating* dan *planning*), tahap pengembangan (fase *modelling*), dan tahap akhir (fase *construction* dan *development*). Pada tahap awal terdiri dari observasi dan wawancara, mengidentifikasi masalah, dan mendapatkan analisis kebutuhan yang terdiri dari analisis kebutuhan pengguna, dan analisis kebutuhan fungsional. Tahapan dilanjutkan pada tahap pengembangan menggunakan alir sistem, DFD, CDM, PDM, dan struktur tabel. Dan tahapan terakhir yang merupakan tahap akhir dalam merancang aplikasi katalog buku terdiri dari kebutuhan sistem, implementasi aplikasi, dan hasil analisis dari implementasi aplikasi yang menggunakan *blackbox testing* [8].

Tahap *Communicating*

Pada tahap *communicating*, merupakan tahap awal memungkinkan pengguna untuk memahami proses bisnis dan mampu menggambarkan secara jelas apa yang dibutuhkan dalam pengembangan aplikasi. Tahapan *planning* sebuah rencana dalam pengerjaan aplikasi sesuai dengan permintaan pengguna yang mencakup teknis dikerjakan [9]. Tujuan dari tahapan ini adalah mendapatkan informasi sebanyak-banyaknya dari sistem atau aplikasi yang akan dibangun. Kegiatan yang dilakukan adalah melakukan Observasi dan wawancara kepada pihak-pihak yang menjadi pengguna utama aplikasi.

Proses observasi dilakukan dengan datang ke lokasi penelitian, untuk mengetahui secara langsung proses pengelolaan data, kendala pada pengelolaan data, serta permasalahan yang dialami. Dalam proses mendapatkan suatu informasi, maka diperlukannya wawancara dengan narasumber yang terpercaya. Bertujuan agar mengetahui proses data informasi yang telah didapatkan, dan menjadi data pendukung dalam menyelesaikan masalah.

Identifikasi Masalah

Setelah melalui proses observasi dan wawancara, didapatkan informasi penting untuk pengembangan aplikasi. Informasi yang didapatkan berhubungan dengan masalah yang dihadapi oleh pengunjung dan petugas perpustakaan. Berikut permasalahan yang teridentifikasi di perpustakaan salah satu Sekolah Menengah Pertama di Sidoarjo yang ditunjukkan pada Tabel 1.

Tabel 1. Identifikasi Masalah

Masalah	Dampak	Solusi
Pendataan buku yang terdapat di lokasi penelitian bersifat manual, yang dimana pendapatan buku masih disimpan buku besar dan beberapa tulisan dibuku besar mulai memudar dan susah untuk dibaca.	Proses tersebut membuat kinerja terhambat.	Membuat sebuah aplikasi pengelolaan buku yang mampu mempermudah petugas perpustakaan dalam membuat dan memeriksa buku yang ada.

Identifikasi Kebutuhan Aplikasi

Setelah menemukan permasalahan, tahap selanjutnya adalah dilakukan analisis pada kebutuhan aplikasi. Identifikasi kebutuhan ini terdiri dari kebutuhan pengguna, kebutuhan fungsional dan kebutuhan non-fungsional. Pada kebutuhan pengguna, analisis yang dilakukan menghasilkan kebutuhan fungsi yang digunakan pada aplikasi katalog yang diperlukan perpustakaan. Detil hasil analisis Tabel 2.

Tabel 2. Analisis Pengguna

Kebutuhan Fungsi	Kebutuhan Data	Kebutuhan Informasi
Mengelola master katalog	- Data Katalog - Data Kategori Buku - Data Kode Klasifikasi - Data Penerbit Buku.	- Daftar Katalog Buku - Daftar Kategori Buku - Daftar Informasi Klasifikasi - Daftar Penerbit

Kebutuhan fungsional menjelaskan secara terperinci mengenai kebutuhan pengguna yang telah diidentifikasikan. Proses ini berperan penting dalam menerapkan fungsi-fungsi yang diperoleh dari analisis kebutuhan pengguna yang sedang berlangsung [10]. Berdasarkan hasil analisis kebutuhan pengguna, maka hasil kebutuhan fungsional aplikasi ini meliputi fungsi *login*, fungsi *dashboard*, fungsi mengelola data buku, fungsi mengelola data kategori buku, fungsi mengelola data klasifikasi, fungsi mengelola data penerbit buku.

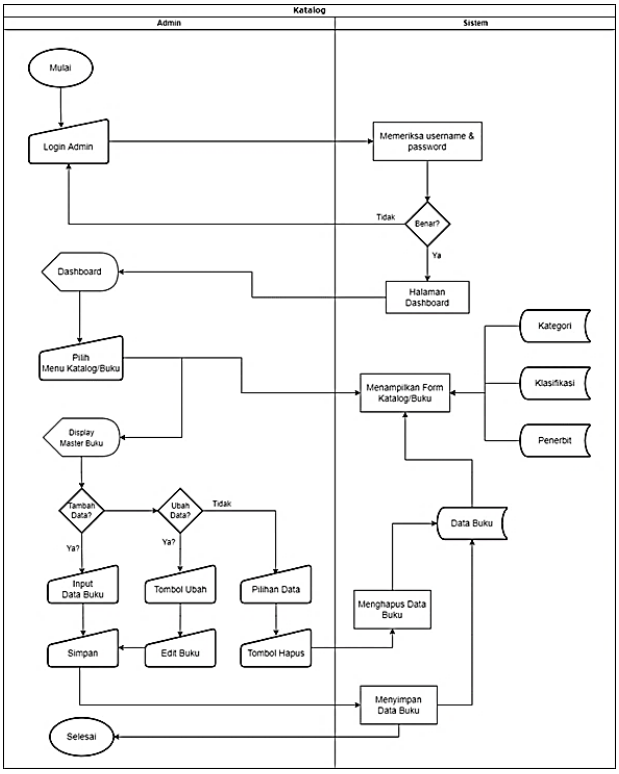
Kebutuhan non-fungsional merupakan kebutuhan yang mendukung jalannya aplikasi. Pada tahap ini berfokus dengan analisis kebutuhan non-fungsional pada sistem yang dirancang, dengan dua kebutuhan utama yang menjadi fokus:

a. Keamanan Sistem: melindungi data dengan autentikasi menggunakan *username* dan *password*.
b. Performa Cepat: mengakses aplikasi dalam waktu kurang dari 10 detik, simpan/ubah/hapus data dalam waktu kurang dari 10 detik.

Tahap *Modeling*

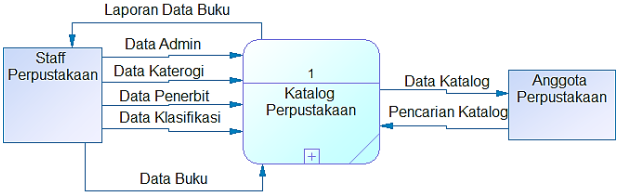
Pada tahap pemodelan ini dibuat berdasarkan proses bisnis yang terjadi pada perpustakaan. Setelah proses bisnis diperoleh, selanjutnya menggambarkan diagram konteks untuk mendapatkan gambaran data yang dibutuhkan oleh para pengguna aplikasi. Tahap ini dilengkapi juga dengan desain data yang digambarkan dengan Model Data Fisik (PDM).

Alir Sistem atau biasanya disebut *System Flow* merupakan bagan yang memperlihatkan tahapan proses sistem yang sedang dikembangkan [11]. Proses alir sistem berperan dalam menggambarkan proses sebuah aplikasi yang dikerjakan. Gambar 2. menjelaskan alur sistem pengelolaan katalog atau buku oleh admin. Pada alur tersebut mempunyai dua actor utama, yaitu siswa dan Admin. Proses ini menentukan koleksi yang akan ditambahkan apabila buku baru itu belum tersedia, menyimpan koleksi lama dari buku besar, ataupun mengubah koleksi yang sudah ada jika data yang disimpan terjadi kesalahan *input*.

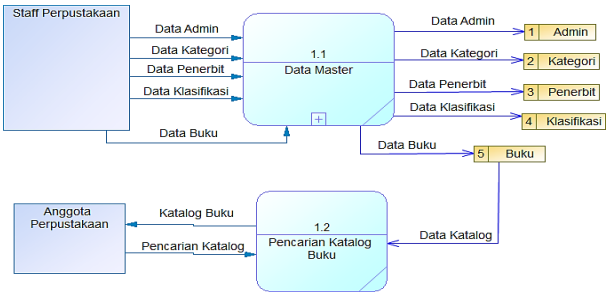


Gambar 2. Sistem *Flow* Katalog

Diagram alir data (DFD) memvisualkan pergerakan data dalam sistem saat beroperasi. Informasi terkait alir data ini dapat dilihat pada Gambar 3, Gambar 4, dan Gambar 5. yang dibuat untuk merancang aplikasi katalog perpustakaan. Gambar 3. menunjukkan diagram alir utama, sekaligus menjelaskan entitas eksternal yang berhubungan terkait rancangan aplikasi katalog. Entitas eksternal tersebut terdiri dari staf atau admin perpustakaan dan anggota perpustakaan.

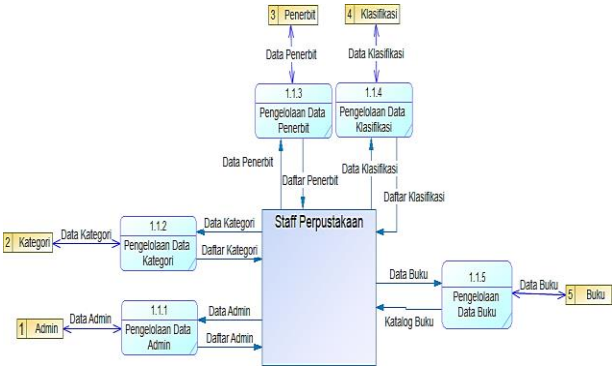


Gambar 3. Diagram Konteks Aplikasi Katalog

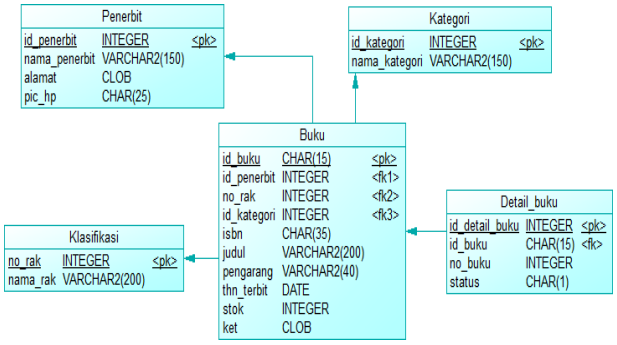


Gambar 4. DFD Level 0

Gambar 4. menampilkan diagram alir data (DFD) level 0 yang merupakan pecahan dari diagram konteks untuk menjelaskan detail fungsionalitas aplikasi katalog perpustakaan. DFD level 0 ini mengidentifikasi dua proses utama antara lain pengelolaan data master yang dilakukan staf perpustakaan dan pencarian katalog buku yang bisa diakses oleh anggota perpustakaan. Sedangkan pada Gambar 5. menjelaskan perincian lebih lanjut terkait pengelolaan data masternya, khususnya untuk data buku. Terdapat lima sub-proses yang teridentifikasi adalah pengelolaan data kategori, pengelolaan data penerbit pengelolaan data klasifikasi, pengelolaan data buku.



Gambar 5. DFD Level 1 Master



Gambar 6. PDM Katalog Perpustakaan

Desain data pada aplikasi ini merupakan desain database Diagram Hubungan Entitas (ERD) yang menggunakan diagram Model Data Konseptual (CDM) dan Model Data Fisik (PDM). Gambar 6. adalah gambaran Model Data Fisik (PDM) aplikasi katalog perpustakaan yang dihasilkan dari men-generate dari Model Data Konseptual (CMD). Dari hasil tersebut, menghasilkan sebuah tabel baru jika sudah menetapakn relasi ke masing-masing tabel. Hasil dari PDM terdapat lima tabel yang berhasil digenerate yaitu tabel buku, kategori, penerbit, klasifikasi, dan daftar buku.

HASIL DAN PEMBAHASAN

Tahap *Construct* dan *Deployment*

Tahap *Construct dan Deployment* merupakan tahap akhir dari proses rancang bangun aplikasi katalog perpustakaan. Tahap ini menjelaskan tujuan membuat perangkat lunak sesuai dengan spesifikasi yang telah dijabarkan pada proses sebelumnya. Tahapan ini berfungsi untuk memperkenalkan fitur-fitur yang terdapat pada aplikasi katalog perpustakaan.

Spesifikasi Sistem

Spesifikasi sistem berperan penting dalam menentukan kebutuhan perangkat lunak dan perangkat keras yang diperlukan untuk menjalankan dan membangun aplikasi katalog. Kebutuhan ini dirincikan dalam Tabel 3. dan Tabel 4. Menetapkan spesifikasi sistem ini sangat diperlukan aplikasi katalog dapat berjalan dengan baik.

Tabel 3. Perangkat Lunak

Aplikasi	Keterangan
XAMPP	*Web server local*
Visual Studio Code	*Text Editor*
MySQL	*Database Server*
Chrome	*Browser*
Windows 10	Sistem Operasi

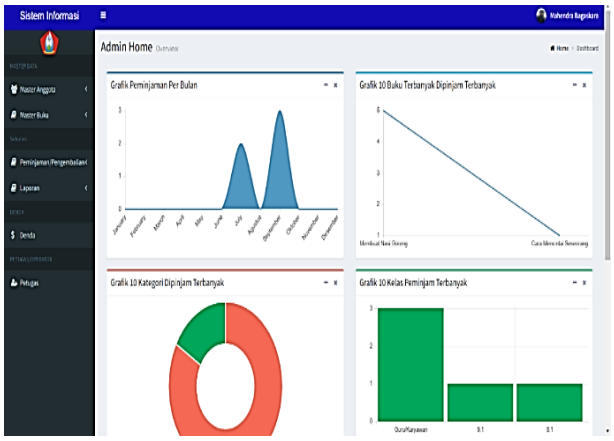
Tabel 4. Perangkat Keras

Komponen	Keterangan
Processor	Intel Core i3
RAM	2GB
Kapasitas Penyimpanan	500GB
Internet	Kecepatan min, 1 Mbps
I/O Devices	Monitor, Mouse, Keyboard

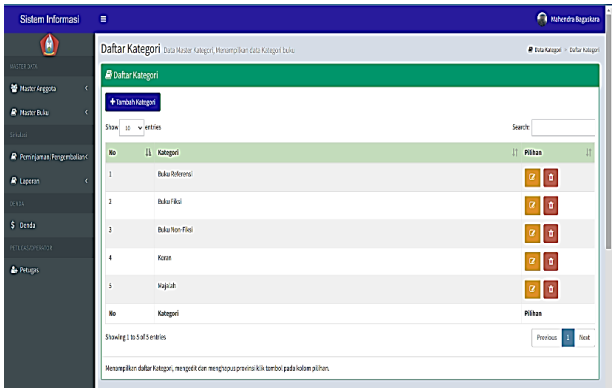
Implementasi Sistem

Implementasi sistem yang dimaksud merupakan gambaran aplikasi yang sudah tercipta dan dijelaskan dengan fungsi tiap-tiap tampilan aplikasi katalog perpustakaan. Setiap fungsi yang dibuat disesuaikan dengan kebutuhan informasi tiap pengguna. Terdapat lima halaman yang ditampilkan, yaitu halaman *Dashboard*, halaman Pengelolaan Kategori, Halaman Klasifikasi, Halaman Penerbit dan Halaman Katalog.

Tampilan pertama dari aplikasi ini adalah Dashboard. Halaman dashboard merupakan sebuah antarmuka yang memberikan informasi penting dan relevan dalam satu tampilan. Pada halaman dashboard, admin dapat melihat informasi tentang jumlah peminjaman per-bulan, 10 buku dengan peminjaman terbanyak, 10 kategori dengan peminjaman terbanyak, dan 10 kelas dengan peminjam terbanyak, yang terlihat pada Gambar 7.

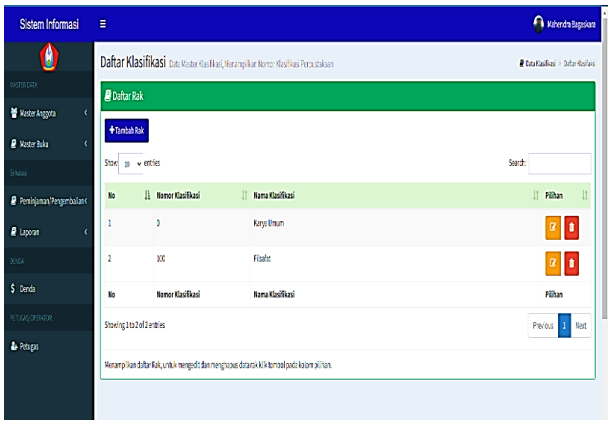


Gambar 7. Halaman Dashboard



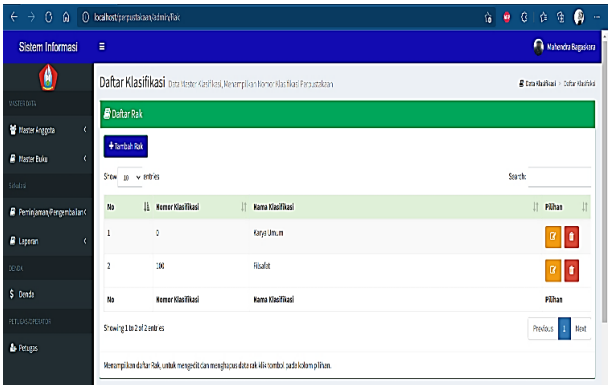
Gambar 8. Tampilan Pengelolaan Kategori

Tampilan berikutnya adalah Halaman Pengelolaan Kategori. Gambar 8. menunjukan halaman pengelolaan kategori, yang difungsikan khusus untuk mengatur jenis koleksi buku diperpustakaan. Di halaman ini, admin memiliki kewenangan penuh untuk menambahkan, mengedit, dan menghapus kategori buku sesuai kebutuhan.



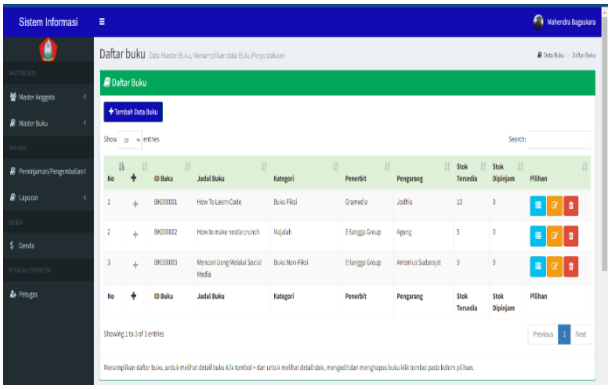
Gambar 9. Tampilan Klasifikasi

Tampilan berikut ini masih digunakan oleh Admin, yaitu Halaman Pengelolaan Nomor Klasifikasi. Halaman klasifikasi merupakan halaman mengelola nomor klasifikasi untuk mengorganisir buku ke dalam kategori numerik berdasarkan subjek dan topiknya. Pada halaman ini juga, admin dapat memasukkan, mengubah atau menghapus data klasifikasi buku. Halaman ini dapat dilihat pada Gambar 9.



Gambar 6. Tampilan Halaman Penerbit

Tampilan berikutnya, masih digunakan oleh Admin, yaitu Halaman Pengelolaan Penerbit. Halaman penerbit merupakan halaman mengelola penerbit untuk mengorganisasi buku berdasarkan masing-masing penerbit buku yang ada. Pada halaman ini juga, admin memiliki kewenangan penuh untuk memasukkan, mengubah ataupun menghapus data penerbit buku. Halaman ini dapat dilihat pada Gambar 10.



Gambar 7. Tampilan Halaman Katalog

Tampilan terakhir pada aplikasi ini yaitu Halaman Katalog. Halaman katalog merupakan sebuah halaman untuk mengelola katalog buku dengan menambahkan katalog dengan pengadaan buku baru ataupun melalui buku besar yang tersedia. Pada halaman ini, admin dapat memasukkan, mengubah ataupun menghapus data katalog buku. Halaman ini dapat dilihat pada Gambar 11.

Hasil Implementasi Aplikasi

Pengujian aplikasi katalog buku dilakukan dengan menguji setiap skenario dari jalannya aplikasi, proses mengelola data admin, hingga proses data *input* katalog buku yang sesuai berlanjut dengan data *input* yang salah. Proses pengujian aplikasi kali ini menggunakan metode *blackbox testing* [8]. Metode ini bertujuan untuk mengevaluasi apakah aplikasi dapat berfungsi dengan baik dan sesuai dengan skenario yang

telah ditentukan. Hasil pengujian aplikasi yang bisa dilihat pada Tabel 5, terlihat bahwa fungsi-fungsi yang terdapat dalam aplikasi berhasil dalam melakukan pengecekan dan memvalidasi data yang telah di masukkan. Sehingga aplikasi berjalan dan berfungsi dengan sesuai pengujian.

Tabel 5. Hasil Pengujian

No	Kasus Pengujian	Skenario	Hasil diharapkan	Hasil Uji
1	Menampilkan halaman *dashboard*	Melakukan *login*	Data yang di*input*kan sesuai dan berhasil masuk ke *dashboard*	Sukses
2	Tampilan data kategori buku	Memilih data kategori buku	Data kategori buku berhasil tampil	Sukses
3	Menambahkan, ubah, dan hapus kategori buku	Menekan tombol tambah, ubah, atau hapus	Data kategori berhasil ditambahakan, berhasil diubah, dan berhasil dihapus	Sukses
4	Tampilan data klasifikasi buku	Memilih data klasifikasi buku	Data klasifikasi buku berhasil tampil	Sukses
5	Menambahkan, ubah, dan hapus klasifikasi buku	Menekan tombol tambah, ubah, atau hapus	Data klasifikasi berhasil ditambahakan, berhasil diubah, dan berhasil dihapus	Sukses
6	Tampilan data penerbit buku	Memilih data penerbit buku	Data penerbit buku berhasil tampil	Sukses
7	Menambahkan, ubah, dan hapus penerbit buku	Menekan tombol tambah, ubah, atau hapus	Data penerbit berhasil ditambahkan, berhasil diubah, dan berhasil dihapus	Sukses
8	Tampilan data katalog buku	Memilih data katalog buku	Data katalog buku berhasil tampil	Sukses
9	Menambahkan, ubah, dan hapus katalog buku	Menekan tombol tambah, ubah, atau hapus	Data katalog berhasil ditambahkan, berhasil diubah, dan berhasil dihapus	Sukses

KESIMPULAN DAN SARAN

Kesimpulan yang didapatkan dari hasil analisis, pengembangan, dan mengimplementasikan aplikasi katalog perpustakaan sebagai berikut. Berdasarkan hasil dari implementasi dan hasil pengujian yang telah dilaksanakan, semua fungsi dari sistem pengelolaan katalog buku perpustakaan berfungsi dengan baik. Perancangan dan pembangunan aplikasi katalog buku perpustakaan berbasis *website* dilaksanakan dengan melakukan analisis masalah dan solusi, pembuatan desain sistem *flow*, diagram konteks, diagram alir data, dan diagram hubungan entitas. Dilanjutkan dengan pembuatan dan implementasi hasil analisa dan desain tersebut menjadi aplikasi.

Saran kepada pengembang aplikasi katalog buku di masa mendatang adalah: Fitur pada halaman master buku (katalog buku) dapat dikembangkan untuk menambahkan fungsi mencetak label buku untuk buku baru maupun buku yang sudah ada. Fitur pencarian katalog buku dapat ditambahkan gambar tiap judul yang ada, dan tampilan jika buku yang sedang dicari itu tidak tersedia ataupun belum ada.

Ucapan Terimakasih

Penulis berterima kasih kepada Program Studi Sistem Informasi Fakultas Teknologi dan Informatika Universitas Dinamika yang telah mengujikan untuk melakukan penelitian ini. Pihak Perpustakaan Sekolah Menengah Pertama yang telah memberikan ijin untuk melakukan penelitian rancang bangun aplikasi ini, dan penulis mengucapkan terima kasih kepada dosen pembimbing yang telah membimbing selama penelitian terlaksana.

DAFTAR PUSTAKA

[1] L. Evawani, “Perpustakaan Sebagai Sumber Belajar Di Madrasah,” *Jurnali Literasiologi Liska Evawani*, vol. 8, no. 1, pp. 136–143, 2022.

[2] P. N. Sari, C. Husadha, R. A. Haryanto, A. Andrian, E. T. Prasetyo, and I. Istianingsih, “Perpustakaan Desa Terhadap Minat Baca Lingkungan Desa Muara Bakti, Kabupaten Bekasi,” *Jurnal Pengabdian kepada Masyarakat UBJ*, vol. 4, no. 1, pp. 17–26, Jan. 2021, doi: 10.31599/jabdimas.v4i1.199.

[3] S. Samsiana, S. Supratno, A. Hasad, S. Marini, R. Sylviana, and Taufiqurahman, “Pelatihan dan Konfigurasi Sistem Informasi Perpustakaan pada Sekolah GRATIS Yayasan Baitul Jihad Bekasi,”

Journal Of Computer Science Contributions (JUCOSCO), vol. 1, no. 1, pp. 95–102, Jan. 2021, doi: 10.31599/jucosco.v1i1.1563.

[4] S. Pratama, E. Karyadiputra, I. Kalimantan, M. Arsyad, and A. Banjari, "Rancang Bangun Sistem Informasi Perpustakaan Berbasis Website Pada SMPN 1 Kertak Hanyar," *Technologia*, vol. 10, no. 2, pp. 68–76, Jun. 2019.

[5] Soetedjo and R. Sidik, "Pengembangan Sistem Informasi Manajemen Layanan Perpustakaan SMK Merdeka Bandung," *Jurnal Teknologi dan Informasi*, vol. 9, no. 2, pp. 115–127, Aug. 2019, doi: 10.34010/jati.v9i2.1793.

[6] H. Deanna Durbin and A. Feni, "Rancang Bangunsistem Informasi Perpustakaanberbasis Webpada Smk Citra Negara Depok," *Jurnal Rekayasa Informasi*, vol. 7, no. 1, pp. 13–22, Apr. 2018, Accessed: Oct. 29, 2024. [Online]. Available: https://ejournal.istn.ac.id/index.php/rekayasainformasi/article/view/272/225

[7] Kartubi and R. W. Arifin, "Sistem Informasi Perpustakaan Berbasis Website Dengan Framework Laravel," *Jurnal Mahasiswa Bina Insani*, vol. 3, no. 2, pp. 213–222, Feb. 2019.

[8] W. Harjono and Kristianus Jago Tute, "Perancangan Sistem Informasi Perpustakaan Berbasis Web Menggunakkan Metode Waterfall," *SATESI: Jurnal Sains Teknologi dan Sistem Informasi*, vol. 2, no. 1, pp. 47–51, Apr. 2022, doi: 10.54259/satesi.v2i1.773.

[9] Sommerville, Software engineering (10th edition). Pearson Education Limited, 2016.

[10] R. S. Pressman, Rekayasa Perangkat Lunak: Pendekatan Praktisi Buku 1. Yogyakarta: Andi, 2015.

[11] Y. D. Wijaya and M. W. Astuti, "Pengujian Blackbox Sistem Informasi Penilaian Kinerja Karyawan Pt Inka (Persero) Berbasis Equivalence Partitions," *Jurnal Digital Teknologi Informasi*, vol. 4, no. 1, p. 22, Mar. 2021, doi: 10.32502/digital.v4i1.3163.

[12] N. Mariyus, N. Purwati, and R. A. Aziz, "Aplikasi Pengolahan Data Puskesmas (Pusat Kesehatan Masyarakat) Desa Margodadi Kab. Tulang Bawang Barat," *SIMADA (Jurnal Sistem Informasi dan Manajemen Basis Data)*, vol. 2, no. 1, pp. 15–25, Jun. 2019, doi: 10.30873/simada.v2i1.1338.

[13] Y. Afrilliai and R. Ramadani, "Penerapan Sistem Informasi Pencarian Tata Letak Buku Pada Perpustakaan Fakultas Ilmu Sosial Dan Ilmu Politik Universitas Malikussaleh," *SAINTEK: Jurnal Sains dan Teknologi*, vol. 2, no. 2, pp. 22–27, Mar. 2021.

[14] R. Zaelani, R. Rakhazona Pamungkas, M. Ramdani, B. Ikrar Bhakti, and N. Ratama, "Pengembangan Sistem Informasi Manajemen Sekolah Berbasis Web," Innovative: *Journal Of Social Science Research*, vol. 3, no. 6, pp. 7040–7048, 2023.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, Bulan Oktober Tahun 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.694

Identifikasi Pengenalan Wajah Berdasarkan Jenis Kelamin Menggunakan Metode *Convolutional Neural Network* (CNN)

Natasya Chalista Imanuela Natun[1], **Maria Angelica Santhia**[2], **Yampi R. Kaesmetan**[3]
[1,2,3]Program Studi/Jurusan Teknik Informatika, STIKOM Uyelindo, Kupang, Indonesia
e-mail: natunnatasya@gmail.com[1], angelsanthia15509@gmail.com[2], kaesmetanyampi@gmail.com[3]
* Penulis Korespondensi: E-mail: admin@siamiruyelindo.ac.id

Abstrak: Wajah merupakan komponen yang paling mudah dikenali dan sering kali menjadi pusat perhatian dalam tubuh manusia. Sering terjadinya kesulitan dalam membedakan dan menganalisis citra wajah dengan jumlah yang banyak secara manual karena banyaknya kemiripan antara laki-laki dan perempuan sehingga memperlambat proses identifikasi jenis kelamin. Penelitian ini bertujuan untuk meningkatkan akurasi dan mengimplementasikan identifikasi jenis kelamin dalam pengenalan wajah melalui penerapan metode *Convolutional Neural Network (CNN)*, sebuah algoritma *deep learning* yang efektif untuk deteksi objek. Dengan mengumpulkan dan memproses dataset citra wajah yang dilabeli berdasarkan jenis kelamin, serta melakukan pra-pemrosesan data yang cermat, model CNN dilatih dan diuji untuk mencapai tingkat akurasi yang signifikan dalam identifikasi jenis kelamin. Hasil penelitian menunjukkan bahwa metode ini dapat meningkatkan kinerja sistem pengenalan wajah berdasarkan jenis kelamin secara praktis. Selain itu, penelitian ini juga memberikan rekomendasi untuk meningkatkan akurasi lebih lanjut dan eksplorasi teknologi terkini dalam upaya mengoptimalkan aplikasi pengenalan gender di masa depan. Abstrak ini merangkum kontribusi utama dan keunikan dari penelitian serta fokus pada elemen-elemen penting seperti penggunaan CNN, pengumpulan dan pelabelan dataset, serta potensi implementasi dan rekomendasi pengembangan lebih lanjut. Meskipun banyak penelitian menggunakan metode CNN, artikel ini mungkin menggunakan versi atau arsitektur CNN yang lebih baru dan dioptimalkan, yang meningkatkan kinerja dalam tugas pengenalan jenis kelamin.

Kata Kunci: CNN; Deteksi Objek; Pengenalan Wajah

Abstract: *The face is the most easily recognized component and is often the center of attention in the human body. It is often difficult to differentiate and analyze large numbers of facial images manually because of the many similarities between men and women, which slows down the gender identification process. This research aims to increase accuracy and implement gender identification in facial recognition through the application of the Convolutional Neural Network (CNN) method, an effective deep learning algorithm for object detection. By collecting and processing a dataset of facial images labeled by gender, as well as performing careful pre-processing of the data, the CNN model was trained and tested to achieve a significant level of accuracy in gender identification. The research results show that this method can practically improve the performance of facial recognition systems based on gender. In addition, this research also provides recommendations for further improving accuracy and exploring the latest technologies in an effort to optimize gender recognition applications in the future. This abstract summarizes the main contributions and uniqueness of the research and focuses on important elements such as the use of CNNs, collection and labeling dataset, as well as potential implementation and recommendations for further development. Although many studies use CNN methods, this article may use a newer, optimized version or architecture of CNN, which improves performance in the gender recognition task.*

Keywords: *CNN; Object Detection; Facial Recognition*

PENDAHULUAN

Wajah sebagai bagian dari tubuh manusia yang paling mudah dikenali dan paling sering dilihat oleh orang lain [1]. Wajah juga merupakan sarana penting dalam berkomunikasi dan mengekspresikan emosi. Dunia teknologi, mengalami perkembangan yang semakin hari semakin canggih, sehingga wajah sering digunakan sebagai identifikasi individu dalam berbagai aplikasi seperti pengenalan wajah .

Dalam beberapa waktu terakhir sistem pengenalan, Identifikasi mulai banyak diciptakan dan dikembangkan sebagai bahan dan menjadi bagian dari sebuah sistem keamanan seperti deteksi wajah yang merupakan tahapan awal dalam melakukan pengenalan wajah [1]. Pada bidang keamanan *Face ID* dirancang dengan tingkat keamanan tinggi, dengan pembobolan yang sangat rendah sehingga kemungkinan seseorang memiliki wajah serupa yang dapat membuka kunci perangkat adalah sangat kecil. Seiring dengan kebutuhan keamanan dan pemantauan yang semakin meningkat, permasalahan utama yang dihadapi yaitu kompleksitas ciri-ciri wajah yang berkaitan dengan jenis kelamin. Dengan melakukan pengambilan data wajah maka dapat diperoleh informasi penting yang dimiliki oleh setiap manusia mencakup bentuk area dari dahi hingga dagu, hidung, mata, mulut, dan kulit di sekitarnya. Oleh karena itu, data wajah menjadi elemen krusial yang dapat digunakan untuk mendeteksi jenis kelamin [2].

Convolutional Neural Network (CNN) merupakan salah satu algoritma *deep learning* yang terdiri dari *neuron 3D* yang tersusun dalam lapisan-lapisan [3]. *Neuron* ini memiliki tiga dimensi, yaitu lebar, tinggi dan kedalaman. Proses utama pada *CNN*

merupakan konvolusi ,menggunakan lapisan konvolusi untuk memproses *input* citra dengan mengambil informasi dari area kecil dalam gambar dan menghitung konvolusi antara area tersebut dan filter atau *kernel* yang digeser melintasi citra untuk menghasilkan fitur -fitur tertentu [4].

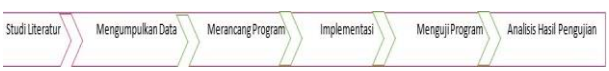
Arsitektur CNN yang digunakan dalam penelitian ini dirancang untuk mengidentifikasi jenis kelamin dari citra wajah dengan resolusi 50x50 piksel, baik dalam format *grayscale* maupun RGB. Pada lapisan *input*, citra wajah ini dimasukkan ke dalam jaringan. Selanjutnya, lapisan konvolusi pertama memiliki 32 filter dengan ukuran 3x3 dan menggunakan fungsi aktivasi ReLU (Rectified Linear Unit) untuk mengekstraksi fitur-fitur dasar seperti tepi dan tekstur [5]. Lapisan konvolusi kedua memiliki 64 filter dengan ukuran 3x3 dan juga menggunakan fungsi aktivasi ReLU untuk mendeteksi fitur yang lebih kompleks. Setelah lapisan konvolusi, lapisan *pooling* digunakan untuk mengurangi dimensi fitur map dan mencegah *overfitting*. Penggunaan arsitektur ini memiliki beberapa alasan kuat. Pertama, CNN memiliki kemampuan luar biasa dalam mengekstraksi fitur visual penting dari citra wajah, seperti tepi, sudut, dan tekstur, yang sangat esensial dalam membedakan wajah laki-laki dan perempuan [6][7]. Kedua, CNN secara otomatis belajar fitur-fitur relevan selama proses pelatihan, sehingga mengurangi kebutuhan akan rekayasa fitur manual. Ketiga, CNN menunjukkan ketahanan yang tinggi terhadap variasi dalam citra, seperti perubahan pencahayaan, pose, dan ekspresi wajah, yang sering menjadi tantangan dalam pengenalan wajah.

Tujuan utama dari Identifikasi ini untuk meningkatkan akurasi dan mengimplementasikan identifikasi jenis kelamin pada pengenalan wajah melalui penerapan metode *Convolutional Neural Network.* Dengan demikian, identifikasi ini diarahkan untuk memberikan kontribusi pada perkembangan sistem pengenalan wajah yang lebih canggih dan handal, memperluas pemahaman terhadap efektivitas metode *CNN* dalam identifikasi jenis kelamin [8].

Hasil dari identifikasi dengan memberikan kontribusi yang signifikan dengan menciptakan model identifikasi jenis kelamin berbasis wajah yang unggul dalam akurasi dan keandalan. Selain itu, dapat diimplementasikan secara praktis dalam berbagai konteks aplikasi, terutama pada sistem keamanan dan pemantauan.

METODE PENELITIAN

Berikut ini beberapa tahapan-tahapan yang dilakukan untuk mengklasifikasikan wajah menggunakan *Convolutional Neural Network* untuk mengetahui jenis kelamin dilihat pada Gambar 1.



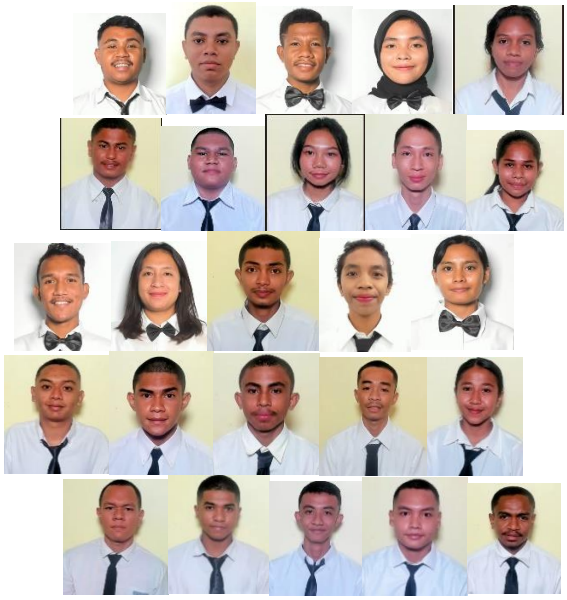
Gambar 1. Metode Penelitian

Studi Literatur

Pada tahap ini akan melakukan pembelajaran dengan membaca dan mempelajari jurnal-jurnal yang terkait dengan topik klasifikasi citra wajah untuk pengenalan jenis kelamin menggunakan metode *Convolutional Neural Network.*

Mengumpulkan Data

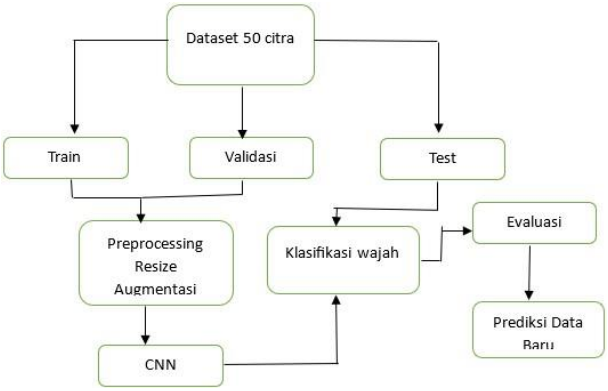
Data yang digunakan penulis merupakan data wajah yang berasal dari *Website* Siamir Stikom Uyelindo Kupang. Setiap gambar dianotasi dengan jenis kelamin sehingga, data yang akan dikumpulkan berupa citra wajah yang akan di ambil dari *Siamir* yang dibuat oleh *angel* dengan nama *gender-detection-face*. Data tersebut memiliki 50 citra wajah yang telah terbagi menjadi 70% data latih,15% data uji, data 15% data validasi.



Gambar 2. Pengumpulan Dataset Citra Wajah Dengan Nama Folder Foto Mahasiswa di Siamir

Merancang Program

Pada tahap ini penulis akan menyusun program yang terstruktur dan sistematis berdasarkan tujuan dari identifikasi dengan bahasa pemrograman *python* dan dari CNN serta untuk klasifikasi citra yang dapat dilihat pada Gambar 2.



Gambar 3. Perancangan Sistem

Dataset wajah yang terdiri dari 50 citra yang telah dibagi menjadi 70% data latih,15% data uji, dan 15% data validasi. Setelah terbagi jumlah *dataset* menjadi 35 data latih, 8 data uji dan 8 data validasi.

Tabel 1. Rincian Dataset

Dataset	Jumlah Citra	Laki-laki	Perempuan
Data Latih	49	25	24
Data Uji	2	1	1
Data Validasi	2	1	1

Preprocessing merupakan proses pengolahan pada data asli sebelum data dilakukan proses ditahap selanjutnya. Fungsi dari *preprocessing* citra adalah untuk memperjelas fitur dari data citra, menentukan bagian citra yang akan diobservasi agar proses di tahap selanjutnya dapat berjalan lancar [9]. *Preprocessing* citra diperlukan sebelum dikakukannya proses klasifikasi *CNN* [10]. Setelah memperoleh *dataset*, langkah berikutnya melibatkan tahapan *preprocessing* seperti perataan, ukuran dan normalisasi gambar wajah, hal ini perlu dilakukan agar sesuai dengan kebutuhan [11][12]. Tahap *preprocessing* dimulai dengan melakukan *cropping*, *resize* ukuran 50 x 50, dan konversi ke *Teachable Machine*. Selanjutnya dilakukan proses *augmentasi* data untuk meningkatkan jumlah sampel data [13][14].

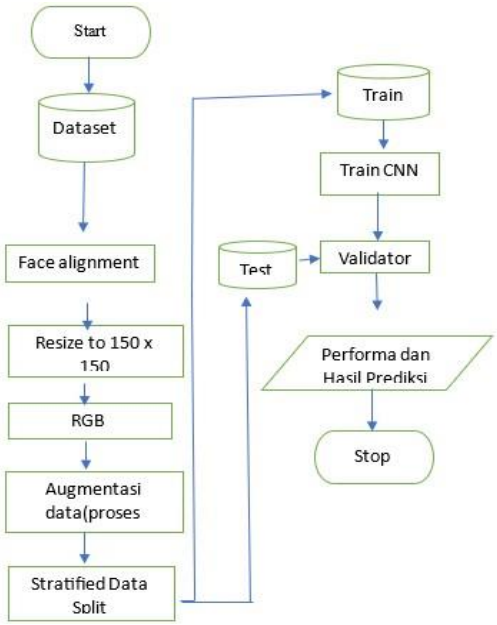
Pelatihan model menggunakan data pelatihan. Model diberikan *input* gambar dan label yang sesuai, dan menggunakan algoritma Adam untuk mengoptimalkan bobot dan bias model agar dapat mempelajari pola dan fitur yang relevan.

Evaluasi performa model menggunakan data pengujian yang tidak pernah dilihat sebelumnya yaitu menggunakan data *test*. Hal ini dilakukan untuk mengukur *accuracy*, *precision*, *recall* dan metrik evaluasi lainnya untuk mengevaluasi sejauh mana model dapat melakukan klasifikasi yang akurat.

Prediksi data baru melibatkan penggunaan model yang telah dibangun untuk melakukan prediksi pada citra wajah yang belum pernah dilihat sebelumnya menggunakan *library* sebagai *interface*, *output* yang akan dihasilkan terdapat persen antara laki-laki dan perempuan.

Implementasi

Pada tahap implementasi, dilakukan perancangan dan pengujian sistem yang digunakan untuk memprediksi jenis kelamin menggunakan metode CNN. Alur kerja program melibatkan beberapa proses. Pertama, dataset akan diproses terlebih dahulu melalui tahap *preprocessing*. Selain itu, dataset akan dibagi menjadi data latih dan data uji dengan perbandingan 70:15. Dataset latih akan digunakan untuk menguji kemampuan CNN dalam mengenali citra wajah, atau gambar untuk membedakan wajah laki-laki atau perempuan menggunakan algoritma CNN.



Gambar 4. Alur Sistem

Menguji Program

Tahap ini program akan di uji menggunakan data set yang telah disediakan sebelumnya yaitu pada folder tes, program akan membandingkan citra wajah laki-laki atau perempuan dalam *Teachable Machine* dan *Visual studio code.*

Analisis Hasil Pengujian

Hasil yang telah di uji coba oleh program akan menghasilkan akurasi, presisi, dan *recall*. Hasil tersebut didapatkan dengan metode *confusion matrix*.*Acurracy* merupakan metrik evaluasi yang digunakan untuk mengukur seberapa akurat model dalam memprediksi label kelas pada data uji. klasifikasi jenis kelamin, kelas yang digunakan dalam *confusion matrix* hanya terdiri dari dua kelas, yaitu laki-laki dan perempuan. Jika terdapat lebih dari dua kelas, seperti disebutkan dalam pertanyaan (laki-laki, perempuan 1, dan perempuan 2), maka hal ini menimbulkan kebingungan karena tidak sesuai dengan tujuan klasifikasi gender yang biasanya hanya membedakan antara dua kelas utama. bahwa hanya ada dua kelas dalam *confusion matrix*: laki-laki dan perempuan. Dalam konteks ini, jika dikatakan bahwa *actual*-prediksi laki-laki = 9, ini berarti terdapat 9 foto laki-laki yang diidentifikasi dengan benar oleh model sebagai laki-laki (*True Positive*).*Precision* merupakan metrik evaluasi yang digunakan untuk mengukur seberapa presisi model dalam memprediksi label kelas positif. *Recall* mengukur kemampuan model untuk mengidentifikasi *instance* positif dengan benar. Rumus akurasi, presisi, dan *recall* yang dapat dilihat pada :

Di sini, *True Positive* (TP) adalah jumlah sampel yang secara benar diprediksi sebagai pria, *False Negative* (FN) adalah jumlah sampel pria yang salah diklasifikasikan sebagai wanita, *False Positive* (FP) adalah jumlah sampel wanita yang salah diklasifikasikan

sebagai pria, dan *True Negative* (TN) adalah jumlah sampel yang benar diklasifikasikan sebagai wanita.

	Prediksi Pria	Prediksi Wanita
Aktual Pria	*True Positive*	*False Negative*
Aktual Wanita	*False Positive*	*True Negative*

Akurasi =Total Sampel/Prediksi Benar

Presisi=*True PositiveTrue Positive+False Positive*Presisi =*True Positive+False PositiveTrue Positive*

Recall=True PositiveTrue Positive+False NegativeRecall=True positive+False NegativeTrue Positive

HASIL DAN PEMBAHASAN

Training Data

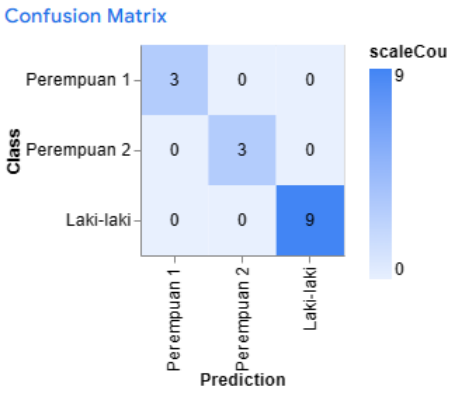
Data pada identifikasi ini terdapat beberapa tahapan yang terdiri dari sebuah *dataset*. Pada tahap *face recognition* memerlukan beberapa rangkap data guna melaksanakan sebuah mekanisme *data training* pada citra. Obyek yang diperlukan untuk melaksanakan mekanisme identifikasi wajah dipecah menjadi beberapa poin ,yakni *data training* ,*data testing* ,dan juga *data evaluation*. Pelatihan pada data diperuntukkan untuk melatih sebuah sistem guna merekam besarnya data pada sistem yang akan dibuat. Pada data *testing* yakni sebuah data yang diperuntukkan untuk sebuah parameter pertimbangan dengan *data training* yang dilatih. Pada proses *training data* ,diambil 50 foto. Berikut lampiran dari *dataset* yang dibagi menjadi 2 bagian untuk klasifikasi jenis kelamin wanita dan pria pada wajah ,data yang telah diklasifikasikan *dataset* pada *teachable machine* dan membuat skrip pada *visual studio code* untuk mendeteksi jenis kelamin, seperti Gambar 5,6 dan 7.

Gambar 5,6 dan 7. Dataset Perempuan dan Laki-laki

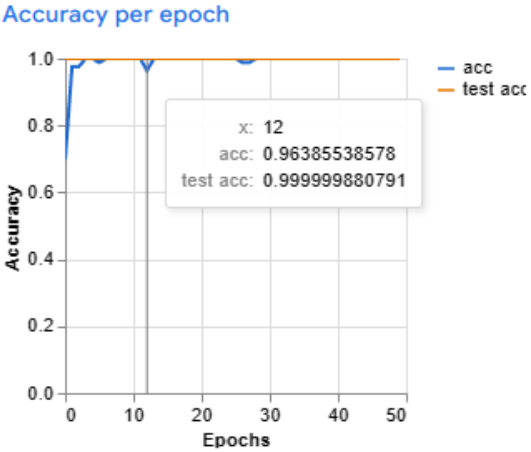
Accuracy per class

CLASS	ACCURACY	# SAMPLES
Perempuan 1	1.00	3
Perempuan 2	1.00	3
Laki-laki	1.00	9

Gambar 8. *Accuracy per class*



Gambar 9 .*Confucion Matrix* Evaluasi *Teachable Machine*



Gambar 10 .Hasil Pelatihan per *Epoch*

Dengan memanfaatkan optimasi Adam dan *learning rate* 0.0001, model akan melakukan penyesuaian pada parameter-parameter dengan tujuan untuk meminimalkan fungsi kerugian pada setiap iterasi pelatihan[].fungsi softmax dengan jumlah 2 kelas yaitu perempuan dan laki-laki ,untukkerugian yang digunakan dalam pelatihan yaitu *categorial_crossentropy.*

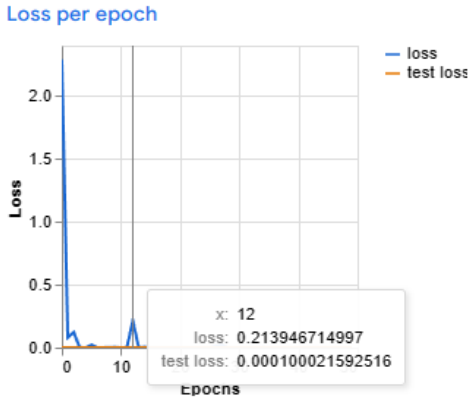
Fungsi *categorical_crossentropy* menghitung kerugian atau perbedaan antara distribusi probabilitas prediksi yang dihasilkan oleh model dan distribusi probabilitas target yang sebenarnya. Secara matematis, fungsi *categorical_crossentropy* dapat dijelaskan sebagai berikut. Jika y adalah distribusi probabilitas yang sebenarnya dan y_hat adalah distribusi probabilitas prediksi, maka fungsi

categorical_crossentropy didefinisikan J(y,y_hat)=−Σ(y∗log(y_hat))

Dalam fungsi ini, algoritma dari probabilitas prediksi yang benar (y_hat) digunakan untuk mengukur perbedaan antara distribusi probabilitas prediksi dan distribusi probabilitas yang sebenarnya. Hasil perkalian tersebut kemudian dijumlahkan dan dikalikan dengan -1 untuk menghasilkan nilai kerugian yang positif.

Hasil pengujian yang telah dilakukan terhadap model *Convolutional Neural Network(CNN)* yang diimplementasikan merupakan evaluasi terhadap kinerja dan efektivitas model dalam menjalankan tugas yang telah ditentukan.

Selama proses pelatihan, model akan melewati serangkaian *epoch*, di mana parameter dan bobot model akan disesuaikan berdasarkan kesalahan atau kerugian yang dihasilkan pada setiap iterasi. Hasil akurasi dan *loss* pada pelatihan tersebut dapat dilihat pada Gambar 11.



Gambar 11. *Loss* Pelatihan Model

Accuracy merupakan persentase prediksi yang benar dari keseluruhan data pelatihan. Pada setiap iterasi, akurasi meningkat dari 0.9638 hingga mencapai 0.9999 pada iterasi ke-12. Hal ini menunjukkan bahwa model secara bertahap meningkatkan kemampuannya dalam memprediksi dengan benar pada data pelatihan. *acc* merupakan persentase prediksi yang benar dari data validasi yang tidak digunakan dalam pelatihan model. Hal ini menunjukkan bahwa model secara bertahap memperbaiki kemampuannya dalam melakukan prediksi yang akurat pada data validasi. Kedua metrik evaluasi menunjukkan peningkatan secara keseluruhan seiring dengan meningkatnya jumlah iterasi. Terdapat beberapa fluktuasi di beberapa iterasi, di mana akurasi atau akurasi validasi mungkin mengalami penurunan sementara sebelum kembali meningkat. Namun, secara umum, terlihat adanya tren peningkatan yang stabil.

Dalam grafik *loss*, terlihat bahwa nilai kerugian atau kesalahan secara keseluruhan menurun seiring dengan bertambahnya iterasi. Hal ini menunjukkan bahwa model semakin baik dalam meminimalkan kesalahan atau kerugian dalam proses pelatihan. Sementara itu, *loss* juga menunjukkan tren penurunan yang stabil, yang menunjukkan bahwa model cenderung menghasilkan prediksi yang lebih akurat pada data validasi seiring dengan bertambahnya iterasi. Pada model dilakukan menggunakan beberapa metrik, seperti akurasi (*accuracy*), presisi (*precision*), dan *recall*.

Dalam kelas laki-laki, presisi sebesar 0.9999 mengindikasikan bahwa 99% dari prediksi yang diklasifikasikan sebagai kelas laki-laki adalah benar. *Recall* sebesar 0.99 menunjukkan bahwa 99% dari data yang sebenarnya merupakan kelas laki-laki dapat diidentifikasi dengan benar oleh model.

Sementara itu, dalam kelas perempuan, presisi sebesar 0.99 menunjukkan bahwa 99% dari prediksi yang diklasifikasikan sebagai kelas *woman* adalah benar. *Recall* sebesar 0.97 menunjukkan bahwa 97% dari data yang sebenarnya merupakan kelas *woman* dapat diidentifikasi dengan benar oleh model. Penjelasan tersebut dapat kita lihat pada Tabel 2.

Tabel 2.Evaluasi Performa *Teachable Machine* pada wajah

	Precission	Recall	Support
0	0.99	0.99	1
1	0.99	0.97	1
Accuracy			2

Face Recognition

Pada tahap pembentukan sebuah alur *face recognition* yang tersusun atas beberapa mekanisme pada saat pengutipan citra yang diaplikasikan oleh kamera hingga saat identifikasi dan klasifikasi pada citra. Pada saat melakukan *preprocessing* pemrosesan data yakni gambar asli dini pada data sebelum diproses oleh Algoritma *Convolutional Neural Network*. Tahap ini mempunyai berbagai macam cara dengan mengubah citra pada gambar berwarna atau biasa dikenal dengan sebutan *(RBG) Red, Blue, dan Green* menjadi hitam putih atau biasa dikenal dengan *greyscale*, pada tahap *preprocessing* biasanya mempunyai beberapa hasil akhir dalam mekanisme seperti menurunkan sampai dengan melesapkan sinyal gangguan, melakukan perubahan pada citra berdasarkan data asli kemudian menjadi selaras yang diperlukan. Tahap selanjutnya merupakan klasifikasi, yakni yang berperan sebagai penghimpunan sebuah citra yang sudah melalui tahap ekstraksi. Sesudah dilakukan ekstraksi pada sebuah citra, bahwa data dihimpunkan beberapa himpunan wajah untuk bisa mendeteksi pria atau wanita.

Implementasi Metode CNN dan Uji Coba Program

Pada tahap ini identifikasi memaparkan mekanisme sebuah *source code* dengan menggunakan sebuah *text editor* atau *code editor* yang bernama *Visual Studio Code*. Kemudian pada mekanisme implementasi metode *Convolutional Neural Network* ini menggunakan sebuah *library* dari Bahasa pemrograman *python* yang bernama *Tensorflow* dan *Keras*. *Library*

tersebut membantu mempermudah pengembang dikarenakan lebih mudah mereplika jaringan tiruan syaraf pada obyek.

Pseudocode untuk pelatihan model deteksi jenis kelamin menggunakan *Convolutional Neural Network* (CNN):

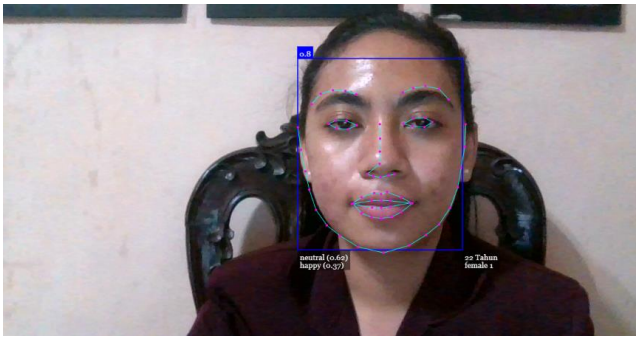
1. Import library yang diperlukan:
   - TensorFlow (ImageDataGenerator, *optimizers*, *layers*, *backend*)
   - sklearn untuk train_test_split
   - matplotlib.pyplot untuk visualisasi
   - *numpy* untuk manipulasi data *array*
   - cv2 untuk manipulasi gambar
   - os dan glob untuk manajemen file
2. Inisialisasi parameter awal:
   - epochs = 100
   - lr (learning rate) = 1e-3
   - batch_size = 64
   - img_dims = (96, 96, 3) # Dimensi gambar (tinggi, lebar, *channel*)
3. Muat *dataset* gambar:
   - data = []
   - labels = []
   - image_files = Ambil semua *path file* gambar dari dataset menggunakan glob
4. *Loop* untuk setiap gambar di image_files:
   - Baca gambar menggunakan cv2.imread()
   - *Resize* gambar ke ukuran img_dims menggunakan cv2.resize()
   - Ubah gambar menjadi *array* menggunakan img_to_array()
   - Tambahkan array gambar ke dalam *list* data
   - Ambil label dari *path* gambar (label "woman" -> 1, label lainnya -> 0)
   - Tambahkan label ke dalam *list labels* dalam format [[label]]
5. *Pre-processing*:
   - Konversi *list* data dan *labels* menjadi *numpy array*
   - Normalisasi data gambar ke rentang [0, 1]
6. Bagi *dataset* menjadi data latih dan data uji menggunakan train_test_split:
   - trainX, testX, trainY, testY = train_test_split(data, labels, test_size=0.2, random_state=42)
7. Augmentasi dataset:
   - Buat ImageDataGenerator untuk augmentasi gambar (rotasi, shift, zoom, dll)
8. Bangun model CNN:
   - Definisi fungsi *build*(width, height, depth, classes) untuk membuat model *Sequential*
   - Tambahkan layer Conv2D, *Activation*, *BatchNormalization*, MaxPooling2D, *Dropout*, dan *Dense*
   - Gunakan *Activation* "*sigmoid*" untuk *output* karena klasifikasi biner
9. Kompilasi model:
   - *Optimizer* menggunakan Adam dengan *learning rate* lr
   - *Loss function* menggunakan "binary_crossentropy"
   - Metrik evaluasi menggunakan "*accuracy*"
10. Latih model menggunakan fit_generator:
   - Gunakan augmentasi data generator untuk data latih
   - Validasi pada data uji
   - Tentukan steps_per_epoch berdasarkan panjang data latih dibagi batch_size
   - Lakukan pelatihan selama epochs
11. Simpan model hasil pelatihan ke dalam file "gender_detection.model"
12. Visualisasi training/validation loss dan accuracy:
   - Plot menggunakan matplotlib.pyplot
   - Simpan plot sebagai "plot.png"

Sesudah mekanisme pengaturan alokasi penempatan direktori, tahapan berikutnya merupakan menyediakan beberapa *dataset* yang telah kita alokasikan untuk data *training* dan data testing dengan tingkat akurasi yang ditentukan oleh *tensorflow* dan juga keras. Pada pengoperasian tahap ini mengaplikasikan sebuah fungsi yang bernama ImageDataGenerator yang terdapat di dalam *library Tensorflow* itu sendiri.

Pada tahap berikutnya melakukan sebuah tahap mekanisme dengan menjalankan data *training* dan *accuracy* serta testing yang telah dibagi menjadi beberapa tahap bagian. Pada saat data selesai dilatih maka diperolehlah sebuah hasil seberapa tepat akurasi dan tingkat kegagalan pada sebuah sistem tersebut yang dapat dilihat pada Gambar 15. Pengujian program menggunakan beberapa foto maupun deteksi secara langsung di layar untuk mengecek apakah prediksi deteksi jenis kelamin sudah sesuai atau belum dengan wajah yang dideteksi.

Pengujian Sistem

Pengujian Setelah proses mekanisme perancangan serta implementasi tahap *source code* telah dilewati, pada tahap ini akan menunjukkan sebuah pengoperasian berhasil atau tidaknya sebuah sistem yang telah di implementasikan dengan metode *Convolutional Neural Network* dengan menjalankan model kamera pada *library* yang terdapat di dalam Tensorflow dan juga Bahasa pemrograman *Python* tersebut. Pada saat kamera diaktifkan maka sistem akan menangkap hasil *recording* kemudian di *capture* menjadi gambar yang terdapat sebuah kernel seperti Gambar 15. Kernel serta *frame* yang dibuat akan otomatis menangkap hasil rekaman layar kamera dan juga identifikasinya.



Gambar 15. Pengujian Sistem Percobaan Pertama dengan Model Pada Tensorflow

Pada pengujian pertama tingkat akurasi kernel pada identifikasi jenis kelamin dengan tingkat akurasi 80% namun obyek wajah, maka perlu dilakukan percobaan kedua agar tingkat *training and accuracy* berjalan sesuai yang dinginkan dan juga yang dibutuhkan. Percobaan kedua pada sebuah sistem tersebut dapat dilihat pada Gambar 16.

Gambar 16. Pengujian Sistem Percobaan Kedua dengan Model Pada Tensorflow

Pada tahap percobaan kedua telah berhasil mendeteksi jenis kelamin dengan sebuah obyek wajah di belakang obyek yang pertama dengan tingkat akurasi jenis kelamin perempuan sebesar 100 % di karena kan obyek wajah pertama mendeteksi perempuan dengan hidung, dan tingkat akurasi obyek wajah kedua juga sama sebesar 100%. Maka sistem berjalan dengan tingkat persentase akurasi yang tinggi, dan tingkat akurasi *error* rendah.

KESIMPULAN DAN SARAN

Identifikasi ini juga memiliki tujuan untuk fitur-fitur penting dalam citra wajah yang berperan dalam klasifikasi jenis kelamin. Dengan memanfaatkan perancangan *CNN* yang kompleks seperti *Teachable Machine* dan *Visual studio code* ,untuk mengidentifikasi menggunakan fitur-fitur yang relevan dalam proses klasifikasi tersebut.

Mekanisme dan implementasi pada penelitian ini dengan menerapkan sebuah Bahasa pemrograman *python* dan sebuah *library* yang bernama *Tensorflow* serta *keras*. *Dataset* pada penelitian mendapatkan perolehan dari beberapa pengujian yang telah dilakukan yaitu memperoleh sebuah persentase tingkat akurasi ketepatan pada sebuah sistem sebesar 99%. Klasifikasi pengenalan ekspresi wajah menggunakan masker atau tidak dengan mengimplementasikan metode *Convolutional Neural Network (CNN)* diperlukan tahapan kecermatan serta ketelitian dan juga kondisi pengoperasian. Kondisi pengoperasian merupakan sebuah waktu yang diperlukan pada saat sistem sedang menjalankan sebuah proses performansi. Berdasarkan beberapa penelitian sebelumnya menjabarkan bahwa metode *Convolutional Neural Network (*CNN*)* benar-benar efektif untuk implementasi sistem klasifikasi pada pengenalan wajah seseorang.

Pada pelatihan data yang dilakukan dengan 50 *epoch* dan *optimizer* Adam telah mencapai pada data uji. Hasil dari penelitian ini menunjukkan bahwa model *CNN* dengan menggunakan fitur-fitur *Teachable machine* dan *Visual studio code* dapat menjadi alat yang efektif dalam mendeteksi dan mengklasifikasikan *dataset* perempuan ataupun laki-laki.

DAFTAR PUSTAKA

[1] Satriawan, M. A., & Widhiarso, W. (2023). Klasifikasi Pengenalan Wajah Untuk Mengetahui Jenis Kelamin Menggunakan Metode Convolutional Neural Network. Jurnal Algoritme, 4(1), 43-52.

[2] Abhirawa, Halprin, Jondri, And Anditya Arifianto. 2017. Pengenalan Wajah Menggunakan Convolutional Neural Networks (Cnn). E-Proceeding Of Engineering 4(3): 4907â€“16.

[3] Aldiani, D., Dwilestari, G., Susana, H., Hamonangan, R., & Pratama, D. (2024). Implementasi Algoritma Cnn Dalam Sistem Absensi Berbasis Pengenalan Wajah. Jurnal Informatika Polinema, 10(2), 197-202.

[4] Arifandi, A. (2022). Identifikasi Dan Prediksi Umur Serta Jenis Kelamin Berdasarkan Citra Wajah Menggunakan Algoritma Convolutional Neural Network (Cnn). Rainstek: Jurnal Terapan Sains & Teknologi, 4(2), 89-96.

[5] Fadli, B. A., & Winarno, E. (2023). Pengenalan Wajah Dengan Face-Api. Js Berbasis Cnn Dan Geolokasi Menggunakan Equirectangular Approximation. Progresif: Jurnal Ilmiah Komputer, 19(2), 935-944.

[6] Firmansyah, R., & Siswanto, S. (2023). Penerapan Algoritma Cnn Untuk Mengenali Jenis Kelamin Yang Berinteraksi Pada Video Advertising. Kresna: Jurnal Riset Dan Pengabdian Masyarakat, 3(2), 166-173.

[7] Hermawan, E. (2021). Klasifikasi Pengenalan Wajah Menggunakan Masker Atau Tidak Dengan Mengimplementasikan Metode Cnn (Convolutional Neural Network). Jurnal Industri Kreatif Dan Informatika Series (Jikis), 1(1), 33-43.

[8] Ilhami, M. F. A., & Supriyanto, A. (2023). Deteksi Wajah Jenis Kelamin Dengan Fitur Hijab Dan Tidak Berhijab Menggunakan Jaringan Saraf Konvolusi. Intecoms: Journal Of Information Technology And Computer Science, 6(2), 693-701.

[9] M.K. Anam, "82 Metode Eigenface / Principle Component Analysis (Pca) Untuk Identifikasi Wajah Manusia," Jutis (Jurnal Teknik Informatika)., Vol. 6, Pp.82-88, Feb. 2020.

[10] M. Verdiansyah And A. Solichin, "Identifikasi Wajah Pada Rekaman Video Zoom Untuk Presensi Kuliah Daring Dengan Metode Haar Cascade Dan Algoritma Convolutional Neural Network," Jurnal Ilmiahinformatika., Vol. 7, Pp. 117-127, Dec. 2022.

[11] Nst, N. F. R. (2018). Metode Convolutional Neural Network Untuk Pengenalan Citra Wajah (Doctoral Dissertation, Universitas Gadjah Mada).

[12] Pradana, A. I., & Wijiyanto, W. (2024). Identifikasi Jenis Kelamin Otomatis Berdasarkan Mata Manusia Menggunakan Convolutional Neural Network (Cnn) Dan Haar Cascade

Classifier. G-Tech: Jurnal Teknologi Terapan, 8(1), 502-511..

[13] Salawazo, V. M. P., Gea, D. P. J., Gea, R. F., & Azmi, F. (2019). Implementasi Metode Convolutional Neural Network (Cnn) Pada Peneganalan Objek Video Cctv. Jurnal Mantik Penusa, 3(1.1).

[14] Sari, Y. (2022). Klasifikasi Jenis Kelamin Menggunakan Metode Convolutional Neural Network (Cnn) Dengan Data Wajah Manusia.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, October 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.637

Operating Systems (OS): An Insight Investigative Research Analysis and Future Directions

Zarif Bin Akhtar[1]
[1]Department of Engineering, University of Cambridge, Cambridge, United Kingdom
e-mail: zarifbinakhtarg@gmail.com[1]
* Corresponding author: E-mail: zarifbinakhtarg@gmail.com

Abstract: *In the realm of technological computing, the pivotal interface of operating systems (OS) governs the orchestration of machinery, orchestrating seamless human interactions with the swiftly advancing array of device peripherals. Over decades, the intricacies of computing have undergone a profound metamorphosis, embracing monumental leaps facilitated by the progressive proliferation of operating system distributions. From the erstwhile colossal processing units to the present-day intricately crafted nano-fabricated microcontrollers, motherboards, and chipsets, all human-computer interactions gravitate towards the nuanced tapestry of OS distributions and intricately woven source-coded programming. This comprehensive research endeavors to undertake a meticulous exploration of the myriad typologies of operating systems, intricately dissecting their distinctive functionalities and performance metrics, with a discerning focus on aligning specific user profiles with the most fitting OS distributions. Moreover, this investigation seeks to unravel the labyrinthine landscape of OS distributions, illuminating the optimal pathways for both seasoned users and neophytes alike.*

Keywords*: Computing; Computer Architecture; Computation Processing; Information Technology; Open-Source; Operating Systems (OS); Security; Technological Computing; System Distributions.*

INTRODUCTION

In the dynamic landscape of the 21st century's tech sector, groundbreaking innovations have significantly shaped the industry's trajectory. From the game-changing advent of the iPhone and Android to the meteoric rise of social media giants like Facebook and Twitter, this era has witnessed a seismic shift in the technological paradigm. The introduction of curved displays and foldable smartphones marked a pivotal moment in addressing the inherent limitations of traditional flat-screen displays. This research investigations precisely delves into the transformative potential of these novel technologies, illuminating how they have revolutionized the technological market since their inception. Notably, tech conglomerates such as Samsung, Honor, and Xiaomi have channeled significant financial resources into the development and deployment of these cutting-edge technologies, propelling the industry towards a new era of innovation and consumer engagement. As of now, Samsung has emerged as the frontrunner, overshadowing its Chinese counterparts, but the relentless investments by various companies are poised to intensify the competitive landscape, promising an expansive array of choices for discerning customers. A key aspect of this research is the comprehensive analysis of the various types of OS shortcomings [1-7], offering valuable insights into which type of device might be most suitable for users. Furthermore, it delves into prevailing consumer trends, providing a forward-looking perspective on the future trajectory of this technology segment [8-11]. By inspecting the merits and demerits of these innovative devices and distribution systems, the research aims to equip users and industry stakeholders with the knowledge necessary to navigate this ever-evolving landscape effectively.

This comprehensive investigation also examines computer operating systems (OS) and offers insightful analysis of emerging trends in this dynamic field [12,13]. OS functionality, serving as an intermediary interface between computer hardware and users, is a critical cornerstone in modern computing systems [14-18]. This research aims to discern the trajectory and projected evolution of OS, unraveling the intricate concepts underpinning its architecture and development over time. The investigative research analysis delves into the intricate components of OS, shedding light on the intricacies of its underlying structures and providing comprehensive insights into security issues associated with various types of OS architectures. Moreover, the exploration offers a compendium of best practices intended to bolster OS security, advocating for a dual-layered security approach [19]. This approach entails fortifying the OS with robust security policies while simultaneously embedding stringent security protocols within the hardware architecture of the OS [20-26]. A significant revelation of the research pertains to the identification of contemporary OS trends, which encompass an array of cutting-edge developments such as IoT OS, Cloud OS, AI-powered OS, Blockchain OS, Hybrid OS, and Container OS [27,28,29]. The analysis also precisely weighs the strengths and weaknesses of these major OS categories, enabling readers to gain a nuanced understanding of their relative merits and limitations. Furthermore, the research espouses a forward-thinking approach, advocating for the conception of a universal OS framework that accommodates diverse architectures, fostering a scalable environment for potential expansion. In an effort to address sustainability concerns, the research highlights the integration of green computing technologies within the design architecture, aimed at

mitigating power consumption issues and fostering environmentally responsible computing practices. By precisely addressing the intricacies of OS architectures and their evolving landscape, this research offers a holistic perspective that not only outlines the current state of the field but also provides invaluable insights into the potential pathways for future advancements and innovation in the realm of operating systems for modern computing systems.

METHODS AND EXPERIMENTAL ANALYSIS

The methodology for this research follows a systematic approach to investigate the impact of operating systems (OS) on distribution systems and computing. Initially, a comprehensive iterative background research explorations with available knowledge investigations was conducted to gather existing knowledge and identify research gaps. Data collection was performed using specific methods such as surveys, experiments, and data mining techniques to ensure relevance and accuracy. All the collected data underwent pre-processing, which included cleaning, normalization, and validation, to maintain quality and relevance. The performance and visualization techniques were evaluated using specific metrics, including system throughput, latency, scalability, resource utilization, and reliability. These metrics provided a benchmark to compare the evaluated techniques against traditional and existing computing approaches. This comparison highlighted the effectiveness and efficiency of the Operating Systems (OS) in enhancing technological computing performance. Results and findings were analyzed and interpreted in line with the research objectives. The analysis discussed the implications of OS advancements on technological computing performance and efficiency, providing insights into the near future developments. The findings demonstrated how OS could enhance computing, influencing the interconnected world of digital device peripherals. Finally, the research summarized the findings, acknowledged the limitations encountered, and suggested future research prospects. These suggestions focused on areas that could benefit from further investigation to enhance the understanding and application of OS in distribution systems and computing. By employing these specific methods and metrics, this methodology ensures a comprehensive exploration of the potential enhancements OS can bring to computing, contributing to improved performance and paving the way for accelerated innovations in the field.

BACKGROUND ITERATIVE RESEARCH AND AVAILABLE KNOWLEDGE

An operating system (OS) is an essential system software that manages computer hardware and software resources, providing fundamental services for computer programs. OSs play a critical role in facilitating communication between applications and the hardware components of a computer system, including memory allocation and input/output functions.

They are mostly found across a wide range of devices, from smartphones and video game consoles to web servers and supercomputers. In the personal computer market as of September 2023, Microsoft Windows maintains a dominant market share of around 68%. macOS by Apple Inc. holds the second position with approximately 20%, and Linux variants, including ChromeOS, collectively account for about 7%. On the other hand, in the mobile sector, Android leads with a share of 68.92%, followed by Apple's iOS and iPadOS with 30.42%, while other operating systems collectively hold 0.6%. In the server and supercomputing sectors, Linux distributions dominate, while other specialized operating systems exist for embedded systems, real-time processing, and security-focused applications. Different types of operating systems cater to varying computing needs. Single-tasking systems can handle only one program at a time, while multi-tasking OSs allow the concurrent execution of multiple programs. This is achieved through time-sharing, where the processor time is divided among different processes. Multi-tasking can be preemptive or cooperative. Single-user systems support only one user but can execute multiple programs simultaneously. Multi-user systems facilitate the interaction of multiple users with the system and allocate resources accordingly.

Distributed systems manage a network of computers, making them appear as a single unit by distributing computations among connected computers. Embedded systems which operate in small, resource-constrained machines like PDAs, embedded OSs are designed for efficiency and compactness. Windows CE and Minix 3 are examples of embedded operating systems. Real-time systems are the OSs that ensure processing of events or data within specific time constraints. They can be single-tasking or multi-tasking, utilizing specialized scheduling algorithms to maintain deterministic behavior. Library operating systems are such systems which provide OS services in the form of libraries, composing with application and configuration code to form a unikernel—a specialized, single address space machine image deployable to cloud or embedded environments [12,13].

The history of operating systems dates back to the 1950s when early computers were programmed for specific tasks. As hardware evolved, basic OS features were developed, leading to the emergence of modern and complex operating systems in the 1960s. Early electronic systems were programmed using mechanical switches and plugboards, with no operating systems. The introduction of high-level languages and machine libraries in the late 1950s paved the way for the modern concept of operating systems. Notable early examples include GM-NAA I/O and the SHARE Operating System [14,15,16]. Mainframe computers in the 1950s pioneered key operating system features such as batch processing, multitasking, and spooling. IBM's OS/360 in the 1960s laid the foundation for a single OS spanning an entire product line.

Other significant mainframe OSs include Burroughs MCP, UNIVAC EXEC, General Electric's GECOS, and the Multiplexed Information and Computing Service (Multics) developed by Bell Labs, General Electric, and MIT. With the advent of microcomputers, minimalistic operating systems like CP/M and MS-DOS were developed in the 1970s and 1980s. The introduction of the Intel 80386 CPU chip in 1985 enabled microcomputers to run multitasking OSs.

The GNU Project, led by Richard Stallman, aimed to create a complete free software replacement for proprietary UNIX. Linus Torvalds's release of the Linux kernel in 1991 marked a significant milestone, leading to the development of the popular Linux OS. Microsoft Windows, initially built on top of MS-DOS, emerged as a dominant family of operating systems, gradually transitioning to the Windows NT kernel. Windows remains widely used, especially on personal computers, although it faces competition from Linux and BSD in the server market [17,18]. In addition to the major operating systems like Unix, Linux, macOS, and Windows, various other systems were once significant but have now become obsolete or niche, including AmigaOS, OS/2, classic Mac OS, BeOS, and others. Some systems like z/OS, OpenVMS, and IBM i continue to be actively used and developed, catering to specific enterprise needs. Academic environments also use specific systems such as MINIX and Singularity for educational and research purposes. The development of these various operating systems has significantly shaped the landscape of modern computing [30].

The operating system, as a crucial layer between user applications and computer hardware, comprises several fundamental components that enable the efficient functioning of a computer system. The kernel, at the core of the operating system, serves as the bridge connecting application software to the hardware components. With the assistance of firmware and device drivers, the kernel exerts control over the computer's hardware devices, managing memory access, allocating resources, and organizing data storage with file systems. It further regulates the CPU's operational states for optimal performance, ensuring the smooth execution of various processes and applications. The execution of application programs involves a systematic process facilitated by the operating system. This process includes the creation of a process by the kernel, where memory space and resources are assigned, and program binary code is loaded into memory. The operating system also sets the priority for the process in multi-tasking environments, ensuring efficient utilization of computing resources. Through this mechanism, application programs interact with users and hardware devices, adhering to predefined rules and procedures incorporated into the operating system [31]. Interrupts, both hardware and software, play a crucial role in the responsiveness and coordination of the operating system. Hardware interrupts enable the CPU to handle asynchronous events efficiently, allowing I/O devices to signal completion without requiring continuous CPU polling.

On the other hand, software interrupts serve as messages to processes, informing them of specific events or errors, such as time slices, error conditions, or user-initiated interruptions. These interrupts ensure the synchronization of processes, aiding in the seamless execution of multiple tasks within the system. Interrupt-driven I/O, a mechanism triggered by user inputs such as keystrokes or mouse movements, facilitates immediate responses to user actions, ensuring real-time interaction between users and the system. Additionally, direct memory access (DMA) allows high-speed data transfer between devices like hard disk drives and memory, circumventing the need for CPU intervention for each data transfer. The operating system manages these processes efficiently, enabling the seamless exchange of data and information between hardware devices and memory. The various components of the operating system work in tandem to provide a stable and efficient platform for the execution of applications, the management of hardware resources, and the seamless coordination of input and output operations. This collaborative functionality ensures a smooth and responsive user experience while harnessing the full potential of the computer's hardware capabilities [32]. The concept of operating systems is integral to the functioning of modern computers. Operating systems facilitate the interaction between the user and the hardware, providing a range of services including memory management, multitasking, disk access, networking, and security. Two key operating modes, user mode and supervisor mode, govern the level of access to resources, with the supervisor mode providing unrestricted access to machine resources and the user mode setting limits on instructions and direct access. Memory management is critical to ensure that programs don't interfere with one another, with memory protection serving as a key mechanism to restrict a process's access to the computer's memory [33,34]. Techniques such as memory segmentation and paging enable the kernel to control the memory accessed by different programs. Virtual memory, a crucial concept, allows the kernel to manage memory effectively by temporarily storing less frequently accessed memory on disks or other media. This makes space available for other programs, giving the illusion of a larger RAM capacity. Multitasking is another vital function that allows for the simultaneous execution of multiple independent computer programs, commonly achieved through time-sharing. Early models of multitasking were cooperative, allowing programs to execute for as long as they wanted, which could potentially lead to system crashes. Modern operating systems implement preemptive multitasking, ensuring all programs receive regular time on the CPU by utilizing a timed interrupt [35].

File systems enable the organization and management of files on a computer, with a hierarchical structure of directories or folders. Various file systems and their characteristics, such as naming conventions and access permissions, present challenges for implementing a single interface for all file systems.

To ensure compatibility, most operating systems provide support for widely used file systems and often require third-party drivers for others. Device drivers play a critical role in enabling interaction with hardware devices. They act as a mediator between the operating system and hardware devices, translating operating system calls into device-specific commands. Networking support in operating systems allows computers to share resources across a network, enabling functions like file sharing and remote access. Security measures within operating systems include authentication, authorization, and auditing, crucial for protecting sensitive data from unauthorized access [36,37]. Operating systems provide user interfaces to interact with computers, ranging from command-line interfaces to graphical user interfaces (GUIs). The evolution of GUIs has seen significant advancements, with many modern systems incorporating GUIs for better user experiences. Real-time operating systems are designed for applications with fixed deadlines, commonly used in embedded systems and industrial control. Operating system development as a hobby has led to the creation of unique systems independent of existing ones, often driven by individuals or small groups with shared interests. Ensuring software portability across various operating systems often requires adaptation or the use of software platforms like Java or Qt, which can minimize the costs of supporting diverse operating systems. Standardization efforts such as POSIX and OS abstraction layers have aimed to reduce the complexities of porting applications across different operating systems. To provide an idea concerning the perspective of the matter figure 1 provides an illustrative representation of the retrospect.

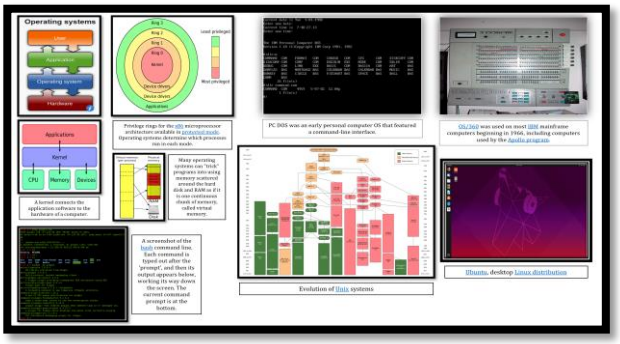


FIGURE 1. An overview of OS and its associated device integrations with distributions

THE DIFFERENT TYPES OF OPERATING SYSTEMS (OS)

Operating systems play a critical role in the functioning of computers and devices, providing the necessary software framework for managing hardware and software resources. Several types of operating systems have been developed to cater to different computing requirements and contexts. Each type has its own unique characteristics, benefits, and drawbacks, catering to a diverse range of applications and user needs.

Batch operating systems, typically utilized in the early days of computing, facilitated the processing of a series of similar jobs grouped together into batches. Despite their effectiveness in managing large workloads and allowing multiple users to share systems, batch systems posed challenges in terms of debugging, job failure impact, and cost, making them less favorable in contemporary computing environments. Distributed operating systems, a recent technological advancement, enable the connection of multiple independent computers through a unified communication channel, providing benefits such as fault tolerance, load distribution, and increased scalability. However, their complex software and high setup costs can present significant challenges, particularly in the event of network failures. Multitasking operating systems, also known as time-sharing systems, allow multiple users to efficiently access the CPU by allocating specific time periods for task execution. While offering equitable access and minimal idle time for the CPU, multitasking systems face issues such as data security concerns and potential communication problems, hindering their seamless operation. Network operating systems manage networking functions within a server-based environment, facilitating resource sharing, security management, and remote access for multiple users. Despite their stability and upgradability, these systems entail high server costs and maintenance requirements, making users heavily reliant on centralized operations. Real-time operating systems cater to time-sensitive applications, with hard real-time systems serving critical, time-constrained operations, and soft real-time systems addressing less stringent time constraints. While delivering optimal resource utilization and memory management, real-time systems are limited by expensive resources, complex algorithms, and restricted task capacity, presenting challenges in thread priority management and task switching. Mobile operating systems are designed for smartphones, tablets, and other mobile devices, allowing users to access various applications on the go. While providing user convenience, some mobile operating systems may exhibit limitations, such as poor battery performance and user interface issues, impacting overall user experience.

In addition to the types of operating systems, distinctions can be made based on single-tasking versus multi-tasking functionalities, desktop versus mobile compatibility, and open-source versus proprietary development models, each catering to specific user requirements and preferences. These diverse types and distinctions reflect the continuous evolution and adaptation of operating systems to meet the diverse needs of modern computing environments.

OS FUNCTIONALITIES, POPULARITY AND MARKET SHARE

Operating systems serve as the fundamental software framework for managing hardware and software resources, enabling the efficient functioning of computers and devices.

Their functions encompass critical tasks such as resource allocation, memory management, device management, user interface management, and security management, ensuring smooth and secure operations for users and applications. Resource allocation and management involve the efficient distribution of CPU, memory, and disk space among various applications, prioritizing tasks based on their importance. Memory management ensures optimal memory utilization and efficient sharing among running programs, facilitating seamless performance. Device management handles input and output devices, ensuring their compatibility and functionality within the system. User interface management provides a graphical user interface, enabling user interactions through windows, menus, and other visual elements. Security management safeguards systems and data through user authentication, firewalls, and antivirus software, protecting against unauthorized access and threats. Among the most widely used operating systems, Windows, macOS, Linux, iOS, and Android cater to diverse user preferences and requirements, offering a range of features and functionalities, from user-friendly interfaces to open-source customizability and robust security measures. Operating system evolution has traversed various generations, each marked by distinct technological advancements. From the earliest vacuum tube-based systems and machine language programming to contemporary AI-driven systems and quantum computing, operating systems have continually evolved to accommodate the growing complexities of computing tasks and user demands. The kernel is the fundamental program at the heart of a computer's operating system, exercising complete control over all system operations and hardware management. It serves as the intermediary between the computer hardware and software applications, ensuring smooth and efficient functionality. There are five distinct types of kernels: microkernel, monolithic kernel, hybrid kernel, exokernel, and nanokernel. Each type has unique characteristics and design philosophies, which influence their functionality, performance, and application. A microkernel architecture is designed to contain only the most essential functions of the operating system, such as low-level address space management, thread management, and inter-process communication (IPC). In a microkernel, user services and kernel services are implemented in separate address spaces. This separation enhances system stability and security, as a failure in user services does not directly affect kernel services. However, the design and implementation of a microkernel are complex, requiring more code and effort. Although microkernels are smaller in size and easier to extend, they typically suffer from lower execution speed due to the overhead of frequent context switches and message passing. Examples of operating systems using microkernels include Mac OS X. Contrastingly, a monolithic kernel runs the entire operating system as a single program in kernel mode, with both user services and kernel services sharing the same address space.

This design simplifies implementation and can lead to higher execution speeds because there is no need for context switching or message passing between user and kernel space. However, monolithic kernels are larger and more challenging to maintain and extend. A failure in any component within a monolithic kernel can potentially lead to a system-wide failure, making them less robust compared to microkernels. Debugging is also more difficult due to the integrated nature of the kernel components. Examples of monolithic kernels include Microsoft Windows 95.

A hybrid kernel combines elements of both microkernel and monolithic kernel architectures. It aims to take advantage of the performance benefits of monolithic kernels while maintaining the modularity and resilience of microkernels. This approach allows for a more balanced trade-off between performance and reliability. An exokernel architecture is designed to give application-level software greater control over hardware resources. By minimizing the abstractions provided by the kernel, exokernels allow applications to directly manage resources, potentially improving performance for specialized tasks. Nanokernels are even more minimalistic than microkernels, providing only the most fundamental services necessary to manage hardware resources. They delegate as much functionality as possible to higher-level software, which runs in user space. Understanding the different types of kernels and their respective architectures is crucial for optimizing system performance, stability, and security. Each kernel type offers distinct advantages and challenges, influencing the design decisions for various operating systems. The choice between microkernel, monolithic kernel, hybrid kernel, exokernel, and nanokernel architectures depends on the specific requirements and constraints of the computing environment.

The progression through multiple generations underscores the continual integration of cutting-edge technologies and the potential for future advancements that could revolutionize computing paradigms and interactions, possibly leading to seamless integration between human cognition and computing systems. To better understand the perspective of the matter figure 2 provides the global stat of OS market share and table 1 provides the illustration of 32bit and 64bit OS variations in terms of parameters.

TABLE 1. 32-bit OS vs. 64-bit OS

Parameter	32-Bit OS	64-Bit OS
Data and Storage	The 32bit OS can store and manage less data than the 64bit OS, as its name would imply. It mainly addresses a total maximum of 4,294,967,296 bytes (4 GB) of RAM in more detail.	In contrast, the 64bit OS has a larger data handling capacity than the 32bit OS. It indicates that a total of 264 memory addresses, or as in 18 quintillion gigabytes of RAM, can be addressed.

Compatibility of System	A 32-bit processor system will run only on 32-bit OS and not on 64bit OS.	A 64-bit processor system can run either a 32-bit or 64-bit OS
Application Support	The 32-bit OS support applications with no hassle.	The 64-bit OS do not support many of the older hardware and software applications.
Performance	Performance of the 32- bit OS is less efficient.	Higher performance than that of the 32-bit processor.
Systems Available	These support Windows 7, Windows XP, Windows Vista, Windows 8, and Linux.	These support Windows XP Professional, Windows 7, Windows 8, Windows 10, Windows Vista, Linux, and Mac OS X.

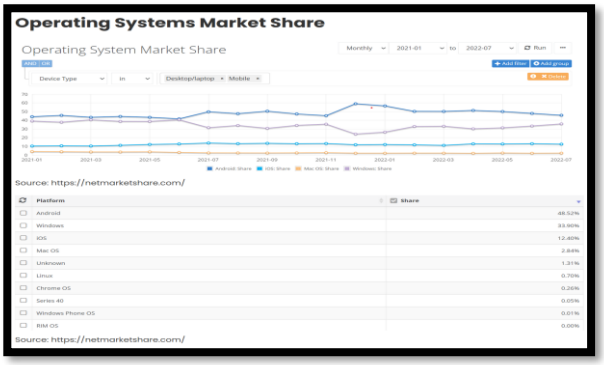


FIGURE 2. OS Market Share around the Globe

OS PROS AND CONS WITH RTOS

Operating systems play a critical role in ensuring the efficient and correct use of a computer's hardware, facilitating the simultaneous operation of various applications, and managing the organization of files and folders. They provide an intuitive user interface, simplifying interactions between users and machines, while also handling security measures to protect against unauthorized access and data breaches. Operating systems efficiently manage system resources, ensuring that hardware components are utilized optimally and that various applications have fair access to necessary resources. Furthermore, they manage printing operations, ensuring that documents and files are printed accurately and efficiently. Operating systems serve as a robust platform for software development, providing a stable and consistent environment for the creation and execution of diverse software applications. Despite their numerous advantages, operating systems are not without their drawbacks. They can be complex and challenging to use, especially for individuals with limited technical expertise, potentially posing barriers to entry for certain users.

The cost associated with purchasing and maintaining operating systems, along with the need for regular updates and maintenance, can impose financial burdens on organizations and individuals. Furthermore, the inherent complexity of operating systems can render them vulnerable to attacks from malicious users, necessitating robust security measures and constant vigilance to safeguard sensitive data and systems from potential threats and breaches.

Real-Time Operating Systems (RTOS) serve as specialized operating systems that execute multi-threaded programs while adhering to stringent real-time deadlines. Unlike the conventional notion of speed, the ***"deadlines"*** in an RTOS pertain to the ability to predict when specific tasks will run before their actual execution. RTOS proves to be an invaluable tool, particularly for complex embedded applications, offering support for task isolation and enabling concurrent operations. Its applications span various critical systems, including defense applications like RADAR, airline reservation systems, systems that demand immediate updating, networked multimedia systems, air traffic control systems, and command control systems. The seamless execution and adherence to real-time constraints make RTOS indispensable in scenarios where timely and accurate data processing and decision-making are of utmost importance.

HOW TO KNOW AND CHOOSE BEST OS

Selecting the most appropriate computer operating system (OS) involves a careful evaluation of several key factors to ensure compatibility and functionality. The price of an OS is an essential consideration, with options ranging from free, like Linux, to paid systems like Windows and macOS. Accessibility is another vital factor, with certain systems, such as macOS and iOS, offering user-friendly interfaces, while others, like Linux, can present a steeper learning curve. The compatibility of an OS with desired applications is crucial, as some systems support a broader range of software compared to others. Additionally, the security features of an OS should be weighed, as certain systems offer stronger security protocols than others.

When selecting an OS, it is critical to prioritize robust memory management, ensuring efficient utilization of system resources. Stability is paramount, especially for individuals relying on their computers for various tasks, whether business-related or for leisure activities like gaming. The OS should be reliable, avoiding frequent crashes and interruptions that could hinder productivity or enjoyment. Support and cost are equally important, with the understanding that paying for an OS does not necessarily guarantee better performance or assistance. Some free OS options offer robust support channels, while certain paid systems may fall short on delivering promised assistance.

Ultimately, the choice of an OS significantly impacts the user experience, influencing the overall functionality and performance of the computer.

Considering factors such as usability, compatibility with preferred applications, and the system's ability to support various tools and customization options are crucial when making this decision. Furthermore, understanding the role of an operating system in organizing and managing files and programs on the computer is fundamental to selecting the most suitable option.

While there may not be a definitive answer to the best OS, careful consideration of individual requirements and preferences will guide users towards making an informed choice and a conclusive decision.

OPERATING SYSTEMS (OS) SECURITY

Security is a critical aspect of any operating system, ensuring the protection of valuable computer resources, including the CPU, memory, disk space, software programs, and data. Authentication is a fundamental component of security, involving the identification of each system user and their association with executing programs. Operating systems employ various authentication methods, such as username/password combinations, user cards/keys, and user attributes like fingerprints or eye retina patterns, to ensure secure access.

One-time passwords enhance security protocols by requiring unique passwords for each login attempt. Implementations of one-time passwords involve the use of random numbers, secret keys generated by hardware devices, or network passwords sent to users' registered mobile or email accounts. Program threats, such as Trojan horses, trap doors, logic bombs, and viruses, pose significant risks by enabling unauthorized access to user credentials, introducing security vulnerabilities, or causing system malfunctions. System threats, including worms, port scanning, and denial of service attacks, can disrupt network performance and compromise system resources, resulting in severe consequences for users and their data.

Computer security classifications, as established by the U.S. Department of Defense Trusted Computer System's Evaluation Criteria, categorize security levels into four types: A, B, C, and D. Each classification represents varying degrees of security measures, with Type A offering the highest level of assurance through formal design specifications and verification techniques. Type B provides a mandatory protection system, while Type C focuses on user accountability and audit capabilities. Lastly, Type D represents the lowest security level, offering minimal protection and often associated with operating systems such as MS-DOS and Windows 3.1. These security classifications serve as vital benchmarks for evaluating and modeling the security of computer systems and their corresponding security solutions.

WINDOWS VS MACOS VS LINUX: AN INVESTIGATIVE ANALYSIS

The comparison of three leading operating systems, namely Windows, macOS, and Linux, reveals distinctive features and suitability for various user preferences and needs. Windows, known for its versatility and widespread usage, is well-suited for general productivity tasks, gaming, software development, and multimedia creation.

With its user-friendly interface and robust security protocols, Windows stands as a popular choice for users who prioritize compatibility, diverse software support, and extensive hardware compatibility. The system's robust security features, including Windows Defender and regular updates, contribute to a secure computing environment, although risks can arise from the installation of potentially malicious software.

MacOS, specifically designed for Apple devices, is celebrated for its seamless integration with the Apple ecosystem, catering to the needs of creative professionals and artists. Recognized for its enhanced security and privacy measures, macOS offers a visually appealing interface and top-notch performance. However, its application compatibility may be limited compared to Windows, and users might encounter challenges in running certain software programs not specifically designed for macOS. The system's superior security measures, such as Gatekeeper and FileVault, contribute to a secure computing experience, safeguarding users' personal data from external threats.

Linux, an open-source powerhouse, offers unparalleled flexibility and customization options, making it a favorite among tech-savvy users, developers, and system administrators. Known for its stability and performance, Linux is widely used for server management, web development, and data analysis tasks. Its strong focus on security and privacy, fostered by its open-source nature, provides users with granular control over system permissions and comprehensive protection against unwanted surveillance. Despite the availability of numerous applications and growing developer support, Linux may present challenges for users accustomed to more mainstream operating systems due to its unique customization options and learning curve for newcomers.

Each operating system offers distinct advantages and serves particular user needs, whether it be Windows' compatibility and user-friendliness, macOS' seamless integration within the Apple ecosystem and enhanced security, or Linux's flexibility, customization, and robust security features.

By understanding the specific requirements and preferences of users, the choice of an operating system can be tailored to individual needs, providing an optimized computing experience and meeting diverse computing demands. For a representative illustration figure 3 provides an insight into the matter.

FIGURE 3. Windows vs. macOS vs. Linux the illustrative representation

RESULTS, FINDINGS, OS CHALLENGES AND FUTURE DIRECTIONS

The realm of modern operating systems is confronted with an array of challenges that predominantly revolve around security, resource management, and device compatibility. Notably, security and privacy concerns have become paramount, as the prevalence of vulnerabilities and exploits, including malware and ransomware, constantly jeopardize the integrity of systems. Additionally, the rising apprehensions regarding user privacy have necessitated the implementation of stringent security measures, in response to issues such as data breaches and unauthorized data collection. Efficient resource management has emerged as a critical facet for optimizing system performance, with memory management techniques such as virtual memory and paging playing a pivotal role in effectively managing limited physical memory. Moreover, the proliferation of hardware devices has posed a significant obstacle for operating systems, demanding the deployment of hardware abstraction layers to ensure uniform management of devices with distinct interfaces. Furthermore, the integration of device drivers has become essential for facilitating communication between hardware and software, while the plug-and-play functionality has streamlined the process of device installation and configuration. Embracing containerization technologies like Docker and Kubernetes has ushered in a new era of lightweight and isolated environments for applications, augmenting scalability and portability within the operating system landscape. The concept of virtualization, encompassing both hardware and software virtualization, has enabled the simultaneous operation of multiple operating systems or instances on a single physical machine, amplifying the versatility of computing environments. Operating systems, functioning as system software, assume the crucial role of managing computer hardware and software resources, while providing fundamental services for computer programs across diverse devices, from cellular phones and video game consoles to web servers and supercomputers.

Notably, Linux distributions have carved a dominant niche in the server and supercomputing sectors, underscoring their versatility and robust capabilities. In light of the burgeoning Internet of Things (IoT) landscape, operating systems are poised to evolve further, necessitating adaptation to accommodate the unique demands of IoT devices. Embedded operating systems, such as FreeRTOS and Zephyr, have garnered significance owing to their prioritization of low resource consumption and real-time responsiveness, enabling seamless integration of IoT devices into networks. As the future unfolds, operating systems will continue to underscore the importance of robust security measures, with secure booting, sandboxing, and secure enclaves emerging as pivotal components in safeguarding system integrity and user privacy. The continued evolution of operating systems is poised to reshape the technological landscape, ushering in more secure, efficient, and adaptable computing environments.

The current structure of modern operating systems, particularly in the Unix/Linux domain, is closely entwined with the C programming language and a shell environment, leading to a complex and often brittle software ecosystem. The weak type systems of both C and the shell, coupled with their reliance on configuration files, make it challenging to verify the correctness of programs and system configurations in advance. Consequently, troubleshooting or configuring a Linux system often involves blindly following online instructions, creating an environment where errors can easily disrupt the system's functionality. This fragility has created an inaccessible hacking environment for many users, contradicting the open-source philosophy that advocates for users to have full control over their machines. The comparison between a program like ***'my_server'*** configured through a file-based approach and one written in a strongly-typed programming language highlights the stark contrast in robustness and predictability. While both scenarios may accomplish the task of running the server, the latter provides a significantly reduced risk of failure and allows for the verification of the program's validity beforehand. This distinction underscores the need for a paradigm shift in system configuration, suggesting a move toward a single declarative-style program for the entire system, akin to what NixOS is pursuing, yet with a stronger emphasis on pre-verification. The notion of content-based addressing and the shift from trusted to untrusted content and infrastructure are also critical elements in the reimagining of a modern operating system. Embracing the principle of identifying programs by the hash of their content fosters a more accessible and streamlined approach to program distribution, promoting the ease of program integration without extensive bureaucratic processes. Leveraging technologies like WebAssembly for cross-platform compatibility and sandboxing programs to isolate them from system resources emerges as a fundamental direction for enhancing security and predictability.

The idea of long-running daemons rather than short-lived programs aligns with the original Unix philosophy of specialized tools accomplishing specific tasks efficiently. This concept emphasizes the need for a shift in perspective, viewing the system as a collection of modules in the context of a programming language rather than isolated programs. Sandboxing technologies, such as Docker containers and virtualization, have arisen as a response to the existing operating system's limitations in isolating programs effectively. However, the challenge remains to ensure a default level of isolation and security within the system, guaranteeing that users can execute programs without jeopardizing the system's stability or their data. Furthermore, redefining the traditional concept of a file system in the context of cloud computing and content-based addressing offers the potential for a simplified and more transparent user experience. Eliminating the dependence on a shared global file system and reimagining the purpose and structure of files could significantly enhance the accessibility and reliability of data management within the system. Ultimately, the pursuit of a modern operating system revolves around creating an accessible, robust, and predictable environment that empowers users to navigate and configure their systems with ease and confidence.

DISCUSSIONS

In the world of operating systems, Windows, macOS, and Linux each come with their own set of specific hardware requirements. While Windows is generally compatible with a wide array of PC configurations, macOS is purpose-built for Apple computers. Linux, known for its broader hardware compatibility, still requires users to verify system requirements before installation. One critical factor to consider is software compatibility. Not all applications are universally compatible across different operating systems. Windows boasts the broadest range of software compatibility due to its widespread use, while macOS caters to a robust selection of creative and multimedia applications. Linux, although increasingly supported by developers, may not have direct alternatives for specialized or exclusive Windows/macOS applications. Cross-platform compatibility is also a consideration. Some Windows software can run on Linux using compatibility layers like Wine or virtualization software, although not all applications function seamlessly. While Linux has made significant strides in user-friendliness, it may still demand a bit more technical know-how compared to the more mainstream Windows and macOS systems. Regarding gaming, Windows is the preferred operating system, providing extensive game libraries and strong driver support. macOS, although offering a smaller gaming selection, still supports several popular titles. Linux has seen substantial progress in the gaming realm, with a growing number of games and compatibility layers like Steam's Proton, enabling gaming on the platform.

In other words, selecting the appropriate operating system depends on individual priorities, needs, and preferences. Windows, macOS, and Linux cater to different user requirements, with each system offering unique features. Windows excels in versatility and gaming capabilities, macOS in elegance and security within the Apple ecosystem, and Linux in flexibility, customization, stability, and security. Therefore, understanding one’s specific requirements is crucial when making an informed decision, as the choice of operating system significantly impacts ones overall digital experience.

CONCLUSIONS

Operating systems play a crucial role in the functioning of computers, managing resources, providing user interfaces, and ensuring security. Windows, macOS, and Linux are three prominent operating systems, each with distinct strengths and weaknesses. Windows stands out for its versatility, extensive software compatibility, and robust gaming capabilities, making it an excellent choice for a wide range of users. macOS, designed exclusively for Apple hardware, offers a seamless, visually appealing interface, robust security, and seamless integration within the Apple ecosystem, making it ideal for creative professionals and artists. Linux, known for its open-source nature, provides unmatched flexibility, customization options, and strong security, making it a preferred choice for tech-savvy users, developers, and system administrators. When selecting an operating system, factors such as price, accessibility, compatibility, security, and support should be considered. Windows enjoys widespread compatibility and a user-friendly interface, catering to diverse user needs. macOS, known for its stringent security measures and seamless integration within the Apple ecosystem, appeals to creative professionals and users valuing elegant design. Linux, with its open-source nature and strong security focus, is suitable for those prioritizing customization, privacy, and stability, particularly in server environments. Each operating system caters to specific user requirements, making it essential to assess individual preferences and needs when making a choice. Considerations such as software compatibility, hardware requirements, and cross-platform usability are crucial. Windows offers the broadest software compatibility, macOS specializes in creative and multimedia applications, while Linux provides a growing selection of software options, often necessitating the use of compatibility layers or virtualization software for cross-platform compatibility. While Windows is widely recognized as the go-to operating system for gaming, macOS and Linux have also made significant strides in gaming support, albeit with a more limited selection of titles. Understanding these factors is essential for making an informed decision, as the choice of an operating system significantly impacts the overall digital experience, productivity, and security of a computer system.

Operating system standards are continually evolving to meet the demands of advancing technologies, security, performance, and user experience. A significant trend in operating system design is the move towards modular and microkernel architectures, aiming to strike a balance between functionality and complexity. While monolithic kernels provide high performance and compatibility, they also raise concerns regarding bugs, crashes, and security vulnerabilities. In contrast, modular or microkernel architectures, as seen in operating systems like Windows NT, Linux, and MINIX, prioritize reducing size, improving reliability, and enhancing security, although they may introduce overhead and communication challenges. Additionally, the increasing prevalence of cloud and edge computing is influencing the future of operating system standards. These technologies facilitate remote data processing and storage, emphasizing scalability and efficiency. Operating systems are required to adapt to the demands of distributed and parallel computing, network and data management, as well as security and encryption. Examples such as Chrome OS, Android, and Azure Sphere demonstrate the integration of these principles. Moreover, the rise of artificial intelligence (AI) and machine learning (ML) with deep learning (DL) is shaping operating system standards by enabling complex pattern analysis, task automation, and optimization. Operating systems must integrate AI and ML capabilities into core functions, providing development tools and frameworks for AI and ML applications. macOS, iOS, and Linux showcase the incorporation of AI and ML features within their systems. User-centric and adaptive design represents another crucial aspect influencing the future of operating system standards. This design philosophy prioritizes improving usability, accessibility, and personalization, catering to changing user needs and preferences. Operating systems achieve this by incorporating features like voice and gesture control, biometric authentication, customization options, and context-awareness. Examples such as Windows 10, Ubuntu, and HarmonyOS exemplify the implementation of user-centric and adaptive design principles. The aspect of security and privacy is of paramount importance in shaping the future of operating systems. Operating systems are increasingly focusing on implementing robust security features such as encryption, authentication, authorization, sandboxing, firewall, antivirus, and update mechanisms. Adherence to data protection regulations such as GDPR and CCPA is also critical. Operating systems like Qubes OS, Tails, and iOS emphasize security and privacy as fundamental components of their standards, ensuring protection against unauthorized access and maintaining user privacy.

ACKNOWLEDGMENTS

The main prospect and the scope for this research was conducted with the idea perspective experimentations along with the manuscript writing was done by the author himself. All the data sources which have been retrieved and resourced for the conduction of this research exploration are mentioned and referenced where appropriate.

REFERENCES

[1] Microsoft. Windows Secure Channel Denial of Service Vulnerability. 2023. Available online: https://msrc.microsoft.com/update-guide/en-US/vulnerability/CVE-2023-21813 (accessed on 1 June 2023).

[2] Research, G.S. Linux (Ubuntu)–Other Users Coredumps Can Be Read via Setgid Directory and killpriv Bypass. 2018. Available online: https://www.exploit-db.com/exploits/45033 (accessed on 1 June 2023).

[3] Gorbenko, A.; Romanovsky, A.; Tarasyuk, O.; Biloborodov, O. Experience report: Study of vulnerabilities of enterprise operating systems. In Proceedings of the 2017 IEEE 28th International Symposium on Software Reliability Engineering (ISSRE), Toulouse, France, 23–26 October 2017; IEEE: New York, NY, USA, 2017; pp. 205–215. [Google Scholar]

[4] Cheikes, B.A.; Waltermire, D.; Kent, K.A.; Waltermire, D. Common Platform Enumeration: Naming Specification Version 2.3; US Department of Commerce, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2011. Available online: https://csrc.nist.gov/publications/detail/nistir/7695/final (accessed on 2 June 2023).

[5] Vander-Pallen, M.A.; Addai, P.; Isteefanos, S.; Mohd, T.K. Survey on types of cyber attacks on operating system vulnerabilities since 2018 onwards. In Proceedings of the 2022 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 6–9 June 2022; IEEE: New York, NY, USA, 2022; pp. 1–7. [Google Scholar]

[6] Kocaman, Y.; Gonen, S.; Baricskan, M.A.; Karacayilmaz, G.; Yilmaz, E.N. A novel approach to continuous CVE analysis on enterprise operating systems for system vulnerability assessment. Int. J. Inf. Technol. 2022, 14, 1433–1443. [Google Scholar] [CrossRef]

[7] Sonmez, F.O.; Hankin, C.; Malacaria, P. Attack Dynamics: An Automatic Attack Graph Generation Framework Based on System Topology, CAPEC, CWE, and CVE Databases. Comput. Secur. 2022, 123, 102938. [Google Scholar]

[8] Niu, S.; Mo, J.; Zhang, Z.; Lv, Z. Overview of linux vulnerabilities. In Proceedings of the 2nd International Conference on Soft Computing in Information Communication Technology, Taipei, Taiwan, 31 May–1 June 2014; Atlantis Press: Dordrecht, The Netherlands, 2014; pp. 225–228. [Google Scholar]

[9] Kaluarachchilage, P.K.H.; Attanayake, C.; Rajasooriya, S.; Tsokos, C.P. An analytical approach to assess and compare the vulnerability

risk of operating systems. Int. J. Comput. Netw. Inf. Secur. 2020, 12, 1. [Google Scholar] [CrossRef]
[10] Siwakoti, Y.R.; Bhurtel, M.; Rawat, D.B.; Oest, A.; Johnson, R. Advances in IoT Security: Vulnerabilities, Enabled Criminal Services, Attacks and Countermeasures. IEEE Internet Things J. 2023, 10, 11224–11239. [Google Scholar] [CrossRef]
[11] Gorbenko, A.; Romanovsky, A.; Tarasyuk, O.; Biloborodov, O. From analyzing operating system vulnerabilities to designing multiversion intrusion-tolerant architectures. IEEE Trans. Reliab. 2019, 69, 22–39. [Google Scholar] [CrossRef] [Green Version]
[12] Stallings (2005). Operating Systems, Internals and Design Principles. Pearson: Prentice Hall. p. 6.
[13] Dhotre, I.A. (2009). Operating Systems. Technical Publications. p. 1.
[14] "Desktop Operating System Market Share Worldwide". StatCounter Global Stats. Archived from the original on 2 October 2023. Retrieved 3 October 2023.
[15] "Mobile & Tablet Operating System Market Share Worldwide". StatCounter Global Stats. Retrieved 2 October 2023.
[16] "VII. Special-Purpose Systems - Operating System Concepts, Seventh Edition [Book]". www.oreilly.com. Archived from the original on 13 June 2021. Retrieved 8 February 2021.
[17] "Special-Purpose Operating Systems - RWTH AACHEN UNIVERSITY Institute for Automation of Complex Power Systems - English". www.acs.eonerc.rwth-aachen.de. Archived from the original on 14 June 2021. Retrieved 8 February 2021.
[18] Lorch, Jacob R.; Smith, Alan Jay (1996). "Reducing processor power consumption by improving processor time management in a single-user operating system". Proceedings of the 2nd annual international conference on Mobile computing and networking. New York, NY, US: ACM. pp. 143–154. doi:10.1145/236387.236437. ISBN 089791872X.
[19] Akhtar,Z.(2024).Securing Operating Systems (OS): A Comprehensive Approach to Security with Best Practices and Techniques. International Journal of Advanced Network, Monitoring and Controls,9(1) 100-111. https://doi.org/10.2478/ijanmc-2024-0010
[20] Javed, F.; Afzal, M.K.; Sharif, M.; Kim, B. Internet of Things (IoT) Operating Systems Support, Networking Technologies, Applications, and Challenges: A Comparative Review. IEEE Commun. Surv. Tutor. 2018, 20, 2062–2100. [Google Scholar] [CrossRef]
[21] Asghar, A.; Farooq, H.; Imran, A. Self-Healing in Emerging Cellular Networks: Review, Challenges, and Research Directions. IEEE Commun. Surv. Tutor. 2018, 20, 1682–1709. [Google Scholar] [CrossRef]
[22] Li, L.; Zhao, G.; Blum, R.S. A Survey of Caching Techniques in Cellular Networks: Research Issues and Challenges in Content Placement and Delivery Strategies. IEEE Commun. Surv. Tutor. 2018, 20, 1710–1732. [Google Scholar] [CrossRef]
[23] Naik, G.; Liu, J.; Park, J.J. Coexistence of Wireless Technologies in the 5 GHz Bands: A Survey of Existing Solutions and a Roadmap for Future Research. IEEE Commun. Surv. Tutor. 2018, 20, 1777–1798. [Google Scholar] [CrossRef]
[24] Mukherjee, M.; Shu, L.; Wang, D. Survey of Fog Computing: Fundamental, Network Applications, and Research Challenges. IEEE Commun. Surv. Tutor. 2018, 20, 1826–1857. [Google Scholar] [CrossRef]
[25] MacHardy, Z.; Khan, A.; Obana, K.; Iwashina, S. V2X Access Technologies: Regulation, Research, and Remaining Challenges. IEEE Commun. Surv. Tutor. 2018, 20, 1858–1877. [Google Scholar] [CrossRef]
[26] Dunkels, A. Full TCP/IP for 8-bit Architectures. In Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, San Francisco, CA, USA, 5–8 May 2003; pp. 85–98. [Google Scholar] [CrossRef]
[27] Al-Boghdady, A.; Wassif, K.; El-Ramly, M. The Presence, Trends, and Causes of Security Vulnerabilities in Operating Systems of IoT's Low-End Devices. *Sensors* 2021, *21*, 2329. https://doi.org/10.3390/s21072329
[28] Bazuku , R., Anab , A., Gyemerah , S., & Daabo , M. I. (2023). An Overview of Computer Operating Systems and Emerging Trends. Asian Journal of Research in Computer Science, 16(4), 161–177. https://doi.org/10.9734/ajrcos/2023/v16i4380
[29] X. Rong, "Design and Implementation of Operating System in Distributed Computer System Based on Virtual Machine," *2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI)*, Ottawa, ON, Canada, 2020, pp. 94-97, doi: 10.1109/ICAACI50733.2020.00024.
[30] Mishra, B.; Singh, N.; Singh, R. (2014). "Master-slave group based model for co-ordinator selection, an improvement of bully algorithm". International Conference on Parallel, Distributed and Grid Computing (PDGC). pp. 457–460. doi:10.1109/PDGC.2014.7030789. ISBN 978-1-4799-7682-9. S2CID 13887160.
[31] Hansen, Per Brinch, ed. (2001). Classic Operating Systems. Springer. pp. 4–7. ISBN 0-387-95113-X. Archived from the original on 11 January 2023. Retrieved 19 December 2020.
[32] "Intel® Microprocessor Quick Reference Guide - Year". www.intel.com. Archived from the original on 25 April 2016. Retrieved 24 April 2016.
[33] Arthur, Charles (5 January 2011). "'Windows 8' will run on ARM chips - but third-party apps will need

rewrite". The Guardian. Archived from the original on 12 October 2016.
[34] "Behind the IDC data: Windows still No. 1 in server operating systems". ZDNet. 26 February 2010. Archived from the original on 1 March 2010.
[35] Hyde, Randall (1996). "Chapter Seventeen: Interrupts, Traps and Exceptions (Part 1)". The Art Of Assembly Language Programming. No Starch Press. Archived from the original on 22 December 2021. Retrieved 22 December 2021. "The concept of an interrupt is something that has expanded in scope over the years. The 80x86 family has only added to the confusion surrounding interrupts by introducing the int (software interrupt) instruction. Indeed, different manufacturers have used terms like exceptions, faults, aborts, traps and interrupts to describe the phenomena this chapter discusses. Unfortunately, there is no clear consensus as to the exact meaning of these terms. Different authors adopt different terms to their own use."
[36] PDP-1 Input-Output Systems Manual (PDF). Digital Equipment Corporation. pp. 19–20. Archived (PDF) from the original on 25 January 2019. Retrieved 16 August 2022.
[37] "Reading: Operating System". Lumen. Archived from the original on 6 January 2019. Retrieved 5 January 2019.

Journal of Technology and Informatics (JoTI)
Vol. 6, No. 1, October 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v6i1.679

Machine Learning Approach for Classification of Cyber Threats Actors in Web Region

Anthony Edet[1], Saviour Inyang[2], Imeh Umoren[3], Ubong E. Etuk[4]
[1,2,3,] Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria.
[4]School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK.
anthonyedet73@gmail.com[1], sinyang5672@gmail.com[2], imehumoren@aksu.edu.ng[3], u.etuk.1@research.gla.ac.uk[4]
* Corresponding author: E-mail: anthonyedet73@gmail.com

Abstract: *In the interconnected scape of today's internet, the dark web emerges as a concealed point, covering a myriad of illicit activities that pose substantial cybersecurity risks. This study investigates the attribution of threats within the dark web environment, leveraging on a machine learning approach to bridge the gap between technical indicators and linguistic and behavioral insights. Through a comprehensive methodology involving web crawling and data gathering, a dataset encompassing key variables such as attack motivation, method, web part, and threat actor was gathered. Principal Component Analysis was employed for feature selection, followed by the application of Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), Random Forest (RF), and CatBoost algorithms for classification. Performance evaluation metrics including precision, recall, and F1-score were utilized to assess the efficacy of each algorithm. Results indicate a notable prevalence of cybercrimes within the dark web, underscoring the necessity for enhanced cybersecurity strategies tailored to address its unique challenges. Furthermore, the comparative analysis demonstrates varying performance levels among the machine learning algorithms, with Multinomial Naive Bayes exhibiting the highest accuracy. This research contributes to advancing threat attribution techniques in the dark web, ultimately aiming to bolster cybersecurity defenses and mitigate future cyber threats.*

Keywords*: Threats;; Cybersecurity; Attribution; Machine Learning*

INTRODUCTION

In today's interconnected world, the dark web stands as a shadowy underbody of the internet, harboring an array of illicit activities that pose significant threats to cybersecurity. The dark web refers to the hidden, encrypted part of the internet that is not indexed by traditional search engines [1]. It is often associated with illegal and illicit activities due to its anonymity and privacy features, making it a breeding ground for cybercriminals. It has become a hub for cybercriminals, offering a haven for activities such as the sale of stolen data, hacking tools [2], illegal substances, and the organization of cyberattacks for financial gain, political motives [3], or other nefarious purposes. The anonymity and encrypted communication channels prevalent on the dark web make it exceptionally challenging to track, identify, and attribute these threats to specific actors or groups [4]. The dark web's elusive nature is compounded by the use of sophisticated complication techniques and the constant adaptation of its users to evade detection [5].

Nevertheless, traditional cybersecurity methods often rely on technical indicators such as IP addresses, malware signatures, and known vulnerabilities for threat identification [6]. These indicators, while valuable, often fall short in providing a complete picture of the threat layout and a target response, as they can be easily manipulated, concealed, or shared among different actors [7]. A targeted response in cybersecurity refers to a specific and tailored action taken to defend against or mitigate a cyber threat [8]. It is based on the attributes and motivations of the threat actor, as well as the insights obtained through threat attribution. Targeted responses can include legal actions [9], enhanced monitoring [10], and customized cybersecurity measures designed to counter the specific tactics used by the threat actor [11]. In response to this evolving cyberspace, the need for a comprehensive cybersecurity system that efficiently maps dark web activities to known threat actors has become increasingly urgent [12]. Therefore, Threat attribution involves the process of determining the origins and identities of cyber threats and threat actors which is paramount for understanding, capabilities, and modus operandi of threat actors [13], such as the source or origin of a cyber threat and identifying the individuals, groups, responsible for a cyberattack or other malicious activities [14]. Accurate threat attribution is critical for developing effective cybersecurity responses [15]. This understanding is essential for devising effective countermeasures, enhancing cybersecurity postures, and safeguarding against future attacks. [16].

Consequently, attributing cyber threats is a great challenge. It requires a combination of technological involvement [17], human intelligence, and cybersecurity expertise [18]. The traditional emphasis on technical indicators, while valuable, is only one piece of the puzzle. Hence, a more comprehensive approach must also account for the linguistic and behavioral aspects of threat actors, as well as the broader contextual information that can be assembled from dark web forums, marketplaces, and chat rooms through dark web crawling and other means [19]. On this note this study proposed a solution to meet the challenge of threat attribution in the dark web environment using Machine learning Approach [20]. Adopting Machine learning techniques will help bridge the gap between technical indicators [21] and linguistic and behavioral insights. By doing so, it seeks to enhance the efficiency and accuracy of the threat attribution process, ultimately leading to more effective and targeted responses.

METHOD

In this study, the method adopted in this research is based framework-based method [22] which is presented in figure 1 which serves as a workflow diagram in the method in this study.

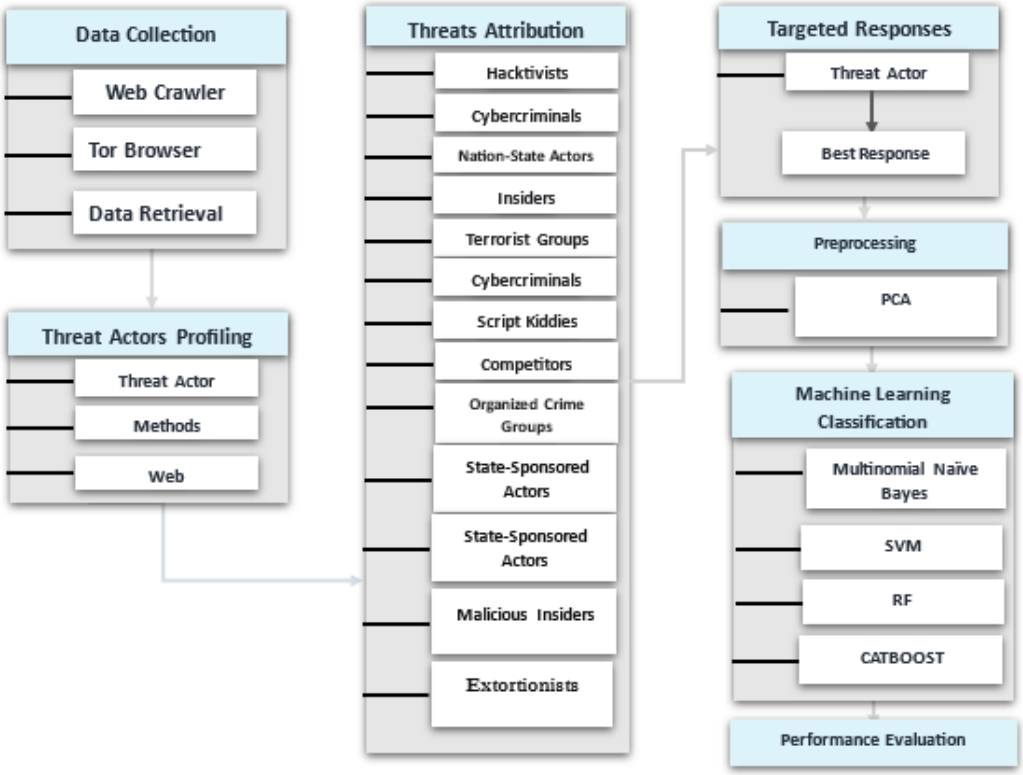


Figure 1: Workflow for Classification of Cyber Threats

A. Data Collection

In this section, the data gathering process is describe as follows;

i. **Web Crawler Tool:** To systematically traverse the Dark Web and collect relevant data for this study, a specialized web crawler tool scrapy based on python was employed. This tool is aiding the navigation through the dark web sites, extracting pertinent information related to cyber threats and their associated actors.

ii. **Tor Browser**: Access to the Dark Web was facilitated through the Tor browser, which ensures anonymity and secure data retrieval. The Tor browser enables the web crawler tool to operate within the Dark Web environment without compromising the security and privacy of the data collection process.

Data Retrieval: This process involved extracting detailed information about threats, threat actors, and their methodologies from the Dark Web. This information includes identities, motivations, capabilities, affiliations, and records of previous cyberattacks. The retrieval process was automated to handle the vast amount of data available on the Dark Web. A total of 10002 dataset was scrabe from the dark web using Scrapy tool which was process and use in the machine learning section of this study for classification of threat actors.

B. Threat Actors Profiling

This involves creating comprehensive profiles of the threat actors found on the dark web. These profiles should include information such as their identities, motivations, capabilities, affiliations, previous cyberattacks, and other relevant details. The aim is to build a clear picture of who these individuals or groups are. Table 1 shows the comprehensive profile of dark web threat actors:

Table 1: Threat Actors, Motivation, Methods and Web

Threat Actor	Motivation	Methods	Web
Hacktivists	Activism	Disruption	Surface Web
Cybercriminals	Financial gain	Malware attacks	Dark Web
Nation-State Actors	Espionage	APTs	Surface Web
Insiders	Malicious intent	Unauthorized access	Surface Web
Terrorist Groups	Disruption	Cyber-terrorism	Surface Web
Script Kiddies	Thrill-seeking	Exploits	Surface Web
Competitors	Corporate espionage	IP theft	Dark Web
Organized Crime Groups	Financial gain	Ransomware	Dark Web

State-Sponsored Actors	Geopolitical influence	Cyber-espionage	Surface Web
Malicious Insiders	Revenge	Data theft	Surface Web
Extortionist	Financial gain	Ransomware	Dark Web
Hacktivist Groups	Activism	Coordinated cyber campaigns	Surface Web
APT Groups	Long-term cyber-espionage	Highly sophisticated attacks	Surface Web
Rogue Employees	Internal dissatisfaction	Exploiting internal access	Surface Web
Data Brokers	Profit from selling stolen data	Acquiring and selling data on the dark web	Dark Web
Hacktivist Collectives	Collective activism	Coordinated cyber campaigns	Surface Web
Mercenary Hackers	Hired for cyber-attacks	Carrying out attacks for others	Dark Web

C. Threat Attribution

Accurate threat attribution is crucial in cybersecurity because it enables organizations, law enforcement, and cybersecurity professionals to understand who is behind an attack. This knowledge is essential for taking appropriate countermeasures, including legal action, if necessary. Here we map each threat in the dark web against an actor to clearly model the best response. It is a combination of the threat actors profile and threats carried out by them. Below is the threat attribution list:

Hacktivists: Malware, Phishing, Denial of Service (DoS) Attack, Distributed Denial of Service (DDoS) Attack, Social Engineering, Form Spoofing, DNS Hijacking, Clickjacking, Hijacking Clicks, Click Fraud

Cybercriminals: Malware, Phishing, Denial of Service (DoS) Attack, Distributed Denial of Service (DDoS) Attack, Man-in-the-Middle (MitM) Attack, SQL Injection, Cross-Site Scripting (XSS), Zero-Day Exploits, Ransomware, Credential Stuffing, Fileless Malware, IoT-Based Attacks, Supply Chain Attacks, Crypto-Malware

Nation-State Actors: Advanced Persistent Threats (APTs), Cyber-espionage, Zero-Day Exploits, Social Engineering, DNS Hijacking, Clickjacking, Watering Hole Attack, Eavesdropping, USB-Based Threats, SIM Card Swapping, Macro-Based Malware, DNS Tunneling.

Insiders: Unauthorized access, Data theft, Sabotage, Social Engineering, Insider Threats, Clipboard Hijacking, File Upload Vulnerabilities, Cross-Site Request Forgery (CSRF), Session Hijacking, Formless Attack

Terrorist Groups: Cyber-terrorism, Phishing, Exploits, Social Engineering, Denial of Service (DoS) Attack, Distributed Denial of Service (DDoS) Attack, Form Spoofing, Clickjacking, Hijacking Clicks, File Upload Vulnerabilities, HTTP Request Smuggling, Fileless Malware, AI-Powered Attacks, ATM Skimming, Formless Attack, Evil Twin Wi-Fi Attack, Juice Jacking, Hijacking Clicks, Shadow IT, Honeypot Attacks, Distributed.

Script Kiddies: Exploits, Malware attacks, Denial of Service (DoS) Attack, Phishing, Social Engineering, Ransomware, File Upload Vulnerabilities, Clickjacking, DNS Tunneling, Formless Attack, Click Fraud

Competitors: IP theft, Phishing, Ransomware, Social Engineering, Form Spoofing, DNS Hijacking, Clickjacking, Watering Hole Attack, USB-Based Threats, Macro-Based Malware, DNS Tunneling, File-Extension Spoofing

Organized Crime Groups: Ransomware, Malware attacks, Data Exfiltration, Social Engineering, Phishing, Denial of Service (DoS) Attack, Distributed Denial of Service (DDoS) Attack, Formjacking, Bluejacking, Whaling, Clickjacking, Watering Hole Attack, Eavesdropping, USB-Based Threats, SIM Card Swapping, Macro-Based Malware, Brute Force Attacks, File Upload Vulnerabilities, Cross-Site Request Forgery (CSRF), Crypto-Malware.

State-Sponsored Actors: Advanced Persistent Threats (APTs), Cyber-espionage, Zero-Day Exploits, Social Engineering, DNS Hijacking, Clickjacking, Watering Hole Attack, Eavesdropping, USB-Based Threats, SIM Card Swapping, Macro-Based Malware, DNS Tunneling, AI-Powered Attacks, EternalBlue Exploit, Formless Attack, Smart Contract Exploits, AI-Enhanced Social Engineering, Packet Sniffing, Backdoor Exploits

Malicious Insiders: Unauthorized access, Data theft, Sabotage, Social Engineering, Insider Threats, Clipboard Hijacking, File Upload Vulnerabilities, Cross-Site Request Forgery (CSRF), Session Hijacking, Formless Attack

Extortionists: Ransomware, Denial of Service (DoS) Attack, Phishing, Social Engineering, Form Spoofing, DNS Hijacking, Clickjacking, Watering Hole Attack, USB-Based Threats, Macro-Based Malware, DNS Tunneling, File-Extension Spoof.

D. Targeted Response

Once the dark web activities have been linked to known threat actors or campaigns and detailed profiles are constructed, organizations can develop more effective and targeted response strategies. This might involve enhancing cybersecurity defenses, sharing threat intelligence with other organizations or authorities, or taking legal action against the threat actors. In this work, responses have equally been matched against individual threat and actors to facilitate targeted response. This study presents the targeted response in table 2.

Table 2: Targeted responses

Threat Actor	Best Responses
Hacktivists	Enhance cybersecurity measures, monitor for DDoS attacks, educate users on phishing prevention.
Cybercriminals	Employ robust antivirus software, conduct regular security audits, implement intrusion detection systems.
Nation-State Actors	Invest in advanced threat intelligence, use network segmentation, conduct regular security assessments.
Insiders	Implement least privilege access, conduct thorough background checks, monitor user activities.
Terrorist Groups	Strengthen cybersecurity infrastructure, collaborate with law enforcement agencies, monitor for unusual activities.
Script Kiddies	Educate users on basic cybersecurity, implement intrusion detection systems, enforce strong password policies.
Competitors	Secure intellectual property, use encryption for sensitive data, conduct regular security training.
Organized Crime Groups	Implement advanced threat detection, conduct penetration testing, secure critical data with encryption.
State-Sponsored Actors	Implement advanced threat intelligence, use network segmentation, conduct regular security assessments.
Malicious Insiders	Implement least privilege access, conduct thorough background checks, monitor user activities.
Extortionists	Regularly backup critical data, use endpoint protection, educate users on phishing prevention.
Hacktivist Groups	Enhance cybersecurity measures, monitor for DDoS attacks, educate users on phishing prevention.
APT Groups	Invest in advanced threat intelligence, use network segmentation, conduct regular security assessments.
Rogue Employees	Implement least privilege access, conduct thorough background checks, monitor user activities.
Data Brokers	Encrypt sensitive data, implement secure data sharing practices, conduct regular security audits.
Hacktivist Collectives	Enhance cybersecurity measures, monitor for DDoS attacks, educate users on phishing prevention.
Mercenary Hackers	Employ advanced threat detection tools, use network segmentation, conduct regular security assessments.

E. Data Collection

This study delves into the profiling of Darkweb threat actors and the attribution of threats to known actors on the surface web. To achieve this objective, a web crawler tool was employed to systematically traverse the Darkweb, gathering specific data features and values. Access to the Darkweb was facilitated through the Tor browser, enabling the retrieval of information pertaining to threats, threat actors, and attack methodologies, which was then stored in a CSV format for subsequent machine learning analysis. Hence, the cross-sectional of the data set gathered figure 2.

INPUT VARIABLES			OUTCOME (CLASS VARIABLE)
attack_motivation	attack_method	web_part	threat_actor
Ransomware	Malware	Surface Web	Mercenary Hackers
Political Ideology	Advanced Persistent Thr	Surface Web	Hacktivist Groups
Financial Gain	Ransomware	Dark Web	Organized Crime Groups
Financial Gain	Advanced Persistent Thr	Surface Web	Extortionists
Financial Gain	Phishing	Surface Web	Extortionists
Criminal Activities	Supply Chain Attacks	Surface Web	Mercenary Hackers
Espionage	Ransomware	Surface Web	APT Groups
Political Ideology	Phishing	Surface Web	Hacktivist Collectives
Internal Disruption	Zero-Day Exploits	Dark Web	Insiders
National Interest	Advanced Persistent Thr	Surface Web	APT Groups
Ransomware	Advanced Persistent Thr	Surface Web	Mercenary Hackers
Advanced Persistent Threa	Social Engineering	Surface Web	APT Groups
Internal Disruption	Zero-Day Exploits	Surface Web	Rogue Employees
Political Ideology	Cryptojacking	Surface Web	Hacktivist Groups
Data theft	Malware	Dark Web	Extortionists
Ideological	Zero-Day Exploits	Dark Web	Terrorist Groups
Financial Gain	Phishing	Dark Web	Organized Crime Groups
Internal Disruption	Social Engineering	Dark Web	Rogue Employees
Data Monetization	SQL Injection	Dark Web	Data Brokers
Political Ideology	Cryptojacking	Dark Web	Hacktivists
Disruption	Zero-Day Exploits	Surface Web	Script Kiddies

Figure 2: Cross Section of the data.

Furthermore, the dataset comprises four key variables, as depicted in the structure outlined in Figure3: attack motivation, attack method, web part, and threat actor. These variables were meticulously selected during the dataset collection process to ensure the dataset encompassed sufficient features conducive to the analytical framework elucidated in this study.

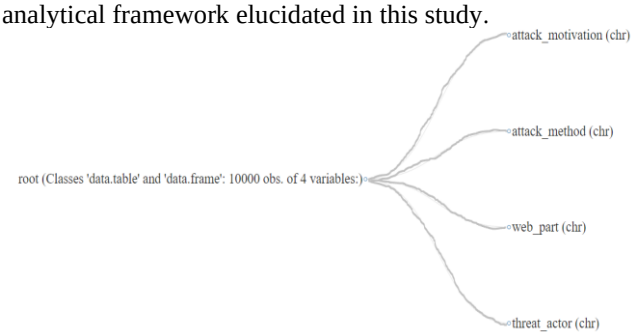


Figure 3: Dataset Structure

F. Preprocessing

Data preprocessing is an important step to prepare the data which will be used for the formation of a model. Data cleaning, data transformation, and feature selection are all key phases in data preprocessing [26]. In this study, Principal component Analysis will be used in feature selection. From the data gathered, PCA was used for feature selection.

First in this study to carryout feature ranking, we will take the hold dataset consisting of $d+1$ dimension. Furthermore, we compute the mean and dimension of every section in the dataset, Again, the computation of the covariance matrix is performed in in eqn 1

$$CO(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})\ (y-\bar{y}) \qquad 1$$

Furthermore, the computation of eigenvectors and their corresponding eigenvectors will be carried out appropriate and we sort the eigenvectors in ascending order. Finally, we, transform the data into the new subspace using this $d * 1$ eigenvector matrix where each feature can be rank as principal components which is shown in figure 4 depicts the different principal components in the datasets and rank the components based on their relevance.

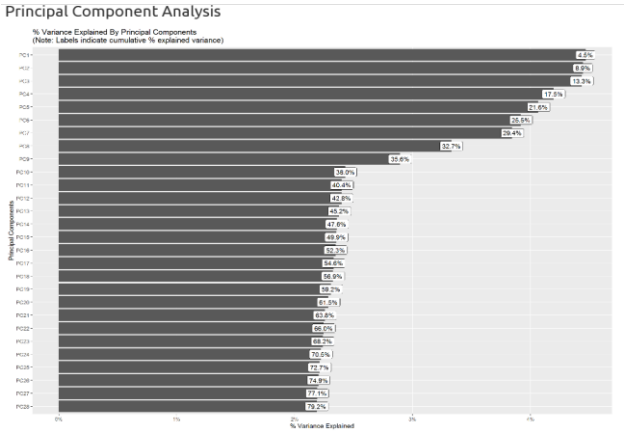


Figure 4: Principal Components ranks

G. Machine Learning Classification

Machine learning classification stands as the fundamental task of sorting input data into distinct classes, drawing upon one or more defining variables [24]. This pivotal process operates under the of supervised learning category, wherein meticulously labeled training datasets are used for algorithms to discern patterns and effectively categorize forthcoming data points [25]. Through a diverse array of algorithms and methodologies, classification endeavors to assign future datasets into relevant categories, thus facilitating informed decision-making and predictive analytics in this study, Multinomial Naive Bayes, Support Vector Machine, Random Forest and CATBOOT algorithms were used for effective classification and comparative result analysis.

H. Performance Evaluation

Performance Evaluation in this study involves the assessment of the machine learning algorithms performances that was use in the classification process in order to ascertain how better each algorithm performs. here confusion matrix will be used for the evaluation of the algorithms while recall, precision and f1score and supports will also aid in showing the performance of each training processes of the ML models.

RESULT AND DISCUSSION

Threat actors play a extensive role, making contributions to the cyber threat data, while others contribute comparatively less. Nevertheless, the figure 5 underscores that all categories of threat actors in this study that consistently remain active, instigating issues and wreaking havoc in various ways. This observation highlights that hackers within these profiles are consistently engaged, conducting daily activities to compromise organizations and execute attacks aligned with their motivations.

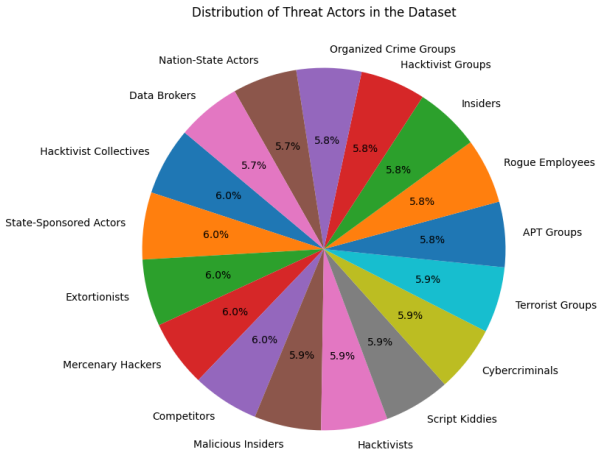


Figure 5: Distribution of Threat Actors

Correlation heatmap is presented in figure 7 illustrating the relationships among the features in the dataset. This visualization provides a clear depiction of the interconnections between each feature within the threat dataset, meaning that any threat actor can employ any type of threat to cause havoc.

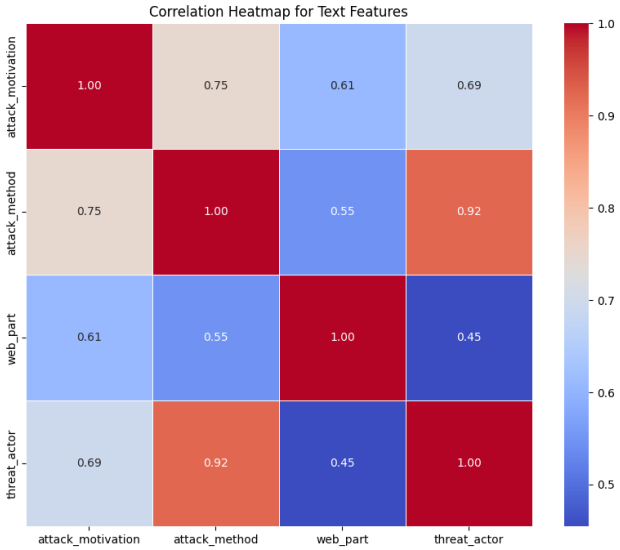


Figure 6: Correlation Heatmap of Features

Furthermore, Machine Learning algorithms was applied in the classification of threat actors based on crucial factors such as motivation, attack method, attack type, and the specific web part utilized. Multinomial Naive Bayes accuracy gave 94%, Support Vector Machine yielded an accuracy of 82%, Random Forest Algorithm gave 82% and CatBoost gave 82% classification results respectively, these results are presented in table 3 to table 5.

Table 3: Multinomial Naïve Bayes

	Precision	Recall	f1-score	support
accuracy			0.94	10000
macro avg	0.91	0.94	0.92	10000
weighted avg	0.91	0.94	0.92	10000

Table 4: SVM

	Precision	Recall	f1-score	support
accuracy			0.82	10000
macro avg	0.75	0.82	0.77	10000
weighted avg	0.75	0.82	0.77	10000

Table 5: Random Forest

	Precision	Recall	f1-score	support
accuracy			0.82	10000
macro avg	0.75	0.82	0.77	10000
weighted avg	0.75	0.82	0.77	10000

Table 6: CatBoost

	Precision	Recall	f1-score	support
accuracy			0.82	10000
macro avg	0.75	0.82	0.77	10000
weighted avg	0.75	0.82	0.77	10000

Furthermore, a comparative analysis is presented in figure 8 comparing all the algorithm and how each performs better than others in the classification process.

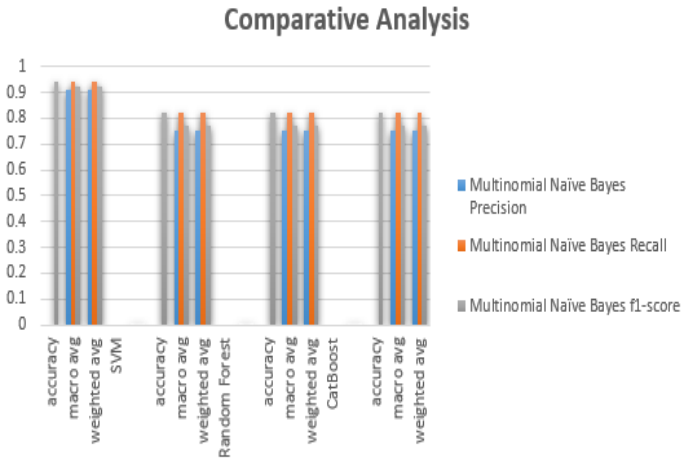


Figure 7: Comparative Analysis

Nevertheless, from the achieve result in this study, we present the findings that is emphasize the relevance of this study as follows:

1. **Consistent Activity of Threat Actors**: Threat actors remain actively engaged, conducting daily activities aimed at compromising organizations and executing attacks aligned with their motivations. This highlights the persistent and evolving nature of cyber threats.
2. **Interconnected Features in Threat Dataset**: The correlation heatmap in Figure 6 illustrates the complex interconnections among features within the threat dataset. This indicates that threat actors can employ a wide range of threat types, adding to the complexity of defending against cyber threats.
3. **Algorithm Performance**: The application of ML algorithms for classifying threat actors yielded varying levels of accuracy, with Multinomial Naive Bayes achieving the highest accuracy at 94%. This suggests that the probabilistic nature of Multinomial Naive Bayes makes it particularly effective for this classification task.

Hence, this study highlights the persistent and multifaceted nature of cyber threats, emphasizing the need for robust and comprehensive cybersecurity measures. The findings also demonstrate the effectiveness of using ML algorithms, particularly Multinomial Naive Bayes, in accurately classifying threat actors based on various features. This study contributes to a deeper understanding of the dynamics of cyber threats and provides valuable insights for enhancing cybersecurity defenses.

CONCLUSION AND SUGGESTION

Insufficient threat attribution and profiling within the dark web pose significant challenges and limitations, largely stemming from the intricate and covert nature of activities in this concealed cyberspace. Addressing these complexities, the present research introduces a novel solution using a Machine learning approach. The application of these Machine Learning algorithms has significantly contributed to our understanding and classification of threat actors based on crucial factors such as motivation, attack method, attack type, and the specific web part utilized. Multinomial Naive Bayes demonstrated the highest accuracy of 94%, followed by SVM, Random Forest, and CatBoost, each yielding 82% During the training and validation of the datasets. Through the systematic analysis of dark web data, the research seeks to provide a comprehensive and accurate understanding of activities occurring in this hidden environment known as the dark web. In the course of this study, we have accomplished a comprehensive analysis encompassing threat attribution, threat actor profiling, and the classification of threat actors. Our endeavors have yielded valuable insights into the intricate domain of cyber threats, particularly within the concealed domain of the dark web. A significant achievement of this work is the successful identification and mapping of cyber threats to threat actors originating from the dark web. Through meticulous analysis and using advanced techniques, we have proficiently correlated these threats in the dark web with known threat actors in the surface web. This not only enhances the general understanding of the diverse motives driving malicious activities but also fortifies our ability to discern and attribute threats emanating from this elusive and often obscured corner of the digital space. The conclusion of our efforts reveals the importance of a holistic approach to cybersecurity, extending beyond surface-level defenses to address the intricacies of the dark web. The successful execution of

threat attribution and actor classification positions this study serves as a valuable contribution to the ongoing pursuit of cyber resilience and the safeguarding of digital space against evolving threats. For further studies, we recommend that focus should be on the inclusion of other web parts other than surface and dark web addressed in this research.

REFERENCES

[1] H. Albasheer, M. Siraj, A. Mubarakali, O. Elsier Tayfour, S. Salih, M. Hamdan, S. Khan, A. Zainal, and S. Kamarudeen, "Cyber-Attack Prediction Based on Network Intrusion Detection Systems for Alert Correlation Techniques: A Survey," Sensors, vol. 22, no. 4, pp. 1494, 2022. [Online]. Available: https://doi.org/10.3390/s22041494.

[2] S. Amit, J. Jay, and K. Gaurav, "Intrusion Detection System: A Comparative Study of Machine Learning-based IDS," Journal of Research Square, vol. 3, no. 2, pp. 1-30, 2022. [Online]. Available: https://doi.org/10.21203/rs.3.rs-1634802/v1.

[3] L. Ashiku and C. H. Dagli, "Network Intrusion Detection System using Deep Learning," Procedia Computer Science, vol. 185, no. 6, pp. 239-247, June 2021. [Online]. Available: https://doi.org/10.1016/j.procs.2021.05.025.

[4] Elsayed, H. A. G., Chaffar, S., & Belhaouari, S. B. (2020). A two-level deep learning approach for emotion recognition in Arabic news headlines. International Journal of Computers and Applications, vol. 44, no. 7, pp. 604-613. doi: 10.1080/1206212X.2020.1851501.

[5] E. A. Emad, E. Wafa, and F. O. Ahmed, "Intrusion Detection Systems using Supervised Machine Learning Techniques: A Survey," in International Conference on Ambient Systems Networks and Technologies, vol. 8, no. 3, pp. 205-212, 2022.

[6] S. B. Erukala, S. R. Mekala, P. Rambabu, R. N. Soumya, and S. Achyut, "A Hybrid Intrusion Detection System against Botnet Attack in IoT using Light Weight Signature and Ensemble Learning Technique," Research Square Journals, vol. 4, no. 2, pp. 1-17, 2022. [Online]. Available: https://doi.org/10.1109/IndiaCom.2014.6828073.

[7] A. D. Evarakonda, N. Sharma, P. Saha, and S. Ramya, "Network intrusion detection: a comparative study of four classifiers using the NSL-KDD and KDD'99 datasets," Research Square Journals, vol. 4, no. 2, pp. 1-17, 2022. [Online]. Available: https://doi.org/10.1088/1742-6596/2161/1/012043

[8] J. D. Gadze, A. A. Bamfo-Asante, J. O. Agyemang, H. Nunoo-Mensah, and K. A. B. Opare, "An Investigation into the Application of Deep Learning in the Detection and Mitigation of DDOS Attack on SDN Controllers," Technologies, vol. 9, no. 1, pp. 14-29, 2021. doi: 10.3390/technologies9010014.

[9] X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu, "An Adaptive Ensemble Machine Learning Model for Intrusion Detection," in IEEE Access, vol. 7, pp. 82512-82521,2019.

[10] L. Guangrui, Z. Weizhe, L. Xinjie, F. Kaisheng, and Y. Shui, "VulnerGAN: a backdoor attack through vulnerability amplification against machine learning-based network intrusion detection systems," *Science China Information Sciences*, vol. 65, no. 1, pp. 1-19, 2022.

[11] K. Gurbani and K. Dharmender, "Classification of Intrusion using Artificial Neural Network with GWO," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 4, pp. 599-606, 2020.

[12] I. A. Hidayat and A. Arshad, "Machine learning based intrusion detection system: an experimental comparison," Journal of Computational and Cognitive Engineering, vol. 3, no. 1, pp. 23-43, 2022.

[13] Ziaul, H., "Cyber Security of the Maritime ICTs, Threat Vectors and Implications on Global Sea Lanes of Commerce (SLOC)," *Global Journal of Science Frontier Research: E Marine Science*, vol. 23, no. 1, pp. 1-10, 2023.

[14] Alanazi, A. T., "Clinicians' Perspectives on Healthcare Cybersecurity and Cyber Threats," Cureus, vol. 15, no. 10, e47026, Oct. 2023. doi: 10.7759/cureus.47026

[15] O. Zahra, Z. El, C. Habiba, and B. Salmane, "Cyber-attack crisis management in the context of energy companies," Web of Conferences, vol. 412, p. 01106, 2023. doi: 10.1051/e3sconf/202341201106

[16] N. Jeffrey, Q. Tan, and J.R. Villar, "A Review of Anomaly Detection Strategies to Detect Threats to Cyber-Physical Systems," Electronics, vol. 12, no. 15, pp. 3283, Aug. 2023. doi: 10.3390/electronics12153283.

[17] A. Jawad, A. S. Syed, L. Shahid, A. Fawad, Z. Zhuo, and P. Nikolaos, "A deep random neural network with particle swarm optimization for intrusion detection in the industrial internet of things," Journal of King Saud University – Computer and Information Sciences, vol. 3, no. 3, pp. 1-10, 2022. [Online].Available:https://pureportal.coventry.ac.uk/files/57466468/Published.pdf

[18] S. B. Erukala, S. R. Mekala, P. Rambabu, R. N. Soumya, and S. Achyut, "A Hybrid Intrusion Detection System against Botnet Attack in IoT using Light Weight Signature and Ensemble Learning Technique," Research Square Journals, vol. 4, no. 2, pp. 1-17, 2022.

[19] K. Wolsing, W. Eric, S. Antoine, and H. Martin, "Breaking up Silos of Protocol-dependent and Domain-specific Industrial Intrusion Detection Systems," in *International Symposium on Research in Attacks, Intrusions and Defenses*, 2020, pp. 1-17. doi: 10.1145/3545948.3545968

[20] I. J Umoren & S. J. Inyang, "Methodical Performance Modelling of Mobile Broadband

Networks with Soft Computing Model," International Journal of Computer Applications, vol. 174, no. 25, pp. 7-21, 2021.

[21] A. E. Edet and G. O. Ansa, "Machine Learning Enabled System for Efficient Classification of Intrusion Severity," Global Journal of Engineering and Technology Advances, vol. 16, no. 3, pp. 41-50, 2023.

[22] S. Inyang and I. Umoren, "From Text to Insights: NLP-Driven Classification of Infectious Diseases Based on Ecological Risk Factors," Journal of Innovation Information Technology and Application (JINITA), vol. 5, no. 2, pp. 154-165, 2023.[Online].Available:https://ejournal.pnc.ac.id/index.php/jinita/article/view/2084.

JoTI

Journal of Technology and Informatics