POLITEKNIK NEGERI BALI

matriX

JURNAL MANAJEMEN TEKNOLOGI DAN INFORMATIKA

Vol. 16 No. 1 | Page 1-61 | March 2026 | ISSN 2580-5630

Editors

Reviewers

PREFACE

We would like to present, with great pleasure, the first issue of Matrix: Jurnal Manajemen Teknologi dan Informatika in Volume 16, 2026. This journal is under the management of Scientific Publication, Research and Community Service Center, Politeknik Negeri Bali, and is devoted to covering the field of technology and informatics management including managing the rapid changes in information technology, emerging advances in electrical and electronics and new applications, implications of digital convergence and growth of electronics technology, and project management in electrical. The scientific articles published in this edition were written by researchers from Universitas Pendidikan Ganesha, STMIK Bandung Bali, Universitas Islam Al-Azhar, Universitas Qamarul Huda Badaruddin Bagu, Universitas Bunda Mulia, Politeknik Negeri Bali, and Universitas Brawijaya. Articles in this issue cover topics in the field of Application of conditional lightweight GAN for retinal fundus image synthesis based on diabetic retinopathy severity levels on the IDRiD dataset, Integration of local wisdom and modern medicine in a treatment recommendation system for toddlers based on the Case-Based Reasoning-Fuzzy Method, Bell's Palsy and stroke face classification using SVM with MediaPipe Face Mesh, User experience testing on Smart Human Capital Dashboard (SHUCADA) from PT Studio Kami Mandiri using User Experience Questionnaire (UEQ), and Model creation for Denial of Service (DoS) attack classification using an ensemble learning approach on multi-dataset network traffic. Finally, we would like to thank the reviewers for their efforts and hard work in conducting a series of review phases thoroughly based on their expertise. We hope that the work of the authors in this issue will be a valuable resource for other researchers and will stimulate further research into the vibrant area of technology and information management in specific, and engineering in general.

Politeknik Negeri Bali, 26 March 2026

Editor-in-chief

Dewa Ayu Indah Cahya Dewi, S.TI., M.T.

ISSN: 2580-5630

Table of contents

.

Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 16 Issue 1 Year 2026 Pages 1-11
DOI: http://dx.doi.org/10.31940/matrix.v16i1.1-11
URL: http://ojs2.pnb.ac.id/index.php/matrix

Application of conditional lightweight GAN for retinal fundus image synthesis based on diabetic retinopathy severity levels on the IDRiD dataset

Acep Taufik Hidayat [1,3]*, I Made Dendi Maysanjaya [2], I Made Gede Sunarya [1], Made Windu Antara Kesiman [1]

[1] *Graduate Program in Computer Science, Universitas Pendidikan Ganesha, Indonesia*
[2] *Information Systems Study Program, Universitas Pendidikan Ganesha, Indonesia*
[3] *Information Systems Study Program, STMIK Bandung Bali, Indonesia*

*Corresponding Author: *taufiqhidayat50737@gmail.com*

Abstract: Diabetic Retinopathy (DR) is a leading cause of preventable blindness, yet the development of automated diagnostic models using Deep Learning is often hindered by the availability of imbalanced medical datasets. This study aims to address this issue by implementing a Conditional Lightweight Generative Adversarial Network (c-LGAN) architecture to synthesize realistic fundus retinal images corresponding to five DR severity levels from the IDRiD dataset. The c-LGAN model was trained on a balanced dataset, and its performance was quantitatively evaluated using Frechet Inception Distance (FID) and Inception Score (IS) metrics. The results demonstrate that the proposed model is capable of generating high-quality images, evidenced by achieving a best FID score of 121.24 at epoch 100. However, further observation identified significant stability challenges in long-term training, marked by a performance collapse after the model reached its optimal point. This phenomenon was attributed to an overpowering discriminator. This study concludes that c-LGAN is a promising approach for data augmentation but emphasizes the critical importance of periodic metric monitoring and model checkpointing strategies to capture peak performance and overcome training stability issues.

Keywords: Conditional GAN, deep learning, diabetic retinopathy, image synthesis.

History Article: Submitted 16 December 2025 | Revised 19 February 2026 | Accepted 3 March 2026

How to Cite: A. T. Hidayat, I. M. D. Maysanjaya, I. M. G. Sunarya, and M. W. A. Kesiman, "Application of conditional lightweight GAN for retinal fundus image synthesis based on diabetic retinopathy severity levels on the IDRID dataset," *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 16, no. 1, pp. 1–11, 2026, doi: 10.31940/matrix.v16i1.1-11.

Introduction

Diabetic Retinopathy (DR) is a microvascular complication of diabetes mellitus and represents a leading cause of visual impairment and blindness among the working-age population worldwide [1], [2]. This disease arises from damage to the fine blood vessels in the retina triggered by chronic hyperglycemic conditions. If not treated at an early stage, such damage can progress to vascular occlusion, fluid leakage, and the growth of abnormal new blood vessels (neovascularization), which may ultimately result in permanent blindness [3], [4]. With the global number of individuals with diabetes projected to continue increasing [5], the development of accurate and efficient early detection methods has become imperative to prevent vision loss [6].

Retinal fundus imaging is considered the gold standard for DR diagnosis due to its ability to provide non-invasive visualization of the internal structures of the eye [7]. To guide diagnosis, the International Clinical Diabetic Retinopathy (ICDR) classification system categorizes DR into five levels of severity: Normal (no abnormalities), Mild Non-Proliferative Diabetic Retinopathy (NPDR), Moderate NPDR, Severe NPDR, and Proliferative Diabetic Retinopathy (PDR), which represents the most critical stage [8].

Along with technological advances, Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), has demonstrated great potential as a fast and accurate method for automated diagnosis [9], [10]. However, the performance of DL models is highly dependent on the

availability of large and diverse training datasets. In many cases, publicly available medical datasets such as the Indian Diabetic Retinopathy Image Dataset (IDRiD) are often insufficient or imbalanced, where the number of images representing severe disease classes is significantly smaller than those of normal conditions [11]. Standard data augmentation techniques, such as rotation and flipping, often fail to address this issue effectively as they merely alter the geometry of existing samples without introducing new pathological feature variations. Consequently, diagnostic models tend to become biased toward the majority class (Normal), resulting in poor sensitivity for critical minority classes (Severe and PDR), which poses a significant risk of missed diagnoses in clinical settings.

To address this challenge, data augmentation through image synthesis has become critically important. Recent advances in the field of Generative Adversarial Networks (GANs), first introduced by Goodfellow et al. [12], have made them a highly promising tool for synthesizing realistic data. GANs consist of two competing neural networks: a generator, which learns to create new data, and a discriminator, which learns to distinguish between real and synthetic data. Previous studies by Zhou et al. [13] introduced DR-GAN to synthesize fine-grained lesions, demonstrating that synthetic data can significantly improve classification accuracy. However, their approach relies on complex architectures that demand substantial computational resources, limiting their applicability in resource-constrained environments often found in developing healthcare systems.

This paper proposes the application of a Conditional Lightweight GAN (c-LGAN) to synthesize retinal fundus images from the IDRiD dataset according to the five specific severity levels. Unlike heavy conventional GANs, the proposed "lightweight" design aims to achieve computational efficiency, enabling model training on hardware with limited specifications [14]. While the synthesis resolution is set to 128 × 128 pixels—a trade-off that sacrifices some fine-grained high-frequency details—this resolution is sufficient to capture global structural pathology and color distribution required to balance class distribution and mitigate model bias. The quality of the synthesized images will be quantitatively evaluated using the Frechet Inception Distance (FID) and Inception Score (IS) metrics. It is expected that the generated images can be utilized for data augmentation to train more robust AI-based classification models and can also serve as supportive diagnostic teaching tools for medical professionals in the future.

Methodology

Experimental Design

The experimental workflow is illustrated in Figure 1. The experiment consists of five main stages: Dataset Collection, Pre-processing, GAN Architecture Selection, Training Process, and Performance Evaluation. This systematic approach ensures that the impact of data augmentation on class balance and image quality can be rigorously assessed.

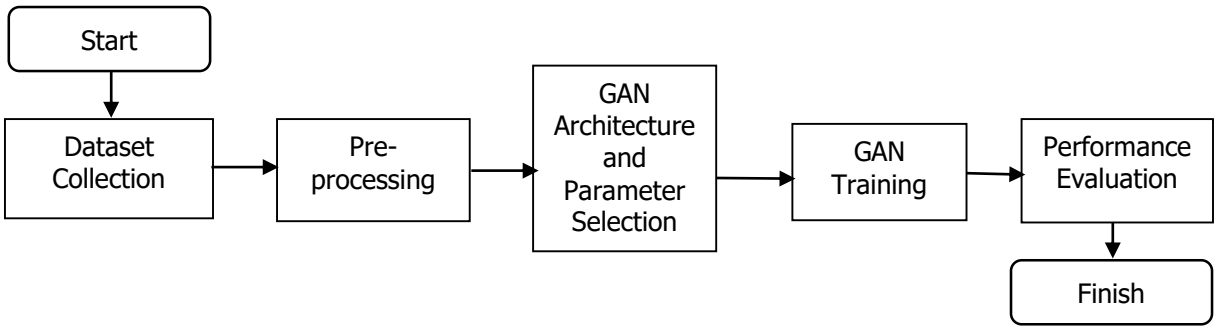


Figure 1. Experimental flow diagram

Dataset Collection

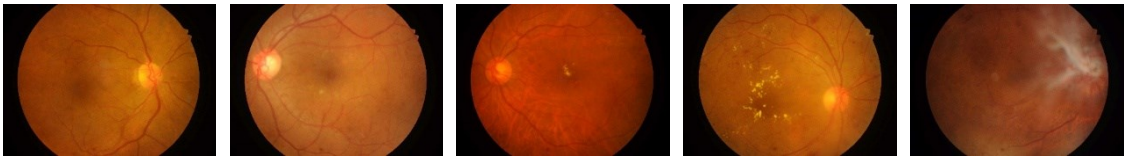
The dataset used in this study is the Indian Diabetic Retinopathy Image Dataset (IDRiD), developed by Porwal et al. [15]. This dataset was specifically designed to support research on the detection and classification of Diabetic Retinopathy (DR) and Diabetic Macular Edema (DME).

The IDRiD dataset is publicly available via the Kaggle platform and provides high-resolution retinal fundus images.

The training subset contains 413 images, each annotated with a DR severity level according to the International Clinical Diabetic Retinopathy (ICDR) scale. As shown in Table 1, the original distribution of the dataset is highly imbalanced, with a significant scarcity of samples in the severe disease classes (Severe NPDR and PDR). This imbalance poses a major challenge for deep learning models, which tend to bias towards the majority class.

Table 1. Distribution of IDRiD dataset before and after balancing

DR Severity Level	Original Count (Images)	Percentage (%)	Target Count After Synthesis
Normal	134	32.4%	500
Mild NPDR	20	4.8%	500
Moderate NPDR	136	32.9%	500
Severe NPDR	74	17.9%	500
Proliferative DR (PDR)	49	11.9%	500
Total	413	100%	2,500



a) Normal *b) Mild* *c) Moderate* *d) Severe* *e) PDR*

Figure 2. Sample IDRiD dataset images based on DR severity levels

Pre-processing

The pre-processing stage was critical to prepare raw data for effective GAN training. The primary objectives were to standardize dimensions, enhance variability, and correct the class imbalance shown in Table 1

Standardization

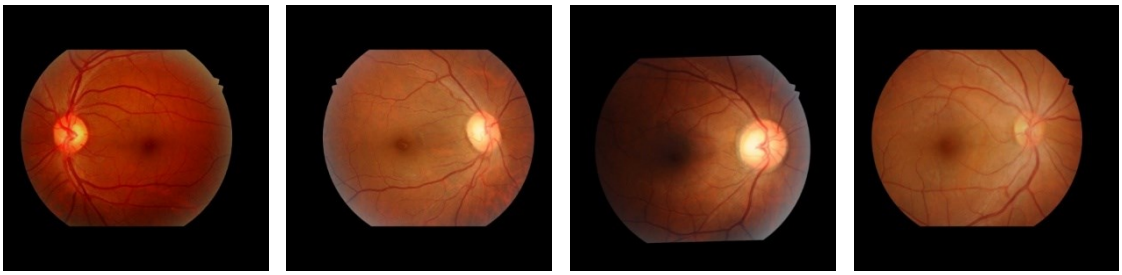
Raw images were first padded to square aspect ratios to prevent distortion and then resized. While the original images are high-resolution, they were resized to 128 × 128 pixels for this study. This resolution was selected as a strategic trade-off to drastically reduce computational load and memory usage, enabling the training of the generative model on standard hardware while still preserving global retinal structures (optic disc, major vessels) required for severity classification.

Augmentation

To enrich the diversity of the training set before synthesis, standard augmentation techniques were applied using the Albumentations library. These included horizontal flipping, brightness/contrast adjustments, color jittering, rotation (+- 10^{o}), and Gaussian blur.

Balancing Strategy

As detailed in Table 1, the final dataset was balanced to contain exactly 500 images per class. This was achieved by combining real augmented images with synthetic images generated by the c-LGAN model, ensuring that the model learns from a uniform distribution of disease severity.



a) Color jitter *b) HFlip* *c) Shift scale rotate* *d) Brightness*

Figure 3. Examples of IDRiD dataset images after pre-processing

GAN Architecture and Parameter Selection

This study employs a Conditional Lightweight GAN (c-LGAN) [16] architecture, adapted from the Lightweight GAN (LGAN) approach proposed by Abd Aziz et al. [14] for retinal fundus image synthesis.

Justification for Lightweight Architecture

Unlike conventional GANs (e.g., DCGAN or StyleGAN2) which rely on deep networks with millions of parameters and require high-end GPU clusters, the c-LGAN is designed for computational efficiency. It utilizes a reduced number of feature maps in the convolutional layers and simpler residual blocks. This design choice makes the model deployable in resource-constrained environments (such as basic medical research facilities in developing countries) and significantly accelerates the training convergence without a massive sacrifice in the semantic quality of the generated medical images.

Generator Structure

The generator is designed to be lightweight yet effective. It takes two inputs, a noise vector z (latent space dimension = 128) and a class embedding label y.

Input Layer

The latent vector and class embedding are concatenated and passed through a dense layer.

Upsampling Blocks

The network consists of a series of transposed convolutional layers (ConvTranspose2d). Instead of deep residual stacks, each block performs efficient upsampling followed by Batch Normalization and ReLU activation.

Output Layer

The final layer uses a Tanh activation function to produce a 3-channel RGB image with values in the range [-1, 1] [17].

Discriminator Structure

The discriminator is a patch-based classifier that also conditions on the class label. It uses Spectral Normalization in its convolutional layers to stabilize training and prevent mode collapse—a common issue in medical image synthesis. The class label is projected and concatenated with the image feature map to ensure the discriminator evaluates both the realism of the image and its correspondence to the target disease severity [18].

Table 2. GAN model training parameters

Parameter	Value	Description
Image size	128 × 128 pixels	Input and output resolution to reduce computation
Latent space dimension	128	Noise vector dimension (generator input)
Optimizer	*Adam*	Used for Generator and Discriminator
Learning rate	0.0002	Learning rate for both networks
β1 (*Adam*)	0.5	First momentum parameter
β2 (*Adam*)	0.999	Second momentum parameter
Maximum epochs	1.000	Total training iterations
Batch size	32	Samples per training iteration
Early Stopping	Yes	Applied to prevent overfitting
Evaluation & checkpointing	Periodic	To monitor and save best model performance

Training parameters were carefully selected to ensure training stability and efficiency. The image size was set to 128 × 128 pixels to reduce computational load, while the latent space dimension was fixed at 128. The Adam optimization algorithm was employed with a learning rate of 0.0002 for both the generator and discriminator, and β1 and β2 values of 0.5 and 0.999, respectively. The model was trained for a maximum of 1,000 epochs with a batch size of 32. In addition, an early stopping mechanism was implemented to terminate training when no significant improvement was observed, thereby preventing overfitting. Periodic evaluation and model checkpointing were also conducted to monitor network performance and stability throughout the training process.

GAN Training

The training process utilized the Adam optimizer (α=0.0002, β1=0.5, β2=0.999) for both networks. The loss function employed was the Binary Cross-Entropy with Logits Loss.

Conditional Training

The model was trained to generate images specific to the five DR grades. This conditional mechanism ensures that the generator learns distinct pathological features (e.g., microaneurysms for Mild DR vs. neovascularization for PDR).

Stability Monitoring

Training was run for a maximum of 1,000 epochs. To address potential instability (e.g., overpowering discriminator), Early Stopping was implemented. If the FID score did not improve for 50 consecutive epochs, training was halted to prevent overfitting and degradation of image quality.

Performance Evaluation

To quantitatively assess the quality and diversity of the synthesized images, two standard metrics were used:

Frechet Inception Distance (FID) [19]

This metric measures the distance between the feature distributions of real and synthetic images. A lower FID indicates that the synthetic images are more similar to real biological data and possess better diversity.

Inception Score (IS) [20]

This metric evaluates the clarity and distinctness of the generated images. A higher IS indicates better image quality.

These metrics were calculated every 10 epochs. The "Best Model" checkpoint was saved not based on the final epoch, but based on the lowest FID score achieved during training, ensuring that the final evaluation uses the most realistic generator.

Results and Discussions

This section presents an in-depth analysis of the experimental results obtained from training the Conditional Lightweight GAN (c-LGAN) model. The discussion focuses on three main aspects: analysis of training progression and stability based on loss data, quantitative evaluation of synthetic image quality using the Frechet Inception Distance (FID) and Inception Score (IS) metrics, and qualitative analysis of the generated images.

Training Progression and Model Stability

The c-LGAN training process was completed over 1,000 epochs with a total computational time of approximately 39.5 hours (142,228 seconds). Analysis of the loss curves (see Figure 4) reveals challenging training dynamics and indicates stability issues.

Loss Dynamics

From the early stages of training, the Discriminator Loss (D_loss) decreased drastically and consistently remained at a very low level, often approaching zero (e.g., 0.0014 at epoch 2). In contrast, the Generator Loss (G_loss) remained high and exhibited significant volatility, fluctuating between 2.0 and peaks above 20.0 (e.g., 20.24 at epoch 921). The condition in which D_loss is extremely low while G_loss is high and unstable is a classic indication of an overpowering discriminator. In this scenario, the discriminator becomes too strong too quickly, providing vanishing gradients that make it difficult for the generator to learn effectively and converge.

Model Stability

This stability issue is further confirmed by the FID evaluation data. Although the model showed improvement up to epoch 100, a drastic performance collapse occurred thereafter. This phenomenon indicates that the model failed to maintain its best performance and entered an unstable training regime known as mode collapse or divergence.

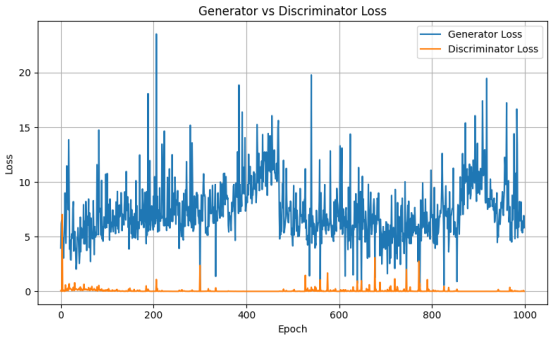


Figure 4. Loss progression curves for the Generator (G) and Discriminator (D) over 1,000 epochs

Synthetic Image Quality

The quality of the synthetic fundus images was quantitatively evaluated every 10 epochs. Key evaluation points highlighting performance trends are summarized in Table 3.

Table 3. Quantitative evaluation results at key epochs

Epoch	FID Score (↓ better)	IS Score (↑ better)	Remarks
10	349.45	1.28 ± 0.04	Early training
40	173.59	1.91 ± 0.18	Significant improvement
70	153.08	1.49 ± 0.19	Approaching peak
90	136.34	1.59 ± 0.18	Near peak
100	121.24	1.56 ± 0.16	Best performance
110	445.85	1.49 ± 0.13	Performance collapse
200	288.68	1.40 ± 0.06	Failed to recover
500	170.16	1.86 ± 0.15	Improved but not recovered
740	156.55	1.40 ± 0.08	Secondary peak (best after collapse)
1000	171.42	1.75 ± 0.27	Final model

The model achieved peak performance at epoch 100, with a minimum FID score of 121.24.

Significance of FID

A lower FID score indicates that the distance between the feature distribution of real and synthetic images is small. This implies that the synthetic images at epoch 100 successfully capture the statistical diversity and global structural features of the real retinal dataset, satisfying the requirement for data augmentation diversity.

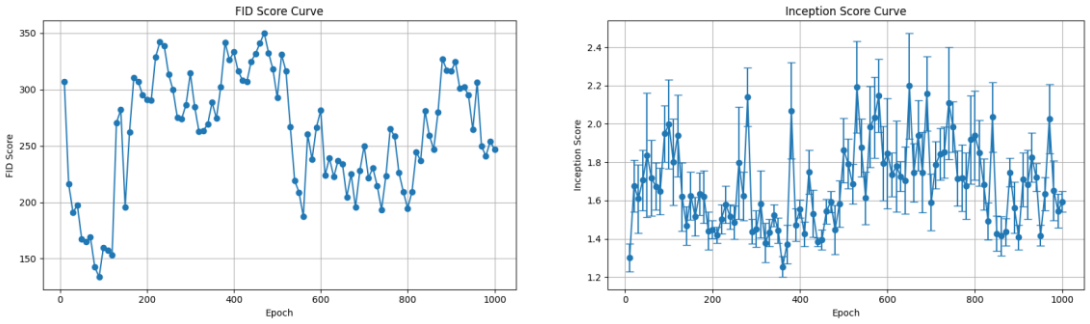
Performance Collapse

However, after this point, a sharp increase in the FID score to 445.85 at epoch 110 signaled a severe quality collapse. Despite partial recovery attempts in later epochs, the model never regained the quality achieved at epoch 100.

Meanwhile, the Inception Score (IS) remained relatively low overall, with the highest value reaching only 1.92 ± 0.20 (at epoch 80).

Interpretation

It is important to note that the IS metric relies on the InceptionV3 network pre-trained on ImageNet (which contains distinct objects like dogs, cars, etc.). Retinal fundus images are visually homogeneous—they all look like orange discs with blood vessels—lacking the distinct object variability found in ImageNet. Consequently, the Inception network does not classify them into "distinct" classes with high confidence, naturally resulting in lower IS values compared to natural image datasets. Therefore, in this specific medical context, the FID score is a more reliable indicator of quality than IS, as FID compares the synthetic distribution directly against the real retinal data distribution.



a) FID curve b) IS curve

Figure 5. Evolution of FID and Inception Score (IS) values during training

Qualitative Analysis

Visual inspection of image samples saved at various training intervals is consistent with the quantitative findings:

Best Images (Epoch 100)

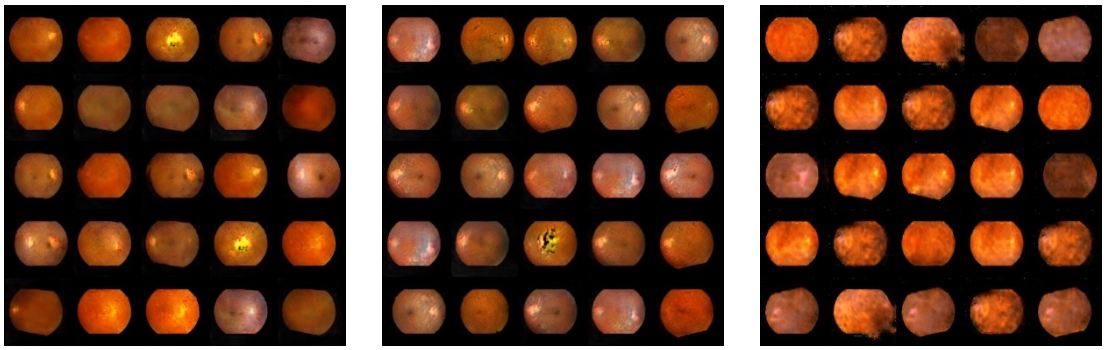
Images generated around epoch 100 exhibit the highest visual quality. Major anatomical structures, such as the optic disc and blood vessels, appear coherent, with natural fundus coloration. The model successfully replicates visual features in accordance with the provided class labels.

Post-Collapse Images (Epoch > 100)

Images generated after epoch 100 show a significant degradation in quality. Numerous visual artifacts, unnatural patterns, and disorganized anatomical structures emerge, corresponding to the deteriorating FID scores.

Final Images (Epoch 1000)

Although better than the complete collapse condition, the visual quality of images at the final epoch does not match that of the images generated at epoch 100.



(a) Samples from epoch 100 (best FID) (b) Samples from epoch 110 (post-collapse) (c) Samples from epoch 1000 (final)

Figure 6. Comparison of synthetic image quality

Discussion

The findings of this study provide two key insights. First, the c-LGAN architecture is capable of generating reasonably realistic retinal fundus images, as evidenced by the minimum FID score of 121.24. This result confirms the potential of GANs as data augmentation tools for imbalanced medical datasets.

Second, and more critically, this study highlights the stability challenges inherent in GAN training. The overpowering discriminator phenomenon (D_loss ≈ 0) emerged as the primary

obstacle, causing the generator to fail to learn stably over extended training periods. The performance collapse after epoch 100 demonstrates that longer training does not necessarily yield better models. This finding underscores the importance of periodic evaluation and model checkpointing during training. Without such strategies, the best-performing model (at epoch 100) would be missed, and the final saved model (at epoch 1,000) would exhibit inferior quality.

This study acknowledges several limitations that need to be addressed in future work:

Resolution

The generated images are limited to 128 x 128 pixels. While sufficient for global structure learning, this resolution restricts the representation of fine retinal micro-lesions (e.g., tiny microaneurysms) crucial for early-stage diagnosis.

Clinical Validation

Currently, the evaluation relies solely on computational metrics (FID/IS). A rigorous clinical validation involving ophthalmologists (e.g., a Visual Turing Test) has not yet been conducted. Expert feedback is essential to verify if the synthetic lesions are medically accurate and not just visually plausible artifacts.

Based on these findings, future research directions may focus on:

Improving Training Stability

Applying advanced regularization techniques to the discriminator, such as Spectral Normalization or Wasserstein GAN with Gradient Penalty (WGAN-GP), to prevent discriminator dominance.

Learning Rate Optimization

Adopting the Two Time-scale Update Rule (TTUR) by employing different learning rates for the generator and discriminator to maintain balanced training dynamics.

High-Resolution Synthesis

Once stability is achieved, scaling up the model to generate images at 512 x 512 pixels to capture fine-grained diabetic retinopathy lesions.

Conclusion

This study successfully implemented and evaluated a Conditional Lightweight GAN (c-LGAN) architecture to synthesize retinal fundus images based on five levels of Diabetic Retinopathy (DR) severity from the IDRiD dataset. The proposed model demonstrated its capability to generate realistic images, as evidenced by the best Frechet Inception Distance (FID) score of 121.24 achieved at epoch 100. These results confirm that generative approaches, particularly c-LGAN, hold substantial potential as data augmentation methods to address data scarcity and class imbalance issues in medical imaging datasets.

Nevertheless, this study also identified significant challenges related to long-term training stability. The occurrence of the overpowering discriminator phenomenon (D_loss ≈ 0) led to a performance collapse after the model reached its optimal point. This finding demonstrates that longer training does not necessarily guarantee better results and underscores the importance of periodic metric monitoring and best-model checkpointing to capture peak performance before instability arises.

Based on the conclusions and limitations identified in this study, several directions for future work can be proposed. Training stability can be improved by mitigating discriminator dominance through the application of more advanced regularization techniques. In particular, incorporating spectral normalization into the discriminator layers or adopting alternative frameworks such as Wasserstein GAN with Gradient Penalty (WGAN-GP) has strong potential to stabilize the adversarial training process. In addition, learning rate optimization represents an important avenue for improvement. Applying different learning rates for the generator and discriminator, a strategy known as the Two Time-scale Update Rule (TTUR), may help maintain a more balanced competitive dynamic between the two networks and prevent one from becoming excessively dominant at an early stage of training. Furthermore, once training stability has been

adequately addressed, future studies should consider increasing the image resolution used during training, for example to 256 × 256 or 512 × 512 pixels. This step is essential for generating synthetic retinal fundus images with richer clinical detail, particularly for the visualization of micro-lesions characteristic of diabetic retinopathy. Finally, clinical validation should be incorporated through qualitative evaluations involving medical experts, such as ophthalmologists, using approaches like the Visual Turing Test. Feedback from domain experts would provide invaluable insight into the clinical realism and potential diagnostic utility of the synthetic images generated by the best-performing model.

This study successfully implemented and evaluated a Conditional Lightweight GAN (c-LGAN) architecture to synthesize retinal fundus images based on five levels of Diabetic Retinopathy (DR) severity from the IDRiD dataset. The proposed model demonstrated its capability to generate realistic images while maintaining computational efficiency, as evidenced by the best Frechet Inception Distance (FID) score of 121.24 achieved at epoch 100. These results confirm that generative approaches, particularly c-LGAN, hold substantial potential as data augmentation methods to address data scarcity and class imbalance issues in medical imaging datasets without requiring high-performance computing infrastructure.

Nevertheless, this study also identified significant challenges related to long-term training stability. The occurrence of the overpowering discriminator phenomenon (D_loss ≈ 0) led to a performance collapse after the model reached its optimal point. This finding demonstrates that longer training does not necessarily guarantee better results and underscores the importance of periodic metric monitoring and best-model checkpointing to capture peak performance before instability arises.

Based on the conclusions and limitations identified in this study, several directions for future work are proposed to enhance both technical robustness and clinical applicability. Future iterations should mitigate discriminator dominance through advanced regularization techniques. Specifically, incorporating Spectral Normalization or adopting the Wasserstein GAN with Gradient Penalty (WGAN-GP) framework has strong potential to stabilize the adversarial training process. Additionally, adopting the Two Time-scale Update Rule (TTUR)—applying different learning rates for the generator and discriminator—may help maintain a balanced competitive dynamic.

To address the limitation of fine-grained detail in the current 128 x 128 output, future studies should aim to increase the synthesis resolution to 256 x 256 or 512 x 512 pixels. This step is essential for capturing micro-lesions (e.g., microaneurysms) characteristic of early-stage diabetic retinopathy, ensuring the data is not just visually similar but diagnostically valuable.

A critical limitation of the current study is the reliance on computational metrics (FID/IS). Future work must incorporate clinical validation through qualitative evaluations involving ophthalmologists, such as a Visual Turing Test. Feedback from domain experts is indispensable to verify the clinical realism and potential diagnostic utility of the synthetic images, ensuring they do not contain misleading artifacts that could confuse diagnostic models.

References

[1] T. E. Tan and T. Y. Wong, "Diabetic retinopathy: Looking forward to 2030," *Front Endocrinol (Lausanne)*, vol. 13, pp. 1–8, Jan. 2023, doi: 10.3389/fendo.2022.1077669.

[2] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic retinopathy fundus image classification and lesions localization system using deep learning," *Sensors*, vol. 21, no. 11, pp. 1–22, Jun. 2021, doi: 10.3390/s21113704.

[3] A. D. Bhatwadekar, A. Shughoury, A. Belamkar, and T. A. Ciulla, "Genetics of diabetic retinopathy, a leading cause of irreversible blindness in the industrialized world," *Genes (Basel)*, vol. 12, no. 8, pp. 1–17, Aug. 2021, doi: 10.3390/genes12081200.

[4] D. S. Fong, L. P. Aiello, F. L. Ferris, and R. Klein, "Diabetic Retinopathy," *Diabetes Care*, vol. 27, no. 10, pp. 2540–2553, Oct. 2004, doi: 10.2337/diacare.27.10.2540.

[5] L. Wu, "Classification of diabetic retinopathy and diabetic macular edema," *World J Diabetes*, vol. 4, no. 6, pp. 290–294, Dec. 2013, doi: 10.4239/wjd.v4.i6.290.

[6] P. Vashist, S. Singh, N. Gupta, and R. Saxena, "Role of early screening for diabetic retinopathy in patients with diabetes mellitus: An overview," *Indian Journal of Community Medicine*, vol. 36, no. 4, pp. 247–252, Oct. 2011, doi: 10.4103/0970-0218.91324.

[7] I. Welch Allyn and Ynjiun Paul Wang, "FUNDUS IMAGING SYSTEM," Aug. 13, 2019

[8] Z. Yang, T. E. Tan, Y. Shao, T. Y. Wong, and X. Li, "Classification of diabetic retinopathy: Past, present and future," *Front Endocrinol (Lausanne)*, vol. 13, pp. 1–18, Dec. 2022, doi: 10.3389/fendo.2022.1079217.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[10] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, pp. 1–26, May 2023, doi: 10.3390/computers12050091.

[11] P. Porwal *et al.*, "IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge," *Med Image Anal*, vol. 59, no. 1, pp. 1–83, Sep. 2019, doi: 10.1016/j.media.2019.101561.

[12] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," Jun. 10, 2014, *arXiv*. [Online]. Available: http://arxiv.org/abs/1406.2661

[13] Y. Zhou, B. Wang, X. He, S. Cui, and L. Shao, "DR-GAN: Conditional Generative Adversarial Network for Fine-Grained Lesion Synthesis on Diabetic Retinopathy Images," *IEEE J Biomed Health Inform*, vol. 24, no. 1, pp. 56–66, Dec. 2022, doi: 10.1109/JBHI.2020.3045475.

[14] N. Abd Aziz, M. Azman Hanif Sulaiman, A. Zabidi, I. Mohd Yassin, M. Syahirul Amin Megat Ali, and Z. Ismael Rizman, "Lightweight Generative Adversarial Network Fundus Image Synthesis," *INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION*, vol. 6, no. 1, pp. 270–277, Mar. 2022, [Online]. Available: www.joiv.org/index.php/joiv

[15] P. Porwal *et al.*, "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research," *Data (Basel)*, vol. 25, no. 3, pp. 1–8, Jun. 2018, doi: 10.21227/H25W98.

[16] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," Nov. 2014, [Online]. Available: http://arxiv.org/abs/1411.1784

[17] P. Zhang *et al.*, "Fundus Image Generation and Classification of Diabetic Retinopathy Based on Convolutional Neural Network," *Electronics (Switzerland)*, vol. 13, no. 18, Sep. 2024, doi: 10.3390/electronics13183603.

[18] T. Miyato and M. Koyama, "cGANs with Projection Discriminator," Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.05637

[19] D. A. Chan and S. P. Sithungu, "Evaluating the Suitability of Inception Score and Fréchet Inception Distance as Metrics for Quality and Diversity in Image Generation," in *CIIS 2024 - 2024 the 7th International Conference on Computational Intelligence and Intelligent Systems*, Association for Computing Machinery, Inc, Feb. 2025, pp. 79–85. doi: 10.1145/3708778.3708790.

[20] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," Sep. 2018, [Online]. Available: http://arxiv.org/abs/1809.11096

Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 16 Issue 1 Year 2026 Pages 12-28
DOI: http://dx.doi.org/10.31940/matrix.v16i1.12-28
URL: http://ojs2.pnb.ac.id/index.php/matrix

Integration of local wisdom and modern medicine in a treatment recommendation system for toddlers based on the Case-Based Reasoning-Fuzzy Method

Yusuf Hendra Pratama [1]*, Hendri Purnomo [2], Recta Olivia Umboro [3]

[1,2] *Computer Science Study Program, Universitas Islam Al-Azhar, Indonesia*
[3] *Pharmacy Study Program, Universitas Qamarul Huda Badaruddin Bagu, Indonesia*

*Corresponding Author: *yshendra.tm@gmail.com*

Abstract: This study aims to develop a treatment recommendation system for toddlers based on Case-Based Reasoning (CBR) combined with Fuzzy Logic, by integrating modern medical knowledge and local wisdom. The system was developed to address the need for adaptive initial diagnosis recommendations, particularly in addressing ambiguous symptoms. At the case representation stage, disease, symptom, and drug data from medical and traditional perspectives are used as the knowledge base. The CBR process serves as the primary mechanism for searching for similar cases, while fuzzy logic is used at the revision stage to provide degrees of symptom intensity so that the diagnosis results are more flexible. System evaluation was conducted through blackbox testing, accuracy measurements, and the System Usability Scale (SUS) method involving 50 respondents. The results showed that all system functions ran as planned, the accuracy level reached 88%, and the average SUS score was 78.4 in the Good Usability category, indicating the system is easy to use and user-acceptable. This study proves that the CBR–Fuzzy integration is effective in providing accurate, adaptive, and culturally relevant initial diagnosis recommendations. For further research, it is recommended to expand the case base, refine fuzzy rules, develop a broader interface, and implement the system on a mobile platform to improve accuracy, ease of access, and wider user acceptance.

Keywords: Case-Based Reasoning, Fuzzy Logic, Local Wisdom, Toddler Treatment, Recommendation System, Usability

History Article: Submitted 26 September 2025 | Revised 11 December 2025 | Accepted 19 December 2025

How to Cite: Y. H. Pratama, H. Purnomo, and R. O. Umboro, "Integration of local wisdom and modern medicine in a treatment recommendation system for toddlers based on the Case-Based Reasoning-Fuzzy Method," *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 16, no. 1, pp. 12–28, 2026, doi: 10.31940/matrix.v16i1.12-28.

Introduction

Toddlers are an age group that is highly susceptible to various diseases due to their immune systems still developing [1]. This situation is further complicated in Indonesia, which is geographically located in a tropical region with a diversity of disease-causing microorganisms [2]. This condition causes toddlers to often suffer from various diseases, including pneumonia, acute respiratory infections (ARI), diarrhea, measles, fever, skin diseases, and influenza [3], [4], [5], [6], [7], [8], [9]. According to UNICEF, diseases such as diarrhea, pneumonia, and malaria are the main causes of infant mortality worldwide, contributing around 29% of the total death rate [10]. In Indonesia, pneumonia even causes the highest medical costs, reaching 8.7 trillion rupiah per year [9].

Apart from the high prevalence of disease, another major challenge is the limited access of people to modern health services, especially in rural areas and 3T (outermost, frontier, and disadvantaged) regions [11]. This condition encourages some people to utilize traditional medicine based on local wisdom [12]. Traditional healing practices have important cultural and historical value, but many of the methods used have not been scientifically verified, thus posing clinical risks [13], [14]. As a result, there is a gap between traditional medicine and modern medicine, both in terms of effectiveness and safety [13]. This shows that there is an urgent need

to present innovations that are able to systematically integrate the advantages of traditional medicine with modern medicine.

As technology advances, artificial intelligence (AI) is increasingly being used in healthcare to support diagnosis and treatment recommendations. The Case-Based Reasoning (CBR) method is considered relevant because it can provide solutions based on similar cases that have occurred previously [15], [16]. Meanwhile, Fuzzy Logic is effective in handling uncertain or ambiguous data, such as subjective symptoms [17], [18]. The combination of these two methods is considered capable of providing a system that is accurate but remains flexible.

Various previous studies have discussed expert systems and AI-based applications for toddler health. For example, Ananda et al. developed a toddler disease diagnosis system using fuzzy logic and forward chaining to facilitate self-diagnosis [18]. Alam et al. developed a system for diagnosing malnutrition in toddlers using the CBR method [15], Meanwhile, Arundy and Mardiani implemented CBR to detect heart disease [16]. Numan et al. also developed a system for diagnosing pneumonia in children [17], and Umami and Wibisono designed an application for early detection of toddler diseases [2].

Christian and Winarsih created an application for processing toddler data [19], Meanwhile, Ulum Fajar et al. developed a screening application for hearing disorders [20]. Some studies apply other algorithms, such as Kharisa et al. [21] who used the weighted Sörrensen algorithm and Nasri et al. [22] who developed the diagnosis of infant illness using the Dempster-Shafer method. On the other hand, Marques et al. discussed the use of traditional medicine [20], Anggresani et al. analyzed the use of traditional and modern medicines in self-medication [21], Ningrum & Purnamasari research the benefits of medicinal plants for maternal and child health [23] , and Azizah & Kurniati developed research on traditional medicine to relieve coughs and colds in toddlers [24] .

Previous research on intelligent systems for diagnosing childhood illnesses has generally used Case-Based Reasoning (CBR) or Fuzzy Logic methods separately, with only a few combining the two. Even then, these methods are limited to modern medical approaches without considering traditional healing practices widely used by the community. This research offers a more comprehensive approach by integrating CBR and Fuzzy Logic with modern medical knowledge and verified local wisdom, enabling the system to provide initial diagnoses and treatment recommendations that are more accurate, adaptive, and culturally relevant. This system is supported by a structured case base containing diseases, symptoms, and medical and traditional solutions, as well as a fuzzy-based revision mechanism to handle symptom ambiguity. Thus, this research contributes to the development of more contextual and applicable health recommendation systems. The main objective of this research is to develop a treatment recommendation system for toddlers that integrates modern medical data and local wisdom within a single artificial intelligence framework to produce more precise and tailored diagnoses and treatment recommendations tailored to the community's conditions. This study used data consisting of four main components: disease data, symptom data, treatment data (modern and traditional), and a collection of cases (case base) which is the core of the Case-Based Reasoning (CBR) process. Overall, there are 10 types of toddler diseases, 20 main symptoms, 25 medical solutions, and 18 treatment solutions based on local wisdom that are used as a knowledge base. From the total data, 95 structured cases were created consisting of 50 cases based on medical data, 30 cases based on local medical practices, and 15 cases resulting from integration through verification by local pediatricians and herbalists. In the evaluation stage, 50 test case scenarios were used to measure the accuracy of the system and its conformity with expert diagnoses.

Methodology

This research is an applied research approach using computational intelligence. The primary method used is Case-Based Reasoning (CBR) to generate solutions based on previous cases, complemented by fuzzy logic as a revision mechanism to address the uncertainty of toddler symptoms. The stages of this research are shown in Figure 1.

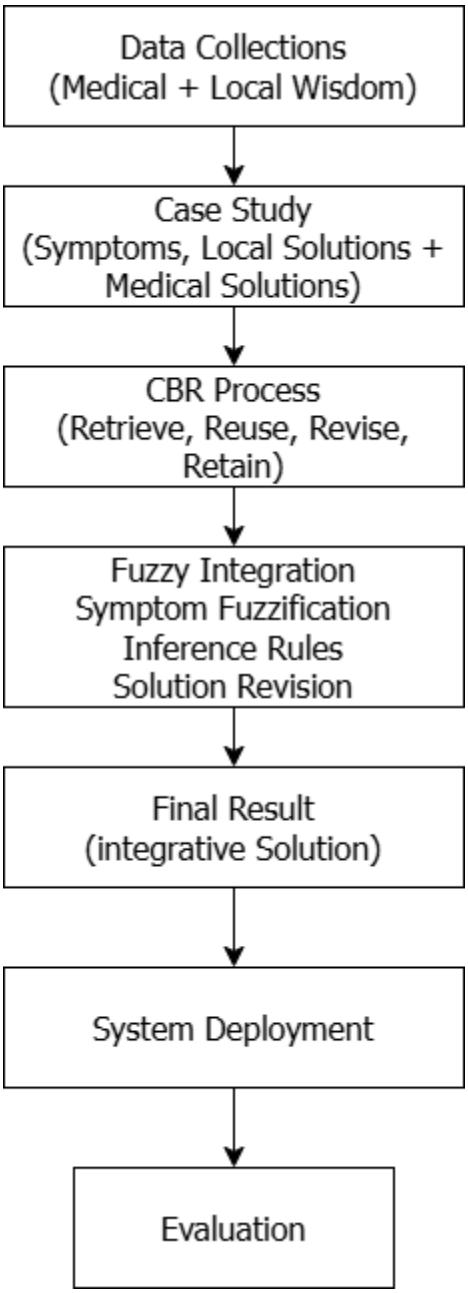


Figure 1. Research stages

Figure 1 shows the research stages carried out in this study, starting from data collection, case representation, CBR process, integration with fuzzy, obtaining integration results, system development to evaluation.

1. Data Collection
   At this stage, medical data related to this research is collected, such as disease data, symptom data, medical treatment data and also local wisdom treatment data.
2. Case Study
   After modern medical data and local wisdom have been successfully collected, the next step is to build a uniform case representation for use in the Case-Based Reasoning system. Each case consists of several important attributes that represent the toddler's condition, such as body temperature, cough, diarrhea, appetite, and other relevant medical history. Each of these attributes is arranged in the form of a case vector, where each vector contains a combination of attribute values and the treatment solution provided, either in the form of modern medical therapy or treatment based on local wisdom. The equation used in the case representation can be seen in Equation (1) [25].

$$Ci = \{a1, a2, \dots, an, S\} \quad (1)$$

a_j = j-th attribute value (e.g. temperature, cough, diarrhea, appetite)
S = treatment solution (local + medical combination)

3. CBR Mechanism
The next stage is the application of the Case-Based Reasoning (CBR) method, which serves as the primary mechanism in the recommendation system. Generally, CBR works through four stages: Retrieve, Reuse, Revise, and Retain. In the Retrieve stage, the system calculates the similarity level between new cases and a collection of old cases in the database. The equation used to calculate the similarity value is the Nearest Neighbor Similarity method, which is written as [26]:

$$Sim(Q, Ci) = \frac{\sum_{j=1}^{n} w_j \times sim(q_j, c_{ij})}{\sum_{j=1}^{n} w_j} \quad (2)$$

Where Q is a new query or case, C_i is an old case, q_j and c_{ij} is the value of the j-th attribute of each case, and w_j is the attribute weight. For numeric attributes, the similarity calculation is performed using Equation (3).

$$Sim(q_j, C_{ij}) = 1 - \frac{|q_j - c_{ij}|}{Range_j} \quad (3)$$

For categorical attributes, similarity is assigned a value of 1 if the cases are the same and 0 if they are different. The Reuse stage then selects the initial solution based on the case with the highest similarity level. The Revise stage is enhanced with fuzzy logic to address uncertainty, while the Retain stage stores the new case and the revised solution in the knowledge base.

4. Fuzzy Integration
The revision stage in CBR is enriched with fuzzy logic integration, particularly to handle continuous or ambiguous attributes, such as toddler body temperature, cough, and others. Each incoming numeric value is mapped to a membership function (fuzzification). The equation used can be seen in Equation (4) [27].

$$\mu_{tri}(x; a, b, c) = \begin{cases} 0, & x \leq a \text{ or } x \geq c \\ \dfrac{x-a}{x-b}, & a < x < b \\ \dfrac{c-x}{c-b}, & b < x < c \end{cases} \quad (4)$$

Information:
x : Input Value
a = starting point (starting from 0)
b = peak point (value μ=1 → meaning 100% membership)
c = end point (value drops back to 0)
The fuzzification results are then processed with inference rules. Examples of rules used are: "If the fever is moderate and the cough is present, then the solution is herbal medicine + compress + monitor hydration," or "If the fever is high and the diarrhea is present, then the solution is medical medication + oral rehydration salts + medical referral." In this way, fuzzy logic can provide finer adjustments to the initial solution obtained from CBR.

5. Final Result
The system's output is a treatment recommendation that combines previous case experience and fuzzy logic-based adjustments. For example, if the system receives input of symptoms of a toddler with a body temperature of 38 °C, cough = yes, and diarrhea = no, then in the CBR stage the system will find the case most similar to the initial solution of herbal medicine + compress. Next, fuzzy logic detects that the temperature of 38 °C is in the "moderate" category with a membership value of 0.67, so the inference rule adds a hydration monitoring

factor to the practice suggestion. Thus, the final recommendation given by the system is herbal medicine + compress + hydration monitoring.

6. System Deployment
   At this stage, system development is carried out to implement a recommendation system developed using a combination of CBR and Fuzzy Logic methods. The developed system has two types of users, each with their own functional needs. These users and their functional needs are depicted in the use case diagram shown in Figure 2.

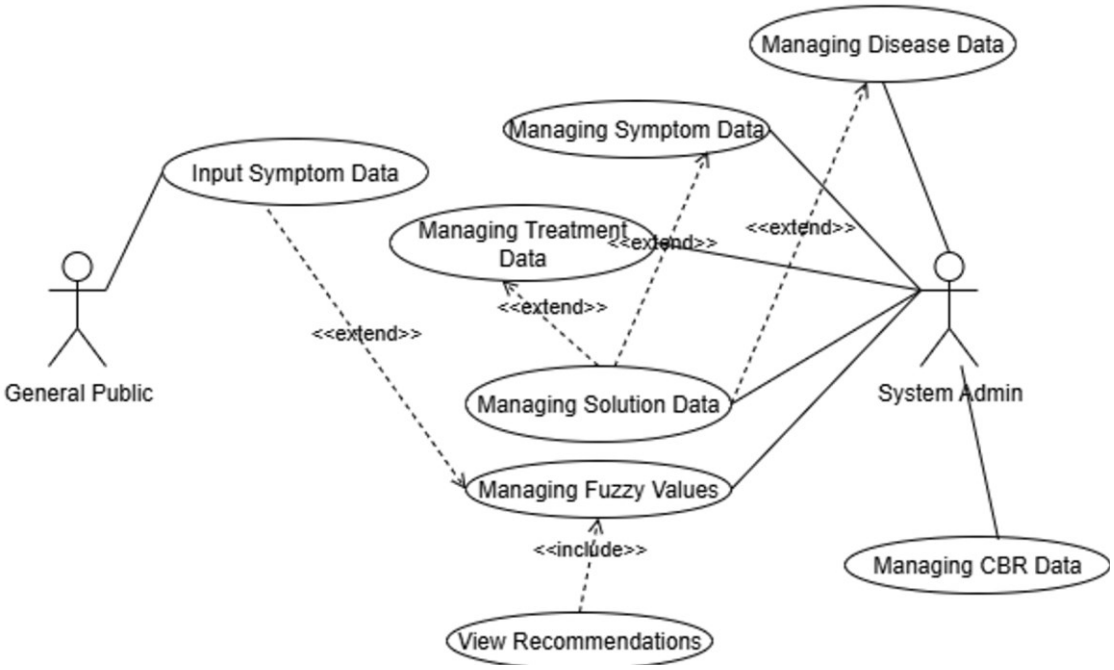


Figure 2. Use case diagram

Figure 2 shows the use case diagram used in the development of this system. The figure shows eight use cases. These use cases identify data requirements that can be used as a reference in developing the system's database. The resulting database design can be seen in Figure 3.

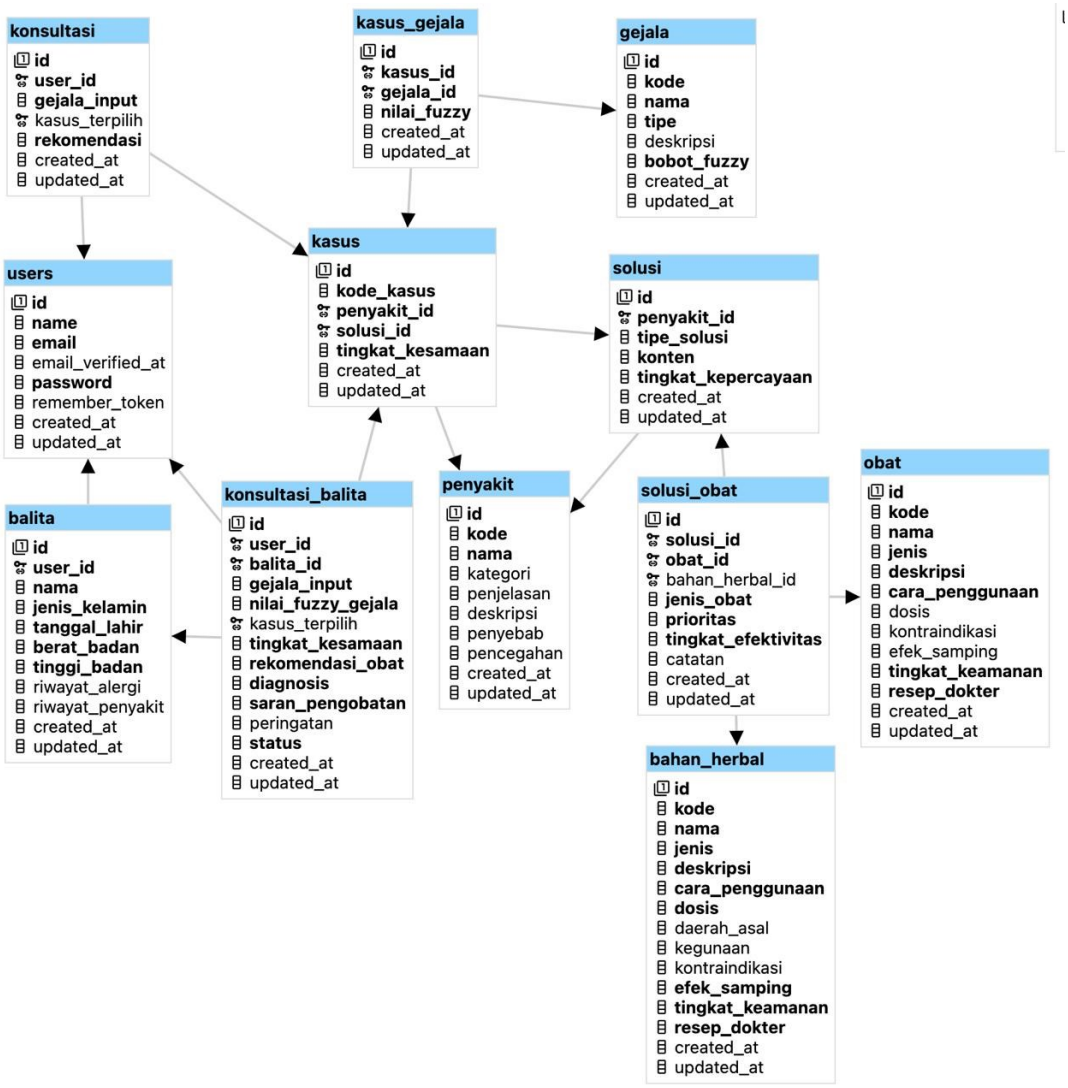


Figure 3. Design database

Figure 3 shows the database design used in the system being developed, in which the design contains twelve tables that will be used later.

7. System evaluation
   At this stage, the developed system is tested using blackbox testing, accuracy testing, and SUS (System Usability Scale) [28].

Results and Discussions

This research focuses on developing a treatment recommendation system for toddlers using the CBR approach and fuzzy logic. The system aims to integrate modern medical knowledge with local wisdom, resulting in recommendations that are not only medically accurate but also culturally relevant to the local community. This research involved seven stages, as explained in the methods section. The results of each stage are as follows:

Results

Data Collection

The master data used in this study included disease data, symptoms, treatments, and solutions. Examples of this data can be seen in Table 1.

Table 1. Disease data

No	Code	Disease	Information
1	P008	Asthma	A respiratory tract infection that causes difficulty breathing.
2	P004	Coughs and Colds	An upper respiratory tract infection that usually heals on its own.
3	P013	Chickenpox	A varicella-zoster virus infection that causes a skin rash and fever.
4	P012	Measles	A contagious disease characterized by fever, cough, runny nose, and a rash all over the body.
5	P001	Dengue Fever	A disease caused by the dengue virus through the bite of the Aedes aegypti mosquito.
6	P006	Diabetes Mellitus	A chronic metabolic disease that requires treatment and blood sugar control.
7	P011	Singapore Flu	A viral infection that affects children, accompanied by a rash and mouth ulcers.
8	P010	Gastritis	Inflammation of the stomach lining, often called an ulcer.
9	P007	Hypertension	High blood pressure that requires ongoing medical treatment.
10	P002	Malaria	A parasitic disease transmitted through the bite of the Anopheles mosquito.

Table 1 shows some examples of disease data used in this study. In addition to disease data, there is also symptom data. Examples can be seen in Table 2.

Table 2. Symptoms data

No	Code	Symptoms	Fuzzy Value	Fuzzy Status
1	G01	High fever (≥38°C)	0.90	Very Important
2	G02	Cough with phlegm	0.70	Important
3	G03	Shortness of breath	0.80	Very Important
4	G04	Diarrhea (>3x/day)	0.75	Important
5	G05	Repeated vomiting	0.70	Important
6	G06	Skin rash	0.65	Important
7	G07	Nasal congestion	0.50	Moderate
8	G08	Sore throat	0.60	Important
9	G09	Decreased appetite	0.55	Moderate

Table 2 shows some examples of symptoms discussed in this study. In addition to symptom data, treatments, both medical and traditional, are also listed. Examples can be seen in Table 3.

Table 3. Treatment

No	Code	Drug	Type
1	BH009	Young Coconut Water	Traditional
2	OBT009	Amlodipine	Medical
3	OBT002	Amoxicillin	Medical
4	OBT005	Cetirizine	Medical
5	BH006	Guava Leaves	Traditional
6	BH007	Papaya Leaves	Traditional
7	BH005	Betel Leaves	Traditional
8	OBT003	Ibuprofen	Medical
9	BH001	Ginger	Traditional
10	BH004	Ginger	Traditional

Table 3 shows examples of treatment data and Table 4 shows examples of treatment solution data used in this study.

Table 4. Solutions

Disease	Solution Type	Solution Content (Summary)	Trust Level	Related Drugs	Status
Dengue Fever (P001)	Medical	1. Give paracetamol for fever 2. Ensure adequate fluid intake 3. Get enough rest 4. If shortness of breath consult a doctor	90.0% (Very High)	Paracetamol, ORS	Very Effective
Dengue Fever (P001)	Traditional	1. Drink young coconut water 2. Guava juice to increase platelets 3. Warm chicken soup 4. Get enough rest	60.0% (High)	Young Coconut Water, Papaya Leaves	Effective
Malaria (P002)	Medical	1. Antimalarial medication according to doctor's prescription 2. Get enough rest 3. Drink plenty of water	95.0% (Very High)	Paracetamol, Amoxicillin	Very Effective
Malaria (P002)	Traditional	1. Drink boiled papaya leaf water 2. Fresh fruit juice 3. Warm nutritious soup 4. Body compress	50.0% (Moderate)	Papaya Leaves	Fairly Effective
Typhus (P003)	Medical	1. Antibiotics as prescribed by the doctor 2. Soft diet that is easy to digest 3. Complete rest 4. Monitor body temperature	90.0% (Very High)	Amoxicillin, Paracetamol	Very Effective
Typhus (P003)	Traditional	1. Warm chicken soup 2. Turmeric boiled water 3. Bananas to bind stool 4. Get enough rest	40.0% (Moderate)	Turmeric	Fairly Effective

Table 4 shows examples of treatment solutions used in this study.

Case Representation

In this study, each case is represented as a knowledge package consisting of the disease identity, associated symptoms, and solution content in the form of modern medical treatments or treatments based on local/traditional wisdom. Case representation is used as the primary knowledge base in the Case-Based Reasoning (CBR) method. An example can be seen in Table 4.

CBR Mechanism

The Case-Based Reasoning (CBR) method works by mimicking human thought patterns, using past experiences (old cases) to solve new problems (new cases). In this study, CBR was used to generate treatment recommendations for toddlers by referring to a case database containing information on the disease, symptoms, medical and traditional solutions, related medications, and the solution's level of confidence. An example is shown below:

1. Retrieve: The system calculates the similarity of the query with the cases of Dengue Fever (P001) and Malaria (P002). Results:
    a. Sim(Query, Dengue Fever) = 0.85
    b. Sim(Query, Malaria) = 0.70
    The most similar case is Dengue Fever (0.85).
2. Reuse: Initial solution → *Paracetamol, Oralit, Get enough rest (Medical, 90% confidence).*
3. Revise: Fuzzy detects temperature 38.5°C → moderate–high category with membership 0.6. Fuzzy rule: IF (moderate–high fever) AND (cough=yes) → add "Monitor hydration + warm compress". Revised solution → Paracetamol + ORS + Rest + Warm compress + Monitor hydration.
4. Retain: This new case is saved to the case base with the revised solution.

Fuzzy Integration and Final Results

This stage involves entering a new case in the form of symptoms experienced by toddlers, namely high fever (≥38°C), diarrhea more than 3 times per day, and decreased appetite. After selecting the symptoms, the user is asked to determine the intensity level of each symptom. The results of the fuzzy intensity input from the user are: high fever = 0.7; diarrhea = 0.5; decreased appetite = 0.6. These values are then processed in the fuzzification stage, so that each symptom is not only expressed binary (present/absent), but with a certain degree of membership in the medium or high category.

The system calculates similarity with the case base through the CBR mechanism. The case with the highest similarity level is Typhus (P003) with a similarity value of 0.78 or 78.3%, while Dengue Fever and Malaria cases have lower similarity values. Based on these results, the system takes the initial solution for Typhus disease contained in the case base, in the form of medical recommendations (antibiotics, paracetamol, soft diet, adequate rest) and traditional recommendations (warm chicken soup, turmeric boiled water, bananas). The revision stage is then carried out using fuzzy integration. Fuzzification applied to a value of 0.7 for high fever places the toddler's condition in the medium-high group, while diarrhea with a value of 0.5 and decreased appetite with a value of 0.6 are categorized as moderate. The active fuzzy rules are:

1. If the fever is moderate and the diarrhea is moderate, the system adds a suggestion to monitor hydration and apply warm compresses.
2. If the appetite decreases moderately, the system adds a solution to provide nutritious soft foods to support the diet.
3. If the fever is approaching high, then medical solutions are given greater priority weight.

The final result was a diagnosis of typhus with a 78.3% confidence level (high). Solutions were divided into two categories: medical solutions with a 90% confidence level (very effective) and traditional solutions with a 40% confidence level (fairly effective).

Details of the recommendations provided include:

a. Medical Solution: use of Paracetamol as a fever-reducing drug (dose 10–15 mg/kgBW every 4–6 hours), Amoxicillin as the main antibiotic (dose 20–40 mg/kgBW/day in 2–3 doses), accompanied by complete rest and a soft diet.

b. Traditional Solution: consume warm chicken soup, boiled turmeric water (1–2 cm of rhizome boiled per glass of water), and consume bananas to bind stool.

System Development

At this stage, the recommendation system is developed. Several system displays can be seen in the following image:

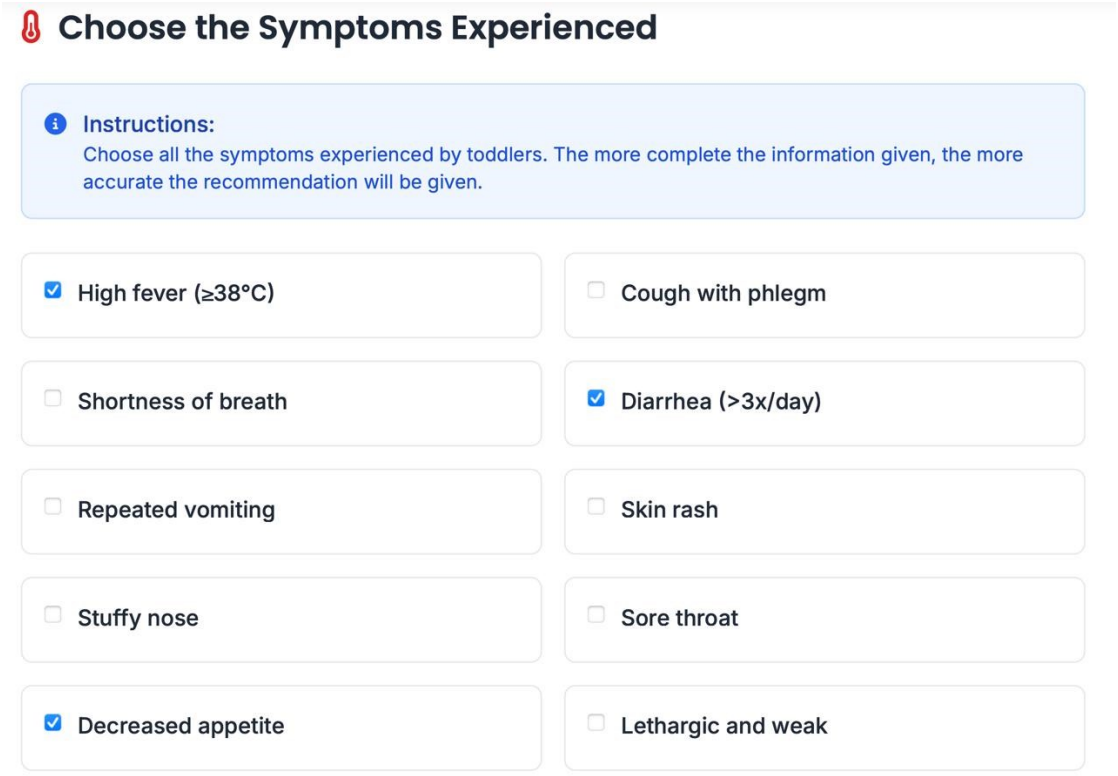


Figure 4. Symptom selection page

Figure 4 shows the symptom selection page that can be used in the developed recommendation system. Users enter the symptoms experienced by their toddler on this page. The user then determines a fuzzy value based on the symptoms experienced. The fuzzy page can be seen in Figure 5.

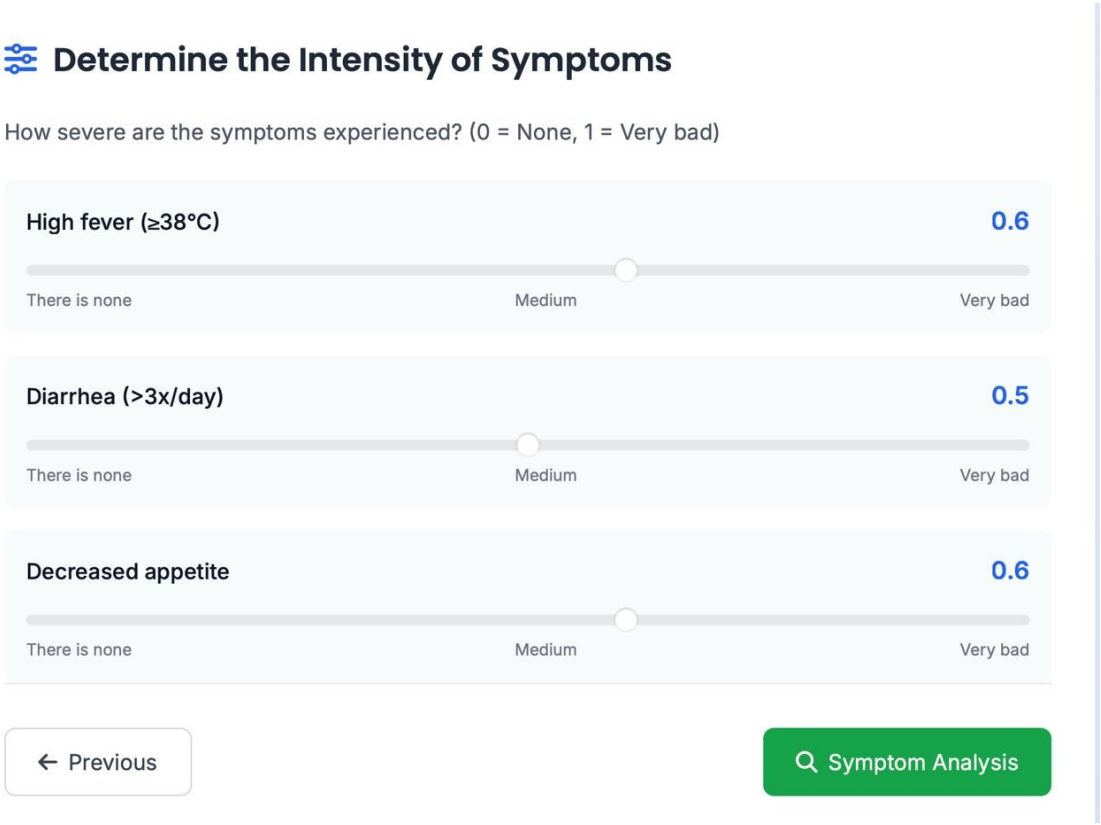


Figure 5. Fuzzy page

Figure 5 shows the page used to determine the fuzzy value by looking at the intensity of the symptoms experienced. The results of the combination of CBR and Fuzzy based on the selected symptoms and intensity can be seen in Figure 6.

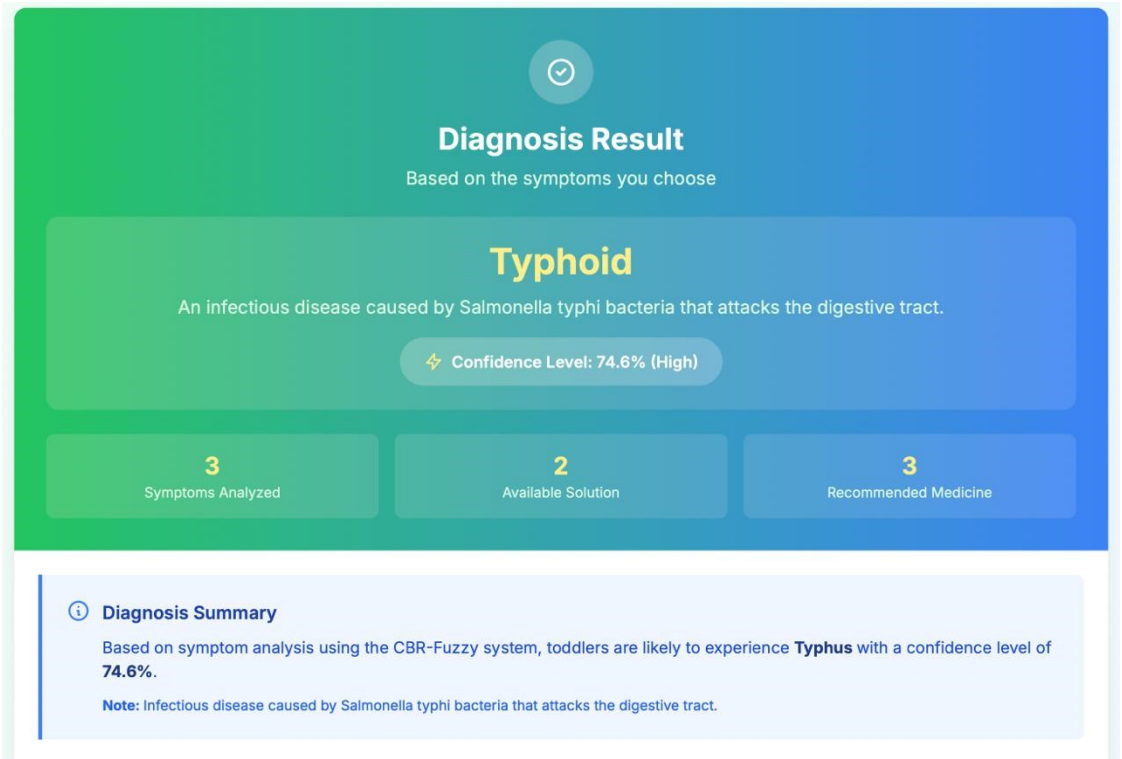


Figure 6. Diagnosis result page

Figure 6 shows the diagnostic results obtained based on the previous input. The image shows that, based on the symptoms and intensity entered, a diagnosis of typhus was obtained. Treatment recommendations can be seen in Figure 7.

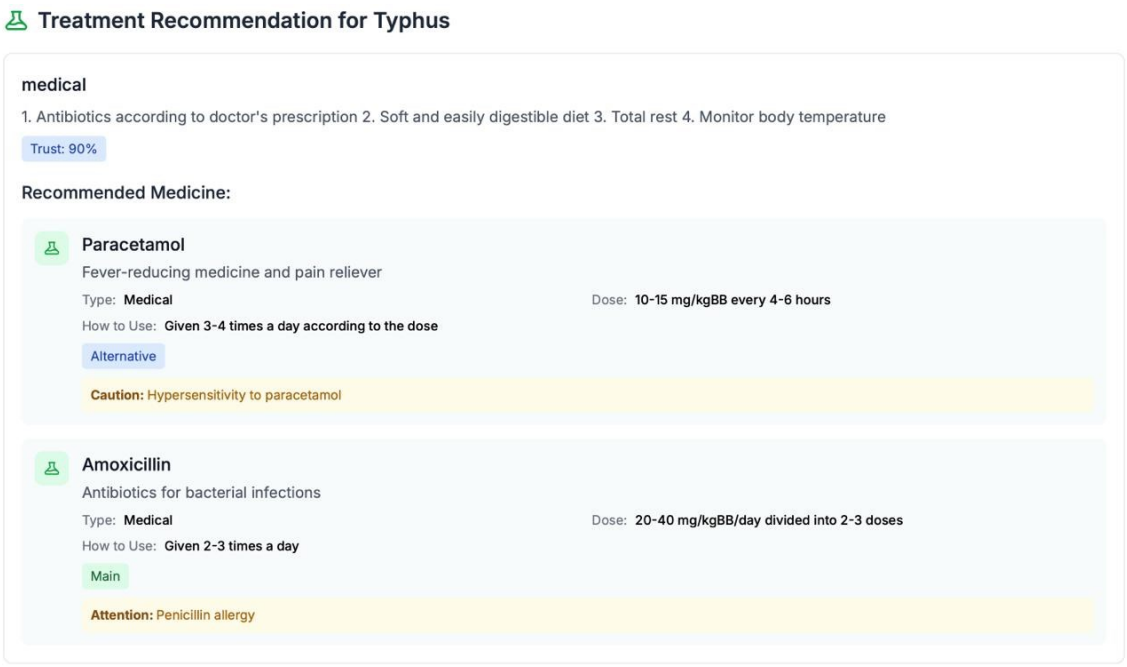


Figure 7. Treatment recommendations page

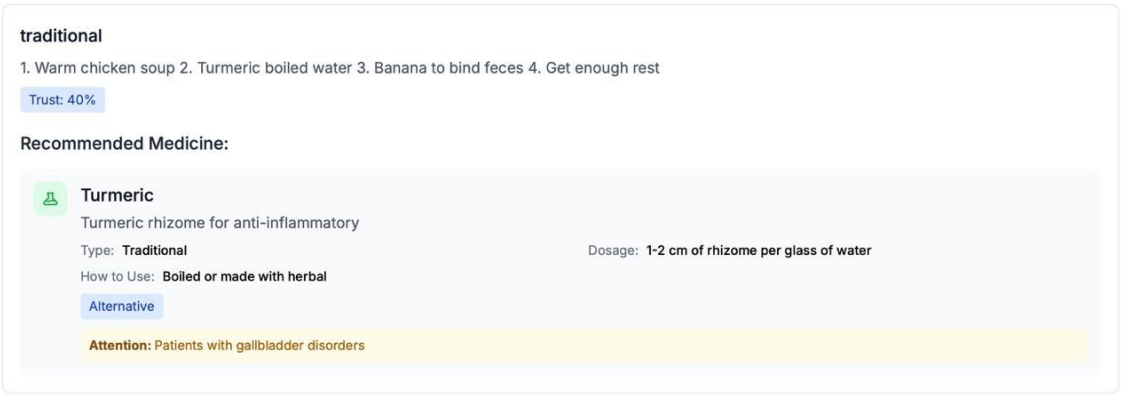


Figure 8. Traditional treatment recommendations page

Evaluation

The system evaluation aimed to assess the performance, reliability, and user acceptance of the CBR-Fuzzy-based medication recommendation system for toddlers. In this study, the evaluation was conducted using three approaches: blackbox testing, system accuracy calculations, and the System Usability Scale (SUS), involving 50 respondents. The results of the blackbox testing are shown in Table 5.

Table 5. Blackbox testing result

Test Scenario	Input Steps	Expected Output	Test Results	Status
Input toddler symptom data	The user selects several symptoms (e.g., high fever, diarrhea, decreased appetite).	The system saves the symptoms to a new case query	Compliant	Successful
Symptom analysis using the CBR method	The user clicks the Symptom Analysis button.	The system calculates similarity with the case base	Compliant	Successful
Revise results using Fuzzy Login	The system receives the symptom intensity (0–1) as input.	The system calculates the fuzzy membership degree and revises the solution	Compliant	Successful
Display diagnosis results	After analysis, the system displays the detected disease along with the confidence level.	The results page displays the disease name + confidence percentage	Compliant	Successful
Recommended medical solutions	When a disease diagnosis is made (e.g., typhoid), the system calls up medical solutions from the knowledge base.	Displays a list of medical drugs + dosage & instructions	Compliant	Successful
Recommended traditional solutions	When a disease is detected, the system also calls up traditional solutions.	Displays herbal ingredients and instructions for use	Compliant	Successful
Alert/user education features	After the diagnosis results are displayed, the admin logs in with a valid account.	The system displays a warning message to consult a doctor	Compliant	Successful
Admin login	The admin adds/changes/deletes disease, symptom, or medication data.	The system displays the admin dashboard	Compliant	Successful
Manage case database (CRUD)	The user clicks the logout button.	New data is saved/changed/deleted in the case base	Test Results	Successful
Admin/user logout	Input Steps	The system returns to the login page	Compliant	Status

Table 5 shows the results of blackbox testing conducted with 10 scenarios. The table shows that all scenarios performed yielded results as expected, resulting in a 100% score for the blackbox testing. The next step is accuracy testing. This testing is conducted by comparing the system's results with those of an expert. The accuracy test can be seen in Table 6.

Table 6. Accuracy test result

No	Main Symptoms (Input)	Systemic Diagnosis	Expert Diagnosis	Status
1	High fever, skin rash, joint pain	Dengue Fever	Dengue Fever	Appropriate
2	High fever, diarrhea >3 times, decreased appetite	Typhus	Typhus	Appropriate
3	Recurring fever, chills, night sweats	Malaria	Malaria	Appropriate
4	High fever, headache, vomiting	Dengue Fever	Malaria	Not Appropriate
5	High fever, weakness, repeated vomiting	Typhus	Dengue Fever	Not Appropriate
….	……	….	……	….
50	Gastritis: stomach pain, nausea	Gastritis	Gastritis	Appropriate

Table 6 shows the accuracy tests performed. Of the 50 scenarios tested, 6 results were inconsistent, resulting in an accuracy rate of 88%. The next test was conducted using SUS [28]. The summary of SUS data can be seen in Table 7.

Table 7. SUS value summary

Resp	Q1	Q2*	Q3	Q4*	Q5	Q6*	Q7	Q8*	Q9	Q10*
R1	4	4	4	4	5	4	4	4	4	4
R2	5	4	5	3	5	3	5	4	5	4
R3	4	3	4	4	4	4	4	3	4	3
R4	3	4	4	3	4	3	4	4	4	4
R5	4	3	5	4	5	4	5	3	5	4
…	…	…	…	…	…	…	…	…	…	…
R50	4	3	5	4	5	3	5	4	5	4

Table 7 shows the results of the SUS summary, where the values were then converted into SUS rules, resulting in an average SUS value of 78.4.

Discussions

The findings of this study indicate that the integration of Case-Based Reasoning (CBR) and Fuzzy Logic in a toddler treatment recommendation system provides significant improvements in both technical performance and user acceptance. This integration enables the system to deliver accurate initial diagnostic recommendations, adapt to uncertainty in symptom data, and remain culturally sensitive through the incorporation of modern medical knowledge and verified local wisdom.

In the case representation stage, the knowledge base was successfully structured to include information on diseases, symptoms, and both medical and traditional treatment solutions. This structured representation allows the system to consistently perform case retrieval and generate recommendations based on previous case experiences with measurable similarity levels. Although the CBR mechanism proved effective in identifying relevant cases, its deterministic nature limits its ability to handle ambiguous symptoms. This limitation is addressed through the incorporation of fuzzy logic in the revision stage, enabling symptoms to be expressed as degrees of membership rather than binary values. As a result, the system can more realistically account for symptom intensity, producing recommendations that are more flexible and aligned with actual patient conditions.

The system achieved an accuracy score of 88%, indicating a high level of reliability in generating initial diagnostic results. Most inconsistencies were found in diseases with overlapping clinical manifestations, highlighting the need for expanding the case base and refining fuzzy membership functions to improve the system's discriminatory capability. Furthermore, blackbox testing confirmed that all system functionalities operate according to design specifications, demonstrating functional feasibility.

From the usability perspective, evaluation using the System Usability Scale (SUS) involving 50 respondents resulted in an average score of 78.4, categorized as *Good*. This shows that the system is easy to use, quick to learn, and provides a satisfying user experience. However, a small number of respondents indicated that some features could still be simplified, suggesting that interface improvements may help achieve an *Excellent* usability rating.

The use of traditional treatment data in this system poses potential biases, as herbal practices are influenced by cultural variations, availability of ingredients, and differing interpretations among practitioners. In addition, several traditional recommendations lack strong medical documentation, which may lead to overgeneralization. To mitigate these issues, this study intentionally selected herbal treatments commonly used across multiple regions and verified by experienced herbal practitioners while still prioritizing standardized medical recommendations. The system also provides warnings for users to seek medical consultation when necessary.

Overall, the results confirm that the integration of CBR and Fuzzy Logic provides a balanced framework for supporting decision-making in toddler treatment recommendations. CBR contributes by leveraging past case experiences, while fuzzy logic enhances the system's ability to manage uncertainty. Although the system demonstrates strong technical performance and good user acceptance, further development is required, particularly in expanding the case base, refining fuzzy rules, and improving the user interface to enhance intuitiveness and broaden system acceptance. It is also important to note that the system has not yet been tested in real field conditions, so practical effectiveness and user behavior in actual healthcare settings remain to be evaluated in future studies.

Conclusion

This research successfully developed a Case-Based Reasoning (CBR)-based treatment recommendation system for toddlers combined with Fuzzy Logic and integrating modern medical knowledge and local wisdom. The system provided initial diagnosis results with an accuracy rate of 88%, obtained a System Usability Scale (SUS) score of 78.4 in the Good Usability category, and blackbox testing results showed that all main functions ran according to design, so the system was considered quite reliable, adaptive, and acceptable to the community. For further research, it is recommended that the system be developed by expanding the case base, refining fuzzy rules, improving the user interface to be more intuitive, and implementing it on a mobile platform so that it can improve accuracy, expand user reach, and achieve an Excellent Usability level in a wider usage context.

Acknowledgments

We would like to express our gratitude to the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia through the Directorate of Research, Technology, and Community Service for funding this research through the 2025 Fiscal Year Beginner Lecturer Grant. We would also like to express our gratitude to the LPPM of Al-Azhar Islamic University of Mataram for facilitating this research activity.

References

[1] Y. H. Pratama, Firmansyah, and A. Arfin, "Development of an Intelligent System for Early Diagnosis of Diseases in Toddlers Using Forward Chaining and Dempster-shaferIntegration," *J. Sist. Cerdas*, vol. 8, no. 1, 2025, doi: https://doi.org/10.37396/jsc.v8i1.469.

[2] N. K. Umami and S. Wibisono, "Deteksi Dini Penyakit Balita Menggunakan Algoritma Sorensen Berbobot," *J. Ilm. Inform.*, vol. 9, no. 02, 2021, doi: 10.33884/jif.v9i02.3744.

[3] S. N. Budihardjo and I. W. B. Suryawan, "Faktor-faktor resiko kejadian pneumonia pada

pasien pneumonia usia 12-59 bulan di RSUD Wangaya," *Intisari Sains Medis*, vol. 11, no. 1, 2020, doi: 10.15562/ism.v11i1.645.

[4] A. Armina and A. Wulansari, "Korelasi Faktor yang Berhubungan dengan Kejadian Pneumonia Balita di Dua Puskesmas Kota Jambi," *J. Ilm. Univ. Batanghari Jambi*, vol. 20, no. 1, 2020, doi: 10.33087/jiubj.v20i1.801.

[5] R. N. Anjaswanti, R. Azizah, and A. Leonita, "Studi Meta-Analisis: Faktor Risiko Kejadian Pneumonia Pada Balita di Indonesia Tahun 2016-2021," *J. Community Ment. Heal. Public Policy*, vol. 4, no. 2, 2022, doi: 10.51602/cmhp.v4i2.65.

[6] A. Amalia, F. Fahdhienie, and F. Fadhlullah, "Faktor-Faktor Yang Berhubungan dengan Kejadian Ispa Pada Balita (1-4 Tahun) di Wilayah Kerja Puskesmas Bandar Kecamatan Bandar Kabupaten Bener Meriah Tahun 2023," *J. Bid. Ilmu Kesehat.*, vol. 14, no. 1, pp. 72–81, 2024, doi: https://doi.org/10.52643/jbik.v14i1.4116.

[7] L. Atamou, D. C. Rahmadiyah, H. Hassan, and A. Setiawan, "Analysis of the Determinants of Stunting among Children Aged below Five Years in Stunting Locus Villages in Indonesia," *Healthc.*, vol. 11, no. 6, 2023, doi: 10.3390/healthcare11060810.

[8] Y. Permanasari *et al.*, "Faktor Determinan Balita Stunting Pada Desa Lokus Dan Non Lokus Di 13 Kabupaten Lokus Stunting Di Indonesia Tahun 2019," *Penelit. Gizi dan Makanan (The J. Nutr. Food Res.*, vol. 44, no. 2, 2021, doi: 10.22435/pgm.v44i2.5665.

[9] A. Muhawarman, "Pneumonia Terus Ancam Anak-anak," Kementrian Kesehatan. Accessed: Feb. 11, 2025. [Online]. Available: https://kemkes.go.id/id/pneumonia-terus-ancam-anak-anak

[10] A. P. Anggraini, "4 Penyakit yang Sering Memicu Kematian Pada Anak," Kompas.com. Accessed: Jun. 30, 2024. [Online]. Available: https://health.kompas.com/read/2022/06/06/080000668/4-penyakit-yang-sering-memicu-kematian-pada-anak?page=all

[11] N. Ahmad, "Ketidakmerataan Fasilitas Kesehatan dan Tenaga Kesehatan di Indonesia," Kompasiana. Accessed: Feb. 11, 2025. [Online]. Available: https://www.kompasiana.com/nabhan2104/64f055574addee637b69e5e2/ketidakmerataan-fasilitas-kesehatan-dan-tenaga-kesehatan-di-Indonesia

[12] D. K. D. I. Yogyakarta, "Penyelenggaraan Pengobatan Tradisional di Indonesia," Dinas Kesehatan Daerah Istimewa Yogyakarta. Accessed: Apr. 10, 2025. [Online]. Available: https://dinkes.jogjaprov.go.id/berita/detail/penyelenggaraan-pengobatan-tradisional-di-indonesia

[13] A. J. Jawan, F. B. Tokan, and D. D. Dhosa, "Kearifan Lokal Masyarakat Adat Dalam Menjaga Ketahanan Pangan Melalui Tradisi Rewa`Ng Plea `(Studi Kasus Desa Di Daniwato Kecamatan Solor Barat Kabupaten Flores Timur)," *J. Educ. Gov. Wiyata*, vol. 3, no. 1, pp. 243–273, 2025.

[14] E. S. N. Sumarlina, Heriyanto, and I. R. Husen, "PENGOBATAN TRADISIONAL BERBASIS KEARIFAN LOKAL NASKAH MANTRA," *J. Pengabdi. Kpd. Masy.*, vol. 1, no. 4, pp. 212–218, 2017.

[15] Sandi Alam and G. widi Nurcahyo, "Sistem Pakar dalam Mendiagnosis Gizi Buruk pada Balita dengan Menggunakan Metode CBR," *J. Sistim Inf. dan Teknol.*, 2022, doi: 10.37034/jsisfotek.v4i4.140.

[16] V. A. Arundy, I. Fitri, and E. Mardiani, "Implementasi Metode Penalaran CBR dalam Mengidentifikasi Gejala Awal Penyakit Jantung menggunakan Algoritma Sorensen Coeffient," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 5, no. 3, pp. 306–313, 2021.

[17] N. Numan, S. Kusumadewi, and N. Muzayyanah, "Sistem Inferensi Fuzzy Untuk Membantu Diagnosis Penyakit Pneumonia Anak," *IT J. Res. Dev.*, vol. 5, no. 1, 2020, doi: 10.25299/itjrd.2020.vol5(1).5088.

[18] P. R. Ananda and S. Sriani, "Sistem Pakar Diagnosis Stunting pada Balita Menggunakan Metode Forward Chaining dan Logika Fuzzy Sugeno," *J. Teknol. Sist. Inf. dan Apl.*, vol. 7, no. 1, pp. 200–216, 2024, doi: https://doi.org/10.32493/jtsi.v7i1.38245.

[19] A. N. S. Christian and S. S. Winarsih, "Aplikasi Antrian Dan Pengolahan Data Balita Berbasis Website (Studi Kasus : Posyandu Arum Dalu Desa Sanggrahan)," *J. Teknol. Inf. dan Komun.*, vol. 12, no. 1, pp. 49–57, 2024, doi: http://dx.doi.org/10.30646/tikomsin.v12i1.810.

[20] B. Marques, C. Freeman, and L. Carter, "Adapting traditional healing values and beliefs

into therapeutic cultural environments for health and well-being," *Int. J. Environ. Res. Public Health*, vol. 19, no. 1, 2022, doi: 10.3390/ijerph19010426.

[21] S. Supriadi, S. Suryani, L. Anggresani, S. Perawati, and R. Yulion, "Analisis Penggunaan Obat Tradisional Dan Obat Modern Dalam Penggunaan Sendiri (Swamedikasi) Oleh Masyarakat," *J. Kesehat.*, vol. 14, no. 2, 2022, doi: 10.24252/kesehatan.v14i2.20347.

[22] R. Anggara, M. H. Nasri, N. Fadli, Fatimatuzzahra, and Y. H. Pratama, "Impelementasi Metode Dempster Shafer Pada Diagnosis Penyakit di Provinsi Nusa Tenggara Barat," *J. Millenial Informatics*, vol. 2, no. 1, pp. 13–22, 2024.

[23] W. M. Ningrum and K. D. Purnamasari, "Tanaman Obat Keluarga untuk Kesehatan Ibu dan Bayi di Desa Sukamulya," *J. Midwifery Public Heal.*, vol. 5, no. 1, pp. 21–26, 2023, [Online]. Available: https://dx.doi.org/10.25157/jmph.v5i1.16180

[24] A. N. Azizah and C. H. Kurniati, "Obat Herbal Tradisional Pereda Batuk Pilek Pada Balita," *J. Kebidanan Indones.*, vol. 11, no. 2, 2020, doi: 10.36419/jkebin.v11i2.370.

[25] S. Park, S. Park, H. Jang, Y. Ahn, and N. Kwon, "Computing green remodeling construction cost for public buildings based on genetic algorithm and case-based reasoning," *Dev. Built Environ.*, vol. 22, no. 100655, 2025, doi: https://doi.org/10.1016/j.dibe.2025.100655.

[26] M. Parola, F. A. Galatolo, G. La Mantia, M. G. C. A. Cimino, G. Campisi, and O. Di Fede, "Towards explainable oral cancer recognition: Screening on imperfect images via Informed Deep Learning and Case-Based Reasoning," *Comput. Med. Imaging Graph.*, vol. 117, no. 102433, 2024, doi: https://doi.org/10.1016/j.compmedimag.2024.102433.

[27] N. Sigani, B. A. Masse, and N. Nurdin, "Sistem Pakar Untuk Mendiagnosa Penyakit Mata Manusia Menggunakan Metode Fuzzy Logic," *J. Elektron. Sist. Inf. dan …*, vol. 2, no. 10, 2022.

[28] R. M. A. Putri, W. G. S. Parwita, I. P. S. Handika, I. G. I. Sudipa, and P. P. Santika, "Evaluation of Accounting Information System Using Usability Testing Method and System Usability Scale," *Sinkron*, vol. 9, no. 1, 2024, doi: 10.33395/sinkron.v9i1.13129.

.

Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 16 Issue 1 Year 2026 Pages 29-38
DOI: http://dx.doi.org/10.31940/matrix.v16i1.29-38
URL: http://ojs2.pnb.ac.id/index.php/matrix

Bell's Palsy and stroke face classification using SVM with MediaPipe Face Mesh

Chelsea Effendi [1], Destriana Widyaningrum [2]*

[1,2] *Informatics Study Program, Universitas Bunda Mulia, Indonesia*

*Corresponding Author: *l0894@lecturer.ubm.ac.id*

Abstract: Stroke and Bell's Palsy share similar manifestations of unilateral facial paralysis, often leading to clinical misinterpretation, particularly in acute cases. Although deep learning approaches have demonstrated strong performance in segmenting facial paralysis regions, these methods primarily focus on area localization rather than directly differentiating Stroke and Bell's Palsy, and typically require large-scale datasets and substantial computational resources. To address this gap, this study proposes an explainable and resource-efficient framework for classifying Stroke and Bell's Palsy using asymmetric facial numeric features extracted from static images. Unlike appearance-based deep learning models, the proposed approach transforms facial landmarks detected by MediaPipe Face Mesh into geometric asymmetry features through Min–Max scaling, Euclidean distance, and angle computation. After class balancing via undersampling, classification was performed using an SVM with an RBF kernel. The 70:30 split achieved the most stable performance, with a testing accuracy of 0.8041 and cross-validation accuracy of 0.8072 ± 0.0069, indicating minimal generalization gap. These findings demonstrate that geometric asymmetry features combined with SVM provide a reliable and interpretable alternative for differentiating BP and ST under limited data and computational constraints.

Keywords: Bell's Palsy, facial asymmetry, machine learning, MediaPipe Face Mesh, stroke, Support Vector Machine

History Article: Submitted 23 January 2026 | Revised 13 February 2026 | Accepted 14 February 2026

How to Cite: C. Effendi and D. Widyaningrum, "Bell's Palsy and stroke face classification using SVM with MediaPipe Face Mesh," *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 16, no. 1, pp. 29–38, 2026, doi: 10.31940/matrix.v16i1.29-38.

Introduction

Stroke (ST) is an acute medical condition caused by vascular injury to the central nervous system [1]. In the simplest terms, stroke occurs when the blood supply to the brain is disrupted, causing areas of the brain that lack oxygen and nutrients to die [2]. Thus, if the affected area involves the region of the brain responsible for facial nerves, one of the symptoms is facial paralysis. Its etiology involves a wide range of risk factors, such as hypertension, obesity, high cholesterol, genetic predisposition, and progression from cardiovascular diseases. On the other hand, Bell's Palsy (BP) is a non-progressive acute medical condition characterized by temporary weakness or paralysis of the facial muscles [3]. These similar symptoms led people with acute unilateral upper and lower facial palsy to believe that they had experienced a stroke, when in fact it was benign Bell's palsy [4]. Furthermore, a study on ischemic stroke reported that 24.8% of cases were stroke mimics, of which 5% were mononeuropathies, including BP [5]. So far, the exact etiology of Bell's Palsy remains uncertain. However, several studies suggest that it may be associated with viral infections, trauma, or specific syndromes, with additional contributing factors such as pregnancy, diabetes, and even environmental conditions [3], [6], [7].

Studies on facial asymmetry have been conducted using deep learning models to segment paralyzed areas within a face [2]. These approaches have demonstrated strong performance, particularly when trained on large-scale and diverse datasets. However, achieving such high performance with deep learning models is closely associated with the availability of large-scale and diverse datasets, as well as substantial computational resources. In addition, privacy and ethical issues surrounding facial data make the data-collection process even more challenging. The study in [2] was conducted using 18,840 facial images of Bell's palsy and private dataset

consisting of approximately 1,435 facial images of Stroke patients, which may not be readily accessible for independent replication or comparative evaluation.

In contrast, the publicly available datasets identified for this study are considerably smaller, particularly for Bell's Palsy cases, for which only 10,754 facial images were obtainable. Furthermore, only 2,279 facial images of Stroke cases were available, resulting in a significant class imbalance that may cause classification bias. To address these data constraints, this study applies an undersampling strategy to mitigate class imbalance and adopts Support Vector Machine (SVM), a supervised machine learning algorithm known for its robustness when trained on limited datasets and suitable for resource-limited environments [8], [9]. The proposed framework is designed to classify Bell's Palsy and Stroke based on asymmetric facial numeric features extracted from static facial images using a Support Vector Machine (SVM) classifier.

A previous evaluation also identified SVM as the more accurate model, achieving 96.59% and outperforming XGB and RF while maintaining high precision, recall, F1-score, and Matthews Correlation Coefficient [10]. A comparative study between CNN and SVM showed that CNN achieved higher classification performance, but required significantly greater computational resources and twice the training time [8]. These findings align with the needs of this study, reinforcing SVM as an appropriate baseline for differentiating Bell's Palsy and Stroke.

Rather than directly analyzing facial appearance, this study employs MediaPipe Face Mesh (MFM) as a geometric abstraction tool to convert facial images into numeric landmark representations. This approach reduces reliance on demographic visual cues such as ethnicity and race, while enabling an explainable analysis based on facial asymmetry. Specifically, MFM is used to detect the human face and extract facial landmark coordinates, which serve as the input for computing asymmetric facial numeric features. MFM is one of the MediaPipe framework libraries capable of determining 468 facial landmark coordinate points of human facial geometry (green dots in Figure 1). MFM is Google's pre-trained library which derives three-dimensional positioning from a two-dimensional image [11]. It combines a lightweight detector model called BlazeFace, which functions to crop the face from the original image for processing by a subsequent landmark generator model that predicts the facial surface and assigns precise positions to the landmarks [12]. In addition, MFM is optimized for extracting facial features efficiently in environments with limited hardware resources [13].

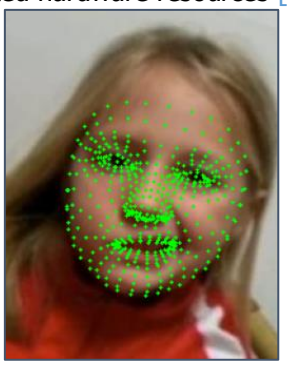


Figure 1. MediaPipe Face Mesh coordinate points

In contrast to prior studies that rely on deep learning architectures, this study proposes an explainable framework based on asymmetric facial numeric features extracted from static images. The proposed approach enables effective differentiation between Bell's palsy and stroke under limited data availability and hardware constraints, while also facilitating interpretable analysis of feature contributions relevant to facial asymmetry.

Methodology

The proposed method is structured into four major stages as presented in Figure 2. The first stage is feature engineering, which consists of facial feature extraction and dataset-based class assignment, where labels are determined according to the source of the images. The second stage is data preprocessing, which includes label encoding, class balancing through

undersampling, and data cleaning. The third stage consists of data splitting and SVM modeling. The fourth stage focuses on model evaluation.

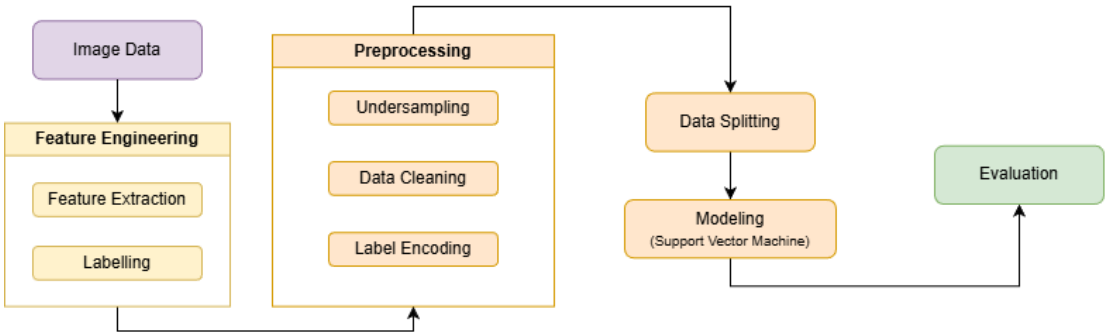


Figure 2. Proposed methodology flow

Dataset Description

The Bell's Palsy dataset was collected from HuggingFace [14], while the Stroke datasets were sourced from Kaggle [15], [16]. The first Kaggle dataset contains 1,029 stroke-labeled images, and the second contains 1,259 images, resulting in a total of 2,283 Stroke images. Meanwhile, the HuggingFace dataset contains 27 videos of bell's palsy patients. To address this limitation, each .mp4 video was converted into .jpg images using a frame-skipping strategy, producing approximately 400 images per video or all available frames if the total frame count was below 400. This process resulted in a total of 10,758 Bell's Palsy images.

Feature Engineering

The image data were converted into numerical features using MFM (see Figure 1). The coordinates of the eyebrows, eyes, nose, and both mouth corners are required to compute facial asymmetry. Therefore, all the coordinates mentioned were extracted along both the X and Y axes using MFM. The extracted coordinates for each class (BP and ST) were then saved into separate CSV files, with the column structure shown in Table 1. Labelling process conducted by adding one extra column in each CSV files. 'BP' for Bell's Palsy CSV and 'ST' for Stroke CSV. The labelling is source-based labelling because each dataset exclusively contained images from a single class. Finally, the labelled 'BP' and 'ST' CSV files were combined and saved into a single CSV file.

Table 1. Numerical features - the extracted coordinates

Index	Column Name
0	file
1	label
2	left_eyebrow_x
3	left_eyebrow_y
4	right_eyebrow_x
5	right_eyebrow_y
6	left_eye_x
7	left_eye_y
8	right_eye_x
9	right_eye_y
10	nose_x
11	nose_y
12	mouth_left_x
13	mouth_left_y
14	mouth_right_x
15	mouth_right_y

To compute the asymmetric facial features, Min-Max scaling, Euclidean distance, and angle calculation were performed. Min-Max scaling is used to normalize all landmark coordinates so that the features become relative to the face rather than global pixel values. Euclidean distance was used to measure the differences in distance between the left and right sides of the face. Finally, angle computation was applied to quantify the angular difference between the two mouth corners.

$$x_{new} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (2)$$

$$\theta = \frac{180}{\pi} tan^{-1}\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \quad (3)$$

Equation (1) shows the Min–Max scaling method used in this study. Equation (2) shows the Euclidean distance calculation, where (x1, y1) is the first coordinate and (x2, y2) is the second coordinate. Equation (3) shows the angle calculation method used in this study. After the computation, the asymmetric facial features were saved into a single CSV file, with the column structure shown in Table 2.

Table 2. Numerical features - asymmetric facial features

Index	Column Name
0	file
1	label
2	delta_eyebrow_y
3	delta_eyebrow_x
4	delta_eye_y
5	delta_eye_x
6	dist_eye_left_to_nose
7	dist_eye_right_to_nose
8	delta_eye_nose
9	dist_mouth
10	mouth_angle x

Preprocessing

The preprocessing of the asymmetric facial feature numerical data consists of three processes. First is the encoding of each label into binary form, where "ST" is encoded as 0 and "BP" as 1. Second is the examination of the number of entries for each label. The label with a greater number of entries is reduced so that it matches the number of entries of the label with fewer samples (undersampling). Third is the removal of the unused column, specifically the "file" column.

Data Splitting

After obtaining a balanced dataset with a 1:1 ratio between Bell's Palsy and Stroke through undersampling, the data were divided into training and testing sets using three splitting ratios (70:30, 80:20, and 90:10). The detailed class distribution for each ratio is presented in Table 3. These three ratios are used to ensure a more reliable evaluation, as relying on a single ratio may introduce bias and fail to represent the model's performance across different data distributions. Stratification was applied to ensure that the class distribution remained proportional in both the training and testing sets.

Table 3. Data splitting details

Ratio	Subset	BP	ST	Total
70:30	Train	1595	1595	3190
70:30	Test	684	684	1368
80:20	Train	1823	1823	3646
80:20	Test	456	456	912
90:10	Train	2051	2015	4102
90:10	Test	228	228	456

Modelling

This study employed Support Vector Machine (SVM) for classification. The extracted facial asymmetry features demonstrate nonlinear patterns, indicating that the classes cannot be separated using a linear hyperplane. Therefore, the Radial Basis Function (RBF) kernel was selected, as it is widely used for handling non-linear data distributions and has demonstrated strong classification performance across various domains [17], [18]. To ensure a fair and reproducible baseline evaluation, the default RBF hyperparameters (C = 1 and gamma = 'scale') were utilized. The objective of this study was to assess the discriminative capability of the proposed asymmetric facial features rather than to maximize classifier performance through extensive hyperparameter tuning. Therefore, systematic optimization procedures such as grid search were considered beyond the scope of the present work.

The Gaussian RBF kernel can be expressed using the following equation:

$$k(x_1 x_2) = \exp(-\gamma \|x_1 - x_2\|^2) \quad (4)$$

Information:

k : kernel
x_1, x_2 : input feature vectors
$\exp$: exponential function (ensures the kernel output is between 0 and 1)
γ : gamma

Evaluation

A confusion matrix was employed to analyze class-level prediction performance. The confusion matrix consists of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), as presented in Table 4. Based on these values, the classification report metrics such as Accuracy, Precision, Recall, and F1-Score can be computed.

Table 4. Confusion matrix components

Component	Column Name
TP	The model correctly predicts Bell's Palsy.
FP	The model incorrectly predicts Bell's Palsy, but the actual label is Stroke.
TN	The model correctly predicts Stroke.
FN	The model incorrectly predicts Stroke, but the actual label is Bell's Palsy.

To further assess the robustness and generalization capability of the proposed SVM model, stratified 5-fold cross-validation was conducted. The training data were divided into five stratified folds, and the model was trained and validated iteratively across them. The average accuracy and standard deviation were reported to provide a more reliable estimate of model performance.

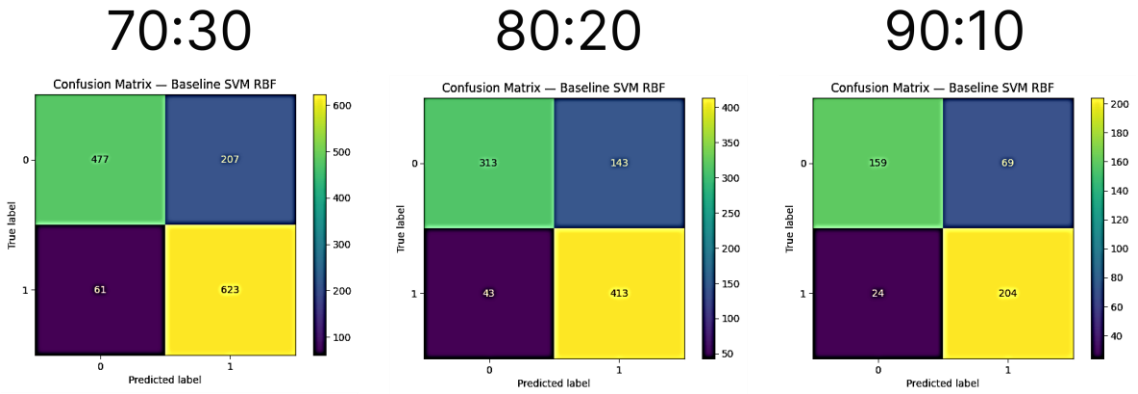
Furthermore, a learning curve analysis was performed to evaluate the model's behavior with respect to varying training set sizes. Learning curve shows the training and validation metric. This analysis helps identify potential overfitting or underfitting by comparing training and validation performance trends as the training size increases.

Results and Discussions

The feature engineering extracted the numerical data from image data using MFM which resulting in 10,754 entries of BP data and 2,279 entries of ST data. There were four images in each of the BP and ST folders where the face could not be detected by MFM, resulting in a total of eight undetected images, as shown in Figure 3.

Figure 3. Undetected facial images

After label encoding, undersampling was performed by reducing the number of BP entries to match the number of ST entries, resulting in 2,279 samples for each class. Next, the "file" column was removed before the data were split into the 70:30, 80:20, and 90:10 ratios as shown in Table 3. The SVM model was trained using the training set, and its performance was evaluated on the unseen testing set. The confusion matrix presented in Figure 4 summarizes the classification results obtained from the testing data, thereby reflecting the model's generalization performance on unseen samples. The ratio-specific results in Table 5 and Table 6 are also provided for further analysis.



(a) (b) (c)

Figure 4. Confusion matrix for 70:30 ratio (a), 80:20 ratio (b), and 90:10 ratio (c)

Table 3. Classification report 70:30 ratio

	precision	recall	F1-score	support
0	0.89	0.70	0.78	684
1	0.75	0.91	0.82	684
Accuracy			0.80	1368
avg Macros	0.82	0.80	0.80	1368
Weighted avg	0.82	0.80	0.80	1368

Table 4. Classification report 80:20 ratio

	precision	recall	F1-score	support
0	0.88	0.69	0.77	456
1	0.74	0.91	0.82	456
Accuracy			0.80	912
avg Macros	0.81	0.80	0.79	912
Weighted avg	0.81	0.80	0.79	912

Table 5. Classification report 90:10 ratio

	precision	recall	F1-score	support
0	0.87	0.70	0.77	228
1	0.75	0.89	0.81	228
Accuracy			0.80	456
avg Macros	0.81	0.80	0.79	456
Weighted avg	0.81	0.80	0.79	456

The 70:30 ratio results show that the baseline model performs well in detecting the positive class as indicated by its high recall value. However, the lower precision suggests that the model still misclassifies some negative (Stroke) samples as positive (Bell's Palsy). This aligns with recent clinical findings showing that around 76% of acute ischemic stroke patients with central facial palsy exhibit upper facial weakness, including the eyebrows [19]. This overlap in asymmetric patterns between Bell's Palsy and certain Stroke cases may reduce class separability, contributing to false positive predictions and thereby affecting precision, while maintaining relatively high recall for detecting facial paralysis. As a result, facial asymmetry in Stroke is not limited to the lower face but also involves the eye and eyebrow regions, making it difficult to distinguish Stroke from Bell's Palsy using static facial images alone. The 80:20 baseline results indicate similar detection ability for the positive class, with the same recall value of 0.91. But its lower precision in positive class and lower recall in negative class makes it less favourable than the 70:30 baseline model. The 90:10 baseline results also show inferior performance compared to the 70:30 ratio, as its accuracy matches that of the 80:20 baseline model, and both its precision and recall are lower than the values obtained with the 80:20 ratio.

Within medical contexts, sensitivity (recall) is a critical metric, as its value can significantly affect treatment outcomes [20]. Based on this consideration, the 70:30 ratio provides the best performance, as it achieves the highest recall among the three ratios. The 80:20 and 90:10 ratios show lower recall values, making them less suitable for medical classification where missing positive cases must be minimized.

Although the confusion matrix provides detailed class-level performance on the testing set, it reflects only a single train–test partition. To ensure that the observed performance is not dependent on a particular data split and to further assess model stability, stratified 5-fold cross-validation was conducted on the training data for each ratio.

Table 6 presents the comparative performance of the proposed SVM model under three different data-splitting scenarios. The 70:30 configuration demonstrates the smallest generalization gap and the highest testing accuracy, indicating the most stable generalization performance. Although the 80:20 configuration yields a slightly higher mean accuracy, its larger standard deviation (0.0239) suggests greater performance fluctuation across folds. Furthermore, the testing accuracy of the 70:30 split is closely aligned with its cross-validation mean, indicating minimal generalization gap and reduced overfitting risk.

Table 6. Performance comparison across data splitting ratios

Split Ratio	CV Mean Accuracy	CV Std Dev	Test Accuracy	Generalization Gap
70:30	0.8072	0.0069	0.8041	0.0031
80:20	0.8110	0.02389	0.7961	0.0149
90:10	0.8084	0.0085	0.7961	0.0123

In addition to cross-validation, learning curve analysis was performed to examine the model's behavior as the number of training samples increased, as presented in Figure 5. For the 70:30 split, both training and validation accuracies increased gradually and converged with a consistently small gap, indicating stable learning and minimal overfitting. The 80:20 split achieved a slightly higher final validation accuracy; however, its training curve showed greater fluctuations, suggesting higher variance. Meanwhile, the 90:10 split exhibited less stable convergence at larger training sizes, with minor performance fluctuations toward the end. Overall, Figure 5 confirms that the 70:30 configuration provides the most stable and balanced generalization performance among the three ratios.

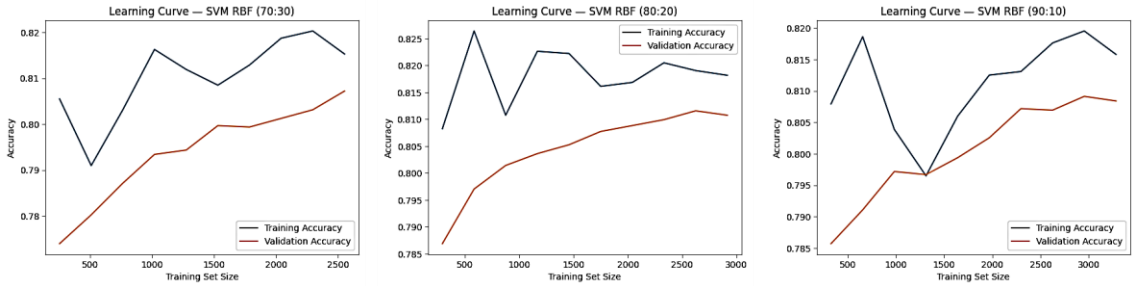


Figure 5 Learning Curve for 70:30 Ratio (a), 80:20 Ratio (b), and 90:10 Ratio (c)

The combined results from cross-validation and learning curve analysis indicate that the proposed SVM model does not exhibit significant overfitting or underfitting. For the 70:30 configuration, the mean cross-validation accuracy was 0.8072 with a low standard deviation of 0.0069, while the corresponding testing accuracy was 0.8041, resulting in a minimal generalization gap of approximately 0.0031. Similarly, the 80:20 and 90:10 splits produced cross-validation means of 0.8110 and 0.8084, with testing accuracies of 0.7961 for both configurations, yielding moderate gaps of 0.0149 and 0.0123, respectively. In the learning curve analysis, the gap between training and validation accuracies remained consistently small (approximately 0.01–0.02) across all splits, with validation performance stabilizing around 0.80–0.81. These findings demonstrate that the model achieves a balanced bias–variance trade-off and maintains stable generalization performance, particularly under the 70:30 configuration.

Conclusion

The proposed SVM-based framework successfully differentiates Bell's Palsy and Stroke using asymmetric facial numeric features extracted from static images, achieving stable generalization performance under limited data and computational constraints. This study proposed an explainable classification framework for differentiating Bell's Palsy (BP) and Stroke (ST) based on asymmetric facial numeric features extracted using MediaPipe Face Mesh. The extracted coordinates were transformed into numerical features through Min–Max scaling, Euclidean distance, and angle computation. The model avoids direct reliance on raw facial appearance while enabling interpretable feature representation.

To address class imbalance in the collected datasets, undersampling was applied to construct a balanced dataset. Among the three evaluated splitting ratios (70:30, 80:20, and

90:10), the 70:30 configuration demonstrated the most stable and reliable performance, achieving a testing accuracy of 0.8041 and a cross-validation mean accuracy of 0.8072 ± 0.0069, with a minimal generalization gap of approximately 0.0031. Learning curve analysis further confirmed stable convergence with a small training–validation gap (≈0.01–0.02), indicating a balanced bias–variance trade-off and minimal overfitting.

From a clinical perspective, the 70:30 configuration is preferable due to its higher recall, which is critical in medical screening scenarios where minimizing missed positive cases is essential. The 80:20 and 90:10 ratios yielded lower recall values, making them less suitable for this diagnostic task. Overall, the findings demonstrate that asymmetric facial geometric features combined with an SVM-RBF classifier can effectively distinguish Bell's Palsy from Stroke under limited dataset availability and hardware constraints. Rather than replacing deep learning approaches, this framework provides an interpretable and resource-efficient alternative suitable for environments where large-scale annotated datasets are unavailable.

Future research may focus on expanding available Bell's Palsy and Stroke datasets to improve data diversity and generalization performance. Incorporating temporal information from video-based facial sequences could enable analysis of dynamic muscle movements rather than relying solely on static images. Additionally, future studies may explore hybrid feature representations that combine geometric asymmetry features with other discriminative descriptors, conduct controlled comparisons with lightweight deep learning architectures, and evaluate alternative SVM kernel configurations to further optimize classification performance. Further analytical extensions, such as quantitative feature contribution analysis (e.g., correlation analysis or feature ranking), may also be performed to provide deeper insight into the relative importance of individual asymmetric features. Expanding the framework to include a healthy control class may also enable more comprehensive multi-class diagnostic modelling.

References

[1] S. J. Murphy and D. J. Werring, "Stroke: causes and clinical features," *Medicine*, vol. 48, no. 9, pp. 561–566, 2020, doi: 10.1016/j.mpmed.2020.06.002.

[2] S. Umirzakova, S. Ahmad, S. Mardieva, S. Muksimova, and T. K. Whangbo, "Deep learning-driven diagnosis: A multi-task approach for segmenting stroke and Bell's palsy," *Pattern Recognit.*, vol. 144, Dec. 2023, doi: 10.1016/j.patcog.2023.109866.

[3] J. Rajangam *et al.*, "Bell Palsy: Facts and Current Research Perspectives," *CNS Neurol. Disord. Drug Targets*, vol. 23, no. 2, pp. 203–214, Mar. 2023, doi: 10.2174/1871527322666230321120618.

[4] B. Boodale, M. Amin, K. Sabetian, D. Quesada, and T. Torrico, "Medial Pontomedullary Stroke Mimicking Severe Bell's Palsy: A Case Report," *Clin. Pract. Cases Emerg. Med.*, vol. 4, no. 3, pp. 380–383, Jul. 2020, doi: 10.5811/cpcem.2020.5.46965.

[5] M. Pohl *et al.*, "Ischemic stroke mimics: A comprehensive review," *Journal of Clinical Neuroscience*, vol. 93, pp. 174–182, Nov. 2021, doi: 10.1016/j.jocn.2021.09.025.

[6] P. N. A. Darmawan, N. M. D. Pratiwi, and I. K. Arimbawa, "Characteristic of Bell's Palsy in Clinical Neurologic at Sanglah Hospital Denpasar Bali Indonesia," *International Journal of Research and Review*, vol. 8, no. 12, pp. 318–322, Dec. 2021, doi: 10.52403/ijrr.20211239.

[7] S. Gupta, M. K. Jawanda, N. Taneja, and T. Taneja, "A systematic review of Bell's Palsy as the only major neurological manifestation in COVID-19 patients," *Journal of Clinical Neuroscience*, vol. 90, pp. 284–292, Aug. 2021, doi: 10.1016/j.jocn.2021.06.016.

[8] Y. I. Royan, P. Pramono, and A. A. K. Asri, "Performance Comparison of Convolutional Neural Networks (CNN) and Support Vector Machine (SVM) Algorithms in Human Face Classification," *G-Tech: Jurnal Teknologi Terapan*, vol. 9, no. 3, pp. 1544–1553, Jul. 2025, doi: 10.70609/g-tech.v9i3.7384.

[9] F. Mahardhika, M. L. Haryanti, and P. Hiskiawan, "Performance Evaluation of Speech Emotion Recognition Using Hybrid Feature Selection and Machine Learning," in *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)*, Kota Cirebon, Indonesia: IEEE, Sep. 2025, pp. 1–7. doi: 10.1109/ICCIT65724.2025.11166879.

[10] Lukman Arif Sanjani, R. Bimo Mandala Putra, and U. Laili Yuhana, "Exploring the Application of Machine Learning for Automatic Inbound Email Classification in CRM System at XYZ Company," *Journal of Technology and Informatics (JoTI)*, vol. 6, no. 1, pp. 1–7, Oct. 2024, doi: 10.37802/joti.v6i1.715.
[11] S. Baul, Md. Ratan Rana, N. Jahan Trisna, and F. Bente Alam, "Development of a Real-time Driver's Drowsiness Detection System Using MediaPipe Face Mesh," *International Journal of Engineering and Manufacturing*, vol. 15, no. 5, pp. 46–57, Oct. 2025, doi: 10.5815/ijem.2025.05.04.
[12] J. Jose, K. Raimond, and S. Vincent, "SleepyWheels: An Ensemble Model for Drowsiness Detection leading to Accident Prevention," Nov. 2022. doi: 10.48550/arXiv.2211.00718.
[13] D. Ciraolo, M. Fazio, R. S. Calabrò, M. Villari, and A. Celesti, "Facial expression recognition based on emotional artificial intelligence for tele-rehabilitation," *Biomed. Signal Process. Control*, vol. 92, Jun. 2024, doi: 10.1016/j.bspc.2024.106096.
[14] Jasir, "jasir/palsynet-data." Accessed: Dec. 06, 2025. [Online]. Available: https://huggingface.co/datasets/jasir/palsynet-data
[15] M. Kaitav, "Facial_Droop_and_Facial_Paralysis_image." Accessed: Dec. 06, 2025. [Online]. Available: https://www.kaggle.com/datasets/kaitavmehta/facial-droop-and-facial-paralysis-image
[16] J. Danish, "Face Images of Acute Stroke and Non Acute Stroke." Accessed: Dec. 06, 2025. [Online]. Available: https://www.kaggle.com/datasets/danish003/face-images-of-acute-stroke-and-non-acute-stroke
[17] M. Alfonso and D. Bhisetya Rarasati, "JISA (Jurnal Informatika dan Sains) Sentiment Analysis of 2024 Presidential Candidates Election Using SVM Algorithm," *JISA (Jurnal Informatika dan Sains)*, vol. 6, no. 2, pp. 110–5, Dec. 2023, doi: doi.org/10.31326/jisa.v6i2.1714.
[18] F. Tampinongkol, "Identifikasi Penyakit Daun Tomat Menggunakan Gray Level Co-occurrence Matrix (GLCM) dan Support Vector Machine (SVM)," *Techno Xplore : Jurnal Ilmu Komputer dan Teknologi Informasi*, vol. 8, no. 1, pp. 08–16, Apr. 2023, doi: 10.36805/technoxplore.v8i1.3578.
[19] N. Fariesya Suhaila Md Sazihan, N. Mu, azzah Abdul Latiff, N. Noordini Nik Abd Malik, S. Mashohor, and F. Taha Al-Dhief, "Evaluating the Effectiveness of Parameter Tuning for Support Vector Machine on Voice Pathology Database," *Elektrika*, vol. 24, no. 2, pp. 118–124, 2025, doi: 10.11113/elektrika.v24n2.628.
[20] D. Asmawati, L. Arif Sanjani, C. Dimas Renggana, C. Fatichah, and T. Mustaqim, "Arrhythmia Classification with ECG Signal using Extreme Gradient Boosting (XGBoost) Algorithm," *Journal of Technology and Informatics (JoTI)*, vol. 6, no. 1, pp. 36–42, Oct. 2024, doi: 10.37802/joti.v6i1.792.

Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 16 Issue 1 Year 2026 Pages 39-48
DOI: http://dx.doi.org/10.31940/matrix.v16i1.39-48
URL: http://ojs2.pnb.ac.id/index.php/matrix

User experience testing on Smart Human Capital Dashboard (SHUCADA) from PT Studio Kami Mandiri using User Experience Questionnaire (UEQ)

I Made Gede Sunia Pradnyantara [1]*, I Wayan Agus Budiarsana [2], I Made Agus Oka Gunawan [3], Gede Indrawan [4]

[1,2,4] *Master Program in Computer Science, Universitas Pendidikan Ganesha, Indonesia*
[3] *Informatics Management Study Program, Politeknik Negeri Bali, Indonesia*

***Corresponding Author: *sunia.pradnyantara.sp@gmail.com*

Abstract: PT Studio Kami Mandiri is a company specialized in software development. PT Studio Kami Mandiri has developed several products or applications that have also been used by large companies. One of the products developed is the Smart Human Capital Dashboard (SHUCADA) which is used to monitor the performance of an employee in a company. User experience testing is needed to find out the possible obstacles that users will face when using the application. Therefore, this research conducted user experience testing on SHUCADA to find out employee perceptions of the application. User Experience Testing of SHUCADA using the User Experience Questionnaire (UEQ) Data Analysis Tool found that aspects of attractiveness with mean value 1.21, dependability with mean value 1.208, and stimulation with mean value 1.125 get above average scores, and aspects of efficiency (mean 1.583) get good scores. The perspicuity aspect (mean 0.604) received a bad score and the novelty aspect (mean 0.604) received a below average score.

Keywords: Data analysis, user experience questionnaire, user experience testing

History Article: Submitted 7 November 2024 | Revised 8 February 2026 | Accepted 17 February 2026

How to Cite: I. M. G. S. Pradnyantara, I. W. A. Budiarsana, I. M. A. O. Gunawan, and G. Indrawan, "User experience testing on Smart Human Capital Dashboard (SHUCADA) from PT Studio Kami Mandiri using User Experience Questionnaire (UEQ)", *Matrix : Jurnal Manajemen Teknologi dan Informatika*, vol. 16, no. 1, pp. 39–48, 2026, doi: 10.31940/matrix.v16i1.39-48.

Introduction

Employee performance is an important factor in the success and growth of a company. Employee performance refers to how well an employee achieves their work objectives and the quality of their work within their role. Employee performance includes not only final results, but also other elements such as cooperation, productivity, attendance, overall quality of results, and quality of work. The progress of a country can be seen from the development and quality of its human resources. The better the development of human resources in a country, the better the development and quality of human resources in that country. The development of the quality of human resources in a country can be seen through the Human Development Index (HDI). Based on data from the Human Development Index (HDI) Indonesia occupies the 6th position in Southeast Asia in the last five years [1].

Based on these problems, in an effort to improve the development of the quality of human resources in Indonesia, one of the companies in Bali engaged in IT created an innovative application that can help monitor the performance of human resources. The application is called Smart Human Capital Dashboard (SHUCADA) developed by PT Studio Kami Mandiri. SHUCADA is a human resource development application or application that can monitor the performance of an employee in a company. This Shucada application is equipped with attendance, work order, key performance indicator (KPI), dashboard, and report features.

Information technology is developing at an extremely fast rate each year. The shift in people's lifestyles that use technology as a tool to make everyday tasks more efficient is one of the advancements in information technology. The development of modern technologies is also

influenced by the internet. Businesses from a variety of industries have introduced mobile applications through the internet to help people in the age of digitization [2].

Research on User Experience Testing at SHUCADA has never been done, despite the fact that literature reviews have been done. Consequently, the business has been unable to assess each employee's performance in a thorough and organized way. This typically occurs when consumers are unable to comprehend or use an application that has multiple menus, which typically leaves users feeling a little perplexed.

One technique to determine whether users are content with the program, whether it is easy for them to use, and how well it can assist them in reaching their objectives is through user experience testing [3]. To determine the potential challenges that users may encounter when utilizing the program, user experience testing is required. In order to address this gap, this study tests SHUCADA's user experience to ascertain how employees feel about the program [4]. Questionnaires and Experience Questionnaire (UEQ) tools were used to test SHUCADA's user experience. Through this approach, this research is expected to provide theoretical contributions in the development of website quality evaluation methods, as well as practical contributions in providing guidance for website managers in designing service improvement strategies based on data and user perceptions.

UEQ is believed to provide further advantages due to its ability to generate comprehensive user experience measurement results [5]. The purpose of this study is to use UEQ to analyze and evaluate the user experience of the SHUCADA application. This study also attempts to provide suggestions and advice for improving the look of the SHUCADA application based on the results of the UEQ evaluation.

Methodology

The research was carried out in a number of phases, including problem analysis, literature review, respondent selection, questionnaire distribution, UEQ measurement, analysis of research results, conclusion, and recommendations [6]. Figure 1 displays the research methodology.

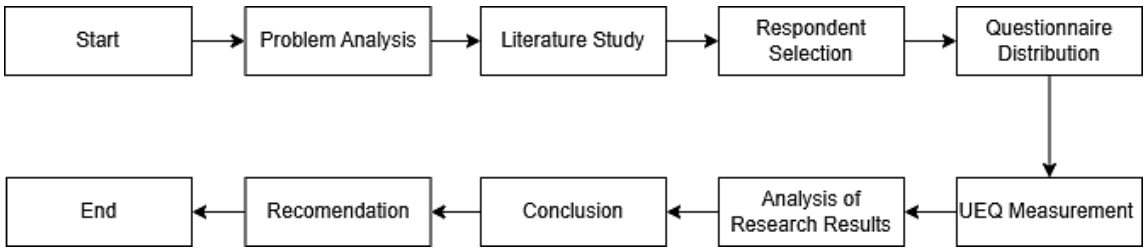


Figure 1. Research methodology

The first step in this research was to identify the problem, which was that new employees at PT Studio Kami Mandiri had difficulty using the SHUCADA application. Therefore, user experience testing was needed to identify the problem, followed by a literature study to find suitable references for the research. After that, respondents were selected to participate in the user experience testing. There were 12 employees from PT Studio Kami Mandiri who participated. Table 1 shows the details of each research participant. The evaluation process using UEQ begins with respondents being asked to use all features available on SHUCADA. The aim is for users to provide good feedback and accurate UEQ results. Users can fill out the UEQ based on their impressions when using the website, including its functionality, colors, font types, layout, and other aspects. Before filling out the questionnaire, users are first instructed to use SHUCADA, then they are instructed to fill out the questionnaire using the form that has been distributed to them. The UEQ will be filled out in accordance with the UEQ guidelines.

Table 1. Respondent's details

Department	Position	Status	Total Employees
Logic	Programmer Staff	Contract	6
Editorial	Content and Media Staff	Contract	2
Office	Marketing Staff	Contract	2
	Manager	Regular	1
	Administration Staff	Regular	1

Following the collection of all questionnaire responses, the data must be processed in order to provide context for the findings. Data from questionnaires must pass through a number of steps, including descriptive analysis, validity and reliability testing, and computer data entry [11]. Then, the user experience measurement of the SHUCADA application is using the User Experience Questionnaire (UEQ). Because the UEQ approach is quick and useful for interacting directly with the application user experience, researchers use it as the foundation for their studies [7]. UEQ tool analysis can be used to process UEQ data [12]. The UEQ measurement findings can serve as a guide to enhance the user interface's quality [8]. There are six scales out of a total of 26 statement items based on the measurement scale included in the UEQ. The six UEQ scales can be categorized into three groups: hedonic quality, pragmatic quality, and attractiveness [9].

Table 2. Item in UEQ

No	Left	Right	Scale
1	Annoying	Enjoyable	Attractiveness
2	Not understandable	Understandable	Perspicuity
3	Dull	Creative	Novelty
4	Difficult to learn	Easy to learn	Perspicuity
5	Inferior	Valuable	Stimulation
6	Boring	Exciting	Stimulation
7	Not Interesting	Interesting	Stimulation
8	Unpredictable	Predictable	Dependability
9	Slow	Fast	Efficiency
10	Conventional	Inventive	Novelty
11	Obstructive	Supportive	Dependability
12	Bad	Good	Attractiveness
13	Complicated	easy	Perspicuity
14	Unlikeable	Pleasing	Attractiveness
15	Usual	Leading edge	Novelty
16	Unpleasant	Pleasant	Attractiveness
17	Not secure	secure	Dependability
18	Demotivating	Motivating	Stimulation
19	Does not meet expectations	Meets expectations	Dependability
20	Inefficient	Efficient	Efficiency
21	Confusing	Clear	Perspicuity
22	Impractical	Practical	Efficiency
23	Cluttered	Organized	Efficiency
24	Unattractive	attractive	Attractiveness
25	Unfriendly	Friendly	Attractiveness
26	Conservative	Innovative	Novelty

ANOVA was selected because it allows the comparison of mean values across multiple UEQ dimensions to determine whether lower-scoring aspects specifically Perspicuity and Novelty differ significantly from higher-scoring UX elements. This helps identify whether these weaknesses are statistically meaningful and therefore require targeted design improvements. Without ANOVA, differences between dimensions would rely solely on descriptive interpretation without statistical

confirmation. Before performing ANOVA, several statistical assumptions were evaluated to ensure the analysis was valid. First, the assumption of independence was satisfied because each respondent completed the UEQ individually. Second, the normality assumption was checked through inspection of the transformed UEQ data, which showed an approximately normal distribution. Third, homogeneity of variance was assumed based on similar variance patterns across UEQ dimensions. Prior UEQ literature supports treating transformed UEQ scores (–3 to +3) as interval data suitable for parametric analysis.

The pragmatic quality component is concerned with usability, efficacy, and perceived benefits. Dependability, effectiveness, and perspicuity are all included in the pragmatic quality component. The hedonic quality component is associated with novelty and stimulation [10]. Table 2 displays the 26 statement items of the UEQ.

Finding the average of each aspect value is how data analysis is done. There are positive and negative values for every item. A scale of 1 to 7, which will subsequently be translated into values from the range of -3 to +3, makes up the negative and positive components of UEQ . It can be said that the results are of good quality if the answer value is between +1 and +2 [15]. According to the UEQ evaluation criteria, a value is deemed negative if it is less than -0.8, neutral if it is between -0.8 and 0.8, and positive if it is greater than 0.8 [11].

Table 3. Benchmark interval

	Attractiveness	**Perspicuity**	**Efficiency**	**Dependability**	**Stimulation**	**Novelty**
Excellent	≥ 1.75	≥ 1.9	≥ 1.78	≥ 1.65	≥ 1.55	≥ 1.44
Good	≥ 1.52 < 1.75	≥ 1.56 < 1.9	≥ 1.47 < 1.78	≥ 1.48 < 1.65	≥ 1.31 < 1.55	≥ 1.05 < 1.4
Above Average	≥ 1.17 < 1.52	≥ 1.08 < 1.56	≥ 0.98 < 1.47	≥ 1.14 < 1.48	≥ 0.99 < 1.31	≥ 0.71 < 1.05
Below Average	≥ 0.7 < 1.17	≥ 0.64 < 1.08	≥ 0.54 < 0.98	≥ 0.78 < 1.14	≥ 0.5 < 0.99	≥ 0.3 < 0.71
Bad	< 0.7	< 0.64	< 0.54	< 0.78	< 0.5	< 0.3

A benchmark test must also be finished in order to compare the value findings on each element in the UEQ Data Analysis Tool. Benchmark test scores are classified into five categories: Bad, Average, Below, Above Average, Good, and Excellent [13]. The processed data was then analysed to obtain user experience information from the SHUCADA application. The results of the analysis are then used to draw conclusions and generate appropriate recommendations for improvements to the SHUCADA application.

Results and Discussions

Results

User Experience (UX) testing on the Smart Human Capital Dashboard (SHUCADA) was carried out by distributing questionnaires via Google Form to respondents. Collected 12 respondents from all employees of PT Studio Kami Mandiri who have used the application and filled out the UEQ questionnaire honestly. The data obtained from the questionnaire were entered into the UEQ data analysis table, then transformed by calculating the values obtained from the UEQ questionnaire minus 4 to obtain positive or negative values for each UEQ item. A value of +3 is the highest positive value and -3 is the lowest negative value. Figure 2 displays the data processing outcomes as well as the average attribute of every respondent.

Items																						
3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
0	1	1	1	2	0	1	1	1	2	0	1	1	1	0	2	1	2	0	1	1	1	1	2
-2	3	3	0	2	2	0	-2	2	2	-3	2	2	2	2	2	2	2	2	2	2	2	2	2
0	0	1	-1	-2	-1	-1	1	1	1	-3	0	0	-2	0	-1	-2	1	-2	0	0	0	1	-1
1	-2	2	0	1	0	2	1	3	2	-1	1	0	1	2	1	1	2	-1	1	2	2	0	1
1	2	2	1	1	1	2	1	2	2	1	1	-1	1	2	1	1	2	2	1	2	1	0	1
2	2	1	1	1	0	1	1	1	2	0	1	-1	1	2	1	1	1	1	1	3	1	1	1
0	1	2	1	1	1	1	1	2	1	-1	0	-1	0	1	0	2	1	1	1	3	2	0	1
2	-1	2	0	2	1	2	2	1	2	-1	0	-1	2	3	1	1	2	2	3	3	3	2	1
1	-1	1	1	2	0	2	2	2	1	-1	1	-1	1	2	0	1	1	2	0	2	1	-1	1
1	-1	1	1	1	0	1	2	1	1	0	1	-1	2	2	1	0	0	1	1	3	2	1	1
3	2	2	1	1	1	3	3	2	3	2	2	-3	3	2	2	2	2	3	3	3	3	1	1
1	1	2	2	3	2	3	2	2	3	2	2	-2	2	2	1	1	3	2	1	1	3	1	1

Figure 2. Transformed data

Figure 2 presents the UEQ transformed scores ranging from −3 to +3 for all 26 items. Positive values indicate favorable user impressions, while negative values indicate unfavorable impressions. The figure visually highlights which items received consistently positive ratings and which items showed variability among respondents. Each aspect's average is generated in the Results table. Reliability testing of the absolute question items is required prior to statistical analysis. The reliability of the questionnaire data was examined using the Cornbach alpha coefficient. The Cronbach alpha coefficient shows that all UX factor scales are consistent. Table 4 shows the Cronbach alpha values for each UEQ component.

Table 4. Cronbach's Alpha values for all UEQ aspects

	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Value	0.94	0.90	0.87	0.92	0.90	0.89

Table 4 shows the Cronbach's alpha values (>0.7) for each admissible UEQ component. These consist of novelty (0.89), dependability (0.92), efficiency (0.87), perspicuity (0.90), Attractiveness (0.94), and stimulation (0.90). According to research conducted by Henim and Sari [3], if the test results show a Cronbach's alpha value of >0.7 for all aspects, this proves that the analysed data is accurate. Table 4 reports the internal reliability of each UEQ scale. All Cronbach's Alpha values exceed 0.87, indicating excellent internal consistency across the six UX aspects. Hedonistic quality, pragmatic quality, and attractiveness are the three main facets of UEQ [3]. The scale values for the three aspects are shown in Table 5.

Table 5. Detailed average scores of each UEQ aspect

Aspects	Value	UEQ Aspects	UEQ Scale Value
Attractiveness	1.21	Attractiveness	1.208
Pragmatic Quality	1.13	Perspicuity	0.604
		Efficiency	1.583
		Dependability	1.208
Hedonic Quality	0.86	Stimulation	1.125
		Novelty	0.604

Annoying/enjoyable, good/bad, unlikeable/pleasing, unpleasant/pleasant, attractive/unattractive, friendly/unfriendly are some of the factors that make up the attractiveness scale. With an average score of 1.21 for this attractiveness. The pragmatic quality factor has an average rating of 1.13. According to the pragmatic quality aspect, which includes the clarity aspect with an average score of 0.604. Additionally, this application has an average dependability score of 1.208. The average value for the efficiency component is 1.583. Hedonic quality is a component of user feelings about a product that influences innovation or new product design, as well as stimulation or motivating pleasure. At 0.86. Hedonic quality is a component of user feelings about a product that influences innovation or new product design, as well as stimulation or motivating pleasure. At 0.86, with an average value of 1.125 for the stimulation and 0.604 for the novelty component of the hedonic quality component hedonic quality has the lowest average.

Table 5 lists the mean scores for all UEQ dimensions categorized into Attractiveness, Pragmatic Quality (Perspicuity, Efficiency, Dependability), and Hedonic Quality (Stimulation, Novelty). The values indicate that Efficiency received the highest positive evaluation, while Perspicuity and Novelty scored the lowest.

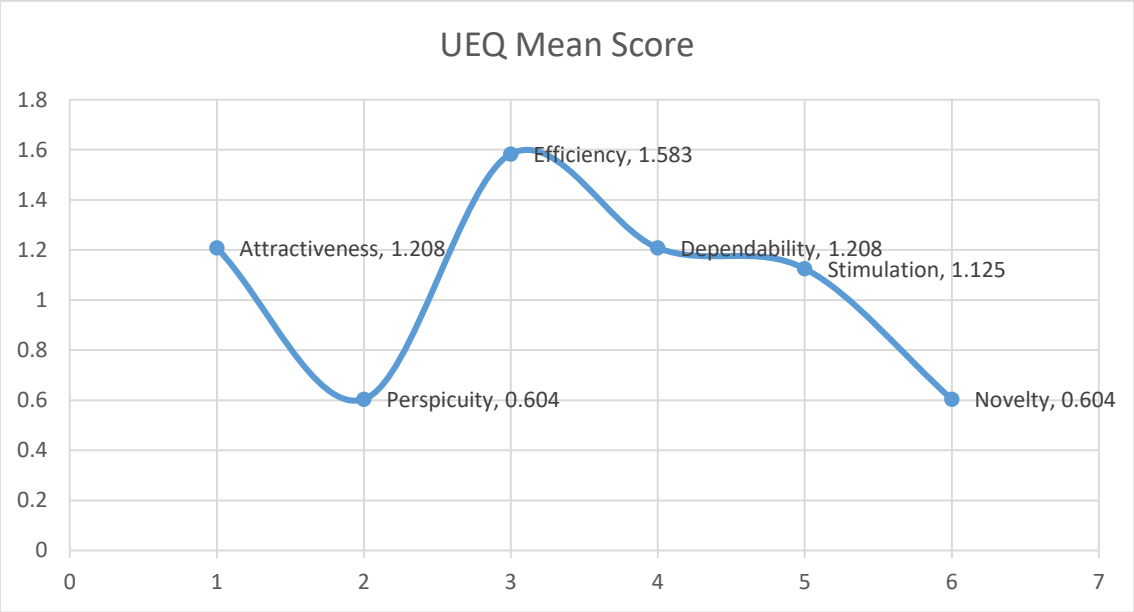


Figure 3. UEQ chart

The UEQ curve in Figure 3 reveals a product with strong functional performance but critical user adoption barriers. The dominant peak at Efficiency (1.583 - Good) confirms the application is highly optimized for fast task completion, fulfilling its pragmatic role effectively. However, the system's utility is immediately undermined by its dual low scores on Perspicuity (0.604 - Bad) and Novelty (0.604 - Below Average). Low Perspicuity is a major red flag, indicating that the interface is difficult to learn and navigate, suggesting a complex information architecture that hampers new user onboarding. Coupled with low Novelty, the application is perceived as uninspiring and outdated. While the mid-range scores for Attractiveness and Stimulation are acceptable, they are insufficient to compensate for the fundamental flaws in clarity and innovativeness, making the application fast but fundamentally unfriendly in its current state.

Table 6. ANOVA between UEQ aspects

Source of Variation	SS	df	MS	F	P-value	F crit
Perspicuity vs Efficiency	5.752604	1	5.752604	7.242623	0.013339	4.30095
Novelty vs Attractiveness	2.190104	1	2.190104	6.243645	0.020424	4.30095
Novelty vs Dependability	2.190104	1	2.190104	7.232995	0.013393	4.30095
Novelty vs Efficiency	5.752604	1	5.752604	18.82184	0.000264	4.30095
Novelty vs Stimulation	1.627604	1	1.627604	5.513232	0.028274	4.30095

Table 6 shows the results of the ANOVA analysis between two aspects of UEQ, namely Perspicuity and Novelty. These two aspects were chosen as comparators because they have lower UEQ Scale Values than other aspects, making it important to further analyze their relationship with other aspects of user experience. The analysis was conducted on five sources of variation, namely Perspicuity vs. Efficiency, Novelty vs. Attractiveness, Novelty vs. Dependability, Novelty vs. Efficiency, and Novelty vs. Stimulation. These five pairs were selected because the p-values obtained were less than 0.05, indicating a significant difference between the compared aspect

pairs. This table presents ANOVA comparisons between low-performing aspects (Perspicuity, Novelty) and other UX dimensions. All pairs listed show p-values < 0.05, indicating statistically significant differences.

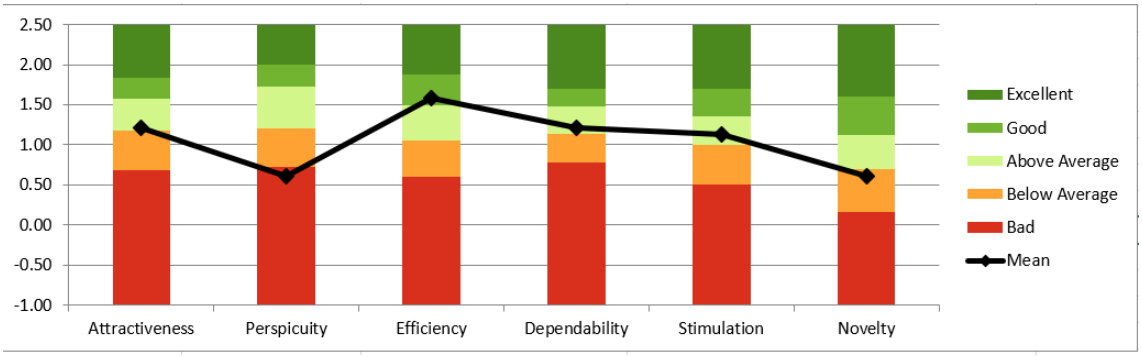


Figure 4. Benchmark result

In Figure 4, we can clearly see each section ranging from bad to excellent and average. Based on the benchmark results, attractiveness, dependability, and stimulation got above average results. Efficiency gets good results. This benchmark visualization compares SHUCADA's UEQ results to international benchmark categories (Bad, Below Average, Above Average, Good, Excellent). Efficiency reaches the Good category, while Attractiveness, Dependability, and Stimulation fall within Above Average. Perspicuity and Novelty fall into Below Average categories.

Discussions

This application is more attractive than usual, it is because this application is an application that is quite fun and encouraging for employees to use, then employees have a good impression of the SHUCADA application because employees feel comfortable when using the application, this application is also attractive and quite friendly to employees when used and this is in accordance with research conducted by Henim and Sari [3], if the attractiveness value has a positive impression, the system can be said to be attractive and comfortable for users. The results align with User Experience theory, which divides UX into pragmatic quality (task-related aspects such as efficiency and perspicuity) and hedonic quality (emotional aspects such as stimulation and novelty). SHUCADA displays strong pragmatic performance in efficiency but shows weakness in perspicuity. Hedonic quality also appears limited due to low novelty, indicating that the system may not feel modern or innovative to users. This theoretical interpretation supports the findings of the statistical analysis.

According to the pragmatic quality aspect, which includes the perspicuity aspect, this program is straightforward to use and clear. However, it offers features and menus that are difficult to grasp. In addition, this app has a good dependability score, indicating that it is safe to use and effective at supporting business operations. It can also meet a company's expectations as an employee management application and in accordance with research conducted by Pangestu et al. [14], if the results of the dependability aspect get a good score, the system can be said to be able to provide support in terms of usage control and a sense of security when users operate the system. The score of the efficiency aspect gets a good score, because only internal company employees use this application, it responds quickly when used to fill out work orders or face attendance forms. It is regarded as efficient, practical, and well-organized because it is simple to use, saves time when monitoring work, and groups its features according to their specific functions. This is in accordance with research conducted by Henim and Sari [3], if the efficiency aspect shows a positive impression then this shows that users can complete tasks quickly when using the system.

With good scores for the stimulation component of the hedonic quality component, this application is beneficial and engaging for businesses to track employee activity in all areas, as well as increasing employee motivation to be more productive in the company because of the help of this application and in accordance with research conducted by Henim and Sari [3], if the score from the stimulation aspect is good, the system is useful for users and motivates users to

use it. However, the novelty aspect scores below average, indicating that the application lacks innovative and modern features, according to research conducted by Fatmawati et al. [16], if the novelty aspect score is below average, users will feel that the application is uninteresting and does not provide a satisfying experience in the long term, because the novelty aspect is related to how the application provides a unique, innovative, and interesting experience for its users. Therefore, the appropriate recommendation for improving the SHUCADA application in order to increase its novelty aspect score is to improve navigation by reorganizing the navigation structure to make it more accessible and user-friendly, and to adopt the latest design trends such as the use of smooth animations and transitions [16].

The ANOVA results confirm that the lower-scoring aspects such as Perspicuity and Novelty are significantly different from the other UX dimensions, indicating that their weaknesses are statistically meaningful rather than the result of random variation. The significant difference between Perspicuity and Efficiency ($p=0.013$) shows that a simple interface does not necessarily lead to faster task completion. Likewise, the significant difference between Novelty and Attractiveness ($p=0.020$) indicates that a system may be visually appealing yet still feel outdated or lacking innovation. The strongest difference found between Novelty and Efficiency ($p=0.00026$) further reinforces that a highly efficient system does not automatically incorporate new or innovative design elements. These findings highlight the priority areas for UX improvement, particularly the need to simplify menus and interface structures to enhance Perspicuity, and to introduce more modern, creative, and up-to-date design elements to improve the Novelty aspect. The ANOVA results show that the significant difference between the aspects of Perspicuity and Efficiency indicates that even though a simple and easy-to-understand interface does not necessarily enable users to complete tasks quickly. The analysis results between the aspects of Novelty and Attractiveness show that the novelty of a system is not automatically considered attractive by users. In the Novelty and Dependability pair, a significant difference emerged because the novelty presented by the system is often accompanied by instability or inconsistency in its functions. In the Novelty and Efficiency pair, where the p-value is very small, it means that innovative systems are not necessarily practical or quick to use, because users need time to understand new functions. Significant differences between Novelty and Stimulation indicate that even though a system is considered new, it does not always provide a pleasant or motivating experience for users.

Therefore, this research makes recommendations to improve the user experience aspect of SHUCADA application based on the result of UEQ. Since they continue to receive below-average and poor benchmark scores, as well as user feedback indicating that this application requires new innovations and feature simplification to make it easier for new hires to learn. The results of this study will be given to PT Studio Kami Mandiri as reference material for them to develop the SHUCADA system.

Conclusion

The results of data processing using descriptive statistics on each attribute that measures user experience with UEQ show that the SHUCADA application gets a neutral evaluation value on the perspicuity aspect with an average of 0.604 and the novelty aspect with an average of 0.604, while getting a positive evaluation value on the attractiveness aspect with an average of 1.21, the efficiency aspect with an average of 1.583, the dependability aspect with an average of 1.208, and the stimulation aspect with an average of 1.125. features of attractiveness, dependability, and stimulation are known to receive above-average scores, while features of efficiency receive good scores, according to benchmark values derived by each UEQ component utilizing the UEQ Data Analysis Tool, it indicates that the application is very efficient in application response, practical and organized features in the application so that it can streamline employee time. However, further evaluation is still needed because it still has some shortcomings in the novelty aspect which gets below average results and the perspicuity aspect which gets bad results. This indicates that this application needs new innovations or creativity in maintaining the lifetime of this application, and it is necessary to pay attention to this application again so that it is not difficult to learn for new employees later because of its complicated features and menus. Therefore, this research makes recommendations to improve the user experience aspect of SHUCADA application based on the result of UEQ. Since they continue to receive below-average

and poor benchmark scores, as well as user feedback indicating that this application requires new innovations and feature simplification to make it easier for new hires to learn, novelty and perspicuity are the suggested areas for improving user experience.

Acknowledgments

This research was made possible by the support and collaboration of various individuals and organizations. First of all, we would like to thank Universitas Pendidikan Ganesha and PT Studio Kami Mandiri for providing the necessary resources and guidance during this research, and also to our supervisors, I Made Agus Oka Gunawan, and Gede Indrawan for their invaluable input, encouragement, and expertise, which has improved the quality of this research. Special thanks to all the participants who contributed their time and insights through surveys and interviews, as their input was crucial to the analysis and findings presented in this paper, and finally, we are grateful for the support of our family and friends for their patience, motivation, and understanding throughout the entire research process.

References

[1] D. Azfirmawarman, L. Magriast, and Yulhendri, "Indeks pembangunan manusia di Indonesia (Kajian perubahan metodologi penghitungan)," *J. Pendidik. dan Konseling*, vol. 5, no. 5, pp. 117–125, 2023.

[2] J. S. Veron, C. H. Primasari, Y. P. Wibisono, T. A. P. Sidhi, and D. B. Setyohadi, "Analisis User Experience (UX) aplikasi virtual reality Gamelan Bonang Barung menggunakan User Experience Questionnaire (UEQ)," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 3, no. 1, pp. 130–141, 2023, doi: 10.24002/konstelasi.v3i1.6626.

[3] S. R. Henim and R. P. Sari, "Evaluasi user experience sistem informasi akademik mahasiswa pada perguruan tinggi menggunakan user experience questionnaire," *J. Komput. Terap.*, vol. 6, no. 1, pp. 69–78, 2020, [Online]. Available: https://jurnal.pcr.ac.id/index.php/jkt/.

[4] T. J. Maulani, Suprapto, and A. R. Perdanakusuma, "Evaluasi user experience menggunakan metode usability testing dan User Experience Questionnaire (UEQ) (Studi kasus: Website Superprof.co.id dan Zonaprivat.com)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 6, pp. 2639–2645, 2021, [Online]. Available: http://j-ptiik.ub.ac.id.

[5] S. Y. R. Marpaung and N. Nuraeni, "Evaluasi user experience website E-Learning My-Elnusa menggunakan User Experience Questionnaire (UEQ)," *Swabumi*, vol. 11, no. 1, pp. 78–84, 2023, doi: 10.31294/swabumi.v11i1.15354.

[6] Y. Wijayanti, S. Suyoto, and A. T. Hidayat, "Evaluasi pengalaman pengguna pada aplikasi seluler Visiting Jogja menggunakan metode User Experience Questionnaire (UEQ)," *J. Janitra Inform. dan Sist. Inf.*, vol. 3, no. 1, pp. 10–17, 2023, doi: 10.25008/janitra.v3i1.169.

[7] A. M. Saleh, H. Y. Abuaddous, I. S. Alansari, and O. Enaizan, "The evaluation of user experience of learning management systems using UEQ," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 7, pp. 145–162, 2022, doi: 10.3991/ijet.v17i07.29525.

[8] I. N. S. W. Wijaya, P. P. Santika, I. B. A. I. Iswara, and I. N. A. Arsana, "Analisis dan evaluasi pengalaman pengguna PaTik Bali dengan metode User Experience Questionnaire (UEQ)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 2, pp. 217–226, 2021, doi: 10.25126/jtiik.2020762763.

[9] S. Elisurya, H. Muslimah Az-Zahra, and N. H. Wardani, "Evaluasi pengalaman pengguna menggunakan usability testing dan User Experience Questionnaire (UEQ) (Studi pada E-Commerce Fashion)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 5, pp. 4327–4332, 2019, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/5193.

[10] M. Ramadhani, Sidharta, and N. P. Budhianto, "User experience evaluation of Surabaya's freeletics community information system using User Experience Questionnaire (UEQ)," *2022 1st Int. Conf. Inf. Syst. Inf. Technol. ICISIT 2022*, no. July, pp. 244–248, 2022, doi: 10.1109/ICISIT54091.2022.9872977.

[11] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Construction of a benchmark for the User Experience Questionnaire (UEQ)," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 4, no.

4, p. 40, 2017, doi: 10.9781/ijimai.2017.445.
[12] P. V. Firstama, M. Galinium, and J. Purnama, "Analysis of The user experience of a mobile application system : A case study of Xyz," *J. Appl. Information, Commun. Technol.*, vol. 4, no. 2, pp. 47–61, 2017, doi: 10.33555/ejaict.v4i2.84.
[13] M. A. Kushendriawan, H. B. Santoso, P. O. H. Putra, and M. Schrepp, "Evaluating user experience of a mobile health application 'Halodoc' using user experience questionnaire and usability testing," *J. Sist. Inf.*, vol. 17, no. 1, pp. 58–71, 2021, doi: 10.21609/jsi.v17i1.1063.
[14] K. K. Pangestu, T. Lathif, M. Suryanto, and A. Pratama, "User Experience Questionnaire (UEQ) sebagai metode pengukuran evaluasi pengalaman pengguna Virtual Campus Tour UPN," *442 J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 7, no. 2, pp. 442–451, 2023, doi: 10.52362/jisamar.v7i2.718.
[15] M. Schrepp, "User experience questionnaire handbook version 8," *URL: https://www.researchgate.net/publication/303880829_User_Experience_Questionnaire_Handbook_Version_2.(Accessed: 31.10. 2024)*, pp. 1–15, 2019, [Online]. Available: www.ueq-online.org.
[16] I. Fatmawati, A. Fajriani, and Z. Razilu, "Evaluasi user experience menggunakan User Experience Questionnare (UEQ) (Studi kasus : Website bombanakab.go.id)", *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 201–211, 2025.

Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 16 Issue 1 Year 2026 Pages 49-61
DOI: http://dx.doi.org/10.31940/matrix.v16i1.49-61
URL: http://ojs2.pnb.ac.id/index.php/matrix

Model creation for Denial of Service (DoS) attack classification using an ensemble learning approach on multi-dataset network traffic

Farhan Ainurrahman [1]*, Hariz Farisi [2], Diva Kurnianingtyas [3]

[1,2] *Information Technology Study Program, Universitas Brawijaya, Indonesia*
[3] *Informatics Engineering Study Program, Universitas Brawijaya, Indonesia*

*Corresponding Author: *farhanainurrahman2147@gmail.com*

Abstract: The rapid advancement of information technology has increased cybersecurity threats, one of which is the Denial of Service (DoS) attack that can disrupt service availability. Most existing studies on DoS attack classification rely on a single dataset and a single machine learning model, which limits the generalizability of their results across different network environments. This study addresses this gap by proposing an ensemble learning-based model for DoS attack classification using multi-dataset network traffic. The datasets used in this research are UNSW-NB15 and TON-IoT, which were combined based on feature compatibility. After the preprocessing stage, a final dataset consisting of 73,302 records was obtained, comprising 64,267 normal traffic instances and 9,035 DoS attack instances. The dataset was then split using stratified sampling with an 80:20 ratio for training and testing data. The ensemble learning methods applied include Random Forest (bagging) and XGBoost (boosting), with training scenarios using both the original dataset and data balanced using the Synthetic Minority Over-sampling Technique (SMOTE). Model evaluation was conducted using a confusion matrix and performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC.The results show that the ensemble learning approach achieves high performance in classifying DoS attacks. However, the application of SMOTE did not improve model performance in this study. The best-performing model was Random Forest trained on the original dataset, achieving an accuracy of 0.9854, precision of 0.9515, recall of 0.928, F1-score of 0.9402, and ROC-AUC of 0.996. These results indicate that the proposed model is effective for DoS attack classification across heterogeneous network traffic data.

Keywords: Attack classification, Denial of Service (DoS) ensemble learning, Random Forest, XGBoost

History Article: Submitted 22 January 2026 | Revised 6 February 2026 | Accepted 16 February 2026

How to Cite: F. Ainurrahman, H. Farisi, and D. Kurnianingtyas, "Model creation for Denial of Service (DoS) attack classification using an ensemble learning approach on multi-dataset network traffic," *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 16, no. 1, pp. 49–61, 2026, doi: 10.31940/matrix.v16i1.49-61.

Introduction

The rapid development of information technology has facilitated data management and processing across various sectors; however, it has also increased the complexity of cybersecurity threats. Indonesia's National Cyber and Crypto Agency (BSSN) reported 3.64 billion cyber attacks or network traffic anomalies between January and July 2025, highlighting the urgent need for comprehensive network protection. Network anomalies refer to activities that deviate from normal traffic patterns, such as abnormal packet spikes, unusual port usage, or irregular communication behavior, making early detection a critical component of network security systems.

One of the most common threats is Denial of Service (DoS) attacks, which aim to overwhelm systems with excessive requests, rendering services inaccessible to legitimate users. According to the Common Vulnerability Scoring System (CVSS), DoS attacks are categorized as high-severity threats because they directly target service availability. The sophistication of these attacks continues to grow, as demonstrated by the mitigation of a 3.47 Tb/s Distributed Denial of Service (DDoS) attack with a packet rate of up to 340 million packets per second handled by Microsoft Azure in 2021 [1]. Such attacks not only degrade system performance but also cause financial losses and reputational damage for organizations.

Various automated detection approaches have been developed, one of which is the application of machine learning techniques. Ensemble learning has emerged as a widely adopted method due to its ability to improve prediction stability and performance by combining multiple base models [2]. This concept is based on the limitation of single models, which may not perform optimally for all problem types, while model combinations can compensate for individual weaknesses and produce more reliable predictions [3].

Previous studies have demonstrated the effectiveness of machine learning techniques for detecting DoS attacks. Primadya *et al.* applied Logistic Regression on the CIC IoT Attacks 2023 dataset and achieved an accuracy of 97% using random undersampling and Recursive Feature Elimination (RFE) [4]. Harto and Basuki employed Random Forest for DDoS detection in Software Defined Networking (SDN) environments and obtained 90% accuracy with a detection time of 0.3 seconds [5]. Ariyanto *et al.* developed a K-Nearest Neighbor-based intrusion detection system using the KDDCUP 1999 dataset and reported an accuracy of 90% [6]. Meanwhile, Firdaus *et al.* evaluated low-rate DDoS detection using Naive Bayes with the CICIDS2017 dataset and achieved an accuracy of 83.45% [7]. Despite these promising results, most existing studies are still limited to single algorithms or single datasets, which may restrict model generalization. as such approaches do not fully capture the diversity and complexity of real-world network traffic and are rarely evaluated under heterogeneous or imbalanced data conditions.

The choice of datasets also plays a crucial role in building generalized detection models. The UNSW-NB15 dataset provides a modern representation of network traffic with diverse attack scenarios, while the TON-IoT dataset introduces additional variations from Industrial IoT environments, which are inherently more complex [8]. Combining these datasets enables the model to learn more comprehensive and realistic attack patterns, thereby improving the robustness and generalizability of the proposed detection model across different network environments. Therefore, this study explicitly addresses the limitations of prior work by proposing an ensemble learning-based DoS attack classification model trained on integrated multi-dataset network traffic and evaluated under both original and balanced data distributions.

Based on these considerations, this study proposes a DoS attack classification model using an ensemble learning approach with the UNSW-NB15 and TON-IoT datasets. The model aims to accurately classify network anomalies and enhance detection reliability. Model performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics derived from the confusion matrix. This research is expected to contribute to the development of more robust intrusion detection systems and support improved network security in the digital era.

Methodology

This study aims to develop a classification model for Denial of Service (DoS) attacks using an ensemble learning approach to classify normal traffic and DoS attack traffic. The model is built using two datasets, namely UNSW-NB15 and TON-IoT. The results of this study are expected to contribute to the development of DoS attack classification models. The overall research workflow is illustrated in Figure 1.

The research procedure consists of data collection, data preprocessing, modeling, and model evaluation. The data used in this study were obtained from two secondary datasets, namely UNSW-NB15 and TON-IoT, with a focus on DoS-labeled and normal traffic data as the basis for model training.

The data preprocessing stage was conducted to ensure that the data were suitable for modeling. This stage included data cleaning to remove irrelevant or inconsistent values, feature selection to identify relevant variables, and data splitting into training and testing subsets. In addition, class imbalance was handled using the Synthetic Minority Over-sampling Technique (SMOTE), and data transformation was applied to prepare the data for modeling.

Modeling was performed using an ensemble learning approach by combining base learners through bagging and boosting techniques. Random Forest was selected as the bagging-based model because previous studies have shown that it outperforms other algorithms such as Support Vector Machine (SVM) in classifying network anomalies in industrial traffic, including systems based on Modbus and OPC UA protocols [9]. Meanwhile, XGBoost was selected as the boosting-based model due to its proven effectiveness in distinguishing between normal and attack traffic

in network environments, particularly based on Simple Network Management Protocol (SNMP) data [10].

The evaluation stage was then carried out using testing data by applying performance metrics including accuracy, precision, recall, F1-score, and AUC-ROC. The model evaluation was used to assess how well the trained model performed in classifying network traffic.

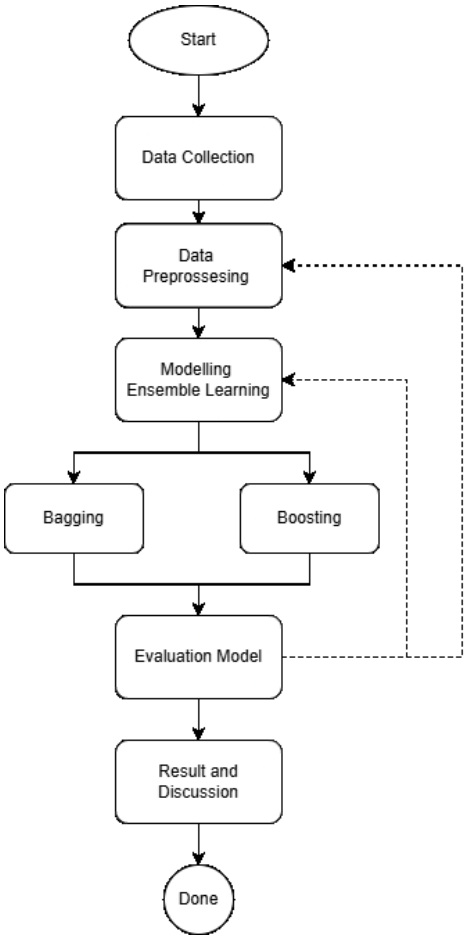


Figure 1. Research workflow

Data Collection and Preprocessing

Data Collection

This study uses two datasets, namely UNSW-NB15 and TON-IoT, both of which were developed by the University of New South Wales (UNSW), Australia. Both datasets contain normal traffic data and Denial of Service (DoS) attack data for model training.

The UNSW-NB15 dataset was generated using IXIA PerfectStorm in the ACCS UNSW Cyber Range Lab. The dataset is available in several formats, and this study uses the CSV format from the provided training and testing sets. Meanwhile, the TON-IoT dataset was developed at the UNSW Canberra IoT Lab and contains data collected from IoT/IIoT sensors. This study utilizes the network traffic subset in CSV format. The multi-source data collection process is illustrated in Figure 2.

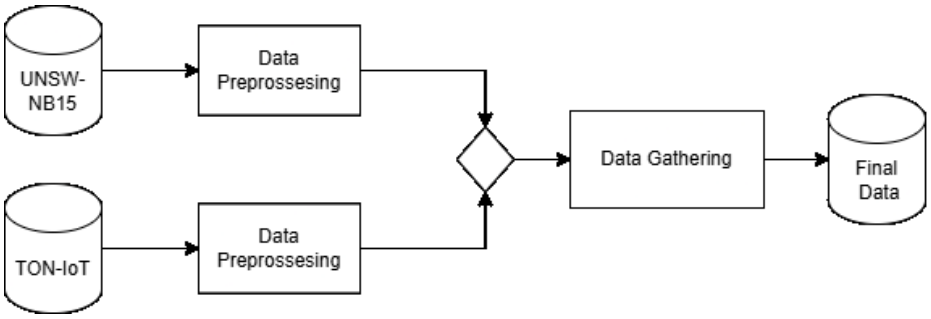


Figure 2. Multi-source data collection

This study employs a multi-source data collection approach using two different datasets, namely UNSW-NB15 and TON-IoT. Both datasets were collected independently, then underwent preprocessing and were integrated through a data integration process to generate the final dataset. The focus of this research is on normal traffic and DoS attack traffic. The UNSW-NB15 dataset provides 56,000 normal traffic instances and 12,264 DoS attack instances, while the TON-IoT dataset contains 50,000 normal traffic instances and 20,000 DoS attack instances. Overall, this study utilizes 106,000 normal traffic records and 32,264 DoS attack records.

The UNSW-NB15 dataset consists of 49 features representing various network traffic characteristics. Meanwhile, the TON-IoT dataset contains 46 features describing network traffic behavior. Features that are representative of DoS attack patterns and normal traffic were selected to optimize the performance of the developed model.

Data Cleaning

The data cleaning stage was conducted to ensure data quality prior to modeling by removing missing values, duplicate records, and invalid data from the UNSW-NB15 and TON-IoT datasets. Missing values in small proportions were removed, while significant missing values were imputed using the median or mean. Duplicate data were eliminated to prevent bias, and illogical values such as negative durations or irrelevant byte values were removed as noise. This process ensured that the dataset used was consistent, valid, and ready for the modeling stage.

Feature Selection

The feature selection stage was conducted to select relevant features from the UNSW-NB15 and TON-IoT datasets so that they could be integrated and analyzed as training and testing data for model development. Features with similar semantic meanings were selected to form a uniform data structure. This approach reduces model complexity, improves computational efficiency, and preserves essential information for detecting DoS attack patterns. As shown in Table 1, the selected features represent the most relevant attributes for the classification task.

Tabel 1. List of selected features

No	Feature Name	Definition
1	dur; duration	Total duration of activity
2	sbytes; src_bytes	Number of bytes transmitted from source to destination
3	dbytes; dst_bytes	Number of bytes transmitted from destination
4	spkts; src_pkts	Number of packets sent from source to destination
5	dpkts; dst_pkts	Number of packets sent from destination to source
6	attack_cat; type	Attack category

Data Gathering

At this stage, feature names from the UNSW-NB15 dataset were renamed to match features with similar meanings in the TON-IoT dataset. he mapping of the original and renamed feature names is presented in Table 2. After the renaming process, both datasets were merged into a single dataset for the analysis of normal traffic and DoS attacks. This integrated dataset was then used for model training in the modeling stage.

Table 2. Renamed features (UNSW-NB15)

No	Original Feature	Renamed Feature
1	dur	duration
2	sbytes	src_bytes
3	dbytes	dst_bytes
4	spkts	src_pkts
5	dpkts	dst_pkts
6	attack_cat	type

The initial integrated dataset consisted of 88,194 records. After data quality verification, 14,892 duplicate records were identified and removed. The final clean dataset contained 64,267 normal traffic records and 9,035 DoS attack records, which were subsequently used for model development.

Data Splitting

At this stage, the dataset was split into training and testing data with an 80:20 ratio using the train_test_split() function from the scikit-learn library. The stratify parameter was applied to maintain the class distribution of normal and DoS traffic in both subsets. This process resulted in 58,641 training samples and 14,661 testing samples, consistent with the original class distribution.

SMOTE was applied only to the training data (80% of the total dataset) to prevent data leakage, with an initial composition of 51,413 normal traffic samples and 7,228 DoS attack samples. SMOTE generates synthetic samples by identifying k-nearest neighbors within the minority class using Euclidean distance.

$$d(x_i, x_j) = \sqrt{\sum_{t=1}^{n} (x_{it} - x_{jt})^2} \quad (1)$$

The new synthetic data were generated using the following formula:

$$x_{new} = x_i + \delta(x_{zi} - x_i) \quad (2)$$

where δ is a random number between 0 and 1. As a result, a total of 43,685 synthetic samples were generated, increasing the number of DoS class samples to 51,413. Consequently, the training dataset became balanced between the normal and DoS classes.

SMOTE for Data Imbalanced

The dataset in this study exhibits an imbalanced class distribution between normal traffic and DoS attacks; therefore, handling this issue is necessary to prevent model bias toward the majority class. The Synthetic Minority Over-sampling Technique (SMOTE) was applied due to its effectiveness in increasing minority class representation and reducing the risk of overfitting [11].

Data Transformation

The final dataset consists of six features with data types as shown in Table 3.

Tabel 3. Feature data types

No	Feature	Data Type
1	Duration	*Float*
2	src_bytes	*Integer*
3	dst_bytes	*Integer*
4	src_pkts	*Integer*
5	dst_pkts	*Integer*
6	Type	*Object*

Five features are numeric, while one feature, namely *Type*, is of object type. Since machine learning models can only process numerical data, the *Type* feature was converted using label

encoding. This process transformed the *normal* category into the value 1 and the *DoS* category into the value 0, following the alphabetical order applied by the label encoder.

Modelling and Evaluation

In the modelling stage, the training data that had undergone preprocessing were used to train the classification models. The trained models were then evaluated using the testing data, which were separated from the training data to prevent overfitting and to ensure an objective assessment of model performance.

In ensemble learning, a base learner refers to the fundamental model that is combined with other models to improve prediction accuracy. In this study, the Decision Tree algorithm was selected as the base learner because it is capable of effectively capturing patterns based on important features and has been proven to deliver strong performance in classification task [12]. he general structure of a Decision Tree is illustrated in Figure 3.

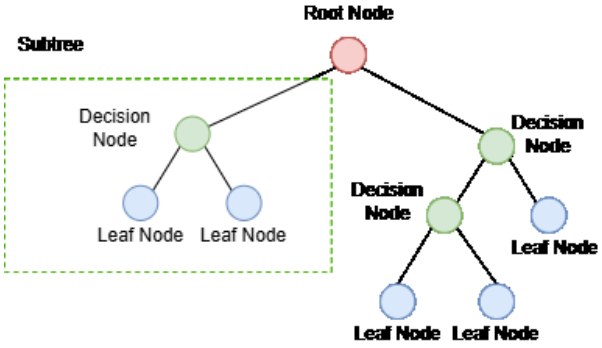


Figure 3. Decision tree

A Decision Tree is a hierarchical model that splits data through a series of tests on feature threshold values. The tree structure consists of nodes and branches, which represent the decision-making process. The selection of the best split is determined using impurity measures such as the Gini Index and Entropy, which quantify the level of uncertainty at each node.The entropy is calculated using the following formula:

$$Entrophy = -\sum_{i}^{C} P_i \log_2 Pi \tag{3}$$

Information Gain is used to determine the most optimal attribute for splitting the data by comparing the decrease in entropy before and after the split. The tree construction process is performed iteratively by selecting the attribute that produces the highest Information Gain until the nodes become homogeneous or meet predefined stopping criteria. The final result is a Decision Tree structure that can be used to classify or predict new data.The Information Gain is defined as:

$$Gain(S, A) = Entrophy(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} \times Entrophy(S_v) \tag{4}$$

Where S is the initial dataset before splitting, A is the tested attribute (feature), $V(A)$ represents all possible values of attribute A, S_v is the subset of $S$for attribute value v.

Random Forest (Bagging)

The Random Forest is an ensemble learning algorithm that utilizes multiple Decision Trees to improve prediction accuracy and stability [13]. Each Decision Tree is constructed from a different subset of the data through bootstrap sampling, which is a random sampling technique with replacement, allowing each data instance to be selected more than once. The overall architecture of the Random Forest algorithm is illustrated in Figure 4.

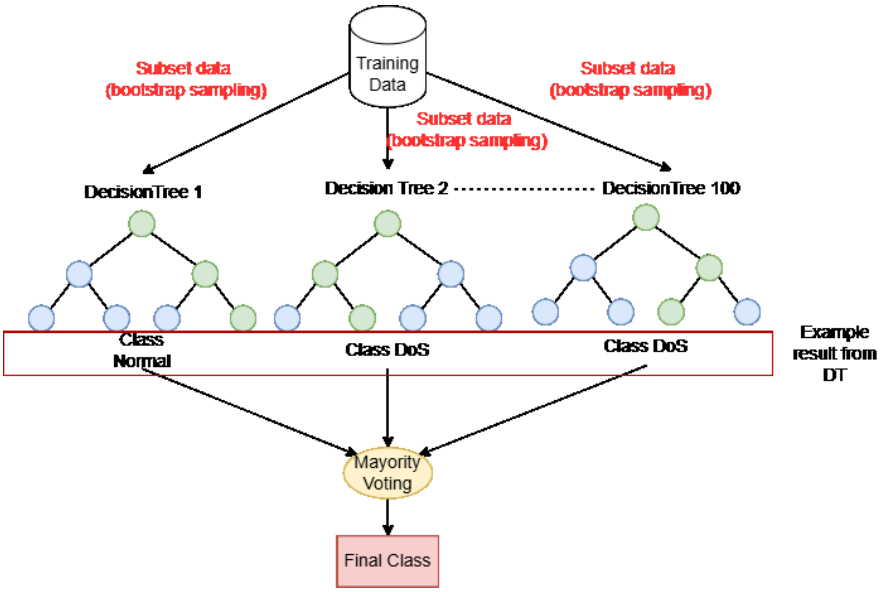


Figure 4. Random Forest Algorithm

During the prediction stage, Random Forest combines the outputs of all trees using a majority voting mechanism, where the class receiving the highest number of votes is selected as the final prediction. This approach is effective in reducing overfitting, which often occurs in a single Decision Tree model, and produces a more robust and reliable model for unseen data. The Random Forest prediction can be expressed as:

$$y(x) = mode(\{h_1(x), h_2(x), \dots, h_T(x)\}) \quad (5)$$

In the implementation of the Random Forest model for DoS attack classification, the model is trained using the prepared training dataset. In this study, the Random Forest model is implemented using the scikit-learn library without applying hyperparameter tuning. Therefore, the model utilizes the default parameter settings provided by scikit-learn for DoS attack classification.

XGBoost (Boosting)

XGBoost (Extreme Gradient Boosting) is an advanced development of the Gradient Boosting algorithm introduced by Dr. Tianqi Chen as a faster, more efficient, and more scalable implementation [14]. This algorithm constructs multiple Decision Trees sequentially, where each new tree is designed to correct the errors (residuals) of the previous tree until the model achieves optimal predictive performance. The general workflow of the XGBoost algorithm is illustrated in Figure 5.

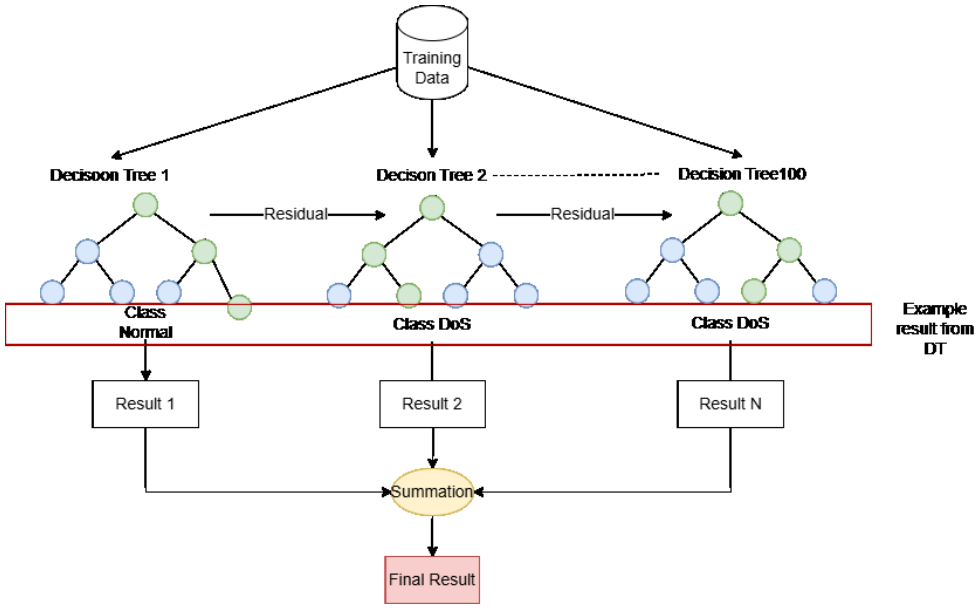


Figure 5. XGBoost Algorithm

In XGBoost, the learning process is optimized through an objective function that combines training loss to measure prediction errors and regularization to prevent overfitting. Residuals are computed at each iteration and are then used to build new Decision Trees through gradient and Hessian calculations. The weight of each leaf node is determined to minimize the objective function, allowing the model accuracy to improve progressively.
The objective function is defined as:

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{6}$$

The final prediction is obtained through a summation process, where the contributions of all constructed Decision Trees are aggregated and a logistic function is applied for classification tasks. In this study, the scikit-learn library is used to implement XGBoost, which provides high computational performance and ease of use. The complete objective function can be expressed as:

$$obj(\theta) = \sum_{i=i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega(f_k) \tag{7}$$

In this study, XGBoost library is utilized to implement the XGBoost model for DoS attack classification. XGBoost is selected because it is an optimized implementation of gradient boosting that is distributed, efficient, flexible, and portable, enabling parallel processing of boosted trees. In this study, the XGBoost model is implemented using the default parameter settings.

Confusion Matrix and ROC-AUC Evaluation

The Confusion Matrix is an evaluation method that provides a comprehensive overview of the number of correct and incorrect predictions for each class. Through this matrix, classification performance metrics can be derived, allowing the evaluation of model performance in distinguishing between normal traffic and DoS attack traffic. An example of a Confusion Matrix is illustrated in Figure 6.

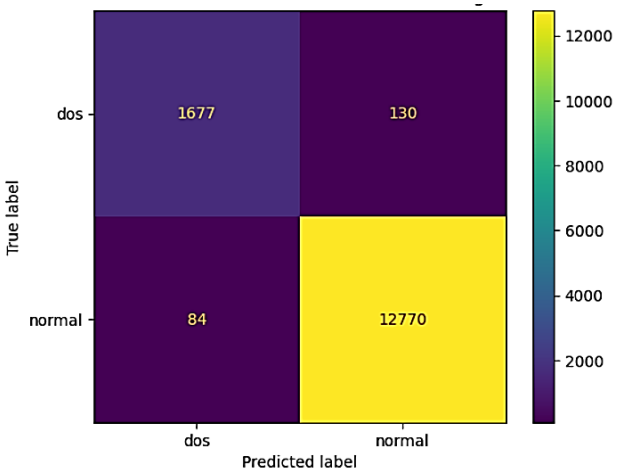


Figure 6. Example of confusion matrix

In addition, this study also employs ROC-AUC as a model evaluation metric. The Receiver Operating Characteristic (ROC) curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different threshold values, thereby representing the trade-off between sensitivity and misclassification errors. The Area Under the Curve (AUC) measures the area under the ROC curve, with values ranging from 0 to 1. A value closer to 1 indicates better model performance in distinguishing between the two classes. ROC-AUC is particularly effective for binary classification problems with imbalanced class distributions. An example of the ROC-AUC curve is illustrated in Figure 7.

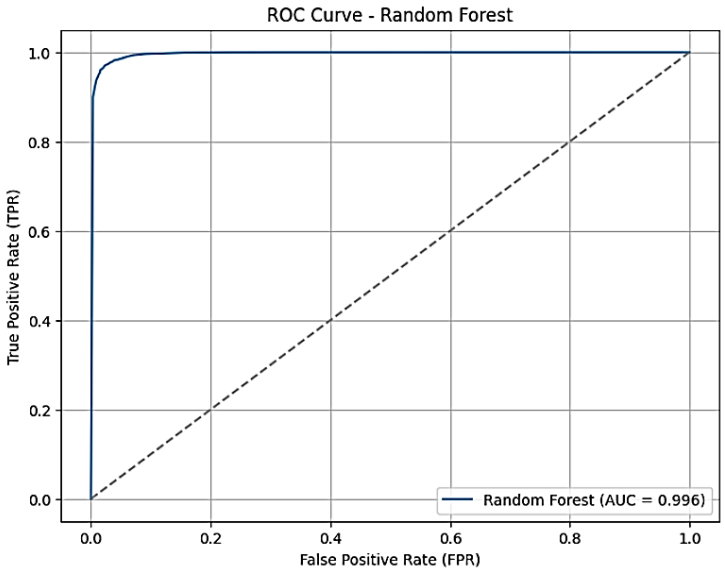


Figure 7. Example of ROC-AUC

Results and Discussions

Results

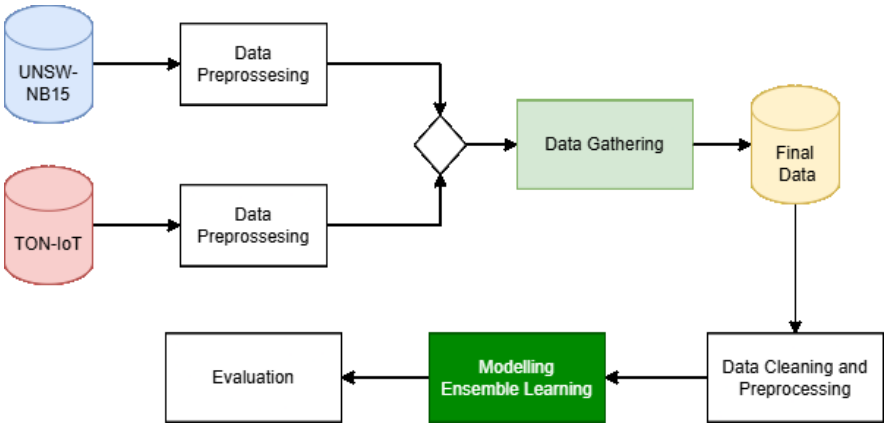
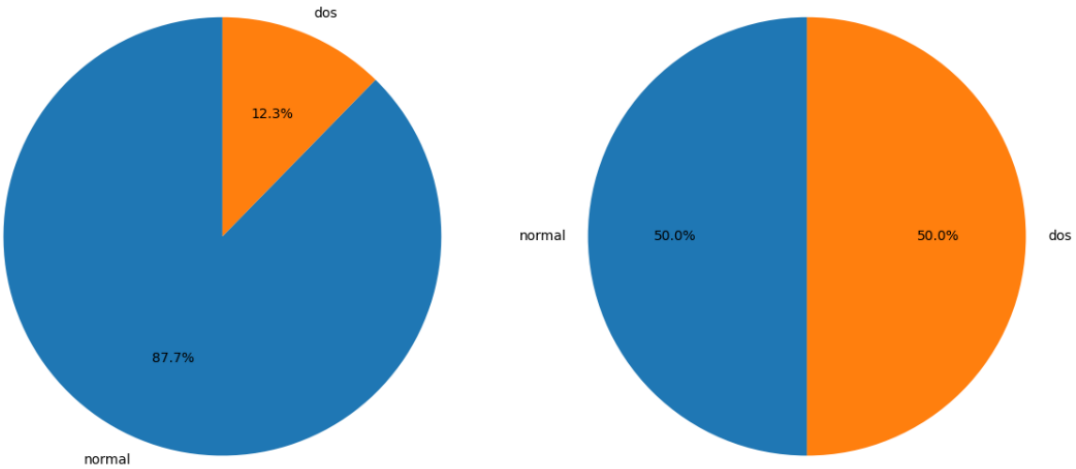


Figure 8. Model development process for DoS attack classification

Based on Figure 8, the UNSW-NB15 and TON-IoT datasets are network traffic datasets used as the basis for developing the DoS attack detection model. Both datasets were combined using a horizontal integration approach, in which only common features were merged to obtain a uniform data structure. Prior to integration, each dataset underwent a preprocessing stage.

The UNSW-NB15 dataset consists of 68,264 records, while the TON-IoT dataset contains 70,000 records. After integration and additional preprocessing, a final dataset of 73,302 records was obtained, consisting of 64,267 normal traffic instances and 9,035 DoS attack instances. The final dataset was split using stratified sampling with an 80:20 ratio, resulting in 58,642 training samples and 14,660 testing samples.

This study analyzes the performance of models trained using both the original imbalanced data and data balanced using the Synthetic Minority Over-sampling Technique (SMOTE). The use of SMOTE is based on previous studies that demonstrated its effectiveness in addressing class imbalance and improving model performance. The distribution of training data classes before and after the application of SMOTE is illustrated in Figure 9.



(a) Before SMOTE (b) After SMOTE

Figure 9. Distribution of training data classes before and after SMOTE

An ensemble learning approach was employed to build the DoS attack detection model due to its ability to improve accuracy, reduce false positives, and adapt to complex attack patterns. Random Forest and XGBoost were implemented using the scikit-learn library. Random Forest was

selected for its stability through majority voting, while XGBoost was chosen for its superior capability in handling complex attack patterns and imbalanced data.

The developed models were evaluated using a confusion matrix to compute performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. The evaluation results are presented in Table 4.

Tabel 4. Model evaluation results

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest (Original Data)	0.9854	0.9515	0.928	0.9402	0.996
Random Forest (SMOTE Data)	0.981	0.956	0.892	0.923	0.996
XGBoost (Original Data)	0.9849	0.9568	0.908	0.9316	0.996
XGBoost (SMOTE Data)	0.972	0.841	0.961	0.961	0.997

Discussions

This study developed four ensemble learning models, namely Random Forest and XGBoost, each trained using both the original dataset and the dataset balanced using the Synthetic Minority Over-sampling Technique (SMOTE), to analyze the impact of class imbalance handling on Denial of Service (DoS) attack detection. The experimental results indicate that the application of SMOTE did not improve model performance and instead reduced precision values, which suggests an increase in false positives. This finding contrasts with previous studies which reported that SMOTE improved classification performance, highlighting that the effectiveness of SMOTE strongly depends on dataset characteristics and the algorithms used [11]. This result may occur because SMOTE generates synthetic samples that do not fully represent real network traffic patterns, causing the model to learn less discriminative features and increasing false positive predictions.

The ensemble learning approach proved effective for DoS attack traffic classification due to its ability to combine multiple base models. Both Random Forest, which applies a bagging approach, and XGBoost, which uses a boosting strategy, demonstrated high performance. These results are consistent with previous studies reporting that ensemble learning significantly enhances network anomaly detection performance [15]. Ensemble models are capable of handling the complexity and variability of DoS attack patterns effectively.

The best-performing model in this study was Random Forest trained on the original dataset, achieving an accuracy of 0.9854 and a ROC-AUC value of 0.996. The balanced precision, recall, and F1-score values indicate that the model can accurately detect DoS attacks with low false positive and false negative rates. The Random Forest model was built using 100 decision trees combined through a majority voting mechanism, resulting in stable and robust predictions across diverse network traffic conditions. The performance evaluation of the best-performing model is illustrated in Figure 10.

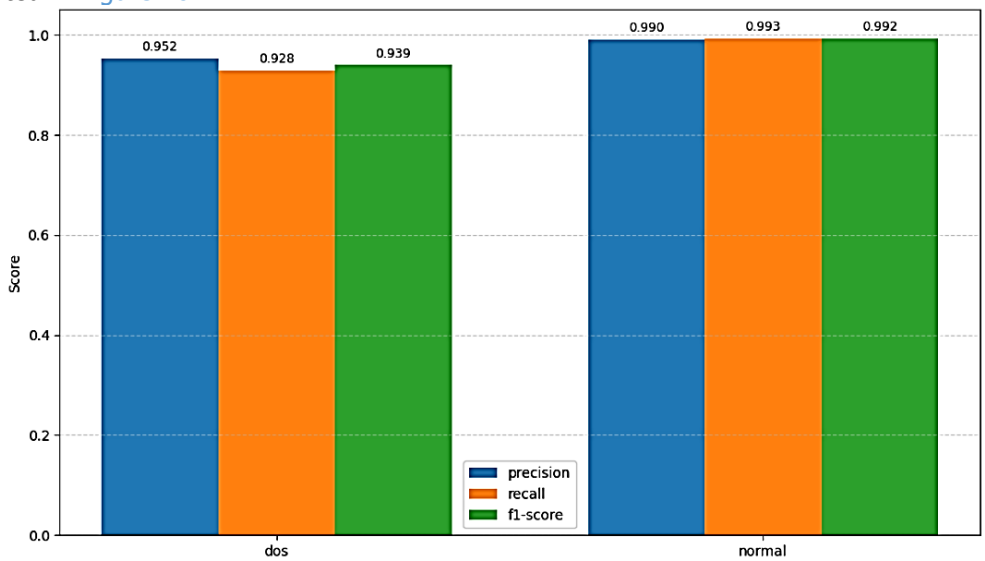


Figure 10. Performance evaluation of the best model

Conclusion

This study utilized network traffic data from the UNSW-NB15 and TON-IoT datasets, which were integrated based on feature compatibility. After the data preprocessing stage, a total of 73,302 records were obtained, consisting of 64,267 normal traffic instances and 9,035 DoS attack instances. The dataset was then split using stratified sampling with an 80:20 ratio into training and testing sets. Four ensemble learning models, namely Random Forest and XGBoost, were trained using both the original dataset and the dataset balanced with SMOTE to analyze the impact of class imbalance handling. The model evaluation stage demonstrated the performance of each developed model.

The ensemble learning approach proved effective for DoS attack traffic classification by combining base learners through Bagging and Boosting mechanisms, thereby improving model performance. In this study, both Random Forest (Bagging) and XGBoost (Boosting) achieved high performance, consistent with previous studies that reported the effectiveness of ensemble methods in handling the complexity of network traffic characteristics. Based on the evaluation results, the best and most stable model was Random Forest trained on the original dataset, achieving an accuracy of 0.9854, precision of 0.9515, recall of 0.928, F1-score of 0.9402, and ROC-AUC of 0.996. These results indicate strong performance in DoS attack classification and outperform the models trained using SMOTE-balanced data.

For future work, this study suggests extending the classification scope to include other types of network attacks beyond Denial of Service (DoS), such as probing, infiltration, and other anomaly-based threats, in order to develop a more comprehensive intrusion detection system. In addition, further experiments should be conducted using alternative class imbalance handling techniques besides SMOTE, such as undersampling methods or hybrid resampling strategies, to investigate their impact on model performance and generalization.

Acknowledgments

The authors would like to express their sincere appreciation to all institutions and researchers who provided the datasets and open-source resources used in this study. Special thanks are also extended to all parties who offered valuable academic guidance and technical support throughout the research process.

References

[1] A. Toh, "Azure DDoS Protection—2021 Q3 and Q4 DDoS attack trends." Accessed: Sep. 24, 2025. [Online]. Available: https://azure.microsoft.com/en-us/blog/azure-ddos-protection-2021-q3-and-q4-ddos-attack-trends/

[2] N. Jeffrey, Q. Tan, and J. R. Villar, "Using Ensemble Learning for Anomaly Detection in Cyber–Physical Systems," *Electronics*, vol. 13, no. 7. 2024. doi: 10.3390/electronics13071391.

[3] J. Vanerio and P. Casas, "Ensemble-learning approaches for network security and anomaly detection," in *Proceedings of the workshop on big data analytics and machine learning for data communication networks*, 2017, pp. 1–6.

[4] N. D. Primadya, A. Nugraha, A. Luthfiarta, and S. Y. Fahrezi, "Optimasi Logistic Regression untuk Deteksi Serangan DoS pada Keamanan IoT," *J. Eksplora Inform.*, vol. 13, no. 2, pp. 245–252, 2024.

[5] A. Harto, M.K. and Basuki, "Deteksi Serangan DDoS pada Jaringan Berbasis SDN dengan Klasifikasi Random Forest," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 1329–1333, 2021, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/8795

[6] Y. Ariyanto, V. A. H. Firdaus, and H. Pramana, "Klasifikasi Jenis serangan DOS dan Probing pada IDS menggunakan metode K-Nearest Neighbor," in *Seminar Informatika Aplikatif Polinema (SIAP)*, 2020.

[7] D. Firdaus, F. Fahira, and R. Rianti, "Deteksi Anomali dan Serangan Low Rate DDOS dalam Lalu Lintas Jaringan Menggunakan Naive Bayes," *Naratif J. Nas. Riset, Apl. Dan Tek. Inform.*, vol. 5, no. 2, pp. 140–148, 2023.

[8] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion

detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6. doi: 10.1109/MilCIS.2015.7348942.

[9] S. D. D. Anton, S. Sinha, and H. D. Schotten, "Anomaly-based Intrusion Detection in Industrial Data with SVM and Random Forests," in *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2019, pp. 1–6. doi: 10.23919/SOFTCOM.2019.8903672.

[10] A. M. A. Rudianto, E. S. Pramukantoro, and D. Kurnianingtyas, "Implementasi Sistem Deteksi Anomali pada Jaringan Komputer dengan Pendekatan XGBoost dan Data SNMP," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 2, 2025.

[11] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "Implementasi algoritma synthetic minority over-sampling technique untuk menangani ketidakseimbangan kelas pada dataset klasifikasi," *Sist. J. Sist. Inf.*, vol. 10, no. 2, pp. 445–459, 2021.

[12] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Comput. Biol. Med.*, vol. 128, p. 104089, 2021.

[13] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[14] T. Chen, "XGBoost: A Scalable Tree Boosting System," *Cornell Univ.*, 2016.

[15] R. Sudiyarno, A. Setyanto, and E. T. Luthfi, "Peningkatan Performa Pendeteksian Anomali Menggunakan Ensemble Learning dan Feature Selection," *Creat. Inf. Technol. J.*, vol. 7, no. 1, pp. 1–9, 2021.

POLITEKNIK NEGERI BALI

9 772580 563008

Redaksi Jurnal Matrix
Gedung P3M, Politeknik Negeri Bali
Bukit Jimbaran, PO BOX 1064 Tuban, Badung, Bali.
Phone: +62 361 701981, Fax: +62 361 701128
e-mail: p3mpoltekbali@pnb.ac.id
https://ojs2.pnb.ac.id/index.php/MATRIX