Jurnal Teknik Informatika Vol. 18 No. 1, April 2025 (157-165)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i1.46064

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Detection of Vulgarity in Anime Character: Implementation of Detection Transformer

Amalia Suciati[1], Dian Kartika Sari[2], Andi Prademon Yunus[3], Nuuraan Rizqy Amaliah[4]

[1,3]Informatics Engineering, Telkom University
[2]Data Science, Telkom University
[2]Centre for Business in Society, Coventry University
[1,2,3]Jl. DI Panjaitan No.128, Karangreja, Purwokerto Kidul, Kecamatan Purwokerto Selatan, Kabupaten Banyumas, Jawa Tengah, Indonesia
[4]Priory St, Coventry CV1 5FB, United Kingdom

Article:
Accepted: February 17, 2025
Revised: October 08, 2024
Issued: April 30, 2025

***Correspondence Address:**
andiay@telkomuniversity.ac.id

ABSTRACT

Vulgar and pornographic content has become a widespread issue on the internet, appearing in various fields include anime. Vulgar pornographic content in anime is not limited to the sexuality genre; anime from general genres such as action, adventure, and others also contain vulgar visual. The main focus of this research is the implementation of the Detection Transformer (DETR) object detection method to identify vulgar parts of anime characters, particularly female characters. DETR is a deep learning model designed for object detection tasks, adapting the attention mechanism of Transformers. The dataset used consists of 800 images taken from popular anime, based on viewership rankings, which were augmented to a total of 1,689 images. The research involved training models with different backbones, specifically ResNet-50 and ResNet-101, each with dilation convolution applied at different stages. The results show that the DETR model with a ResNet-50 backbone and dilation convolution at stage 5 outperformed other backbones and dilation configurations, achieving a mean Average Precision $AP_{50:95}$ of 0.479 and AP_{50} of 0.875. The other result is dilated convolution improves small object detection by enlarging the receptive field, applying it in early stages tends to reduce spatial detail and harm performance on medium and large objects. However, the primary focus of this research is not solely on achieving the highest performance but on exploring the potential of transformer-based models, such as DETR, for detecting vulgar content in anime. DETR benefits from its ability to understand spatial context through self-attention mechanisms, offering potential for further development with larger datasets, more complex architectures, or training at larger data scales.

Keywords : *anime; detection transformer; object detection; transformer; vulgarity.*

Jurnal Teknik Informatika Vol. 18 No. 1, April 2025 (157-165)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i1.46064

1. INTRODUCTION

Anime is an animation made in Japan that is currently a popular spectacle. Anime is an art that does not escape the art of depicting real-time moving animation with a story that uses symbolism and metaphor in delivering message [1]. This is in line with the anime industry that is getting soaring attention in various countries. From 2000 to 2020, anime production reached over 4,000 releases on the internet, indicating the massive development of anime [2].

The anime industry has visually progressed in line with technological developments. Enhanced visuals such as sharper image resolution, added visual effects and smoother motion reflect its growing quality [3]. However, anime visual content today is often inseparable from symbolic or graphic elements that lean toward vulgarity and pornography. The use of mini costumes and the display of intimate and explicit body parts indicate intentional content direction. This can have a bad impact that causes deep exploration and more terrible addiction. Pornography causes many negative impacts, such as academic performance, sexual risk behavior, aggression, and affects many in social life [4].

Parents who have children that consumes anime assume that anime is a spectacle for minors, and not the consumption of teenage-adult children. They are also unaware of the vulgar and pornographic content in anime. This makes anxiety and confusion about how to overcome or filter the content in anime so that it is safe to watch [5]. The number of illegal sites that provide anime content makes anime easily accessible for all ages. However, in fact, illegal sites that are closed will continue to be present and difficult to handle.

The amount of enthusiasm for watching anime has led to the proliferation of genres, triggering changes and expansions in existing genres [6]. This includes several genres that can be referred to as sexuality content: (1) Harem – male surrounded by female; (2) Reverse Harem – female surrounded by men; (3) Josei – female adult main characters; (4) Seinen – male adult main characters; (5) Ecchi – sexual and vulgar to semi-pornographic content; (6) Hentai, sexual to pornographic content; (7) Yuri – sexual perversion between women; (8) Yaoi – sexual perversion between men.

Popular anime with its innate vulgar and pornographic content can lead to a bad impact if not handled further. The focus of this research is the creation of a model that becomes the initial gate in preventing vulgar and pornographic content in anime. Therefore, the detection of vulgarity in anime is an early step in detecting vulgar content which can later be continued in further research for the development of the system. This research will create a model with the help of deep learning technology. Deep learning was introduced by Geoffrey Hinton, et al in 2006 where deep learning is an algorithm that adapts the concept of artificial neural network (ANN) [7].

The method of this research is implementation of Detection Transformer (DETR). DETR is a deep learning model in object recognition in images developed in 2020 by Facebook AI Research (FAIR). DETR uses Convolutional Neural Network (CNN) coupled with transformer architecture. The CNN in DETR is used for feature extraction, then proceeds to encoder and decoder transformers that use self-attention with the results in the form of bounding boxes and object class labeling [8]. Self-attention relies on feature extraction obtained from the CNN extraction process. This research uses one object class, namely vulgarity.

In this research, we strongly believe that this research contributes:

1. To help identify anime with vulgar content in society, which can be a foundation for automatic vulgarism censorship.
2. A model using the DETR method is presented as a reference for other research related to the development of vulgarity detection systems using this model.
3. A novel dataset for anime vulgarism detection.

1.1. Literature Review

Zainal Abidin Achmad et al. conducted a study involving parents and anime with pornographic content on the internet. The results revealed that many parents assume anime is solely for children, unaware it may contain vulgar or pornographic elements. This lack of awareness allows teenagers to freely access inappropriate anime content. Moreover, the absence of age-based restrictions and the

difficulty of blocking such content online contribute to its continued exposure [5].

The research conducted by Nicolas Carion et al. introduced the Detection Transformer or DETR method. Departing from problems related to how to model object detection with direct set prediction and without post-processing, unlike other object detection methods. This is in line with the goal of this research, which is to use DETR that implements the Transformer mechanism without the need for a post-processing mechanism. This research introduces the DETR method with its unique bipartite matching loss characteristic indirect detection. There is a Transformer mechanism and CNN assistance as feature extraction on the backbone of DETR. This research shows that the DETR method achieves competitive results compared to the Faster R-NN method on quantitative evaluation in the COCO dataset [8].

Research by Huaqi Zhao et al. (2025) proposes an improvement in the DETR object detection method for autonomous driving systems. The feature extraction method is enhanced by a multi-scale feature extraction approach that integrates residual partition units and a coordinate attention module, aiming to capture more detailed features and improve localization. The experimental results demonstrate that the proposed model achieves improvements of 3.3%, 4.5%, and 3% in average precision on the COCO, PASCAL VOC, and KITTI datasets, respectively. On the COCO dataset, the model achieves an AP of 0.552, an AP_{50} of 0.741, and AP_s of 0.336 [9].

The research by Yian Zhao et al. (2024) proposes an enhancement to DETR, called Real-Time DETR (RT-DETR). The results show that RT-DETR supports flexible speed with adjustable decoder layer configurations without the need for retraining. The experiments conducted on the COCO dataset show an AP of 53.1% / 54.3%. The study also demonstrates that RT-DETR outperforms YOLO [10].

Research by Xizhou Zhu et al. (2021) proposes an advanced development of the DETR method, called Deformable DETR. The attention module of Deformable DETR only focuses on a small set of key sampling points around the reference. Numerous experiments have been conducted, and the results show that the best model on the COCO dataset is Deformable DETR with a ResNeXt-101 backbone + DCN. This model achieved an AP of 52.3, AP_{50} of 71.9, AP_{75} of 58.1, AP_s of 34.4, AP_m of 54.4, and AP_l of 65.6 [11].

Research by Attila Biró et al. (2022), titled *Visual Object Detection with DETR to Support Video-Diagnosis Using Conference Tools*, compared real-time detectors—YOLOv4, Detectron2, and DETR—combined with OCR for textual object detection. The results showed DETR performed better in object detection. The study produced a DETR model capable of extracting textual predictions from bounding boxes, which were then categorized and processed for real-time multilingual translation [12].

The research conducted by Endang Suherman et al. (2023) implements ResNet-50 to enhance the performance of object detection in images. The results show that the use of ResNet-50 improves the performance of the DETR model and can detect objects with an accuracy of around 90%. The model demonstrated fast object detection and can be applied to real-time systems [13].

a. Anime and Vulgarity

Anime is a Japanese animation style that began in 1963, recognized for its distinctive character designs and roles [1]. It offers unique artistic visuals, storytelling, and genre. Human and humanoid forms are often modified to suit various genres, including hentai and pornographic anime, which link emotions with extreme body modifications [6].

Vulgarity refers to inappropriate behavior or attire, such as clothing that exposes intimate parts. This includes depictions of children or adolescents in revealing outfits, emphasizing certain body parts, or performing sexually suggestive movements. Pornography involves explicit depictions of the human body or sexual acts aimed at arousing desire [4].

b. Transformer

Transformer, a model architecture introduced by Vasmani, et al. Transformer is an architecture designed to overcome the limitations of sequence modeling such as RNN or convolution. Transformers rely more on parallelization and training efficiency. Transformers use attention mechanisms as the main component for modeling in data dependencies between inputs and outputs [14].

Jurnal Teknik Informatika Vol. 18 No. 1, April 2025 (157-165)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i1.46064

c. Detection Transformer

Detection Transformer (DETR) is a method that adopts the architecture of the encoder and decoder of a Transformer. DETR is designed for direct-set predictions or detecting set objects directly. DETR predicts all objects in one prediction, trained by an end-to-end loss function that will perform bipartite matching between the predicted and ground truth of detected objects. The main feature used in DETR is the matching loss function (bipartite matching loss), where this loss function is not affected by changes in the order or arrangement of predictions so that it can be performed in parallel [8].

1. Backbone

The backbone in DETR is used to extract a concise representation of features in the image of the object to be detected from the input. The backbone of DETR can use a Convolutional Neural Network (CNN) such as ResNet in feature extraction from dataset images. The resulting feature extraction is then continued for the encoder process. The backbone function is also to reduce the dimension of the input data or flattened [8].

2. Encoder

The encoder in DETR uses a transformer architecture with multi-head self-attention and feed-forward networks (FFN). Input features from the backbone are flattened and added with positional encoding before entering the encoder. Each encoder layer captures different parts of the feature map, followed by FFN, and normalized through ADD and NORM layers. The encoder transforms backbone features into richer feature representations in the form of feature vectors [8].

3. Decoder

The decoder in DETR also adopts a transformer architecture, similar to the encoder, with multi-head self-attention. However, unlike the autoregressive decoder in the original Transformer by Vaswani et al., DETR's decoder processes N objects in parallel. It uses input embeddings called object queries (a random vectors with positional encodings) as inputs to each attention layer. These object queries represent object representations that guide the decoder to predict bounding boxes and class labels through a feed-forward network (FFN) [8].

4. Prediction Feed-Forward Networks (FFNs)

The prediction FFNs in DETR are three-layer perceptrons with ReLU activation, hidden dimensions, and a linear projection layer. They predict normalized center coordinates, height, and width of the bounding boxes. The projection layer uses softmax to predict class labels, including a special "no object" (∅) label to represent the absence of a detected object [8].

5. Bipartite Matching Loss

DETR features a bipartite matching loss mechanism that characterizes DETR itself. Bipartite matching will calculate the optimal match between the predicted object and the ground truth, which will then optimize the loss of the specific object or bounding box. Bipartite matching uses the Hungarian algorithm. The matching cost considers the predicted class as well as the fit of the predicted box to the ground-truth. Each element i of the ground truth is a vector that represents the coordinates of the center of the ground truth box with a height and width that are relatively dependent on the image size. This matching is done by one-to-one matching for direct set prediction without any duplicates. The second part of matching cost and Hungarian loss is bounding box matching, where a linear combination of loss and IoU loss is used. Both losses are normalized by the number of objects in the batch [8]. The equation of this loss is in Eq.1 and Eq.2.

$$\hat{\sigma} = \underset{\sigma \epsilon\, \tau N}{\operatorname{argmin}} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) \quad (1)$$

$$\mathcal{L}_{\text{Hungarian}}(y_i, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + 1_{\{c_i \neq \emptyset\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right] (2)$$

Where $\hat{\sigma}$ represents the optimal permutation as defined in Eq.1. The $\mathcal{L}_{\text{match}}$ denotes the matching cost function, which measures the cost between the ground truth y_i, and the corresponding prediction $\hat{y}_{\sigma(i)}$. This matching cost is computed as a combination of classification loss (cross-entropy loss) and bounding box regression loss, which includes both L1 loss and Generalized IoU (GIoU) loss.

The total loss after the matching step is defined in Eq.2, which uses the Hungarian loss. Here, $\hat{p}_{\hat{\sigma}(i)}(c_i)$ represents the predicted probability that the $\hat{\sigma}(i)$-th prediction

corresponds to the ground truth class $c_{i,}$. The term $-\log \hat{p}_{\hat{\sigma}(i)}(c_i)$ indicates that a lower predicted probability results in a higher classification loss. The indicator function $1_{\{c_i \neq \emptyset\}}$ equals 1 when the object is not background. Meanwhile, $\mathcal{L}_{\text{box}}(b_{i,} \hat{b}_{\hat{\sigma}}(i))$ computes the bounding box loss between the ground truth $y_{i,}$ and the matched prediction $\hat{b}_{\sigma(i)}$.

d. ResNet

Residual Network (ResNet) is a convolutional neural network (CNN) architecture designed to address the vanishing gradient problem in deep neural networks. ResNet utilizes residual blocks, which allow information to bypass several layers. Shortcut connections are used to link the input from previous layers directly to deeper layers, preventing the loss of crucial information during training. This enables the training of deeper networks without compromising performance [15][16].

ResNet-50 is a version of ResNet with 50 convolutional layers, offering faster training and inference times, making it computationally efficient. **ResNet-101** is a deeper version of ResNet with 101 convolutional layers. The increased number of layers in ResNet-101 allows the network to learn more detailed and complex feature representations, but it requires longer training and inference times [15][16].

e. Evaluation Method

Mean Average Precision (mAP) is the average of Average Precision (AP) values, used as an evaluation metric to measure the performance of object detection models [11]. It serves as a key indicator of how accurately and reliably a model can detect objects without errors.

1. Intersection over Union (IoU)

Intersection Over Union (IoU) is an evaluation metric used to measure the accuracy of object detection [11].

$$\text{IoU} = \frac{\hat{B} \text{ and B overlapping area}}{\hat{B} \text{ and B union area}} \quad (3)$$

Where the predicted bounding box is represented by $\hat{B}$, and the ground-truth bounding box is represented by B, which can be expressed in Eq.3.

2. Precision

Precision measures the proportion of true positive predictions out of all positive predictions. It is calculated as the number of true positives (TP) divided by the total number of detections [11].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

3. Recall

Recall measures the proportion of actual positive values that are correctly predicted. In other words, it evaluates how many object predictions are accurately detected. Recall is the number of true positives (TP) out of all ground truth values [11].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

4. Average Precision (AP)

The AP value is determined by calculating the average precision on the Precision-Recall (PR) curve with respect to recall [11].

$$\text{AP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots 1\}} P_{\text{inter}}(r) \quad (6)$$

$$P_{\text{inter}}(r) = \max_{p:p \geq r} p(\tilde{r}) \quad (7)$$

Where $P_{\text{inter}}(r)$ is the maximum interpolated precision for all recall values $\tilde{r} \geq r$, with r being the recall at each of the 11 points ranging from 0.0 to 1.0 and it is used to smooth the PR curve with 11 recall points.

5. Mean Average Precision (mAP)

The mAP value is determined from evaluation metrics such as IoU, confusion matrix, precision, and recall. mAP is the average result of AP, so the mAP value can be calculated as in Eq.8 [11].

$$\text{mAP} = \frac{\sum_{i \in \text{classes}} \text{AP}_i}{\text{Total no. of classes}} \quad (8)$$

Where total number of classes is 2 include no object in this study.

Jurnal Teknik Informatika Vol. 18 No. 1, April 2025 (157-165)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i1.46064

2. METHODS

The research is segmented into four primary phases: data acquisition, a comprehensive analysis of the Transformer model, an investigation of DETR, and an assessment of model performance using metrics mAP. Each phase is discussed in detail below.

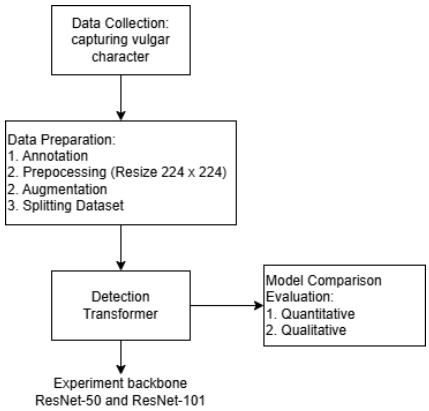

Figure 1. *Proposed method for vulgarity detection*

2.1. Data Collection and Preprocessing

The dataset was collected by capturing vulgar characters from popular anime series based on specific criteria, resulting in a total of 800 images and one class labeled as "vulgar" (single class. The dataset was then annotated on the vulgar parts using Roboflow.

Table 1. *Number of annotations per image*

Annotation/Image	Total
1	651
2	135
3	14
Total	**800**

Figure 2. *Sample of annotation dataset*

The data was preprocessed by resizing the images to 224 × 224 dimensions, referring to the standard size for CNNs. The data was then augmented to enrich the dataset and split into train, validation set and testing set.

Table 2. *Splitting dataset*

Splitting Set	Percentage	Total Image
Train Set	88%	1689
Valid Set	4%	80
Test Set	8%	157
Total		**1926**

Before the augmentation process, the data was split into 70% training, 10% validation, and 20% test sets. Augmentation was applied only to the training set, resulting in a significantly higher number of images in the training set compared to the valid and test sets.

2.2. Detection Transformer Modelling

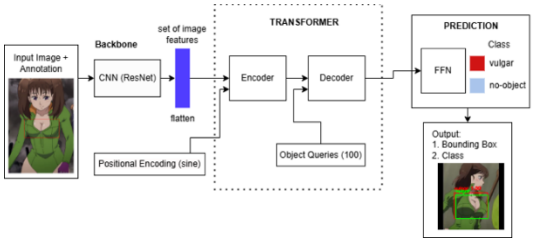

Figure 3. *Task expected from detection transformer model*

2.3. Experiment

The experiment in this study uses DETR as the base method for detecting vulgarity in anime. The model training uses the pre-trained model from Facebook Hugging Face. DETR employs a bipartite matching mechanism that utilizes the Hungarian algorithm. In pre-trained model, the bipartite matching calculations are already configured; however, the losses used for this mechanism defined as follows:

1. **Loss ce,** classification loss, which measures the error in predicting the object or target
2. **Loss_bbox,** or bounding box loss, which measures the error in predicting the bounding box compared to the ground truth.
3. **Loss_giou,** or Generalized IoU Loss, which calculates the discrepancy between the predicted and ground truth bounding boxes by evaluating their overlap.

Hungarian matching is a computation mechanism that requires the above-defined losses. These losses are then calculated through the Hungarian matching process to obtain more accurate predictions.

The detailed parameters used listed in Table 3. for architecture parameters and Table 4. for training parameters.

Table 3. *Architecture parameters of DETR model pre-trained*

Parameter	Value
Number Queries	100
Hidden Dimension	256
Number Heads	8
Number Encoder Layer	6
Number Decoder Layer	6
Position Embedding	sine
Dropout	0.1

Jurnal Teknik Informatika Vol. 18 No. 1, April 2025 (157-165)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i1.46064

Table 4. *Training parameters model DETR pre-trained*

Parameter	Value
Learning Rate Backbone	1×10^{-5}
Learning Rate	1×10^{-4}
Batch Size	4
Weight decay	1×10^{-4}
Epoch	100
Optimizer	AdamW
Learning Rate Backbone	1×10^{-5}

The pre-trained DETR model uses the default ResNet-50 backbone. In this study, experiments will be conducted by replacing the DETR backbone with ResNet-101 and comparing the results with ResNet-50. For each ResNet experiment, dilated convolution will be applied to specific stage implement.

3. RESULTS AND DISCUSSION

3.1. Model Training Comparison

Experiments with different backbones were conducted to evaluate the performance of each backbone in the context of the dataset and method used in this study, as well as to determine which backbone yields the best results.

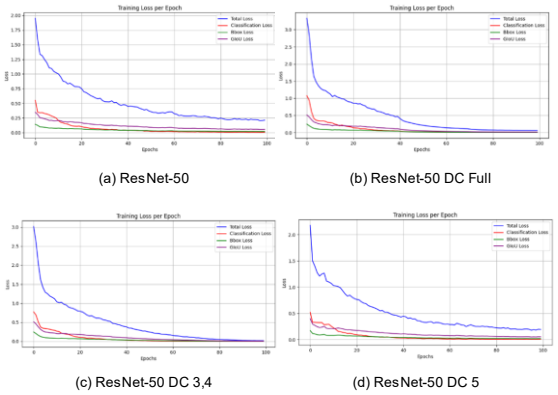

Figure 4. *Training loss ResNet-50 backbone*

The experimental results of the DETR model with the ResNet-50 backbone in four variations, as shown in Fig. 4, demonstrate that the model is capable of learning effectively. In particular, the ResNet-50 versions with dilated convolutions applied to stage 3 and 4, as well as to all stage (full dilation), show a significant decrease in the loss curve. This contrasts with the standard ResNet-50 (without dilation) and the version with dilation applied only to the final stage, which exhibit a more gradual and consistent decline in loss.

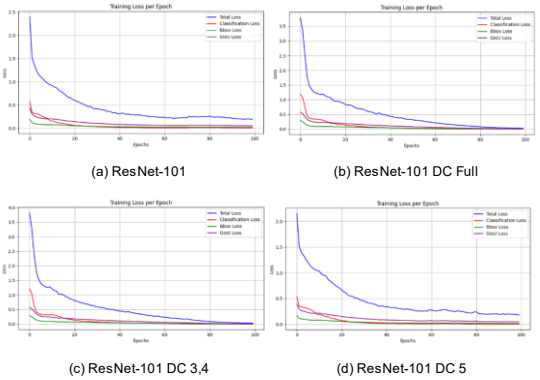

Figure 5. *Training loss ResNet-101 backbone*

The experimental results using the ResNet-101 backbone in four different variations (Fig.5.) also yielded outcomes that were largely consistent with those observed in the ResNet-50 backbone experiments. In this case, applying dilated convolutions to a greater number of stage tended to result in a more significant decrease in training loss compared to models without dilated convolutions or those with dilation applied only to the final stage.

3.2. Evaluation

a. Quantitative

In the quantitative evaluation, several commonly used metrics for assessing DETR-based methods were employed. These include $AP_{50:95}$, which calculates the average precision over IoU thresholds ranging from 50 to 95; AP_{50} and AP_{75}, which represent precision at IoU thresholds of 50 and 75 respectively; AP_s, AP_m, and AP_l, which measure average precision for small, medium, and large objects; as well as AR (Average Recall), which reflects the model's recall performance.

Table 5. *DETR model evaluation for backbone experiments*

Backbone	Dilated	$AP_{50:95}$	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR
ResNet-50	-	0.462	0.877	0.395	0.277	0.471	**0.628**	0.552
	1-4	0.450	0.855	0.455	0.343	0.462	0.526	0.541
	3,4	0.424	0.841	0.375	0.357	0.435	0.354	0.514
	5	**0.479**	0.875	**0.485**	0.299	**0.491**	0.518	0.556
ResNet-101	-	0.452	0.851	0.426	0.360	0.461	0.464	0.535
	1-4	0.449	0.862	0.401	**0.367**	0.458	0.434	0.540
	3,4	0.450	0.859	0.362	0.366	0.466	0.341	0.532
	5	0.477	**0.883**	0.476	0.291	**0.491**	0.553	**0.557**

The evaluation results indicate that ResNet-50 with dilated convolution applied at stage 5 outperforms other variations in three

Jurnal Teknik Informatika Vol. 18 No. 1, April 2025 (157-165)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i1.46064

metrics: $AP_{50:95}$ with a score of 0.479, AP_{75} with a score of 0.485, and AP_m with a score of 0.491. Similarly, ResNet-101 with dilated convolution at stage 5 achieves the highest performance in three metrics: AP_{50} (0.883), AP_m (0.491), and AR (0.557). These results suggest that ResNet-50 with stage 5 dilated convolution performs better in terms of average precision (AP) metrics.

As shown in Table 5, applying dilated convolutions to a greater number of stages results in a noticeably higher AP_s score compared to other backbone configurations. This improvement can be attributed to the ability of dilated convolutions to increase the receptive field without reducing resolution, allowing the model to capture more contextual information around small objects [12]. Dilated convolution stage preserve feature resolution in the backbone stages, helping to retain fine details of small objects that might otherwise be lost.

Backbones with dilated convolutions applied to stages 3 and 4, or all stages tend to perform worse than those with dilation only at stage 5 or without dilation. This is likely due to the loss of important local information when dilation is applied too early. While dilated convolutions expand the receptive field without increasing parameters, they also increase the sampling gap between pixels, leading to less dense feature representations and reduced sensitivity to fine spatial details [13]. As a result, the hierarchical structure of features is disrupted, negatively impacting the detection of medium and large objects. Although AP_s improves due to better context capture for small objects, AP_m, AP_l, and overall mAP scores decrease.

b. Qualitative

The results of the two models with the highest evaluation metrics are presented in Fig.6. and Fig.7.

Figure 6. *Vulgar detection with DETR backbone ResNet-50 dilated convolution stage 5*

Figure 7. *Vulgar detection with DETR backbone ResNet-101 dilated convolution stage 5*

The object detection results of the DETR model with a ResNet-50 (Fig.5.) and ResNet-101 (Fig.6.) backbone using dilated convolution at stage 5 demonstrate highly promising visual outcomes. The predicted bounding boxes accurately localize vulgar objects, with confidence scores reaching 0.99 and 1.00. Despite the relatively modest evaluation metric scores, this model exhibits strong detection capabilities in practice, particularly in terms of object localization and confidence in predictions.

CONCLUSION

The research presents a sophisticated approach to network anomaly detection through advanced machine learning methodologies, demonstrating significant insights into computational cybersecurity strategies. By employing a meticulously preprocessed synthetic dataset comprising 40,000 data rows with 25 unique features, the study effectively addresses critical challenges in network security analysis. The Naive Bayes algorithm emerged as the most outstanding classification technique, achieving an exceptional classification accuracy of 94.8% and consistently outperforming alternative machine learning approaches such as k-Nearest Neighbors and Support Vector Machine.

The comparative analysis of machine learning algorithms revealed nuanced performance variations, with Naive Bayes demonstrating superior capabilities in handling network log/event datasets. The algorithm's effectiveness, despite its fundamental assumption of feature independence, highlights the complexity of anomaly detection in network security contexts. The Stochastic Gradient Descent analysis further enriched the research by identifying critical parameters influencing anomaly detection, such as source ports, attack signatures, and monitoring alerts, thereby

providing a multidimensional understanding of network security dynamics.

Ultimately, the study underscores the evolving landscape of cybersecurity research, emphasizing that no single algorithmic approach provides a universal solution. Instead, the findings advocate for a sophisticated, context-aware methodology that integrates diverse classification techniques and continuous computational refinement. By minimizing false positives and false negatives with only a 5.2% prediction error rate, the research offers a promising framework for early threat detection, potentially mitigating significant potential cybersecurity risks across complex network environments.

REFERENCES

[1] D. Brou, "OpenSIUC Searching for Freedom : An Investigation of Form in Japanese Storytelling and Animation," no. Fall, 2023.

[2] "From the Holy Land to the Homeland : The Impact of Anime Broadcasts on Economic Growth Waseda INstitute of Political EConomy," no. April, 2024.

[3] G. Liu, "Influence of Digital Media Technology on Animation Design," *J. Phys. Conf. Ser.*, vol. 1533, no. 4, 2020, doi: 10.1088/1742-6596/1533/4/042032.

[4] F. W. Paulus, F. Nouri, S. Ohmann, E. Möhler, and C. Popow, "The impact of Internet pornography on children and adolescents: A systematic review," *Encephale*, vol. 50, pp. 649–662, 2024, doi: 10.1016/j.encep.2023.12.004.

[5] Z. Achmad, S. Mardliyah, and H. Pramitha, "The Importance of Parental Control of Teenagers in Wacthing Anime with Pornographic Content on the Internet," vol. 138, no. IcoCSPA 2017, pp. 81–84, 2019, doi: 10.2991/icocspa-17.2018.22.

[6] C. Massaccesi, "Anime: A Critical Introduction, by Rayna Denison," *Alphav. J. Film Screen Media*, no. 17, pp. 259–264, 2019, doi: 10.33178/alpha.17.25.

[7] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 6, pp. 1–20, 2021, doi: 10.1007/s42979-021-00815-1.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12346 LNCS, pp. 213–229, 2020, doi: 10.1007/978-3-030-58452-8_13.

[9] A. Biró, K. T. Jánosi-Rancz, L. Szilágyi, A. I. Cuesta-Vargas, J. Martín-Martín, and S. M. Szilágyi, "Visual Object Detection with DETR to Support Video-Diagnosis Using Conference Tools," *Appl. Sci.*, vol. 12, no. 12, Jun. 2022, doi: 10.3390/app12125977.

[10] L. K. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, "Attention Is All You Need," *Int. Conf. Inf. Knowl. Manag. Proc.*, no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.

[11] R. Padilla, S. L. Netto, and E. A. B. Da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," *Int. Conf. Syst. Signals, Image Process.*, vol. 2020-July, no. July, pp. 237–242, 2020, doi: 10.1109/IWSSIP48289.2020.9145130.

[12] J. Contreras, M. Ceberio, and V. Kreinovich, "Why dilated convolutional neural networks: A proof of their optimality," *Entropy*, vol. 23, no. 6, 2021, doi: 10.3390/e23060767.

[13] B. Liu, F. He, S. Du, J. Li, and W. Liu, "An advanced YOLOv3 method for small object detection," *J. Intell. Fuzzy Syst.*, vol. 45, no. 4, pp. 5807–5819, 2023, doi: 10.3233/JIFS-224530.