VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i4.4717

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

PERFORM COMPARATION OF DEEP LEARNING METHODS IN GENDER CLASSIFICATION FROM FACIAL IMAGES

Yosefina Finsensia Riti [1*]; Ryan Putranda Kristianto[2]; Dionisius Reinaldo A[3]

Information Science[1,2,3]
Universitas Katolik Darma Cendika, Indonesia[1,2,3]
https://siakad.ukdc.ac.id/[1,2,3]
yosefina.riti@ukdc.ac.id[1*], ryan@ukdc.ac.id[2], dionisius.reinaldo@student.ukdc.ac.id[3]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Identifying gender through facial images is a crucial aspect in various life contexts. Biometric technology, such as facial recognition, has become an integral part of various applications, including fraud detection, cybersecurity protection, and consumer behavior analysis. With the advancement of technology and the progress in artificial intelligence, especially through the use of Convolutional Neural Networks (CNNs), computers can now identify gender from facial images with a high level of accuracy. Although there are still some challenges, such as variations in pose, facial expressions, and different lighting conditions, CNNs can overcome these obstacles. This study uses the CelebA dataset, which consists of 122,000 facial images of both men and women. The dataset has been processed to maintain a balanced number of samples for each gender class, resulting in a total of 101,568 samples. The data is divided into training, validation, and test sets, with 80% used for training, and the remaining 20% split between validation and testing. Eight different CNN architectures are applied, including VGG16, VGG19, MobileNetV2, ResNet-50, ResNet-50 V2, Inception V3, Inception ResNet V2, and AlexNet. Although previous research has shown the potential of CNN architectures for various classification tasks, these studies often encounter issues of overfitting on large datasets, which can reduce model accuracy. This study applies dropout techniques and hyperparameter tuning to address overfitting issues and optimize model performance. The training results indicate that ResNet-50, ResNet-50 V2, and Inception V3 achieved the highest accuracy of 98%, while VGG16, VGG19, MobileNetV2, and AlexNet achieved accuracies of 95% and 97%, respectively. Performance evaluation using confusion matrices, precision, recall, and F1-score demonstrates excellent performance.*

Keywords*: Deep Learning, Classification, Gender, CNN.*

Intisari— *Mengenali jenis kelamin melalui citra wajah merupakan aspek yang vital dalam berbagai aspek kehidupan. Teknologi biometrik seperti pengenalan wajah telah menjadi bagian integral dari berbagai aplikasi, termasuk deteksi kecurangan, perlindungan keamanan siber, dan analisis tren perilaku konsumen. Dengan kemajuan teknologi dan perkembangan dalam kecerdasan buatan, khususnya dengan pemanfaatan Convolutional Neural Network (CNN), komputer saat ini memiliki kapabilitas untuk mengidentifikasi jenis kelamin dari citra wajah dengan tingkat akurasi yang tinggi. Meskipun masih terdapat beberapa tantangan, seperti variasi pose, ekspresi wajah, dan kondisi pencahayaan yang beragam, CNN mampu mengatasi kendala-kendala tersebut. Penelitian ini menggunakan dataset CelebA yang terdiri dari 122.000 citra wajah laki-laki dan perempuan. Dataset ini telah diproses untuk menjaga keseimbangan jumlah sampel pada setiap kelas jenis kelamin, sehingga menghasilkan total 101.568 sampel. Data dibagi menjadi data latih, validasi, dan uji, dengan 80% digunakan untuk pelatihan, dan 20% sisanya dibagi untuk validasi dan pengujian. Delapan arsitektur CNN yang berbeda, termasuk VGG16, VGG19, MobileNetV2, ResNet-50, ResNet-50 V2, Inception V3, Inception ResNet V2, dan AlexNet. Meskipun penelitian sebelumnya menunjukkan potensi arsitektur CNN dalam berbagai tugas klasifikasi, sering kali penelitian tersebut menghadapi masalah overfitting pada dataset besar, yang dapat menurunkan akurasi model. Penelitian ini menerapkan teknik dropout dan penyesuaian hyperparameter untuk mengatasi masalah overfitting dan mengoptimalkan kinerja model. Hasil pelatihan*

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.4717

menunjukkan bahwa ResNet-50, ResNet-50 V2, dan Inception V3 mencapai akurasi tertinggi sebesar 98%, sementara VGG16, VGG19, MobileNetV2, dan AlexNet masing-masing mencapai akurasi 95% dan 97%. Evaluasi performa menggunakan matriks kebingungan, presisi, recall, dan F1-score menunjukkan kinerja yang sangat baik.

***Kata Kunci**: Deep Learning, Klasifikasi, Jenis Kelamin, CNN.*

INTRODUCTION

In human life, gender or sex is an important factor in determining the identity of males and females. This identity serves several purposes, such as shaping how individuals interact with others, understanding social roles, social expectations, equality and justice, education, and employment. To distinguish gender, it can be observed through facial image information, as each human face possesses distinct characteristics [1] [2]. The identification of gender from facial features is necessary for certain occupations using biometric technology. Biometric technology enables systems to recognize individuals and verify their identities based on their biological characteristics [3].

With technological advancements, many applications utilize biometric technology in human-computer interactions for various purposes, including fraud detection, cybersecurity, banking, healthcare, and even customizing application behavior based on the user's gender. Furthermore, this technology can be employed to gather data for trend analysis and provide product recommendations tailored to the user's gender.

Identifying gender based on human perception is relatively straightforward, but teaching a computer to see as accurately as humans is a technological challenge. Through the use of Artificial Intelligence (AI) technology, gender identification from facial images can be achieved by the computer through automated feature extraction, analysis, classification, and automatic understanding of information within an image. Within a computer system, facial recognition systems analyze features within an image and compare them with other images.

To perform this feature analysis, extraction and classification are essential. Deep learning technology is a scope of Artificial Intelligence which excels at image classification because it has the ability to recognize patterns and accurate information from unstructured or unlabeled data [6]. Deep Learning enables computers to learn autonomously from provided data using multiple layers within a deep neural network to recognize patterns and make predictions or decisions, such as classification [2]. To perform classification, an appropriate method is required. Data can be categorized into meaningful categories, making it easier to understand and enabling more effective analysis. Gender classification is a binary classification problem, where an individual is predicted as either male or female based on a given image [2].

One of the deep learning algorithms that can be applied for object classification is the Convolutional Neural Network (CNN). Theoretically, this study is grounded in deep learning approaches, particularly the Convolutional Neural Network (CNN), which has been widely applied for visual data classification due to its hierarchical feature extraction capabilities. CNN is an artificial neural network composed of one or more convolutional layers, often accompanied by subsampling layers, as well as one or more fully connected layers similar to those in standard neural networks [2].

CNN has been widely used in numerous studies related to image classification due to its ability to extract spatial features and effectively process visual data. While CNN-based approaches have demonstrated effectiveness in prior research, there remains a theoretical and methodological gap in comparing multiple architectures under consistent, large-scale, and balanced datasets. Most studies either limit their evaluations to a single model or utilize relatively small and imbalanced datasets, limiting generalizability.

Several previous studies have applied CNN in classification tasks and have demonstrated promising performance. Studies conducted by Rajendra et al. [4] predicting the gender of males and females using the CNN algorithm with an accuracy of approximately 70%. Tilki et al [5] Using CNN and the AlexNet architecture in gender classification yielded accuracy results of 92.40% and 90.50%.

However, challenges in facial identification continue to evolve, such as variations in pose, facial expressions, lighting conditions, hairstyles, facial accessories, data quality, data imbalance, and complex backgrounds. Therefore, the approach used in this study is built upon the theory of Convolutional Neural Networks (CNN) and is methodologically strengthened by applying data augmentation techniques and dropout regularization to address overfitting issues and

improve the model's generalization capability. Several previous studies have indeed shown promising results, but they have not comprehensively compared various CNN architectures using a large-scale and balanced dataset that considers diverse facial variations under real-world conditions. Although previous studies have demonstrated the potential of CNN architectures, they often encounter overfitting issues with large datasets, which can reduce model accuracy.

To address this problem, this research applies dropout techniques and data augmentation using methods such as rotation, shifting, cropping, and horizontal flipping. These techniques aim to reduce overfitting and improve model accuracy. Another innovation of this research is the use of a larger dataset, totaling 101,568 images, with balanced data for each gender.

Additionally, by comparing various CNN architectures, this research provides insights into the most optimal CNN architecture for addressing challenges posed by diverse facial data, including different lighting conditions, poses, and facial attributes. The application of optimized hyperparameters, such as learning rate, batch size, and the Adam optimizer, significantly enhances model convergence and training efficiency, demonstrating adaptability in handling complex and diverse datasets.

MATERIALS AND METHODS

A. Dataset

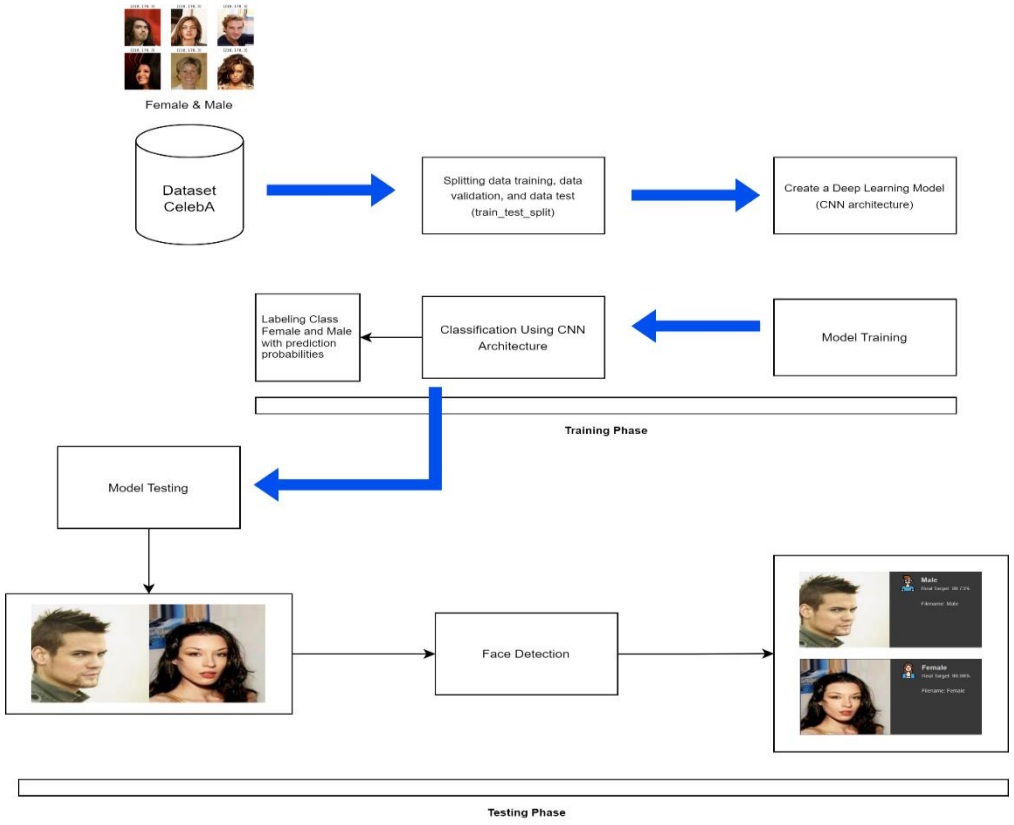
Data collection in the form of facial images is obtained from the Kaggle database [6]. The dataset comprises a total of 122,000 facial images, divided into two classes: 50,784 images of males and 71,216 images of females. There are 41 facial attributes, including eyes, eyebrows, mustache, facial shape, and so on [7].

B. Implementation Phase

1. Data Preparation
   Fetching, extracting, and displaying the CelebA dataset from the Kaggle data source using Python for subsequent processing and analysis in the next stage [8].
2. Data Preprocessing
   The extracted data is unbalanced and random, thus requiring a balancing process to ensure an equal number of datasets for each gender class, male and female. The balancing process is discharge by aligning the dataset with the class that has the fewest number of datasets. In this case, there are 50,784 datasets for males and 71,216 datasets for females, so the number of datasets in the female class will be reduced to match the number of datasets in the male class, which is 50,784 thus, the total dataset used in this study amounts to 101,568 samples. Additionally, a dataset check is needed to avoid redundant data that could lead to overfitting during model training. If duplicate data is detected during the check, the data will be joined into a single dataset [7].
3. Data Processing
   In this stage, the data is separated into training data, validation data, and test data [9]. In this study, the dataset was divided such that 80% (81,254 samples) were used as the training set to train the model, and the remaining 20% (20,314 samples) were designated as the testing data. From this testing data portion, 33% (6,704 samples) were then allocated as validation data, while the remaining 67% (13,610 samples) continued to be used as testing data. The training data is necessary to train the model used in this research. The model will learn to recognize patterns and relationships between attributes and the desired target output. Validation data is required to measure the model's performance during the training process. This process can help monitor whether the model is experiencing overfitting (memorizing training data) or underfitting (unable to follow the data patterns). On the other hand, testing data is used to assess the final performance of the model after the training process.
4. Algorithm Implementation
   In this stage, it is divided into several steps, including architecture modeling, training, evaluation, and data testing. The first step involves deep learning modeling using CNN architecture. After modeling, the training process is conducted to train the model to classify gender based on facial images. The next step is model evaluation using a confusion matrix to calculate accuracy, precision, recall, and F1-score. In the data testing stage, the model is tested to determine the accuracy level of predictions compared to the actual data.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.4717



Source: (Research Results, 2024)

Figure 1. Data Management Phases

C. Deep Learning Architecture

1. VGG-16

VGG-16 is a CNN architecture consisting of 6 layers. The model undergoes 5 convolution processes with a size of 3x3. Based on testing with the ImageNet database, which comprises 1000 classes and 1.3 million image samples, it achieved an accuracy of 92.7% in the top-five accuracy [10], [11]. The model training phase consists of 13 convolutional layers, 3 fully-connected layers, 5 max-pooling layers, along with ReLU and softmax activation layers. The architecture comes with pre-trained weight models on the ImageNet dataset, enabling the use of transfer learning during model training, which saves time and training resources [12] .

2. VGG-19

VGG-19 is a CNN architecture consisting of 16 convolutional layers and 3 fully-connected layers. It utilizes 3x3 filters with a stride and padding of 1 [13]. VGG-19 is a CNN architecture with the highest number of layers and depth, which helps reduce parameters. This model has undergone training on specific portions of the ImageNet27 dataset, which was used in the large-scale ImageNet Visual Recognition Challenge (ILSVRC). During training, VGG-19 processed over a million images and was able to classify these images into one of the 1000 diverse object categories. As a result of this training process, the model successfully developed rich feature representations for various types of images. Similar to VGG-16, this architecture comes with pre-trained weights, enabling the application of transfer learning, which saves training time and resources.

3. MobilenetV2

MobileNetV2 is a CNN architecture designed for mobile and embedded systems with resource constraints. [14]. This architecture utilizes depth-wise separable convolutions to reduce the number of parameters. Depth-wise convolutions gather spatial features separately, significantly reducing the required number of parameters. By reducing the number of parameters, the number of computational operations decreases, making it suitable for devices with limited resources [15], [16]. This architecture is designed for efficiency, yet it maintains a good level of accuracy. It also comes with pre-trained weights, making it possible to apply transfer learning to save training resources and time.

4. Alexnet
   AlexNet is another CNN algorithm that won the ImageNet competition in 2012. This architecture comprises 8 convolutional layers, pooling layers, local normalization, and fully connected layers . AlexNet is a large network composed of approximately 650,000 neurons and has 60 million parameters [17]. AlexNet is known for its use of the Rectified Linear Unit (ReLU) as an activation function and the application of dropout techniques to reduce overfitting. AlexNet has more filters in each layer, and the stacked convolutional layers consist of a total of 25 layers.

5. Resnet
   ResNet, short for Residual Network, is another CNN architecture. This architecture uses residual blocks that allow for deeper architectures without suffering from the degradation in performance that can occur in traditional deep networks [11]. ResNet-50 has 50 layers, including convolutional layers, pooling layers, fully connected layers, and residual blocks. [13], [18]. The layers in ResNet-50 consist of convolutional layers, pooling layers, and fully connected (fc) layers, along with the addition of skip connections in some convolutional layers. Towards the end, ResNet-50 employs the softmax function as the activation function. The use of shortcut connections helps address the gradient vanishing/exploding problem that often occurs in very deep networks. This makes training large-scale models more efficient and manageable.

6. GoogleNet
   GoogleNet or Inception is a CNN architecture created by Google. Inception v1 is the architecture used in GoogleNet, which was first introduced by the Google team. GoogleNet is the name given to the architecture that utilizes the Inception module. . In this architecture, the Inception module is utilized, which consists of convolutions with filters of various sizes run in parallel. This approach allows GoogleNet to efficiently extract features from data and diminish the number of parameters required in the architecture. By using this approach, GoogLeNet can achieve high accuracy levels despite having a more compact architecture. Inception-V3 is an update of the Inception-V1 model. To enhance the model's adaptability, Inception-V3 employs various methods to optimize its network [11].

7. Inception ResNet V2
   It is a CNN network architecture that combines features from two different architectures, namely Inception (GoogleNet) and ResNet. The Inception module is the result of combining various types of convolutions, pooling layers, and all feature maps, which are merged into a single vector in the output section [19]. In general, this module uses 5×5, 3×3, and 1×1 filters to extract both local and global features from image data. [9]. Meanwhile, ResNet is renowned for its shortcut connections, which efficiently gather feature information from the previous layers to the next layers. This strengthens these features and ultimately improves accuracy. On the other hand, Inception-ResNet-v2 demonstrates superior performance compared to other networks.

D. Model Evaluation

Model performance measurement is crucial as it can depict the performance of a model in classifying the available data . In this research, the measurement method used is the confusion matrix. The Confusion Matrix is a machine learning tool for evaluating a model that contains information about the model's predictions and the actual data[20]. In measuring with the confusion matrix, there are four (4) terms that represent the results of the model's classification. True Positive (TP) is when positive data is correctly predicted as positive. On the other hand, if negative data is incorrectly predicted as positive, it falls into the category of False Positive (FP). True Negative (TN) is when negative data is correctly predicted as negative. If positive data is incorrectly detected as negative by the model, it is classified as False Negative (FN).

In this research, the male class represents positive data, and the female class represents negative data. True Positive occurs when actual data is male and is correctly predicted as male by the model. True Negative occurs when actual data is female and is correctly predicted as female by the model. False Positive is when actual data is female, but the model predicts it as male. Conversely, if actual data is male but predicted as female by the model, it falls under False Negative. These values are then represented as shown in the following table.

Table 1. *Confusion Matrix*

Object	Classification	
	Positive	Negative
Positive	TP	FN
Negative	FP	FN

Source: (Research Results, 2024)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.4717

The Confusion Matrix is an important aspect for calculating evaluation metrics such as Accuracy, Precision, Recall (Sensitivity), F1 Score, and Support. Accuracy describes how often the model provides correct predictions for both classes, positive as well as negative [6]. The accuracy value is calculated using the following equation (1) .

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

Precision aims to measure the model's performance in correctly classifying positive data, and Equation (2) is used to calculate precision. .

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

Meanwhile, Recall (Sensitivity) aims to measure the model's performance in detecting all actual positive data. , As found in Equation (2).

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

To measure the balance between precision and recall, you can use the F1-Score as shown in Equation (3) below. .

$$F1 - score = 2\frac{(Precision*Recall)}{(Precision+Recall)} \quad (5)$$

Furthermore, Support refers to the number of instances in each class used to calculate Precision, Recall, and F1-Score. This aims to understand the class distribution within the dataset, which is crucial for analyzing evaluation results by considering the number of significant instances. It can help in analyzing whether the evaluation results reflect the majority of classes present or only a small portion of them.

RESULTS AND DISCUSSION

This study utilizes the CelebA dataset, which encompasses a wide range of facial attributes, including 202,599 images that display features such as black hair, blonde hair, and arched eyebrows. However, for this research, the dataset used consists of 101,568 images. The majority of the dataset comprises images of celebrities from around the world. Nevertheless, the unique facial characteristics of Indonesians, especially those who wear hijabs, may not be fully represented in this dataset. Despite this, the results of this research provide insights into constructing an accurate gender classification model.

This research compares eight different CNN models for gender classification from facial images, including VGG-16, VGG-19, ResNet50, ResNet50 V2, Inception ResNet V2, MobileNet V2, Inception V3, and AlexNet. Based on experimental results, the Resnet50 V2, Inception V3, and Inception ResNet V2 models showed the most superior performance, each achieving 98% in precision, recall, and F1 score. Compared to a similar study by Janahiraman T on gender classification from Asian facial images using deep learning, which compared VGG-16, Resnet 50, and Mobilenet algorithms and found that VGG-16 had the highest accuracy of 88%. Although this research achieved high accuracy levels above 90%, the VGG-16 model still could not match the accuracy of the Resnet50 V2, Inception V3, and Inception ResNet V2 models. This difference can be explained by several factors such as the difference in dataset size. Janahiraman's study used a total of 1000 images with 500 male and 500 female images, whereas this study used a more varied facial image dataset with 101,568 images, consisting of 50,784 male and 50,784 female images.

Additionally, there was a difference in the optimizers used, Janahiraman's study used SGD (Stochastic Gradient Descent), updating the model weights based on the gradient of the loss function with a fixed learning rate. Although effective for large and simple datasets, SGD may be less flexible than Adam in handling high complexity and variation in datasets. In contrast, this study used the Adam optimizer with the ability to adaptively adjust the learning rate for each parameter [21], which proved very effective in enhancing model convergence and training efficiency in this study scenario. Furthermore, the Adam optimizer was very helpful in addressing challenges such as limited dataset size and sample variation with its flexibility in handling diverse and complex datasets.

A. Training Model

The training was performed for the eight (8) models with a training data size of 81,254, 15 epochs, and a batch size of 128 [37], [38]. The training was carried out using Google Colab, running a Python program on a device with specifications of 12GB RAM and a 15GB T4 GPU . Additionally, Adam was used as the optimizer to control parameters, and the minimum learning rate was set to 0.001. The validation process, involving 6,704 validation samples, is conducted in approximately 52 steps per epoch. This allows for a comprehensive and efficient evaluation of the model without overburdening the memory. The results from this training reveal significant insights regarding the model's performance and affirm that

the approach used in this study is capable of handling large datasets very effectively.

Based on the research results, it was found that the ResNet-50, ResNet-50 V2, and Inception V3 architectures achieved the highest training accuracy, reaching 98%. Furthermore, MobileNet V2 achieved a training accuracy of 97%, while AlexNet reached 95%, VGG-16 reached 94%, and VGG-19 reached 93%, as listed in Table 2 below. Subsequently, testing was conducted on the trained models to assess their ability to classify images using a different dataset from the training dataset.

Table 2. Result Training Model Architecture CNN

Model	Image Size	Epochs	Number of Parameters	Training Model Accuracy	*Loss*
VGG-16	218 x 178	15	18,731,490	94%	0,1366
VGG-19	218 x 178	15	24,041,186	93%	0,1575
Resnet 50	218 x 178	15	111,950,114	98%	0,0559
Resnet 50 V2	218 x 178	15	111,927,202	98%	0,047
Inception Resnet V2	218 x 178	15	86,076,034	98%	0,0465
Mobilenet V2	218 x 178	15	2,592,578	97%	0,0548
Inception V3	218 x 178	15	23,851,784	98%	0,0369
Alexnet	218 x 178	15	159,220,482	95%	0,1167

Source: (Research Results, 2024)

B. Model Evaluation

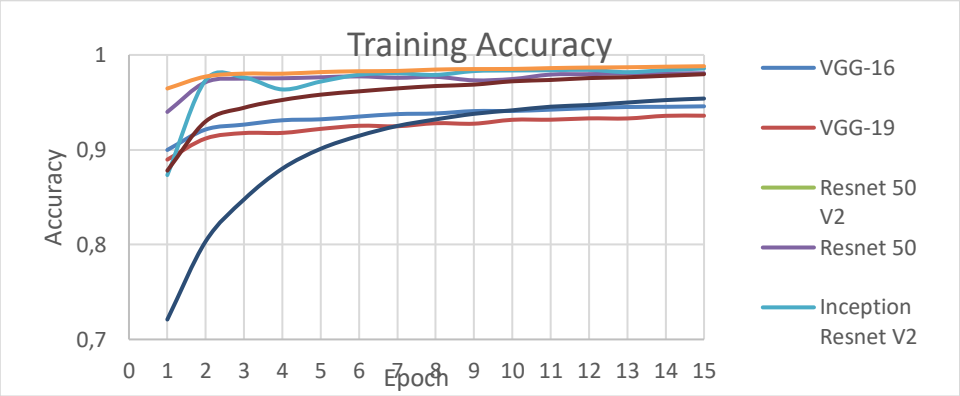
Based on the training results of the model with 15 epochs using input data from the training data, it resulted in visualizations of the total loss decreasing and the accuracy graph increasing as the iterations progressed during the training process . This is shown in Figure 2 and Figure 3. Figure 2 shows the loss graph during the training phase of the model, where the loss value refers to the degree of deviation between the model's predictions and the actual data. A decrease in the loss graph indicates that the model is learning and improving in making accurate predictions. Meanwhile, Figure 3 displays the accuracy graph, with accuracy referring to the proportion of correct predictions for both male and female classes. An increase in the accuracy graph signifies that the model's performance in classifying male and female classes is becoming better and more precise. After the training process, the model was then tested using 13,610 test data. Subsequently, a confusion matrix calculation was performed, producing values as shown in Tables 3, 4, 5, 6, 7, 8, 9, and 10 below

Source: (Research Results, 2024)

Figure 2. CNN Architecture Training Loss Graph

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.4717



Source: (Research Results, 2024)

Figure 3. CNN Architecture Training Accuracy Graph

Table 3. Result of *Confusion Matrix for All* Model

Model		Positive	Negative
VGG-16	Positive	6394	411
	Negative	232	6573
VGG-19	Positive	6521	284
	Negative	444	6361
Resnet 50	Positive	6464	315
	Negative	119	6712
Resnet 50 V2	Positive	6706	99
	Negative	121	6684
Inception Resnet V2	Positive	6757	48
	Negative	212	6593
Mobilenet V2	Positive	6649	156
	Negative	197	6608
Inception V3	Positive	6682	123
	Negative	97	6708
Alexnet	Positive	6529	276
	Negative	197	6608

Source: (Research Results, 2024)

Based on the results of the confusion matrix calculations, The models presented in this study demonstrate remarkable capability in classifying images based on gender, with all models achieving evaluation scores above 90%. This underscores their effectiveness and accuracy in classification tasks. For instance, the confusion matrix analysis of the VGG-16 model indicates a very high level of precision. This model recorded a precision of 95% in identifying both male and female gender images, signifying significant consistency in gender classification by the VGG-16 model. Furthermore, when tested for its recall ability, the results were similar. The VGG-16 showed the capability to re-identify male gender images with 95% accuracy and achieved the same result for female gender images. This recall accuracy, also at 95%, reinforces the effectiveness of the VGG-16. Thus, based on the precision and recall measurements, the VGG-16 model demonstrates high and stable average accuracy, at 95%, in gender classification. This consistent performance proves the reliability of the VGG-16 model in handling gender classification tasks with high accuracy. The precision, recall, F1-score, and support values for the model testing are as follows:

Table 4. Model Calculation Results

Model	*Precision*	*recall*	*F1-score*	*Support*
VGG-16	95%	95%	95%	95%
VGG-19	95%	95%	95%	95%
Resnet 50	97%	97%	97%	97%
Resnet 50 V2	98%	98%	98%	98%
Inception Resnet V2	98%	98%	98%	98%
Mobilenet V2	97%	97%	97%	97%
Inception V3	98%	98%	98%	98%
Alexnet	97%	97%	97%	97%

Source: (Research Results, 2024)

C. Data Prediction

While the majority of models perform well in classifying gender based on facial images, it's worth noting that, at a low level, the trained models also make prediction errors in some cases. In this research, if actual female data is detected as male, it can be classified as False Positive. Conversely, if actual male data is predicted as female, it can be classified as False Negative. The table below displays prediction errors in the CNN architecture models that have been trained.

Table 5. Wrong Data Prediction

Model	FP		FN	
VGG-16				

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i4.4717

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

Table 5. Wrong Data Prediction (Continue)

Model	FP		FN	
VGG-19				
ResNet-50				
ResNet-50V2				
Inception-Resnet V2				
Mobilenet V2				
Inception V3				
Alexnet				

Source: (Research Results, 2024)

CONCLUSION

The research results of the models implemented with TensorFlow and the Convolutional Neural Network (CNN) structure indicate excellent performance in image classification. The precision, recall, and F1-score levels are consistently high across all tested models. The main challenge faced in this research is the dataset's limitation, which may not fully reflect the diversity of Indonesian faces, especially those wearing hijabs. Nevertheless, this study provides important insights into developing more accurate gender classification models.

The novelty of this research lies in the in-depth comparison of eight different CNN models with a larger and more diverse dataset. Additionally, the use of the Adam optimizer has proven highly effective in improving model convergence and training efficiency, demonstrating flexibility in handling complex and diverse datasets. Architectures such as ResNet-50, ResNet-50 V2, Inception ResNet V2, and Inception V3, executed using TensorFlow, achieved the highest training accuracy, reaching up to 98%. This reflects their outstanding ability to recognize patterns and features in the training dataset. MobileNet V2 also achieved a high training accuracy of 97%, demonstrating excellent performance. VGG-16, VGG-19, and AlexNet showed an increase in accuracy as the model testing progressed, indicating good stability and a lack of overfitting issues.

After the training phase, the models were tested using a dataset consisting of 13,610 data points. The results, expressed through a confusion matrix, yielded metrics for precision, recall, F1-score, and support, consistently showing high performance. This uniformity indicates that these models have achieved a good balance between correctly identifying positive class instances and detecting all actual positive class examples.

REFERENCE

[1] M. Galih Pradana and H. Khoirunnisa, "Analisis Performa Algoritma Convolutional Neural Networks Menggunakan Arsitektur Lenet Dan Vgg16," *Indonesian Journal of Business Intelligence*, vol. 6, no. 2, 2023, doi: 10.21927/ijubi.v6i2.3765.

[2] R. Firdaus, Joni Satria, and B. Baidarus, "Klasifikasi Jenis Kelamin Berdasarkan Gambar Mata Menggunakan Algoritma Convolutional Neural Network (CNN)," *Jurnal CoSciTech (Computer Science and*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.4717

Information Technology), vol. 3, no. 3, pp. 267–273, Dec. 2022, doi: 10.37859/coscitech.v3i3.4360.

[3] N. Adisaputra Sinaga, R. Rosnelly, K. kunci-CNN, and J. Kelamin, "Analisis Penggunaan Model EfficientNetV2 Dalam Memprediksi Jenis Kelamin Pada Wajah Pengguna Masker," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 9, no. 3, 2022, [Online]. Available: http://jurnal.mdp.ac.id

[4] G. Rajendra, K. Sumanth, C. Anjali, G. Pardhasai, and M. Supraja, "Gender Prediction using Deep Learning Algorithms and Model on Images of an Individual," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Aug. 2021. doi: 10.1088/1742-6596/1998/1/012014.

[5] S. Tilki *et al.*, "Gender Classification using Deep Learning Techniques Healthy-Heart-an intelligent personalised visual model for risk analysis and lifestyle learning to prevent Cardiovascular Diseases View project Gender Classification using Deep Learning Techniques," 2021. [Online]. Available: https://www.researchgate.net/publication/351564423

[6] Y. Finsensia Riti and S. S. Tandjung, "Klasifikasi Covid-19 Pada Citra CT Scans Paru-Paru Menggunakan Metode Convolution Neural Network," *Jurnal Ilmiah Komputer*, vol. 18, 2022.

[7] G. George, S. Adeshina, and M. Mahamat Boukar, "International Journal of Intelligent Systems And Applications In Engineering Development of Android Application for Facial Age Group Classification Using TensorFlow Lite," *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE*, vol. 2023, no. 4, pp. 11–17, 2023, [Online]. Available: www.ijisae.org

[8] L. Goyal, A. Dhull, A. Singh, S. Kukreja, and K. K. Singh, "VGG-COVIDNet: A Novel model for COVID detection from X-Ray and CT Scan images," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 1926–1935. doi: 10.1016/j.procs.2023.01.169.

[9] H. Zhang *et al.*, "Recurrence Plot-Based Approach for Cardiac Arrhythmia Classification Using Inception-ResNet-v2," *Front Physiol*, vol. 12, May 2021, doi: 10.3389/fphys.2021.648950.

[10] A. I. Mansour and S. S. Abu-Naser, "Age and Gender Classification Using Deep Learning-VGG16," 2022. [Online]. Available: www.ijeais.org/ijaisr

[11] M. Mujahid, F. Rustam, R. Álvarez, J. Luis Vidal Mazón, I. de la T. Díez, and I. Ashraf, "Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network," *Diagnostics*, vol. 12, no. 5, May 2022, doi: 10.3390/diagnostics12051280.

[12] C. Sitaula and M. B. Hossain, "Attention-based VGG-16 model for COVID-19 chest X-ray image classification," *Applied Intelligence*, vol. 51, no. 5, pp. 2850–2863, May 2021, doi: 10.1007/s10489-020-02055-x.

[13] A. Victor Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 375–381, Nov. 2021, doi: 10.1016/j.gltp.2021.08.027.

[14] W. Hastomo, E. Hadiyanto, and D. Sutarno, "Klasifikasi Covid-19 Chest X-Ray Dengan Tiga Arsitektur Cnn (Resnet-152, Inceptionresnet-V2, Mobilenet-V2)," 2021.

[15] G. Jin, Y. Liu, P. Qin, R. Hong, T. Xu, and G. Lu, "An End-to-End Steel Surface Classification Approach Based on EDCGAN and MobileNet V2," *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23041953.

[16] P. N. Srinivasu, J. G. Sivasai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with mobilenet v2 and lstm," *Sensors*, vol. 21, no. 8, Apr. 2021, doi: 10.3390/s21082852.

[17] K. M. Hosny, M. A. Kassem, and M. M. Fouad, "Classification of Skin Lesions into Seven Classes Using Transfer Learning with AlexNet," *J Digit Imaging*, vol. 33, no. 5, pp. 1325–1334, Oct. 2020, doi: 10.1007/s10278-020-00371-9.

[18] Q. A. Al-Haija and A. Adebanjo, "Breast cancer diagnosis in histopathological images using ResNet-50 convolutional neural network," in *IEMTRONICS 2020 - International IOT, Electronics and Mechatronics Conference, Proceedings*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020. doi: 10.1109/IEMTRONICS51293.2020.9216455.

[19] Y. Chen *et al.*, "Classification of lungs infected COVID-19 images based on

inception-ResNet," *Comput Methods Programs Biomed*, vol. 225, Oct. 2022, doi: 10.1016/j.cmpb.2022.107053.

[20] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf Sci (N Y)*, vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.

[21] D. Irfan and T. Surya Gunawan, "Comparison Of SGD, Rmsprop, And Adam Optimation In Animal Classification Using CNNs," 2023.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri