Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

Enhancing Diagnostic Accuracy of Polycystic Ovary Syndrome Classification in Ultrasound Images Using a Hybrid Deep Learning Model of VGG16 and AlexNet

Hj. Maisarah*[1], M. Arief Soeleman[2], Pujiono[3], Iqbal Firdaus[4], Gusti Aditya Aromatica Firdaus[5]

[1,2,3]Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia
[4,5]Information Technology, Institut Bisnis dan Teknologi Kalimantan, Banjarmasin, Indonesia

Email: [1]maisarah@ibitek.ac.id

Received : Jun 23, 2025; Revised : Oct 25, 2025; Accepted : Nov 13, 2025; Published : Apr 15, 2026

Abstract

Diagnosis of Polycystic Ovary Syndrome (PCOS) using ultrasound (USG) imaging still faces a major challenge in the form of inter-observer variability, which can lead to inconsistent diagnostic outcomes and increase the risk of misclassification. This limitation highlights the urgent need for an automated artificial intelligence (AI)–based system capable of performing ultrasound image classification with greater objectivity, accuracy, and consistency. This study aims to develop an automated PCOS classification model based on a hybrid Convolutional Neural Network (CNN) architecture that integrates VGG16 and AlexNet through a feature concatenation mechanism, following preprocessing and data augmentation steps to enhance model generalization. The model's performance was evaluated using accuracy, precision, recall, F1-score, and specificity as key metrics. Experimental results demonstrate that the VGG16–AlexNet hybrid model achieved the best performance, with an accuracy of 98.26%, precision of 97.90%, recall of 97.90%, F1-score of 97.90%, and specificity of 98.52%. These results outperform other hybrid configurations such as VGG16–MobileNetV2, VGG16–ResNet50, and VGG16–InceptionV3, each of which achieved accuracies above 96%. These findings confirm that combining the feature depth of VGG16 with the computational efficiency of AlexNet enables more comprehensive extraction of spatial and textural patterns in ultrasound images. Consequently, the proposed hybrid model offers a promising AI-driven diagnostic support system that not only enhances the accuracy of PCOS detection but also assists clinicians in making faster, more objective, and consistent medical decisions.

Keywords : *Convolutional Neural Network (CNN), Deep Learning, Hybrid CNN, Polycystic Ovary Syndrome (PCOS), Ultrasound Imaging, VGG16-AlexNet*

1. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrine disorders in women of reproductive age, with a global prevalence estimated at 8–20% depending on diagnostic criteria. The true prevalence may be higher because many cases remain undiagnosed [1], [2], [3]. PCOS is characterized by three primary components: ovulatory dysfunction, hyperandrogenism, and polycystic ovarian morphology (PCOM), which can be observed through ultrasound imaging (US) [4], [5], [6]. Genetic factors, insulin resistance, and low-grade chronic inflammation are believed to play significant roles in PCOS pathogenesis [7], [8]. Clinically, the condition often manifests as irregular menstruation, acne, hirsutism, abdominal obesity, and scalp hair loss. In the long term, PCOS increases the risk of type 2 diabetes, hypertension, cardiovascular disease, and endometrial cancer [9], [10].

PCOS diagnosis generally follows the Rotterdam criteria, which require at least two of the following: oligo/anovulation, clinical or biochemical hyperandrogenism, and polycystic ovarian morphology detected via ultrasound [11]. However, ultrasound examination results are highly operator-

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

dependent, often leading to variations in interpretation and inconsistent diagnosis. This limitation underscores the need for artificial intelligence (AI)-based diagnostic systems to improve accuracy and objectivity [12].

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable effectiveness in medical image analysis, including breast cancer detection [13], pulmonary disease classification [14], and liver organ segmentation [15]. In the context of PCOS, several CNN-based approaches have been proposed for ultrasound image analysis. For example, the PCONet model achieved over 96% accuracy when compared to popular architectures such as VGG and ResNet [16]. Similarly, F-Net, which integrates YOLOv8-based follicle detection with CNN and GLCM features, reached an accuracy of 97.5% [17]. Another study introduced CsyNet, a multilevel thresholding and ensemble classifier, which achieved 97.5% accuracy along with high sensitivity, specificity, and AUC [18]. Comparative studies of pretrained CNN architectures such as AlexNet, VGG16, ResNet50, and InceptionV3 have also reported accuracies around 95% for ovarian ultrasound image classification [19]. Despite these promising results, most approaches remain limited to single architectures or require complex segmentation steps, which increase computational costs and reduce efficiency [20].

Among widely used CNN architectures, VGG16 is known for its ability to extract deep spatial features, while AlexNet offers the advantage of being lightweight and computationally efficient [21], [22]. Combining these two networks could leverage the feature depth of VGG16 and the processing efficiency of AlexNet. Such hybrid CNN approaches have successfully improved performance in other medical imaging domains, including liver ultrasound classification [23], COVID-19 detection [24], cervical cancer screening [25], lung image detection and classification [26], and breast cancer classification [27]. These studies indicate that integrating CNN architectures can achieve higher accuracy and better generalization compared to single-model approaches.

To date, no study has directly combined VGG16 and AlexNet for PCOS ultrasound image classification, despite the potential of such a hybrid model to overcome the limitations of single architectures, simplify training, and improve diagnostic accuracy on real-world data [16], [28], [23], [27]. Therefore, this study proposes a hybrid VGG16–AlexNet model for PCOS classification based on ultrasound images and compares its performance with other combinations such as VGG16–MobileNetV2, VGG16–ResNet50, and VGG16–InceptionV3. Evaluation results show that the VGG16–AlexNet hybrid achieved the highest performance, reaching an accuracy of 98.26%, outperforming previous approaches [17], [16]. These findings highlight the strong potential of hybrid CNNs to enhance classification accuracy while offering a practical and efficient solution to support clinical decision-making in PCOS diagnosis.

2. METHOD

This study employs a quantitative experimental approach using a computational deep learning method to classify ovarian ultrasound images into two categories: infected and non-infected. The primary objective is to compare the performance of several VGG16-based hybrid models in order to identify the best architecture with the highest classification accuracy. The proposed research workflow is illustrated in Figure 1, which outlines the sequential stages of data preprocessing, model training, and evaluation.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

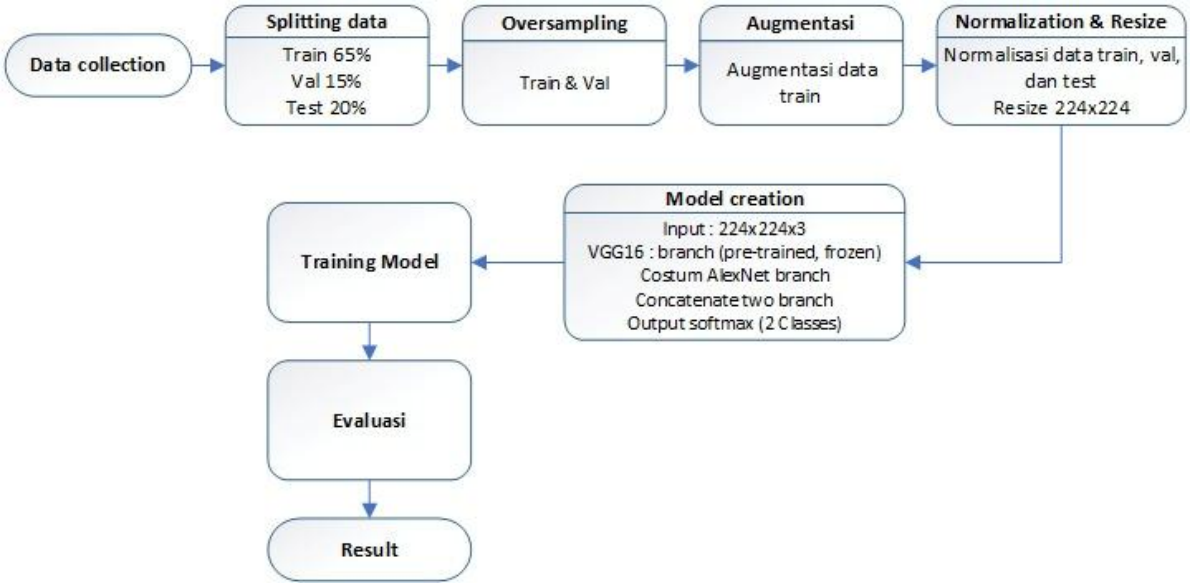
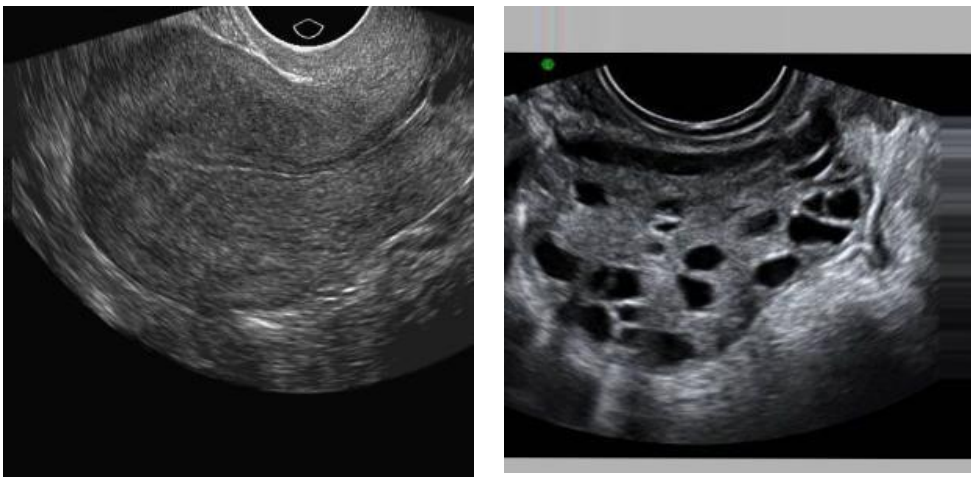


Figure 1. Flowchart of the Proposed Method

2.1. Dataset

The dataset used in this study was collected from two publicly available sources on the Kaggle platform, namely the Kaggle Delhi Ultrasound Dataset [29] and the PCOS-XAI Ultrasound Dataset [30]. These datasets were combined to obtain a total of 13,798 ovarian ultrasound images. Annotation of the images was carried out by a team of radiologists and gynecologists to ensure label validity, with only images approved by at least two independent medical experts retained for analysis. Inclusion criteria required images to have adequate resolution, clear visual quality, and verified medical diagnoses, while images that were blurred, visually degraded, or lacked expert consensus were excluded. All data were stored in JPEG format with varying resolutions and were randomly split into 65% training, 15% validation, and 20% testing sets using a random shuffle procedure to maintain class balance. The directory structure was organized according to the requirements of the Keras ImageDataGenerator for efficient model training. Representative samples from each class, illustrating the visual diversity and quality of the dataset, are presented in Figure 2, which shows examples of both infected and non-infected ovarian ultrasound images used in this study.



noninfected infected

Figure 2. Examples of images in the dataset

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

2.2. Data Preprocessing

Preprocessing was conducted to enhance the quality of the data before training the model. One of the key initial steps was class balancing, as the original dataset exhibited an uneven distribution of images between the infected and non-infected categories. To address this issue, oversampling was applied to the minority class, specifically within the training and validation subsets. This process involved randomly duplicating minority-class images until the number of samples across classes became balanced [31]. Although oversampling is effective for mitigating class imbalance, it carries the potential risk of overfitting due to repeated samples. To counter this, additional strategies such as data augmentation and dropout regularization were implemented to prevent the model from memorizing duplicated images. The initial class distribution prior to oversampling is illustrated in Figure 3, which clearly shows the dominance of the majority class, while the balanced distribution achieved after oversampling is presented in Figure 4, demonstrating the effectiveness of the applied balancing procedure.

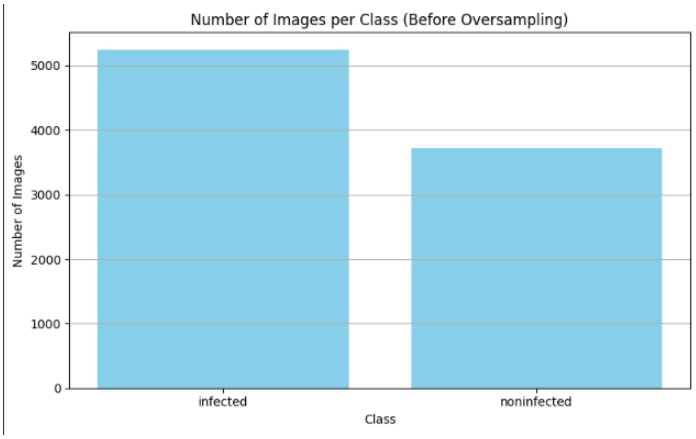


Figure 3. Class distribution before oversampling on infected and non-infected datasets

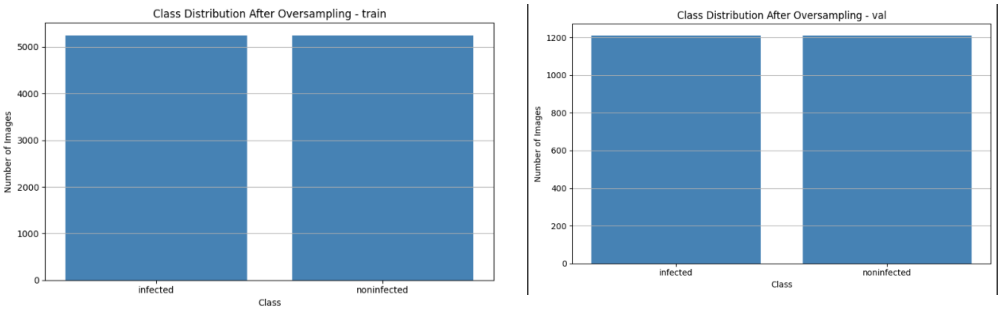


Figure 4. Distribution of train and val classes after oversampling on infected and non-infected datasets

The next preprocessing steps involved data augmentation and normalization to further enhance the dataset quality and improve model generalization. All ultrasound images were first resized to 224 × 224 pixels to match the standard input dimensions of the CNN architectures used in this study, such as VGG16 and AlexNet. Following resizing, pixel-value normalization was performed by dividing each RGB pixel by 255, thereby scaling the data to the range of [0, 1]. This normalization step helps accelerate model convergence during training and reduces numerical irregularities that may affect gradient stability [33]. For the training subset, data augmentation was applied to introduce greater variability and to mitigate the risk of overfitting caused by the limited dataset size. Augmentation techniques included random rotation up to 20 degrees, horizontal and vertical translations of up to 10%, and horizontal flipping. All augmentation processes were executed randomly using the ImageDataGenerator utility in

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

TensorFlow, ensuring that transformations were applied exclusively to the training subset. This approach follows the recommendations outlined in Data Augmentation: A Comprehensive Survey of Modern Approaches [32], which emphasize that augmentation should be confined to training data to improve generalization while maintaining the validation and test subsets in their original form (with only normalization) for unbiased evaluation. The augmentation workflow is illustrated in Figure 5, providing a visual representation of the sequence of transformations applied to the training data.

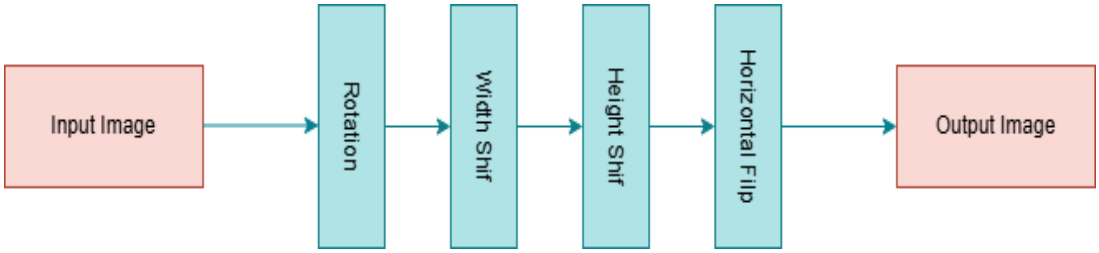


Figure 5. Flowchart of image augmentation process on training data subset

An example of the augmented images is presented in Figure 6, illustrating how the applied transformations increase dataset variability while preserving the essential morphological characteristics of the ovaries. Through this approach, the model can be trained on a broader range of image variations without losing critical diagnostic features. Subsequently, the image labels were transformed using one-hot encoding to enable the model to produce probabilistic predictions across the two target classes.

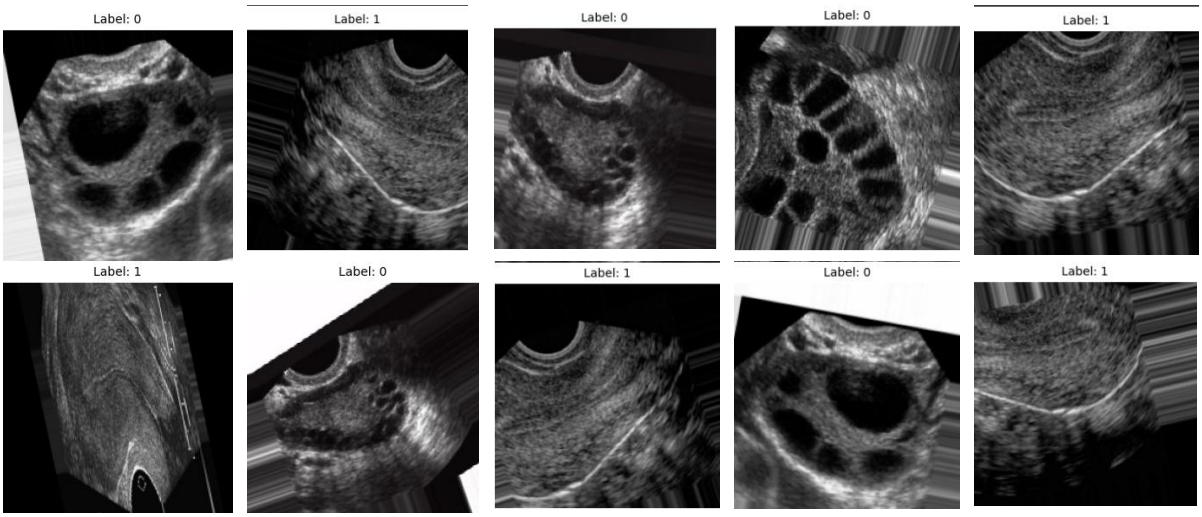


Figure 6. The resulting image after augmentation

2.3. Model Architecture

The model architecture employed in this study is a hybrid Convolutional Neural Network (CNN) that combines two popular architectures, VGG16 and AlexNet. VGG16 was selected as the primary feature extractor due to its strong ability to capture fine-grained details in medical images [21]. All convolutional layers of VGG16, pretrained on the ImageNet dataset, were frozen to retain their initial weights. The output was then processed through a Global Average Pooling layer, followed by a Dense layer with 128 units and ReLU activation, along with a 30% Dropout to prevent overfitting [33].

AlexNet was chosen as the complementary branch because of its relatively lightweight architecture, use of varying kernel sizes, and capability to extract spatial features distinct from VGG16

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

[22]. This combination was expected to produce more comprehensive feature representations. The outputs from both branches were fused using feature concatenation [34], followed by a Dense layer with 128 units and ReLU activation, a 50% Dropout, and finally a Dense Softmax layer with two neurons for binary classification into infected and non-infected classes.

The baseline and hybrid architectures used in this study are illustrated in Figure 7 and Figure 8, respectively. In the baseline models (Figure 8), each network functions as a single feature extractor either VGG16 with pretrained weights or a custom AlexNet which is then connected to a fully connected layer for binary classification. In contrast, the hybrid architecture (Figure 7) merges the outputs of VGG16 and AlexNet via feature concatenation before passing them to the fully connected layers. This approach leverages the deep spatial feature extraction of VGG16 alongside the parameter efficiency of AlexNet, resulting in richer feature representations and improved performance compared to individual baseline models.

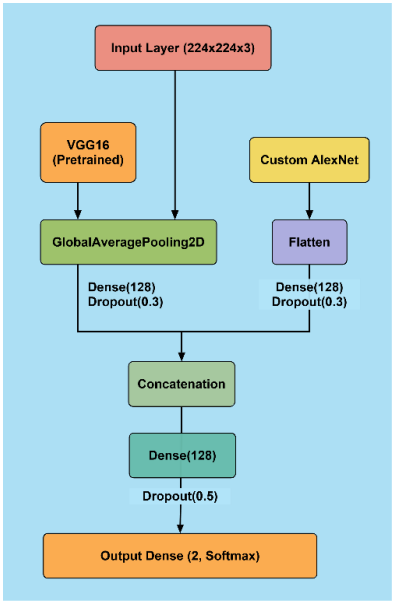


Figure 7. Hybrid Convolutional Neural Network (CNN) architectural model (VGG16 and AlexNet)

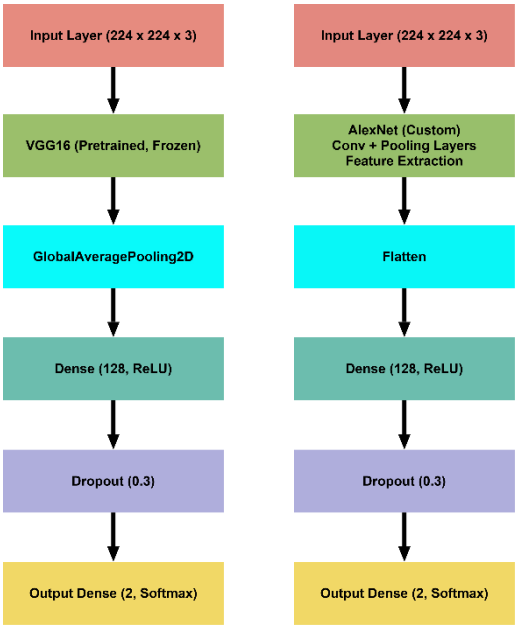


Figure 8. Baseline Convolutional Neural Network (CNN) model architecture

2.4. Training Parameters

The model was compiled using the Adam optimization algorithm with an initial learning rate of 1×10^{-4}, a batch size of 32, and a maximum of 30 epochs. The binary crossentropy loss function was employed to match the binary classification objective (infected vs. noninfected). Adam was chosen due to its capability to handle sparse gradients and accelerate convergence in deep neural networks, as demonstrated in the development of Adam variants for neural network optimization [35]. To prevent overfitting, an EarlyStopping strategy was applied by monitoring the validation loss, while ReduceLROnPlateau was used to reduce the learning rate by 50% whenever the validation loss failed to improve for three consecutive epochs. The best-performing model was automatically saved using ModelCheckpoint based on the lowest validation loss. This optimization strategy has been proven effective in enhancing convergence stability and classification efficiency in medical image analysis [36], [37]. The experimental setup is summarized in Table 1, while the computing environment used for training included Google Colab Pro+ with an NVIDIA Tesla T4 GPU (16 GB VRAM) and a 25 GB system RAM, ensuring adequate resources for deep learning model training.

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

2.5. Model Evaluation

Model evaluation was performed on a test set that was completely separated from the training and validation data to ensure objective performance assessment. The evaluation employed five key metrics accuracy, precision, recall (sensitivity), F1-score, and specificity all of which were calculated based on the confusion matrix using four fundamental components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These metrics are widely adopted in medical image classification studies because they provide a comprehensive understanding of model performance, capturing not only overall accuracy but also the balance between type I and type II errors [38], [39]. The formulas for each metric are presented in Table 1.

Table 1. Evaluation metric calculation formula

Metrics	Formula
Accuracy	$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$ (1)
Precision	$Precision = \frac{TP}{TP+FP}$ (2)
Recall	$Recall = \frac{TP}{TP+FN}$ (3)
F1 score	$F1\ score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$ (4)
Specificity	$Specificity = \frac{TN}{TN+FP}$ (5)

Furthermore, to validate the superiority of the proposed hybrid approach over the baseline models, a statistical test was performed using F1-score values on the validation dataset. A paired t-test was applied to compare the performance between models, and the results revealed a statistically significant difference ($p < 0.05$). This finding confirms that the hybrid model provides a substantial improvement over the baseline architectures.

From an ethical perspective, although the datasets used in this study are publicly available, all medical ultrasound images were fully anonymized by the dataset providers and therefore contain no patient-identifiable information. The annotation process followed established medical research ethical standards and was verified by qualified domain experts. Since this study did not involve direct data collection from patients, it does not raise any additional ethical concerns.

3. RESULT

3.1. Dataset and Preprocessing

The combined dataset used in this study consisted of 13,798 ovarian ultrasound images, categorized into two classes: infected and noninfected. After data splitting, 8,968 images (65%) were allocated for training, 2,070 images (15%) for validation, and 2,760 images (20%) for testing. The initial distribution revealed a class imbalance, prompting the application of oversampling on the training and validation subsets. After oversampling, the training subset contained 8,969 images evenly distributed between the infected and noninfected classes, as shown in Figure 4, while the validation subset comprised 2,072 images, also balanced across both classes.

Representative ultrasound images from each class are presented in Figure 2, whereas the results of the augmentation process are illustrated in Figure 6. The augmentation procedures including rotation,

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

horizontal/vertical flipping, and zooming were carefully applied to preserve the morphological characteristics of the ovary, enabling the model to train on a more diverse dataset without losing critical clinical information.

3.2. Visualization of Model Training Results

The training processes of the four hybrid models were visualized using accuracy and loss curves for both the training and validation sets. Figures 9 to 12 present the performance of each hybrid combination, namely VGG16–MobileNetV2, VGG16–InceptionV3, VGG16–ResNet50, and VGG16–AlexNet, respectively. Each graph illustrates the dynamic changes in training and validation accuracy as well as loss over the course of 30 epochs, allowing evaluation of convergence behavior and the potential occurrence of overfitting in each model.

As shown in Figure 9, the VGG16–MobileNetV2 hybrid demonstrates a steady increase in accuracy up to approximately the 15th epoch, after which the curve begins to plateau. The validation accuracy reaches 97.76%, while the training accuracy stabilizes around 95%, indicating strong generalization capability. The validation loss remains lower than the training loss, suggesting that the applied regularization and augmentation techniques effectively prevented overfitting. Furthermore, the small gap between the training and validation loss confirms that the model did not suffer from underfitting, highlighting the stability and efficiency of the VGG16–MobileNetV2 hybrid architecture during the training process.

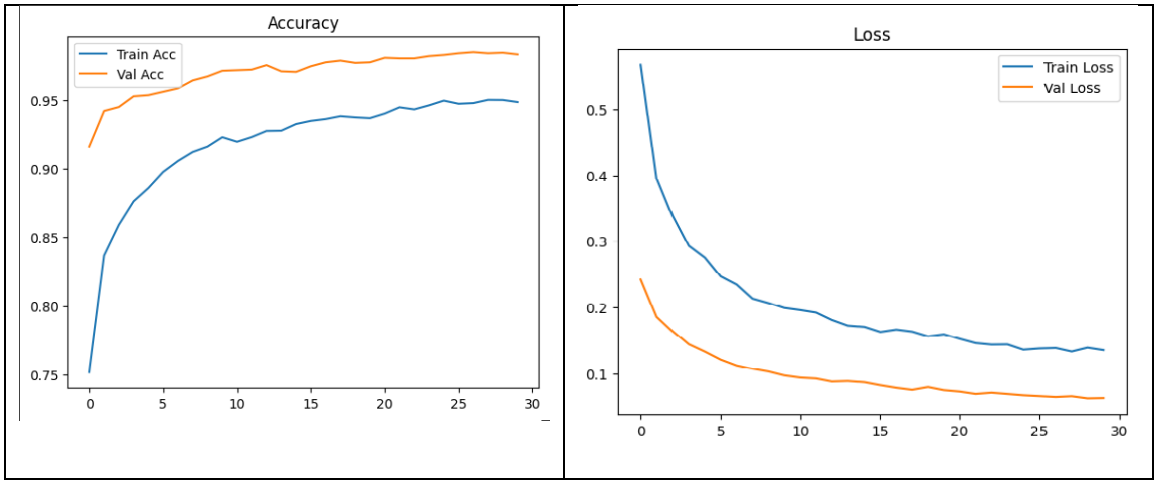


Figure 9. Accuracy and loss graph of the hybrid VGG16 and MobileNetV2 models

As illustrated in Figure 10, the VGG16–InceptionV3 hybrid model exhibits a significant improvement in training accuracy, approaching 96%, while the validation accuracy peaks at approximately 96.81% around the 20th epoch. The stable validation accuracy trend closely follows the training curve, confirming the model's strong generalization capability. Both training and validation loss decrease consistently, with the most notable reduction occurring during the first 10 epochs and only a minimal gap between the two curves. These results indicate that the applied augmentation and regularization strategies effectively prevented overfitting while maintaining stable learning. Overall, although the model achieves a high level of performance, it remains slightly behind the best-performing hybrids in this study.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

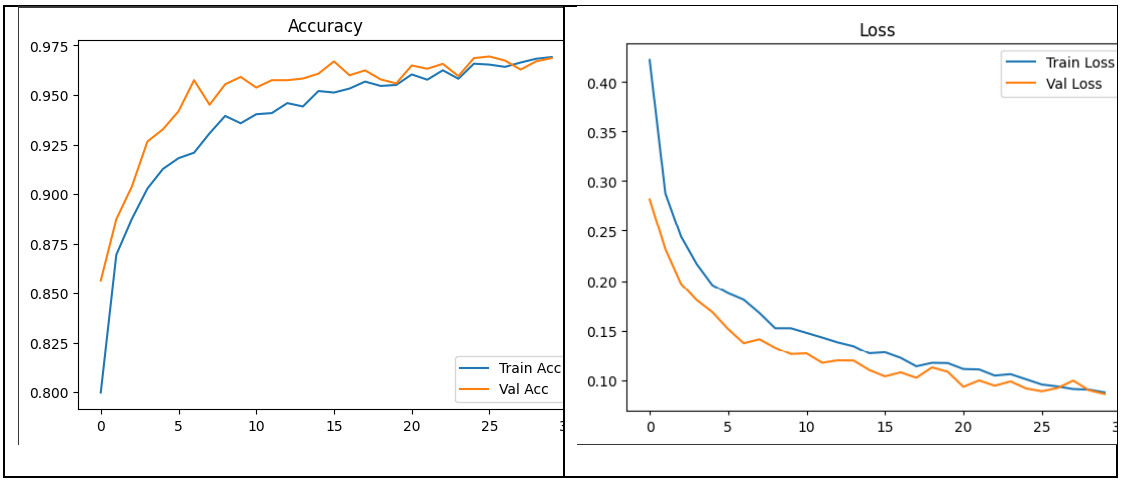


Figure 10. Accuracy and loss graph of hybrid VGG16 and InceptionV3 models

Figure 11 presents the training results of the VGG16–ResNet50 hybrid, which shows a sharp increase in accuracy during the early epochs and stabilizes after approximately the 10th epoch. The validation accuracy reaches 97.28%, with training accuracy nearing 97%, demonstrating excellent generalization. The loss curves for both subsets decline consistently, with validation loss slightly lower than training loss. This pattern confirms that the model experiences neither overfitting nor underfitting, and that the VGG16–ResNet50 hybrid achieves stable, convergent, and optimal training behavior for ovarian ultrasound image classification.

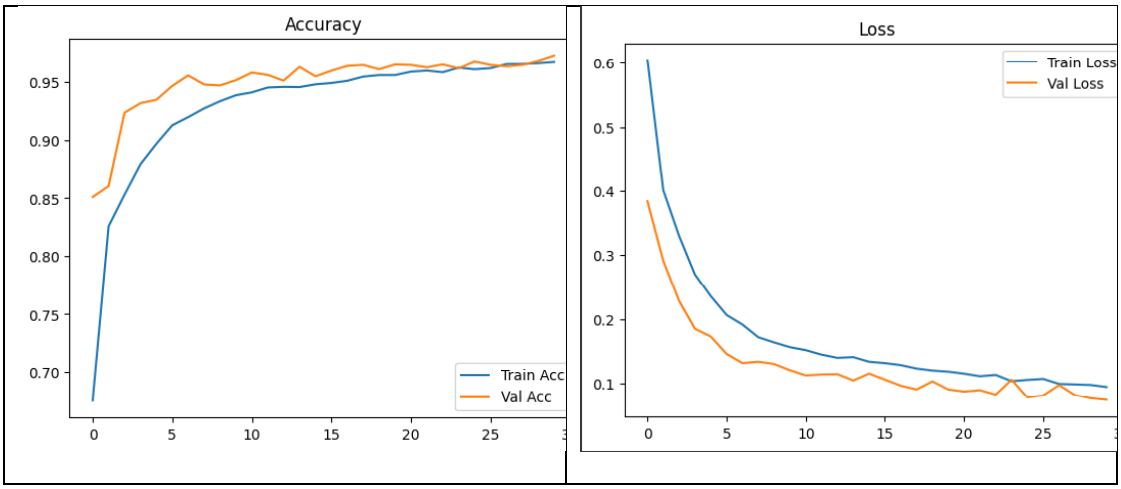


Figure 11. Accuracy and loss graph of the hybrid VGG16 and ResNet50 model

As shown in Figure 12, the VGG16–AlexNet hybrid delivers the strongest training performance among all models. Training accuracy steadily increases from approximately 78% to over 97%, while the validation accuracy peaks at 98.26%, reflecting outstanding generalization ability. The training and validation loss curves decrease steadily, with validation loss slightly lower than training loss, further confirming the absence of overfitting or underfitting. This superior performance demonstrates that combining the deep feature extraction of VGG16 with the computational efficiency of AlexNet successfully enhances both accuracy and efficiency in classifying ovarian ultrasound images.

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

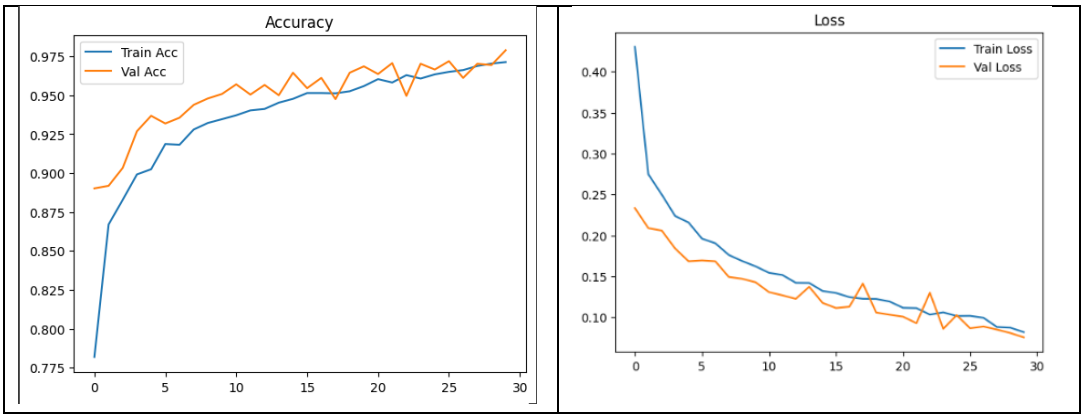


Figure 12. Accuracy and loss graph of the hybrid VGG16 and AlexNet model

3.3. Model Performance Evaluation

To quantitatively evaluate model performance, confusion matrices and several key metrics including accuracy, precision, recall, F1-score, and specificity were used. All metrics were calculated using the independent test set, which was never involved in training or validation, ensuring that the reported results reflect the true generalization ability of each model.

3.3.1. Confusion Matrix

The confusion matrix provides a detailed breakdown of correct and incorrect predictions for each class, allowing assessment of the models' discriminative power. The evaluation results confirm that all hybrid CNN models achieved high and consistent prediction accuracy, demonstrating that integrating two CNN architectures can significantly enhance the classification of ovarian ultrasound images.

As shown in Figure 13, the VGG16–MobileNetV2 hybrid correctly classified 1,584 infected and 1,116 noninfected images, with only 33 false negatives and 29 false positives. This low error rate indicates strong accuracy and a well-balanced trade-off between sensitivity and specificity.

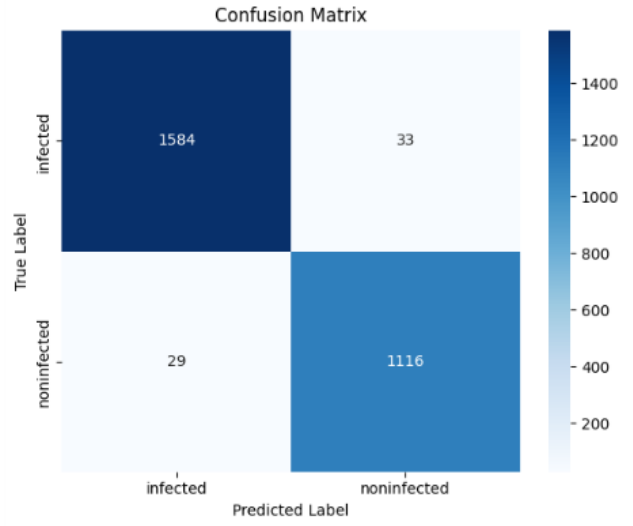


Figure 13. Confusion Matrix VGG16 and MobileNetV2

Figure 14 presents the confusion matrix for the VGG16–InceptionV3 hybrid, which correctly classified 1,562 infected and 1,112 noninfected images. However, it produced 55 false negatives and 33 false positives, resulting in slightly lower performance compared to the other hybrid combinations.

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

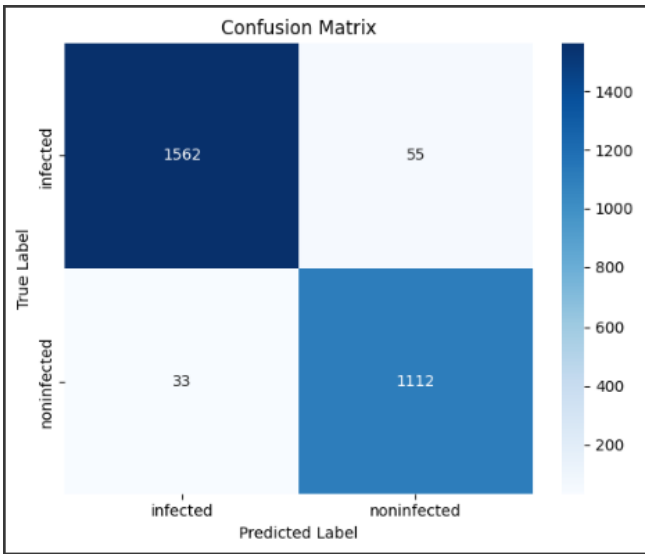


Figure 14. Confusion Matrix VGG16 and InceptionV3

The results for the VGG16–ResNet50 hybrid are displayed in Figure 15, where the model correctly classified 1,581 infected and 1,106 noninfected images. Although the number of errors remained relatively low (36 false negatives and 39 false positives), its overall accuracy was slightly below that of the top-performing model

.

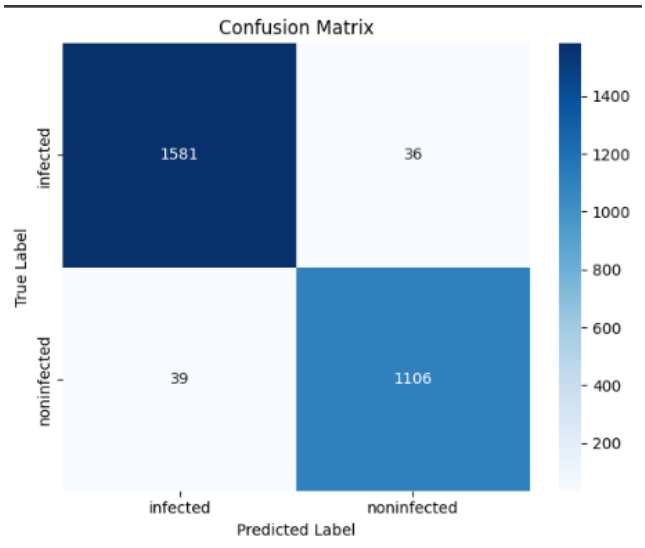


Figure 15. Confusion Matrix VGG16 and ResNet50

Finally, Figure 16 highlights the superior performance of the VGG16–AlexNet hybrid. This model achieved the highest accuracy by correctly classifying 1,339 infected and 973 noninfected images, with only 19 false negatives and 27 false positives. These results underscore the advantage of combining VGG16's deep spatial feature extraction with AlexNet's computational efficiency, enabling a balanced optimization of both precision and recall.

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

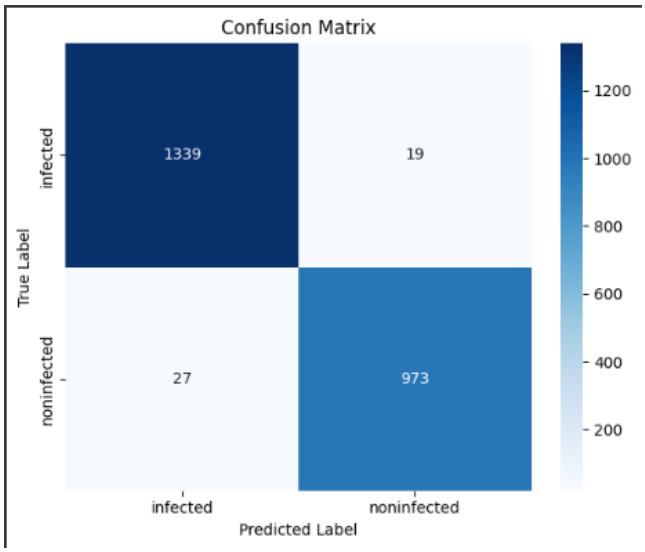


Figure 16. Confusion Matrix VGG16 and AlexNet

3.3.2. Comparison of Classification Metrics

The quantitative evaluation across all hybrid models is summarized in Table 2, which compares accuracy, precision, recall, specificity, and F1-score. Among the four combinations, the VGG16–AlexNet hybrid achieved the best overall performance, recording an accuracy of 98.26%, precision of 97.90%, recall of 97.90%, specificity of 98.52%, and F1-score of 97.90%.

The VGG16–MobileNetV2 and VGG16–ResNet50 hybrids followed closely, with accuracies of 97.76% and 97.28%, respectively, indicating competitive performance though slightly below that of VGG16–AlexNet. In contrast, the VGG16–InceptionV3 hybrid yielded the lowest metrics, reaching an accuracy of 96.81%, precision of 95.29%, recall of 97.12%, specificity of 96.60%, and an F1-score of 96.19%.

To verify whether these performance differences were statistically significant, an ANOVA test was performed using the F1-scores from each cross-validation fold. The one-way ANOVA revealed a significant difference among models ($p < 0.05$). Subsequent paired t-tests showed that the performance of VGG16–AlexNet differed significantly from VGG16–InceptionV3 ($p < 0.05$). However, no significant differences were found between VGG16–AlexNet and the VGG16–MobileNetV2 or VGG16–ResNet50 hybrids ($p > 0.05$).

Table 2. Comparison of performance parameters for different networks

Classifier	Accuracy	Precision	Recall	Specificity	F1 score
VGG16 & MobileNetV2	97,76	97,13	97,47	97,96	97,30
VGG16 & InceptionV3	96,81	95,29	97,12	96,60	96,19
VGG16 & ResNet50	97,28	96,85	96,59	97,77	96,72
VGG16 & AlexNet*	98,26	97,90	97,90	98,52	97,90

These findings confirm that, while all hybrid architectures deliver high classification accuracy, the VGG16–AlexNet combination provides a notable improvement over VGG16–InceptionV3 and remains statistically comparable to the other two high-performing hybrids.

3.4. Analysis of Misclassification

Although the proposed hybrid models achieved high overall performance, a small number of misclassifications were still observed. Detailed analysis revealed that false negatives (FN) generally

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

occurred in infected ultrasound images with low resolution and poor contrast, where characteristic features such as multiple cysts were difficult for the model to detect. Conversely, false positives (FP) were mostly found in non-infected images containing patterns that resembled abnormal ovarian structures, causing the model to incorrectly interpret normal tissue textures as indicators of PCOS.

Representative misclassified cases are shown in Figure 17. On the left, a non-infected ovary image was predicted as infected (false positive) due to shadow-like patterns that mimic cystic regions. On the right, an infected ovary image was predicted as non-infected (false negative), likely caused by low image resolution or cyst distributions that lacked sufficient contrast with surrounding tissue.

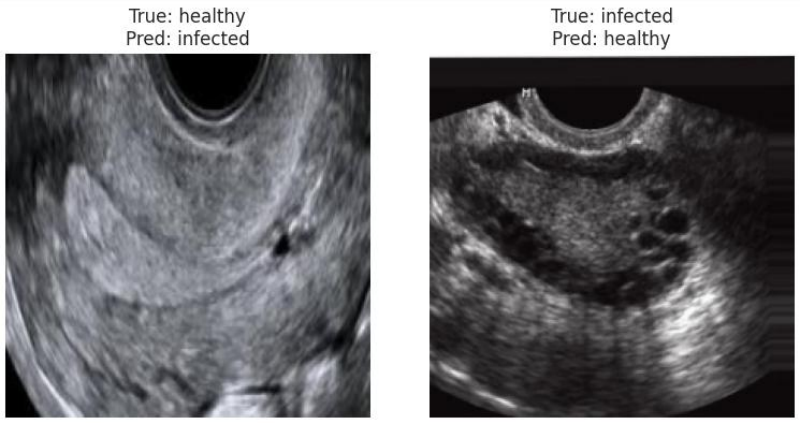


Figure 17. Example of model misclassification case: (left) non-infected image predicted as infected (false positive), (right) infected image predicted as non-infected (false negative).

This analysis highlights the strong influence of ultrasound image quality on model performance. Imaging artifacts, low contrast, or cyst-like textures can trigger classification errors. Therefore, implementing advanced preprocessing techniques such as adaptive contrast enhancement, edge sharpening, or denoising could be an important step to reduce misclassification rates and further improve the reliability of the hybrid CNN models.

4. DISCUSSIONS

The results of this study demonstrate that the hybrid VGG16–AlexNet model achieved the best performance in ovarian ultrasound image classification for the detection of polycystic ovary syndrome (PCOS), attaining an accuracy of 98.26%, precision of 97.90%, recall of 97.90%, and specificity of 98.52%. This performance slightly surpasses PCONet, as reported in [16], which achieved an accuracy of 98.12%, yielding an improvement of 0.14% despite employing a lighter architecture. Such an increase, although seemingly small, is meaningful in medical image classification, where even marginal improvements can significantly affect clinical outcomes. Moreover, AlexNet requires lower computational resources compared to deeper architectures such as ResNet or Inception, making it well-suited for clinical environments with limited computational infrastructure.

These findings are consistent with [19], which demonstrated that hybrid CNNs can improve the sensitivity of PCOS detection. Comparative results from other studies further support this conclusion. For instance, [41] reported a ResNet50 model achieving an accuracy of 97.02%, but at the cost of significantly longer training time. Similarly, [42] employed InceptionV3 for medical image classification and obtained an accuracy of 96.85%, which is comparable to the 96.81% achieved by the VGG16–InceptionV3 combination in this study. Susanto et al. [43] reported an accuracy of 94.2% using MobileNetV2 for medical image tasks, such as COVID-19 detection from CT scans. This aligns with the present findings, in which the VGG16–MobileNetV2 hybrid achieved an accuracy of 97.76% in

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

ovarian ultrasound classification. Furthermore, [44] highlighted that hybrid models consistently yield 1–2% higher accuracy than single-architecture networks, reinforcing the evidence that architectural combinations can improve both accuracy and computational efficiency.

From a computer science perspective, this research contributes by demonstrating that combining a deep model (VGG16) with a lightweight model (AlexNet) provides an optimal balance between accuracy and computational efficiency. This design is particularly relevant for edge computing applications and clinical systems operating under resource constraints, where large models such as ResNet50 or InceptionV3 are difficult to deploy in real time.

The performance differences among the models can be explained by their architectural complexity. InceptionV3, with its multi-branch structure, is advantageous for large-scale datasets but may introduce redundant feature extraction and increase overfitting risk on smaller datasets. ResNet50, although capable of mitigating network degradation through deep residual blocks, imposes a higher computational burden without offering significant performance gains on medium-sized datasets. In contrast, the VGG16–AlexNet hybrid achieves a balance between deep spatial feature extraction and computational efficiency, enabling high accuracy with lower resource requirements outperforming more complex combinations in this study.

From a clinical informatics perspective, the proposed model holds significant potential for practical deployment. Integration into a Picture Archiving and Communication System (PACS) would allow automatic predictions immediately after ultrasound images are uploaded. In hospitals with limited access to radiology specialists, the model could function as a second reader, assisting general practitioners in preliminary diagnosis before expert verification. With near–99% accuracy and low computational cost, this system could accelerate diagnosis, reduce radiologists' workload, and improve access to healthcare in underserved regions.

However, this study has several limitations. First, the dataset remains relatively small and homogeneous, which may introduce demographic bias and limit the model's ability to generalize across different ultrasound devices or patient populations. Second, the analysis relies solely on single-modality imaging data, without incorporating complementary information such as electronic medical records, laboratory tests, or hormonal profiles, which could enrich the classification process. Third, the interpretability of the model's predictions has not been fully explored, making it difficult to transparently explain classification decisions to medical professionals.

Future research could address these limitations through several key directions. (1) The integration of explainable AI (XAI) methods, such as Grad-CAM, to improve model transparency and clinical trust. (2) Multicenter validation using larger and more diverse datasets to enhance model generalizability. (3) Incorporation into clinical decision-support systems connected to Electronic Health Records (EHR), allowing predictions to be combined with patient histories and laboratory data. (4) Exploration of multimodal approaches, such as combining ultrasound images with medical records or hormonal test results, to provide a more comprehensive diagnostic framework.

5. CONCLUSION

This study successfully developed and evaluated several hybrid CNN architectures for the detection of polycystic ovary syndrome (PCOS) using ultrasound images. Among the four tested combinations, the VGG16–AlexNet model achieved the best performance, with an accuracy of 98.26%, precision of 97.90%, recall of 97.90%, F1-score of 97.90%, and specificity of 98.52%. These results demonstrate that combining the deep feature extraction capability of VGG16 with the computational efficiency of AlexNet produces a classification system that is not only highly accurate but also lightweight and resource-efficient. The implications of these findings extend beyond medical computer vision to the field of reproductive health, where early detection of PCOS is crucial as it is one of the

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

leading causes of infertility in women of reproductive age. The high recall and specificity values highlight the model's potential as an AI-based diagnostic support tool, capable of reducing misclassification risks, accelerating diagnosis, and improving healthcare accessibility particularly in facilities with limited specialist resources.

Despite these promising outcomes, the study has several limitations that warrant consideration. The limited dataset size and potential bias in the data sources may affect the model's generalizability, and real-world clinical validation has not yet been conducted. Future research should therefore focus on external validation using multi-center datasets from diverse healthcare facilities to ensure consistent performance in real clinical settings. Furthermore, integrating multimodal data such as ultrasound images, hormonal biomarkers, and patient medical histories could enhance diagnostic accuracy by providing richer clinical context. The development of explainable AI (XAI) approaches is also critical to ensure that model predictions are transparent and interpretable by healthcare professionals. Additionally, exploring integration with ovarian segmentation techniques may further enrich the spatial analysis of ultrasound images.

Equally important, deploying the model within Electronic Health Record (EHR)-based decision support systems could provide immediate clinical benefits by allowing automated classification results to be seamlessly incorporated into diagnostic workflows. In summary, this research demonstrates that a hybrid CNN approach, particularly the VGG16–AlexNet combination, represents a promising solution for AI-assisted PCOS diagnosis. With further development, the proposed model has the potential to make a significant contribution to medical informatics, enhance early detection, and support women's reproductive health in the digital era.

REFERENCES

[1] R. Deswal, V. Narwal, A. Dang, and C. S. Pundir, "The Prevalence of Polycystic Ovary Syndrome: A Brief Systematic Review," *J. Hum. Reprod. Sci.*, vol. 13, no. 4, pp. 261–271, 12 2020, doi: 10.4103/jhrs.JHRS_95_18.

[2] H. J. Teede *et al.*, "Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome," *Fertil. Steril.*, vol. 110, no. 3, pp. 364–379, Aug. 2018, doi: 10.1016/j.fertnstert.2018.05.004.

[3] K. M. Jakubowska-Kowal, K. J. Skrzynska, and A. M. Gawlik-Starzyk, "Prevalence and diagnosis of polycystic ovary syndrome (PCOS) in adolescents what's new in 2023? Systematic review," *Ginekol. Pol.*, vol. 95, pp. 643–649, 2024, doi: 10.5603/gpl.98849.

[4] Y. Hasegawa *et al.*, "Impact of the difference in diagnostic criteria for adolescent polycystic ovary syndrome excluding polycystic ovarian morphology," *J. Obstet. Gynaecol. Res.*, vol. 50, no. 8, pp. 1289–1294, Aug. 2024, doi: 10.1111/jog.15975.

[5] A. Ghafari, M. Maftoohi, M. E. Samarin, S. Barani, M. Banimohammad, and R. Samie, "The last update on polycystic ovary syndrome(PCOS), diagnosis criteria, and novel treatment," *Endocr. Metab. Sci.*, vol. 17, p. 100228, Mar. 2025, doi: 10.1016/j.endmts.2025.100228.

[6] J. P. Christ and M. I. Cedars, "Current Guidelines for Diagnosing PCOS," *Diagnostics*, vol. 13, no. 6, p. 1113, Mar. 2023, doi: 10.3390/diagnostics13061113.

[7] D. Fahs, D. Salloum, M. Nasrallah, and G. Ghazeeri, "Polycystic Ovary Syndrome: Pathophysiology and Controversies in Diagnosis," *Diagnostics*, vol. 13, no. 9, p. 1559, Apr. 2023, doi: 10.3390/diagnostics13091559.

[8] T. Baba, "Polycystic ovary syndrome: Criteria, phenotypes, race and ethnicity," *John Wiley Sons Aust. Ltd Behalf Jpn. Soc. Reprod. Med.*, p. 12, Jan. 2025, doi: 10.1002/rmb2.12630.

[9] J. Parker, C. O'Brien, C. Yeoh, F. Gersh, and S. Brennecke, "Reducing the Risk of Pre-Eclampsia in Women with Polycystic Ovary Syndrome Using a Combination of Pregnancy Screening, Lifestyle, and Medical Management Strategies," *J. Clin. Med.*, vol. 13, no. 6, p. 1774, Mar. 2024, doi: 10.3390/jcm13061774.

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

[10] G. E. Colombo, S. Pirotta, and A. Sabag, "Diet and Exercise in the Management of Polycystic Ovary Syndrome: Practical Considerations for Person-Centered Care," *Semin. Reprod. Med.*, vol. 41, no. 01/02, pp. 026–036, Mar. 2023, doi: 10.1055/s-0043-1777116.

[11] The Rotterdam ESHRE/ASRM-sponsored PCOS consensus workshop group, "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS)," *Hum. Reprod.*, vol. 19, no. 1, pp. 41–47, Jan. 2004, doi: 10.1093/humrep/deh098.

[12] D. J. Patel, K. Chaudhari, N. Acharya, D. Shrivastava, and S. Muneeba, "Artificial Intelligence in Obstetrics and Gynecology: Transforming Care and Outcomes," *Cureus*, July 2024, doi: 10.7759/cureus.64725.

[13] F. Nasir, R. Shanur, and N. Nasir, "Breast Cancer Detection Using Convolutional Neural Networks: A Deep Learning-Based Approach," *Springer Nat.*, vol. 17, no. 5, May 2025, doi: 10.7759/cureus.83421.

[14] I. Nurcahyati, T. Hamonangan Saragih, A. Farmadi, D. Kartini, and Muliadi, "Classification of Lung Disease in X-Ray Images Using Gray Level Co-Occurrence Matrix Method and Convolutional Neural Network," *J. Electron. Electromed. Eng. Med. Inform.*, vol. 6, no. 4, pp. 332–342, Aug. 2024, doi: https://doi.org/10.35882/jeeemi.v6i4.457.

[15] C. Yin, H. Zhang, J. Du, Y. Zhu, H. Zhu, and H. Yue, "Artificial intelligence in imaging for liver disease diagnosis," *Front. Med.*, vol. 12, p. 1591523, Apr. 2025, doi: 10.3389/fmed.2025.1591523.

[16] A. K. M. S. Hosain, M. H. K. Mehedi, and I. E. Kabir, "PCONet: A Convolutional Neural Network Architecture to Detect Polycystic Ovary Syndrome (PCOS) from Ovarian Ultrasound Images," Oct. 02, 2022, *arXiv*: arXiv:2210.00407. doi: 10.48550/arXiv.2210.00407.

[17] S. S., S. Umapathy, O. Alhajlah, F. Almutairi, S. Aslam, and A. R. K., "F-Net: Follicles Net an efficient tool for the diagnosis of polycystic ovarian syndrome using deep learning techniques," *PLOS ONE*, vol. 19, no. 8, p. e0307571, Aug. 2024, doi: 10.1371/journal.pone.0307571.

[18] P. Moral, D. Mustafi, A. Mustafi, and S. K. Sahana, "CystNet: An AI driven model for PCOS detection using multilevel thresholding of ultrasound images," *Sci. Rep.*, vol. 14, no. 1, p. 25012, Oct. 2024, doi: 10.1038/s41598-024-75964-3.

[19] P. Chitra, K. Srilatha, M. Sumathi, F. V. Jayasudha, T. Bernatin, and M. Jagadeesh, "Classification of Ultrasound PCOS Image using Deep Learning based Hybrid Models," in *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India: IEEE, Mar. 2023, pp. 1389–1394. doi: 10.1109/ICEARS56392.2023.10085400.

[20] L. Wang, "Mammography with deep learning for breast cancer detection," *Front. Oncol.*, vol. 14, p. 1281922, Feb. 2024, doi: 10.3389/fonc.2024.1281922.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[23] F. Alshagathrh *et al.*, "Hybrid Deep Learning and Machine Learning for Detecting Hepatocyte Ballooning in Liver Ultrasound Images," *Diagnostics*, vol. 14, no. 23, p. 2646, Nov. 2024, doi: 10.3390/diagnostics14232646.

[24] M. Abdullah, F. B. Abrha, B. Kedir, and T. Tamirat Tagesse, "A Hybrid Deep Learning CNN model for COVID-19 detection from chest X-rays," *Heliyon*, vol. 10, no. 5, p. e26938, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26938.

[25] A. Hanzala, T. Akter, and Md. S. Rahman, "A hybrid approach for cervical cancer detection: Combining D-CNN, transfer learning, and ensemble models," *Array*, vol. 27, p. 100434, Sept. 2025, doi: 10.1016/j.array.2025.100434.

[26] S. M. Shafi and S. K. Chinnappan, "Hybrid transformer-CNN and LSTM model for lung disease segmentation and classification", doi: DOI%2010.7717/peerj-cs.2444.

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

[27] M. Kaddes, Y. M. Ayid, A. M. Elshewey, and Y. Fouad, "Breast cancer classification based on hybrid CNN with LSTM model," *Sci. Rep.*, vol. 15, Feb. 2025, doi: doi.org/10.1038/s41598-025-88459-6.

[28] J. D. Miller, V. A. Arasu, A. X. Pu, L. R. Margolies, W. Sieh, and L. Shen, "Self-Supervised Deep Learning to Enhance Breast Cancer Detection on Screening Mammography," Mar. 2022, doi: 10.48550/arXiv.2203.08812.

[29] D. Gambhir, D. Aggarwal, and Sadhan02, "KaggleDelhi : PCOS Detection From Ultrasound Images." [Online]. Available: https://www.kaggle.com/datasets/divyangambhir1/kaggledelhi

[30] ibadeus, "PCOS-XAI Ultrasound Dataset." [Online]. Available: https://www.kaggle.com/datasets/ibadeus/pcos-xai-ultrasound-dataset)

[31] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[32] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, Dec. 2022, doi: 10.1016/j.array.2022.100258.

[33] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.

[34] S. R. P M, V. S, S. R P, A. S, R. K, and R. K, "Advances in Multimodal Fusion of EHR and Medical Imaging Data Using deep learning techniques for advanced treatment of brain cancer," in *Proceedings of the 2024 8th International Conference on Algorithms, Computing and Systems*, Hong Kong NB China: ACM, Oct. 2024, pp. 104–109. doi: 10.1145/3708597.3708613.

[35] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Comput. Appl.*, vol. 35, no. 23, pp. 17095–17112, Aug. 2023, doi: 10.1007/s00521-023-08568-z.

[36] A. J. Aiya *et al.*, "Optimized deep learning for brain tumor detection: a hybrid approach with attention mechanisms and clinical explainability," *Sci. Rep.*, vol. 15, no. 1, p. 31386, Aug. 2025, doi: 10.1038/s41598-025-04591-3.

[37] D. Muller, I. Soto-Rey, and F. Kramer, "An Analysis on Ensemble Learning Optimized Medical Image Classification With Deep Convolutional Neural Networks," *IEEE Access*, vol. 10, pp. 66467–66480, 2022, doi: 10.1109/ACCESS.2022.3182399.

[38] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

[39] B. Kocak *et al.*, "Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations," *Eur. J. Radiol. Artif. Intell.*, vol. 3, p. 100030, Sept. 2025, doi: 10.1016/j.ejrai.2025.100030.

[40] A. Marvellous, B. Matthew, M. Pezzè, and S. Abrahão, "Statistical Significance Testing in ML Model Comparisons: Beyond p-values and t-tests," *researchgate*, June 2025, [Online]. Available: https://www.researchgate.net/publication/392727623

[41] D. Singh, V. Kumar, Vaishali, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks," *Eur. J. Clin. Microbiol. Infect. Dis.*, vol. 39, no. 7, pp. 1379–1389, July 2020, doi: 10.1007/s10096-020-03901-z.

[42] X. Li, X. Shen, Y. Zhou, X. Wang, and T.-Q. Li, "Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet)," *PLOS ONE*, vol. 15, no. 5, p. e0232127, May 2020, doi: 10.1371/journal.pone.0232127.

Vol. 7, No. 2, April 2026, Page. 760-777
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.4932

[43] P. E. Susanto, A. Kurniawardhan, D. H. Fudholi, and R. Rahmadi, "A Mobile Deep Learning Model on Covid-19 CT-Scan Classification," *Int. J. Artif. Intell. Res.*, vol. 6, no. 2, July 2022, doi: 10.29099/ijair.v6i1.257.

[44] Y. Wang, X. Liao, D. Qiao, and J. Wu, "A Hybrid Classification Method of Medical Image Based on Deep Learning," Aug. 24, 2021. doi: 10.21203/rs.3.rs-836474/v1.