IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 June 2022, accepted 29 July 2022, date of publication 5 August 2022, date of current version 26 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3196647

RESEARCH ARTICLE

Optimizing CNN Hyperparameters for Blastocyst Quality Assessment in Small Datasets

IRMAWATI[1], **(Student Member, IEEE), RIFAI CHAI**[2], **(Senior Member, IEEE), BASARI**[1,3], **(Member, IEEE), AND DADANG GUNAWAN**[1], **(Senior Member, IEEE)**

[1]Department of Electrical Engineering, Universitas Indonesia, Depok, Jawa Barat 16424, Indonesia
[2]School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Hawthorn, VIC 3122, Australia
[3]Biomedical Engineering Department, Universitas Indonesia, Depok, Jawa Barat 16424, Indonesia

Corresponding author: Dadang Gunawan (guna@eng.ui.ac.id)

This research is supported by Universitas Indonesia through the Hibah Publikasi Terindeks Internasional (PUTI) Kolaborasi Internasional (KI) (2Q2) Scheme under contract number NKB-782/UN2.RST/HKP.05.00/2020, in which Dadang Gunawan is the principal investigator (corresponding author).

ABSTRACT Morphological assessment of blastocyst quality is one of the most significant challenges in the IVF process because the current assessment is based on evaluation by an embryologist; thus, it is still manual and subjective and lacks precision. Artificial intelligence (AI) plays a role in overcoming the limitations of the manual assessment system, and its use is expected to increase implantation rates in IVF. This study aims to optimize the convolutional neural network (CNN) model using the grid search method and to evaluate the effectiveness of different machine learning models in classifying the blastocyst quality in a small dataset. The reliability of the proposed model will be compared with that of other machine learning methods, such as logistic regression (LR), support vector machine (SVM), k-nearest neighbors (KNN), the boosting algorithm, and with the addition of the Canny operator as a segmentation process and principal component analysis (PCA) as a feature extraction approach. We evaluated the results using various performance measures, such as the precision, recall, F1-measure, accuracy, and area under the curve of the receiver operating characteristic curve (AUC-ROC). The final results showed that our proposed CNN model achieves a validation accuracy of 84.00%, a test accuracy of 83.33%, and an AUC of 0.844. McNemar's statistical test results support that our CNN model outperforms the other classifiers.

INDEX TERMS IVF, human blastocyst, CNN, augmentation, hyperparameters.

I. INTRODUCTION

One procedure performed to overcome infertility problems is called IVF. The process is reserved for cases in which other methods, such as fertility drugs, surgery, and artificial insemination, have not worked. Blastocysts have a higher implantation potential than embryos at the cleavage stage (embryonic day 3) [1]. Research has shown that continuing embryo culture up to day 5 results in a higher chance of successful delivery [2]. Therefore, grading the embryo on day five is crucial. Grading of embryos on day 5 (blastocyst) has been based on the Gardner system, in which the grade is determined by the quality of the inner cell mass (ICM) and the trophectoderm (TE) [3]. Blastocysts with good grades will be transferred to the uterus so that pregnancy can be expected, thereby avoiding repeated IVF cycles that incur additional costs. The IVF process may produce more than one embryo, but not all embryos have implantation potential. Transferring more than one embryo can increase the chance of pregnancy but also increases the likelihood of pregnancy complications for both the mother and baby. One solution to minimize multiple pregnancies is to transfer only one embryo, although this will reduce the probability of pregnancy [4]; thus, accurate embryo grading is necessary. Morphological grading of embryos is one of the challenges associated with IVF, which is currently still determined based on the embryologist's

The associate editor coordinating the review of this manuscript and approving it for publication was Mouloud Denai.

assessment with a microscope. It is still manual, subjective, and lacks precision. Artificial intelligence (AI) plays a role in helping overcome the limitations of the manual scoring system, and its use is expected to increase the implantation rate in IVF. The inaccuracy of manual assessment is caused by the blastocyst texture image, in which it is difficult to distinguish between the ICM and TE texture shapes, and by the blurring of the image of the blastocyst edge due to high noise levels. The blastocyst images used in this study were taken with Hoffman modulation contrast (HMC) [5] imaging, which is a light microscopy or optical microscopy technique. HMC imaging is routinely used in IVF clinics to capture images of developing embryos. In performing pattern recognition based on blastocyst images, this work uses segmentation and SVM [6], [7] as classifiers. However, determining the optimal classifier engine is still a problem, especially in small datasets, because this is a challenge in pattern recognition studies. The use of a pattern learning feature is necessary during the machine learning preprocessing stage. The correct way to solve this problem is to implement a deep learning technique using a CNN model.

TABLE 1. Summary of related work on blastocyst classification.

Work	Dataset Size	Method	Validation Accuracy (%)	Test Accuracy (%)	AUC
[9]	1,100	Inception v3		89.01	
[10]	3,469	CNN, Genetic Algorithms		75.30	
[11]	50,000	Inception			0.98
[13]	8,836	Feed Forward Deep Learning			0.93
[14]	2,440	CNN	49.17		
[15]	742	CNN	90.00		
[17]	171,239	ResNet50		75.36	
[18]	7,002	CNN, Discriminant Classifier		97.62	
[19]	186	ANN		72.7	
Our	249	CNN	84.00	83.33	0.84

Deep learning methods, especially CNNs, are currently being used in the IVF field to evaluate embryo morphology, embryo quality, implantation potential, and system quality control. Dimitriadis *et al.* [8] proposed a CNN model that was trained and tested with a dataset of 3,469 embryos to classify between 2PN embryos and non-2PN embryos. Their model classified embryos using a test dataset of 947 images with an accuracy of 91.86%. Using the Inception v3 architecture, Irene Dimitriadis *et al.* [9] classified two types of embryos: blastocysts and nonblastocysts. After training on a dataset of 1,100 embryos, the model could classify 182 test embryos with an accuracy of 89.01%. Hariton *et al.* [10] developed a CNN model by combining genetic algorithms that can select the best quality blastocyst. The CNN model was trained using a dataset of 3,469 images, and the resulting test accuracy was 75.3%. Hariton *et al.* and Khosravi *et al.* [11] proposed a framework based on a deep neural network with a dataset of 50,000 time-lapse embryos to select high-quality embryos. Based on the Inception model, the framework can predict blastocyst quality with an AUC > 0.98. Thirumalaraju *et al.* [12] proposed an AI system to evaluate fertilization and blastocyst development and used it on 947 images of embryos. The resulting coefficient of variation in measuring the best grade of blastocyst quality was 10.9%. To predict the case of fetal heart pregnancy [13], they proposed a deep learning model that can automatically predict this without assessing the blastocyst morphology. The resulting model can predict fetal heart pregnancy with an AUC of 0.93. Thirumalaraju *et al.* [14] proposed a multilayered CNN model to differentiate embryos based on their morphological qualities. Using data from 2,440 embryos, the developed model distinguished between blastocyst and nonblastocyst embryos with a validation accuracy of 49.17%. Bormann *et al.* [15] proposed a CNN model trained with 742 embryos; an accuracy of 90% in selecting the best-quality embryos was achieved. To automatically predict the value of ICM and TE in blastocysts [16], they proposed a deep learning model that can assess blastocyst quality. Chen *et al.* [17] proposed an automatic scoring system for embryo assessment using a dataset of 171,239 embryo images and training data using the ResNet50 model; the average predictive accuracy was 75.36% for the three blastocyst assessment categories. Dirvanauskas *et al.* [18] used 7,002 embryo images to develop a combined CNN with the discriminant classifier model for evaluating and predicting embryo quality. The proposed model can predict the embryo quality with 97.62% accuracy. Bori *et al.* [19] developed an ANN-based AI model to predict live births using the blastocyst morphology. Using data from 186 embryo images, the total accuracy in predicting live births was 72.7%.

However, they used large datasets from time-lapse microscopy and provided limited information about the neural network itself, and the effect of hyperparameters on assessing the embryo morphology with a small dataset is not clear. Several researchers have also researched implementing deep learning methods on small medical datasets, but this is not the case for blastocyst image datasets. The authors [20] used a transfer learning method based on a CNN to evaluate a limited number of magnetic resonance imaging (MRI) datasets. In the current work, the aim is to increase the efficiency and effectiveness in image recognition. Our previous paper [21] using the transfer learning method in the blastocyst quality classification task resulted in a test accuracy of only 64.29%; thus, the accuracy of the method needs to be further improved. In Table 1, detailed related work is presented along with the methods, datasets used, and results.

The problem of applying deep learning or machine learning techniques with little training data is overfitting. Overfitting occurs when the trained model fits the training data but does not generalize well. Adding dropout and regularization

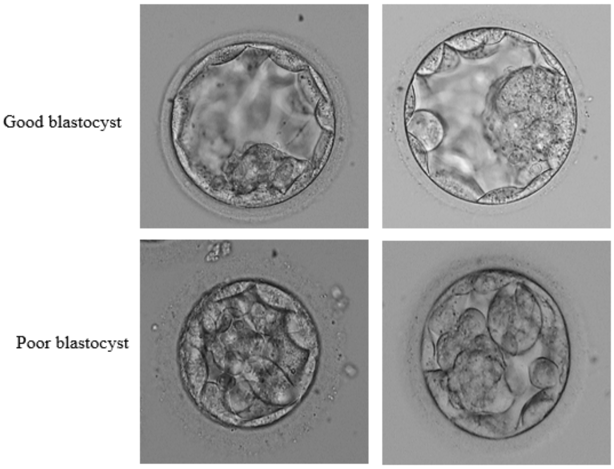


FIGURE 1. Representative HMC human blastocyst quality images.

strategies [22], [23] [24] to a deep learning network can reduce the occurrence of overfitting. This hyperparameter evaluation study on a CNN network used images of blastocysts because of their importance in the IVF process. Some researchers have focused their work on embryonic development [14], [15] [16], [17]. Therefore, the specific purpose of this task was to evaluate the effects of the types of hyperparameters on the CNN when using a small dataset of blastocyst images in classifying two quality classes (good and poor) based on their morphology and to decide which hyperparameter is the best for our dataset.

II. MATERIALS AND METHODS

A. DATASET

We used a publicly available blastocyst dataset to evaluate the performance of our proposed CNN model. The dataset contains 249 images of blastocysts from HMC microscopy and is accessible via ''https://vault.sfu.ca/index.php/login'' upon request [25]. The dataset includes two grades of images, good-quality and poor-quality [11], and an expert embryologist at the Pacific Centre for Reproductive Medicine (PCRM) graded each blastocyst. The blastocyst images in the dataset have various pixel sizes, and we resized them to 224 × 224 pixels. The blastocyst images were obtained using an Olympus IX71 inverted microscope with Nomarski optics (DIC). As the embryo develops from day 1 to day 5, it undergoes many cell divisions. The essential structure of the blastocyst on day 5 is shown in Fig. 1, which also shows the two quality levels of blastocyst images we used.

B. TRAINING, VALIDATION, AND TESTING

To train and evaluate the proposed CNN model, we used 249 blastocyst images. The dataset was divided into three subsets for training (70%), validation (20%), and testing (10%). We performed an augmentation process at the training stage to avoid overfitting due to a lack of training data. In the augmentation process, we sheared, zoomed and flipped the training data. In all experiments on the proposed CNN model, we used Python programming language with Jupyter Notebook as an IDE (Integrated Development Environment). We also used Keras [28] as a framework with the TensorFlow backend.

The performance of our CNN model was compared with that of conventional machine learning. In previous studies, the classification task was generally applied using SVM [29], KNN [30], LR [31], and gradient boosting [32].

The frameworks commonly used in machine learning include image acquisition, preprocessing, feature extraction, and classification. In this study, for the classification that used conventional machine learning, we performed preprocessing and feature extraction steps.

In a previous study [33], we used the Canny operator and achieved the best detection for blastocyst images. As a feature extraction method, we used principal component analysis (PCA) [34], a feature extraction method that uses an appearance-based approach that attempts to identify blastocyst components using global representations based on the whole image instead of only on local features of the blastocyst. Using the same input as for the CNN, all blastocyst image datasets were also used in other classification engines. Fig. 2 shows our proposed framework for human blastocyst quality classification.

We classified blastocyst quality using conventional machine learning methods on the same dataset as for the proposed CNN model. We split the dataset into 80% for training and 20% for testing. In the first step, we converted the blastocyst image from RGB to grayscale and resized the image from the average dimensions of 424 x 378 to 224 x 224 so that our machine learning model could be trained faster on smaller images. The second step consisted of processing the blastocyst image with Canny edge detection to quickly determine the boundaries of the objects in the image. Canny operators have the advantage of better detection, especially under noise conditions, compared to other operators. In the third step, PCA was used to reduce the dimensions of the feature vector and remove the less essential features. We implemented the last feature vector at the blastocyst quality classification step. Finally, we applied four conventional classifiers to the classification task, SVM, KNN, LR, and gradient boosting, using the Keras library in Python. We optimized the parameters used in the conventional classifier using the grid search method; the list of optimized parameters is shown in Table 2. We evaluated the classification results and compared them with the results obtained using the proposed CNN model.

C. HYPERPARAMETER OPTIMIZATION

Hyperparameter optimization is the process of determining the best hyperparameter combination to use. It is done to find the hyperparameter values that can produce the best-performing model. One way to determine the best combination of hyperparameters is to use a grid search method. Grid search is the strategy most frequently used to optimize

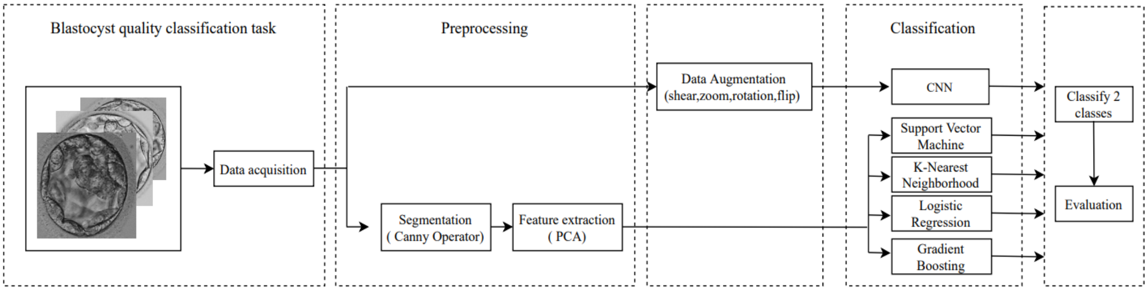


FIGURE 2. **Proposed framework for human blastocyst quality classification. The proposed approach aims to optimize the performance of the model on a small dataset.**

TABLE 2. **Hyperparameters of the conventional classifiers that are tuned.**

Classifiers	Parameter Name	Default Value	Value Range
SVM-RBF	Sigma	0.5	[0.1,0.9]
	Cost	1	[0.25,0.4]
KNN	Number of neighbors	1	[1,17]
	Weights	uniform	['uniform', 'distance']
LR	Penalty	$l2$	[$l1$,$l2$,none]
Gradient Boosting	Learning Rate	0.1	[0.0, inf]
	Max Depth	3	[1, inf]

hyperparameters [35] because it can be easily parallelized [36], and hyperparameter model optimization can improve the model accuracy. Grid search works by combining the hyperparameter values input into the model, searching for all combinations and choosing the best combination based on the highest score. This work used grid search on CNN, SVM, KNN, LR, and gradient boosting models. Scikit-learn was used to perform a grid search, where gridsearchcv performs a search across all parameter sets in the grid. The tuned hyperparameters of the conventional classification model are given in Table 2.

D. BLASTOCYST QUALITY STATE CLASSIFICATION

The deep learning performance can be affected by CNN models. In this work, we configure the CNN model to obtain better performance. The CNN is built with multiple 2D (two-dimensional) convolutions, maximum pooling, fully connected neurons, and dropout [26]. In each convolution layer, the output form can be calculated according to the following equation (1) [27].

$$Convolution\ output = \frac{(i - f + (2 * p))}{s} + 1 \tag{1}$$

$$Pooling\ output = \frac{(i - f)}{s} + 1 \tag{2}$$

where i is the input dimensions of the image, f indicates the size of the filter or kernel in the 2D convolution layer, p is the padding provided as additional data outside the input, and s is the stride, which is a parameter that determines how much the filter shifts. We use a stride value of 1; thus, filter convolution will shift the filter 1 pixel horizontally and then vertically.

This optimization process uses the following network structure hyperparameters: number of filters per layer, kernel size in each layer, dropout rate and L2 regularization. We use the most common technique, known as L2 regularization, which aims to minimize the square of the weights. Because our image data are too complex to be modeled accurately, L2 is a better choice because it can learn the patterns inherent in the data. A weight regularizer is added to each layer in the Keras model with a value of 0.01.

The details of the hyperparameters of our proposed method are shown in Table 2. There is nothing to learn in the input layer or layer 0; the input image is given and reshaped to dimensions of 224×224. We perform the augmentation process using ImageDataGenerator. The stride value defines the number of kernels that convolve the blastocyst image. In this model, we choose a stride value of 1, and the convolution produces an output that is usually called an activation map. The smaller the stride value is, the more detailed the information we obtain from the input, but a small stride value requires more computation than a large stride. However, the use of a small stride will not always result in good performance. The activation map process resembles an extraction process, as in a handcrafted feature extraction process. The kernel weights are randomly initialized, and the output of the convolution operation has a separate activation map for each filter. In the first layer, convolution is performed inside the kernel and filters, leading to a new activation map. The resulting activation map is wrapped with the kernel, and the process is repeated. The second layer is the pooling layer used to reduce the dimensions of the activation map; the operations can use maximum or average values. In this case, we use the maximum value, in which the value for each activation map patch is calculated. This reduces the number of parameters to be studied and the amount of computation performed in the network. The convolution process is continued for the next-to-the-last layer in our CNN architecture. The activation map generated from the feature extraction layer is still in the form of a multidimensional array, so we have to flatten or reshape the activation map into a vector to use it as input from the fully connected layer. In the last layer, the number of

TABLE 3. Network structure hyperparameters of our CNN model.

Layer	Type	Activation Function	Output Shapes	Kernel Size	Number of Filters	Stride
0	Input					
1	2D Convolution	ReLU	222	3	32	1
2	2D Max Pooling	ReLU	220	3	32	1
3	2D Convolution	ReLU	110	3	32	1
4	2D Max Pooling	ReLU	108	3	32	1
5	2D Convolution	ReLU	54	3	32	1
6	2D Max Pooling	ReLU	52	3	64	1
7	Flatten	-	26	-	64	-
8	Dense	-	43264	-	-	-
9	Dropout (0.5)	-	-	-	64	-
10	Fully Connected	-	1	-	-	-

TABLE 4. Network training hyperparameters of our CNN model.

Hyperparameter	Range
Optimizer	['Adam','SGD','RMSprop']
Learning rate	[0.0001,0.0005,0.001,0.005,0.01,0.1]
Momentum	[0.8, 0.9, 0.95]

output values is one, where each image has one output value, namely, its label is 0 or 1 with the previously added dropout. Because we have a small dataset, to limit overfitting and speed up the learning process, we use a dropout process by assigning a value of 0.5. In the CNN model, we use a rectified linear activation function or ReLU with the sigmoid classifier because binary image classification provides better accuracy than the combination of activation and other classifiers [28]. We also optimize the network training hyperparameters, such as the optimizer, learning rate, and momentum. Table 3 shows the training hyperparameters to be optimized; in each case, the range of values is shown in square brackets.

E. DATA AUGMENTATION

To obtain optimal performance, deep learning requires more data than other machine learning algorithms. We have only 249 blastocyst images; 164 of these are good-quality blastocyst images, and 85 are poor-quality blastocyst images. This amount of data is insufficient to obtain optimal performance. Therefore, we need to perform a data augmentation process.

Data augmentation creates additional training data that artificially expands the training set with a label preserving transformation [37]. Data augmentation aims to generate virtual data samples that can be used to improve the training dataset and reduce overfitting. In this study, we add data only to the training dataset and not to the validation or test datasets; this is different from preparing data using image resizing, which requires consistency across all datasets that interact with the model. Using ImageDataGenerator, a function of Keras, we perform random transformations on the training dataset using shear, zoom, rotation and flip augmentation techniques and then change the parameters in the function. The parameters we declare in the ImageDataGenerator function are shear_range=0.2, which shifts the image by 20%, zoom_range=0.2, which zooms in and out by 20%, rotation_range=45, and horizontal_flip=True. After declaring the parameters of the ImageDataGenerator function, we create an iterator that fetches the image and loops in batches by streamlining the image into an ImageDataGenerator object. To stream images, we use the flow_from_directory method, which takes the directory path and generates an additional dataset. When the iterator has been created, it can be used to train our CNN model by calling the fit_generator() function.

F. EVALUATION OF THE PROPOSED CNN MODEL

Evaluating the model created is essential in developing good deep learning and machine learning models. In this section, we will discuss evaluation of the model specifically with respect to the classification of blastocyst quality. The model evaluation process is conducted after the model training is completed. The model evaluation uses evaluation data that cannot be the same as the data used to train the model. Testing with these evaluation data will provide the actual accuracy of a model that has been trained. However, accuracy (3) is not the only parameter considered in conducting evaluations because high accuracy values can be deceptive due to dataset imbalances [38]. Therefore, more comprehensive evaluation metrics such as the confusion matrix, precision (4), recall (5), and F1-score (6) are needed.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - Score = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{6}$$

In binary classification, four parameters can be considered in evaluating the prediction results of a model. These four parameters are true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Comparisons of the accuracy results when using different parameters are applied in this work, i.e., the difference in the accuracy between different filters, kernel sizes, optimizers, and kinds of machine learning.

III. RESULTS

A. IMPACT OF FILTERS ON ACCURACY

The accuracy of the CNN model can depend on the number of kernels in the convolution. A larger number of filters in each convolution will cause overfitting when the dataset is small. We conduct simulations to determine the impact of the number of filters on the results obtained using our proposed CNN model. To determine the optimum number of filters, we explored and evaluated four sets of filters, viz., 32, 64, 128, and 256. After assessing the impact of the number of filters on the accuracy in the classification of blastocyst quality, it was found that the use of 32 filters yielded the greatest impact on increasing the accuracy (0.759±0.128)

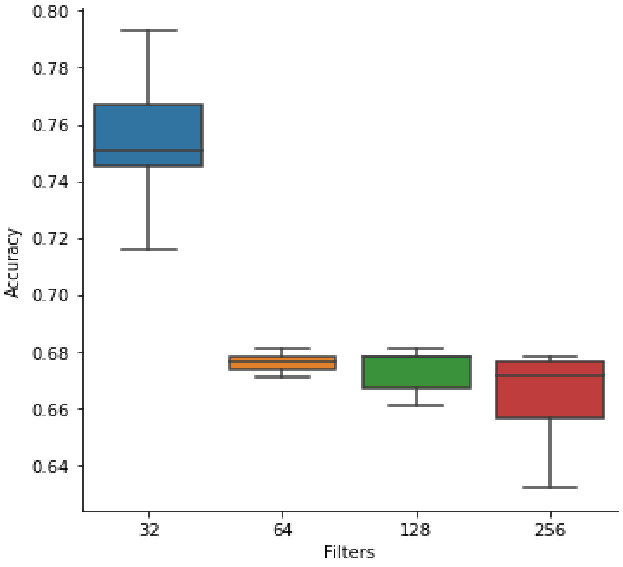


FIGURE 3. Correlation of the number of filters in each layer with the accuracy of blastocyst quality classification.

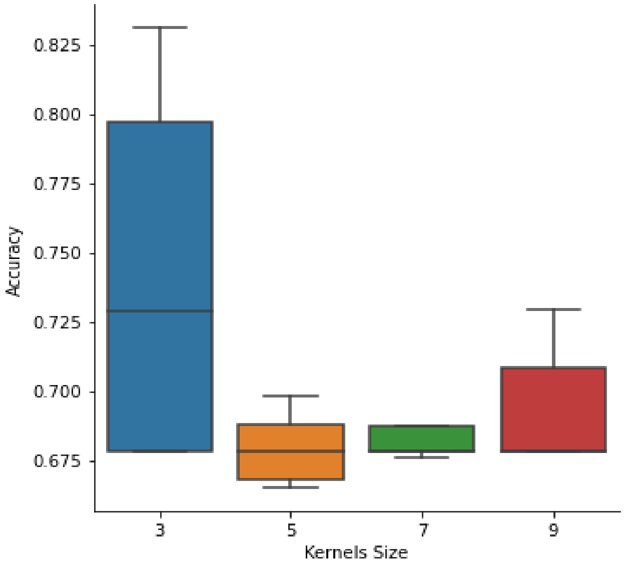


FIGURE 4. Correlation of the kernel size in each layer with the accuracy of blastocyst quality classification.

compared to the number of filters of 64 (0.669±0.043), 128 (0.678±0.075) and 256 (0.665±0.024). This experiment shows that the use of 32 filters has a more favorable impact on increasing the accuracy of the classification than the use of 64, 128 or 256 filters. The use of 256 filters had the most negligible impact on improving the accuracy of our model, as shown in the boxplot diagram in Fig.3.

B. IMPACT OF KERNEL SIZE ON ACCURACY

The kernel size on the CNN determines the receptive field. It provides information about the number of input image pixels that can be seen on activation in the network. Additionally, the use of a small kernel rather than a fully connected network benefits from weight sharing and reduced computational costs. In the experiment with kernel sizes, we determined whether the use of a small or a large kernel size affects the accuracy of blastocyst image classification. Based on the experimental results, we found that a kernel size of 3 gives the highest accuracy (0.743±0.119), followed by a kernel size of 5 (0.679±0.047), a kernel size of 7 (0.681±0.022), and a kernel size of 9 (0.694±0.024). From these results, it can be concluded that kernel size does not significantly affect the accuracy. Nevertheless, the kernel size jeopardizes the training and testing process time. More detailed information about the distribution of accuracy values with respect to the kernel size is presented in Fig.4.

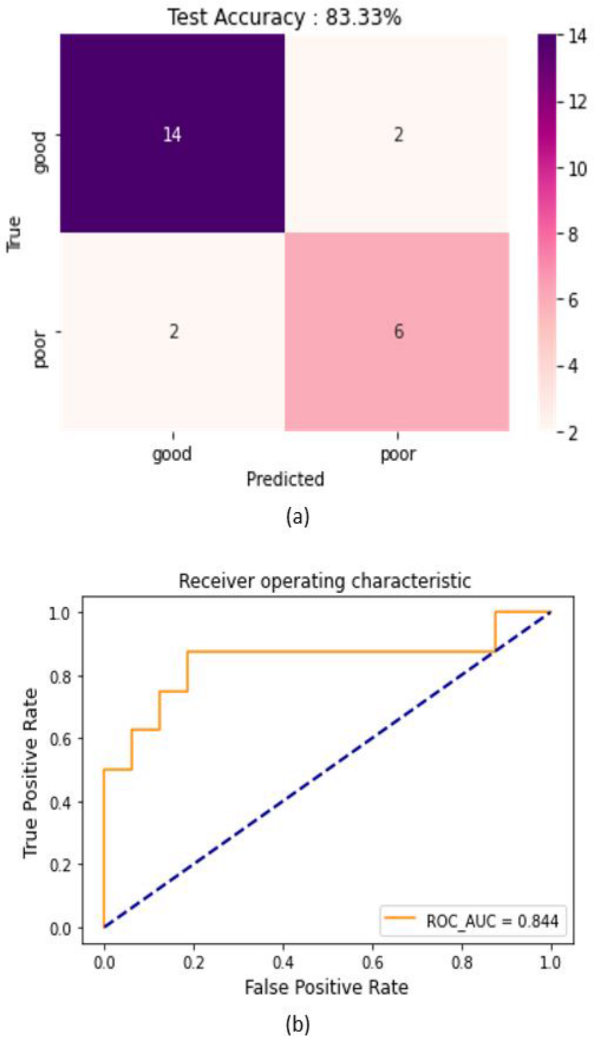
C. EVALUATION OF THE CNN ARCHITECTURE

The best correlation can be achieved by using optimization strategies or algorithms called optimizers. The optimization algorithm is responsible for reducing losses and providing the most accurate results possible. Our proposed CNN model used several optimizers, such as adaptive moment estimation (Adam) [39], stochastic gradient descent (SGD) [40] and RMSprop, and the results were evaluated. The following Table 2 shows how the use of an optimizer increases the accuracy of our CNN model. Table 2 shows that the best validation accuracy of 86.00% is achieved using the Adam optimizer with a learning rate of 0.001. We used the early stopping function to obtain the number of epochs. The early stopping function can halt the training of neural networks at the optimal time. The early stopping callback function can monitor the loss or accuracy value. When the loss value is being monitored, the training will stop if the loss value increases. If the accuracy is being monitored, then training is stopped when there is a decrease in the accuracy. We obtained the best number of epochs, which is 36. After obtaining the best CNN model, we tested it on a dataset consisting of 24 blastocyst images. The test results on these testing data yielded an accuracy of 83.33% with a loss of 0.6141.

We evaluated the accuracy of the blastocyst quality selection model and determined the area under the curve (AUC) and receiver operating characteristic (ROC) values. The resulting AUC in predicting the blastocyst quality in the test dataset is 0.844. Fig.5 shows the confusion matrix visualization of the test accuracy and the AUC of the ROC curves.

D. COMPARISON TO THE OTHER MACHINE LEARNING METHODS

This paper evaluated binary classification of blastocyst quality using the proposed CNN, SVM, KNN, LR, and gradient boosting on a dataset of 249 human blastocysts. We performed analysis and evaluation of the best performance of our CNN model. Fig. 6 shows the variation in the accuracy of each method.



(a)

(b)

FIGURE 5. Evaluation of the performance of the blastocyst classification model based on the (a) confusion matrix, (b) ROC and AUC.

TABLE 5. Training and test result accuracies of the proposed CNN model.

Optimizer	Learning rate	Momentum	Train Acc (%)	Train Loss	Val Acc (%)	Val Loss
Adam	**0.001**	**-**	**89.99**	**0.41**	**84.00**	**0.46**
SGD	0.001	0.8	65.70	0.65	66.00	0.64
RMSprop	0.001	-	82.23	0.65	79.55	0.62

The confusion matrix in binary classification represents predictions with actual conditions from the data generated by the trained model. The confusion matrix uses the parameters TP, FP, TN, and FN in its representation.

For validation, we used 50 blastocyst images. Based on the confusion matrix results, we can see that our CNN model produces a maximum accuracy of 84.00%. The SVM method produces a maximum accuracy of 82.00%, the KNN method produces a maximum accuracy of 74.00%, LR has a top accuracy of 82.00%, and gradient boosting has a maximum accuracy of 64.00%. The CNN model has the highest accuracy, and gradient boosting produces the lowest accuracy.

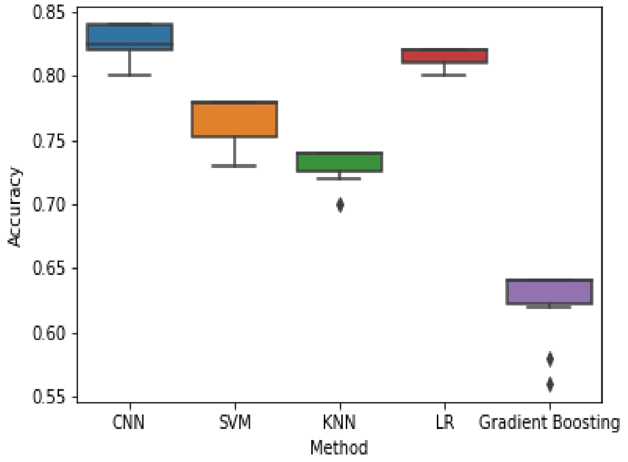


FIGURE 6. Descriptive statistics of the accuracies obtained using all methods (CNN, SVM, KNN, LR, gradient boosting) with segmented input using Canny operators and PCA for feature extraction. A boxplot diagram represents all features combined.

Testing of all methods on the blastocyst image testing data based on the confusion matrix indicated that the model performance was not good enough to classify blastocyst images with ‘‘poor’’ quality. This problem is due to the very low amount of input data and to the unbalanced dataset regarding the number of blastocyst images, where the number of ‘‘good’’ quality blastocyst images is far greater than the number of ‘‘poor’’ quality blastocyst images.

E. MODEL PERFORMANCE MEASUREMENT

Based on the confusion matrix, we can calculate the recall, precision, and F1-scores. CNN has the best recall value for classifying good-quality and poor-quality blastocysts, with average values of 0.89 for good-quality blastocysts and 0.76 for poor-quality blastocysts. In second place is SVM with a recall value of 0.97 for good-quality blastocysts and 0.51 for poor-quality blastocysts. The recall scores and the F1-score of LR outperform those of other techniques. The KNN method has the lowest precision, with a value between 0.50 and 0.66. Based on the recall value, the gradient boosting classifier obtains the lowest score, and KNN has the lowest F1-score. In addition to evaluating the model through the accuracy, the precision, recall, and F1-scores can show the classification performance, as shown in Fig. 8, which shows that all of the classifier machines offer different performances.

F. STATISTICAL SIGNIFICANCE TEST

To measure the statistical significance of our results, we used the McNemar test. The McNemar test is still widely used in

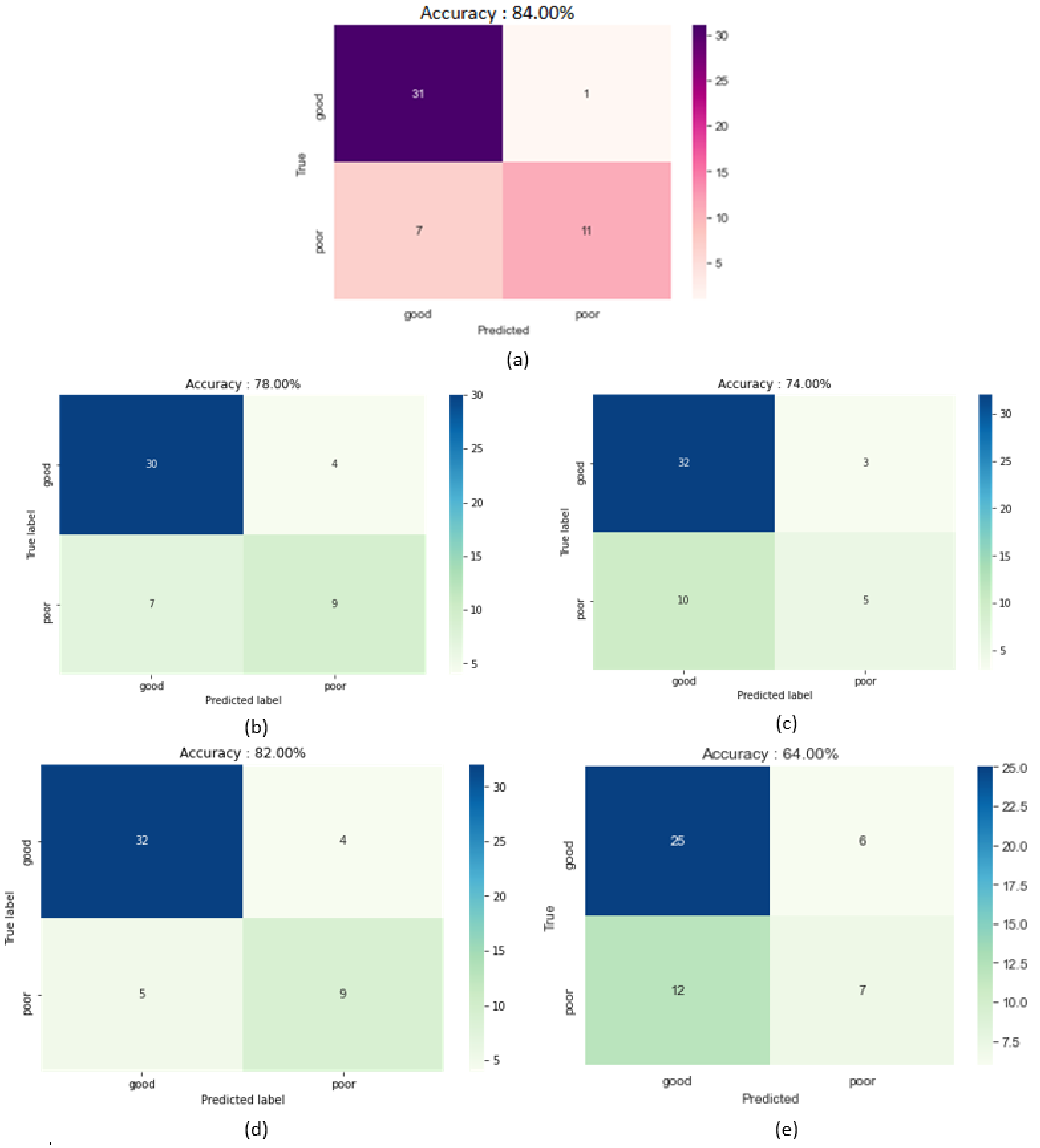
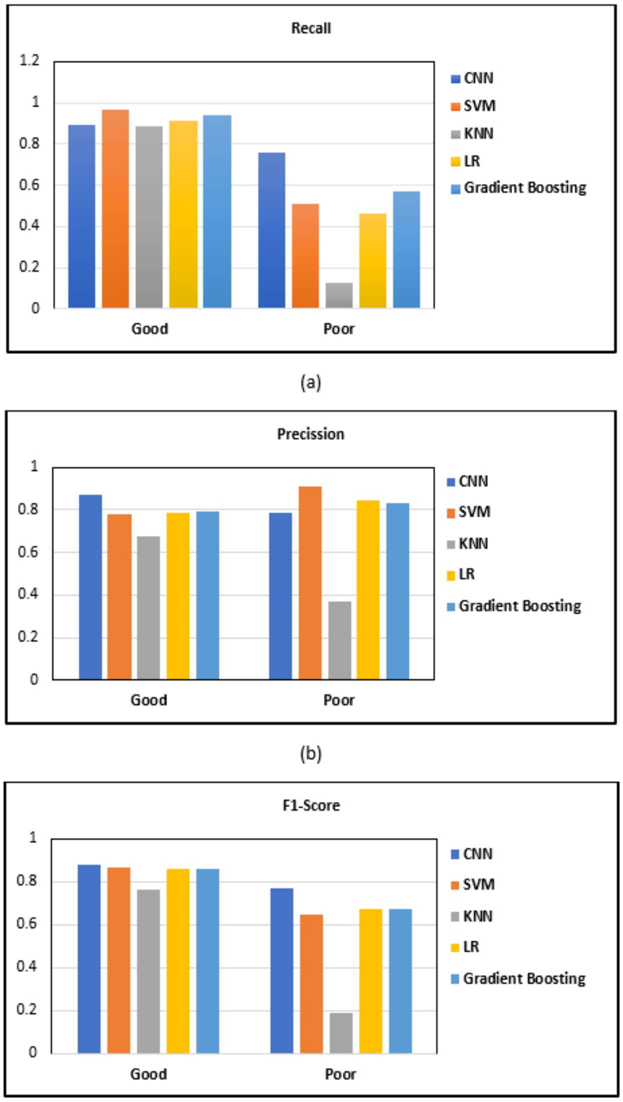


FIGURE 7. Confusion matrix visualization of 50 validation images for (a) CNN, (b) SVM, (c) KNN, (d) LR and (e) gradient boosting.

machine learning studies. In machine learning, McNemar's test can be applied to compare model accuracy and select the best classifier. We implemented the McNemar test in Python using the Mcnemar () Statsmodels function. The function takes a contingency table as an argument and returns a calculated test statistic and a p-value. The classification accuracy of the CNN that has been optimized is 84.00%, with a p-value of 0.034. Here, a statistical test is used to validate the accuracy results. The accuracy results and statistical hypothesis testing confirmed that the optimized CNN model was the best classifier for predicting blastocyst quality. The p-value obtained in the hypothesis test can be interpreted as failing to reject H0 if $p>\alpha$, and if $p<=\alpha$, H0 is rejected, where there is a significant difference. The results of the McNemar statistical test and the p-values shown in Table 6 include the p-value of the CNN classification model, which rejected H0, where the



(c)

FIGURE 8. **Comparison of the performance of all methods (CNN, SVM, KNN, LR, and gradient boosting) based on the two classes of good quality and poor quality. The performance was measured based on the (a) recall, (b) precision, and (c) F1-score.**

model has a different error proportion. Other classification models (SVM, KNN, LR, and gradient boosting) produce the same proportion of errors or fail to reject H0. Table 6 presents the p-values and T-statistics obtained using McNemar's test to test whether the performance of the CNN model significantly differs from that of the other classification models.

TABLE 6. **McNemar's test result.**

Classification Models	T-Statistic	p-value
CNN	4.500	0.034
SVM	0.818	0.366
KNN	3.769	0.052
LR	0.111	0.739
Gradient Boosting	2.000	0.157

IV. DISCUSSION

In this study, we optimized the CNN to assess the quality of blastocysts produced during the IVF process. The optimization procedure was conducted using a grid search due to its good performance in general and its ease of implementation. We also propose using weight decay and dropout regularization techniques to reduce overfitting. The input data for this research are raw images of human blastocysts from HMC microscopy. In this study, we used four conventional classifiers to compare our CNN model and performance measures such as the accuracy, precision, recall, and F1-score. The essential part of deep learning is the convolution layer, which uses some filters. In general, the use of a large number of filters affects the accuracy. Our experimental results show a significant difference in the results obtained using different numbers of filters. The use of 32 filters has the greatest influence on the accuracy compared to the use of 64, 128 or 256 filters. Although this research contributes to knowledge, particularly for small blastocyst images, it has some limitations. One limitation of the present study is that the blastocyst images produced by the inverted microscope are affected by artifacts and noise; another limitation is the imbalanced number of blastocyst datasets in the "good" and "poor" quality classes. These limitations prevent the optimized CNN from providing satisfactory results, and the imbalance in the classes causes the model to not make sufficient observations in the very few data class. The classification performance can be further improved by building models based on larger datasets and using techniques to address imbalanced classes in medical images [41]. The latest study [42] enhances the research on blastocysts through the use of a convolutional neural network to image the blastocyst combined with an elemental layer for maternal age. With an accuracy of 75%, this study shows potential in determining the probability of live birth. Vaidya *et al.* [43] used a combination of CNN and LSTM models to automatically assess embryos in time-lapse images. This study obtained a 100% accuracy validation result without performing an accuracy test. Based on the latest research, it has been shown that our proposed CNN model has the potential to be improved for predicting the probability of a live birth, whether time-lapse images are used or not.

V. CONCLUSION

Our study examined the performance of the CNN model in assessing human blastocyst quality in the case of small datasets. The optimization process yielded good results: the highest validation accuracy was 84.00%, and the test accuracy was 83.33%, with an AUC value of 0.844. Based on the accuracy and the AUC results, the classification of blastocyst quality described in this paper has excellent potential to assist embryologists in assessing and selecting the best-quality blastocysts. McNemar's statistical significance test proved that the CNN model scores high in prediction compared to the other classifiers. The confusion matrix study results show that the accuracy of the proposed model enable classification of blastocysts as ''good'' or ''poor'' quality.

In future work, we will explore grid search methods and expand the boundaries of the CNN hyperparameter space used for optimization [44]. Finally, we would also like to extend our case study to time-lapse images of blastocysts. More test subjects are needed to ensure that the results obtained are statistically significant and that the proposed approach can be applied as a general tool for assessing blastocyst quality.

REFERENCES

[1] S. Tandulwadkar, M. Patil, and S. Naik, ''Optimising the outcome of embryo transfer,'' *EMJ*, vol. 5, no. 1, pp. 110–119, Aug. 2019.

[2] D. Glujovsky, C. Farquhar, Q. R. Am, A. S. Cr, and D. Blake, ''Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology,'' *Cochrane Library*, vol. 5, no. 6, 2016.

[3] D. K. Gardner, M. Lane, J. Stevens, T. Schlenker, and W. B. Schoolcraft, ''Blastocyst score affects implantation and pregnancy outcome: Towards a single blastocyst transfer,'' *Fertility Sterility*, vol. 73, no. 6, pp. 1155–1158, Jun. 2000.

[4] D. Bodri, S. Kawachiya, M. D. Brucker, H. Tournaye, M. Kondo, R. Kato, and T. Matsumoto, ''Cumulative success rates following mild IVF in unselected infertile patients: A 3-year, single-centre cohort study,'' *Reproductive Biomed.*, vol. 28, no. 5, pp. 572–581, May 2014.

[5] R. Hoffman and L. Gross, ''The modulation contrast microscope,'' *Nature*, vol. 254, no. 5501, pp. 586–588, Apr. 1975.

[6] A. Chavez-Badiola, A. Flores-Saiffe Farias, G. Mendizabal-Ruiz, R. Garcia-Sanchez, A. J. Drakeley, and J. P. Garcia-Sandoval, ''Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning,'' *Sci. Rep.*, vol. 10, no. 1, pp. 1–6, Dec. 2020.

[7] E. S. Filho, J. A. Noble, M. Poli, T. Griffiths, G. Emerson, and D. Wells, ''A method for semi-automatic grading of human blastocyst microscope images,'' *Hum. Reproduction*, vol. 27, no. 9, pp. 2641–2648, Sep. 2012.

[8] I. Dimitriadis, C. L. Bormann, M. K. Kanakasabapathy, P. Thirumalaraju, R. Gupta, R. Pooniwala, I. Souter, S. T. Rice, P. Bhowmick, and H. Shafiee, ''Deep convolutional neural networks (CNN) for assessment and selection of normally fertilized human embryos,'' *Fertility Sterility*, vol. 112, no. 3, p. e272, Sep. 2019.

[9] I. Dimitriadis, C. L. Bormann, P. Thirumalaraju, M. Kanakasabapathy, R. Gupta, R. Pooniwala, I. Souter, J. Y. Hsu, S. T. Rice, P. Bhowmick, and H. Shafiee, ''Artificial intelligence-enabled system for embryo classification and selection based on image analysis,'' *Fertility Sterility*, vol. 111, no. 4, p. e21, Apr. 2019.

[10] E. Hariton, I. Dimitriadis, M. K. Kanakasabapathy, P. Thirumalaraju, R. Gupta, R. Pooniwala, I. Souter, S. T. Rice, P. Bhowmick, L. B. Ramirez, C. L. Curchoe, J. E. Swain, L. M. Boehnlein, C. L. Bormann, and H. Shafiee, ''A deep learning framework outperforms embryologists in selecting day 5 euploid blastocysts with the highest implantation potential,'' *Fertility Sterility*, vol. 112, no. 3, pp. e77–e78, Sep. 2019.

[11] P. Khosravi, E. Kazemi, Q. Zhan, J. E. Malmsten, M. Toschi, P. Zisimopoulos, A. Sigaras, S. Lavery, L. A. D. Cooper, C. Hickman, M. Meseguer, Z. Rosenwaks, O. Elemento, N. Zaninovic, and I. Hajirasouliha, ''Deep learning enables robust assessment and selection of human blastocysts after *in vitro* fertilization,'' *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–9, Dec. 2019.

[12] P. Thirumalaraju, M. K. Kanakasabapathy, R. Gupta, R. Pooniwala, H. Kandula, I. Souter, I. Dimitriadis, C. L. Bormann, and H. Shafiee, ''Automated quality assessment of individual embryologists performing ICSI using deep learning-enabled fertilization and embryo grading technology,'' *Fertility Sterility*, vol. 112, no. 3, p. e71, Sep. 2019.

[13] D. Tran, S. Cooke, P. J. Illingworth, and D. K. Gardner, ''Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer,'' *Hum. Reproduction*, vol. 34, no. 6, pp. 1011–1018, Jun. 2019.

[14] P. Thirumalaraju, M. K. Kanakasabapathy, C. L. Bormann, R. Gupta, R. Pooniwala, H. Kandula, I. Souter, I. Dimitriadis, and H. Shafiee, ''Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality,'' *Heliyon*, vol. 7, no. 2, Feb. 2021, Art. no. e06298.

[15] C. L. Bormann, M. K. Kanakasabapathy, P. Thirumalaraju, R. Gupta, R. Pooniwala, H. Kandula, E. Hariton, I. Souter, I. Dimitriadis, L. B. Ramirez, C. L. Curchoe, J. Swain, L. M. Boehnlein, and H. Shafiee, ''Performance of a deep learning based neural network in the selection of human blastocysts for implantation,'' *eLife*, vol. 9, pp. 1–14, Sep. 2020.

[16] M. F. Kragh, J. Rimestad, J. Berntsen, and H. Karstoft, ''Automatic grading of human blastocysts from time-lapse imaging,'' *Comput. Biol. Med.*, vol. 115, Dec. 2019, Art. no. 103494.

[17] T.-J. Chen, W.-L. Zheng, C.-H. Liu, I. Huang, H.-H. Lai, and M. Liu, ''Using deep learning with large dataset of microscope images to develop an automated embryo grading system,'' *Fertility Reproduction*, vol. 1, no. 1, pp. 51–56, Mar. 2019.

[18] D. Dirvanauskas, R. Maskeliunas, V. Raudonis, and R. Damasevicius, ''Embryo development stage prediction algorithm for automated time lapse incubators,'' *Comput. Methods Programs Biomed.*, vol. 177, pp. 161–174, Aug. 2019.

[19] L. Bori, F. Dominguez, E. I. Fernandez, R. Del Gallego, L. Alegre, C. Hickman, A. Quiñonero, M. F. G. Nogueira, J. C. Rocha, and M. Meseguer, ''An artificial intelligence model based on the proteomic profile of euploid embryos and blastocyst morphology: A preliminary study,'' *Reproductive Biomed.*, vol. 42, no. 2, pp. 340–350, Feb. 2021.

[20] D. R. Clymer, J. Long, C. Latona, and S. Akhavan, ''Applying machine learning methods toward classification based on small datasets: Application to shoulder labral tears,'' *J. Eng. Sci. Med. Diagnostics Therapy*, vol. 3, pp. 1–10, Feb. 2020.

[21] D. Gunawan, ''Classification of human blastocyst quality using wavelets and transfer learning,'' in *Advances in Computer, Communication and Computational Sciences*. Cham, Switzerland: Springer, 2021, pp. 965–974.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in *Proc. Neural Inf. Process. Syst. Conf.*, 2012, pp. 1–1432.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' *Popul. Health Manag.*, vol. 18, no. 3, pp. 186–191, 2014.

[24] G. H. Srivastava, ''Dropout: A simple way to prevent neural networks from overfittin,'' *Phys. Lett. B*, vol. 299, nos. 3–4, pp. 345–350, 2014.

[25] P. Saeedi, D. Yee, J. Au, and J. Havelock, ''Automatic identification of human blastocyst components via texture,'' *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2968–2978, Dec. 2017.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, pp. 1–3.

[27] B. B. Traore, B. Kamsu-Foguem, and F. Tangara, ''Deep convolution neural network for image recognition,'' *Ecol. Informat.*, vol. 48, pp. 257–268, Nov. 2018.

[28] N. Ketkar, *Deep Learning With Python*. Berkeley, CA, USA: Apress, 2017, pp. 95–109.

[29] K. Chauhan and S. Ram, ''Image classification with deep learning and comparison between different convolutional neural network structures using tensorflow and keras,'' *Int. J. Adv. Eng. Res. Develop.*, vol. 5, no. 2, pp. 533–538, 2018.

[30] S. U. Khan, N. Islam, Z. Jan, K. Haseeb, S. I. A. Shah, and M. Hanif, ''A machine learning-based approach for the segmentation and classification of malignant cells in breast cytology images using gray level co-occurrence matrix (GLCM) and support vector machine (SVM),'' *Neural Comput. Appl.*, vol. 34, no. 11, pp. 1–8, 2021.

[31] Y. Wang, Z. Chen, H. Shao, and N. Wang, ''A KNN-based classification algorithm for growth stages of haematococcus pluvialis,'' in *Proc. IEEE 4th Adv. Inf. Manage., Communicates, Electron. Autom. Control Conf. (IMCEC)*, Jun. 2021, pp. 6–9.

[32] R. Sciorio, D. Thong, K. J. Thong, and S. J. Pickering, ''Clinical pregnancy is significantly associated with the blastocyst width and area: A time-lapse study,'' *J. Assist. Reproduction Genet.*, vol. 38, no. 4, pp. 847–855, Apr. 2021.

[33] Q. Liao, Q. Zhang, X. Feng, H. Huang, H. Xu, B. Tian, J. Liu, Q. Yu, N. Guo, Q. Liu, B. Huang, D. Ma, J. Ai, S. Xu, and K. Li, "Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring," *Commun. Biol.*, vol. 4, no. 1, pp. 1–9, Dec. 2021.

[34] B. Irmawati and D. Gunawan, "Automated detection of human blastocyst quality using convolutional neural network and edge detector," in *Proc. 1st Int. Conf. Cybern. Intell. Syst. (ICORIS)*, Aug. 2019, pp. 181–184.

[35] M.-A. Zöller and M. F. Huber, "Benchmark and survey of automated machine learning frameworks," *J. Artif. Intell. Res.*, vol. 70, pp. 409–472, Jan. 2021.

[36] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[37] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1542–1547.

[38] S. Straube, "How to evaluate an agent's behavior to infrequent events?—Reliable performance estimation insensitive to class distribution," *Frontiers Comput. Neurosci.*, vol. 8, no. 1, pp. 1–6, 2014.

[39] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[40] R. Milewski, A. Kuczyńska, B. Stankiewicz, and W. Kuczyński, "How much information about embryo implantation potential is included in morphokinetic data? A prediction model based on artificial neural networks and principal component analysis," *Adv. Med. Sci.*, vol. 62, no. 1, pp. 202–206, Mar. 2017.

[41] A. Bria, C. Marrocco, and F. Tortorella, "Addressing class imbalance in deep learning for small lesion detection on medical images," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103735.

[42] Y. Miyagi, T. Habara, R. Hirata, and N. Hayashi, "Deep learning to predicting live births and aneuploid miscarriages from images of blastocysts combined with maternal age," *Int. J. Bioinf. Intell. Comput.*, vol. 1, no. 1, pp. 10–21, 2022.

[43] G. Vaidya, S. Chandrasekhar, R. Gajjar, N. Gajjar, D. Patel, and M. Banker, "Time series prediction of viable embryo and automatic grading in IVF using deep learning," *Open Biomed. Eng. J.*, vol. 15, no. 1, pp. 190–203, Dec. 2021.

[44] A. Lemay, K. Hoebel, C. P. Bridge, B. Befano, S. De Sanjosé, D. Egemen, A. C. Rodriguez, M. Schiffman, J. P. Campbell, and J Kalpathy-Cramer, "Improving the repeatability of deep learning models with Monte Carlo dropout," 2022, *arXiv:2202.07562*.

IRMAWATI (Student Member, IEEE) was born in Jakarta, Indonesia, in 1977. She is currently pursuing the Ph.D. degree with Universitas Indonesia, Indonesia. Her research interests include image processing, signal processing, and artificial intelligence.

RIFAI CHAI (Senior Member, IEEE) received the B.Eng. degree from Krida Wacana Christian University, Jakarta, Indonesia, in 2000, and the Ph.D. degree in engineering from the University of Technology Sydney (UTS), in 2014. From 2000 to 2010, he worked as a Product Development Engineer, a Research and Development Engineer, and a Project Engineer with companies in Indonesia and Australia. From 2012 to 2018, he worked as an Associate Lecturer and a Lecturer with the University of Technology Sydney. Currently, he is working as a Senior Lecturer with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, Australia. His research interests include brain–computer interfaces, medical signal processing and devices, artificial intelligence, and autonomous systems. He serves as an Associate Editor for *Electronics Letters*.

BASARI (Member, IEEE) was born in Tegal, Central Java, Indonesia, in November 1979. He received the B.E. degree in electrical engineering from Universitas Indonesia, Jakarta, Indonesia, in 2002, and the M.E. and D.E. degrees in electrical engineering from Chiba University, Japan, in 2008 and 2011, respectively. He worked at Radio Network Planning of PT Indonesian Satellite Corporation Tbk (Indosat Company Ltd.) and Radio Network Operation of PT Telkomsel, Indonesia, from 2003 to 2004. He is currently an Assistant Professor (Lektor) with the Faculty of Engineering, Universitas Indonesia, Indonesia. He is the author of more than 140 publications, including proceedings and journals. His current research interests include the areas of biomedical engineering (informatics, diagnosis, and therapy), microwave medical imaging, MRI systems, microstrip antennas, planar arrays, microwave medical devices, radar applications, metamaterials, MIMO antennas, reconfigurable antennas, RFIDs, UWB antennas, nanosatellites, microwave circuits, reflect arrays, and mobile satellite antennas.

He was the Secretary of the IEEE MTT/AP Indonesia Chapter, from 2012 to 2015. He was a recipient of the IEEE AP-Society Japan Chapter Young Engineer Award, Who's Who in the World, the Dean Award of Chiba University, the APRASC-URSI Young Scientist Award, and the 2015 QiR Best Paper Award, in 2010, 2011, 2011, 2013, and 2015. In 2016 and 2017, he was the Vice Chair of the IEEE MTT/AP-S Joint Chapter Indonesia Section. From 2018 to 2019, he served as the Chapter Chair of the IEEE MTT/AP-Society Indonesia Section. He served as the General Chair of the International Symposium on Biomedical Engineering (ISBE), in 2016, and the IEEE Region 10 Humanitarian Technology Conference (HTC), in 2019. He was the Principal Editor for the IEEE Publication of Quality in Research (QiR), in 2017. He also serves as a referee at some technical conferences and for a national journals in the area of engineering and technology. He was a Managing Editor of the *Makara Journal of Technology*, from 2016 to 2018. Since March 2018, he has been serving as the Head of the Biomedical Engineering Program with the Department of Electrical Engineering, UI, for undergraduate and master's degrees. He is currently the Editor-in-Chief of the *Makara Journal of Technology* (Universitas Indonesia).

DADANG GUNAWAN (Senior Member, IEEE) received the bachelor's degree in electrical engineering from the University of Indonesia, in 1983, the master's degree from Keio University, Japan, in 1989, and the Ph.D. degree from the University of Tasmania, Australia, in 1995. He is currently with the Department of Electrical Engineering, Universitas Indonesia, Kampus Depok, Depok, Jawa Barat, Indonesia. He is also a Professor with the Department of Electrical Engineering, Universitas Indonesia. He has published 100's of academic papers in international journals and proceedings as a first author or coauthor. His research interests include wireless and signal processing technology.