SUPPLEMENTARY MATERIALS

Definitions

Specificity [1]: Specificity, also known as the "true negative rate", measures the proportion of true negative predictions among all actual negative instances in the dataset. It indicates how well a model can correctly identify the negative cases. Specificity is calculated using the formula:

Specificity = True Negatives / (True Negatives + False Positives)

Sensitivity (Recall or True Positive Rate)[1]: Sensitivity measures the proportion of true positive predictions among all actual positive instances in the dataset. It indicates how well a model can identify all the positive cases. Sensitivity is calculated using the formula:

Sensitivity = True Positives / (True Positives + False Negatives)

Precision [2]: Precision is a metric that measures the proportion of true positive predictions among all positive predictions made by a model. It tells how many of the items predicted as positive are true positives. Precision is calculated using the formula:

Precision = True Positives / (True Positives + False Positives)

Recall (Sensitivity or True Positive Rate) [2]: Recall is a metric that measures the proportion of true positive predictions among all actual positive instances in the dataset. It indicates how well a model can identify all the positive cases. Recall is calculated using the formula:

Recall = True Positives / (True Positives + False Negatives)

F1 Score [3]: The F1 score is a harmonic mean (the reciprocal of the arithmetic mean of the reciprocals) of precision and recall. It provides a balanced measure that considers false positives and false negatives. The F1 score is useful when you want to find a balance between precision and recall. It is calculated using the formula:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Area Under the ROC Curve (AUC) [4]: The ROC curve (receiver operating characteristic curve) is a graphical representation of the performance of a classification model at different classification thresholds. AUC measures the area under this curve and provides an aggregated measure of a model's ability to discriminate between positive and negative instances. AUC values range between 0 and 1, with higher values indicating better performance.

Negative Predictive Value (NPV) [3]: The Negative Predictive Value is a metric that represents the proportion of true negative predictions among all negative predictions made by a model. NPV is useful for understanding a model's performance in identifying true negatives. It is calculated using the formula:

NPV = True Negatives / (True Negatives + False Negatives)

Supplementary Figures

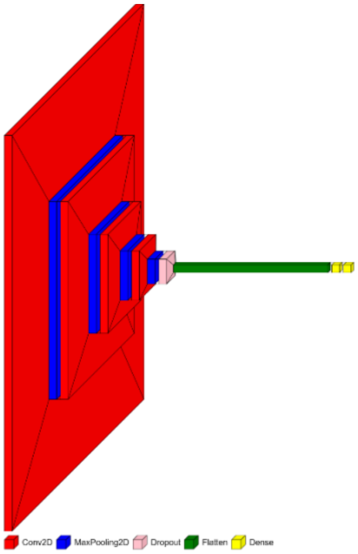

Figure S1: Schematic representation for visualization of the custom CNN Model. The color map defines the different layer types in the visualization (convolution, max pooling, dropout, flatten, and dense layers).

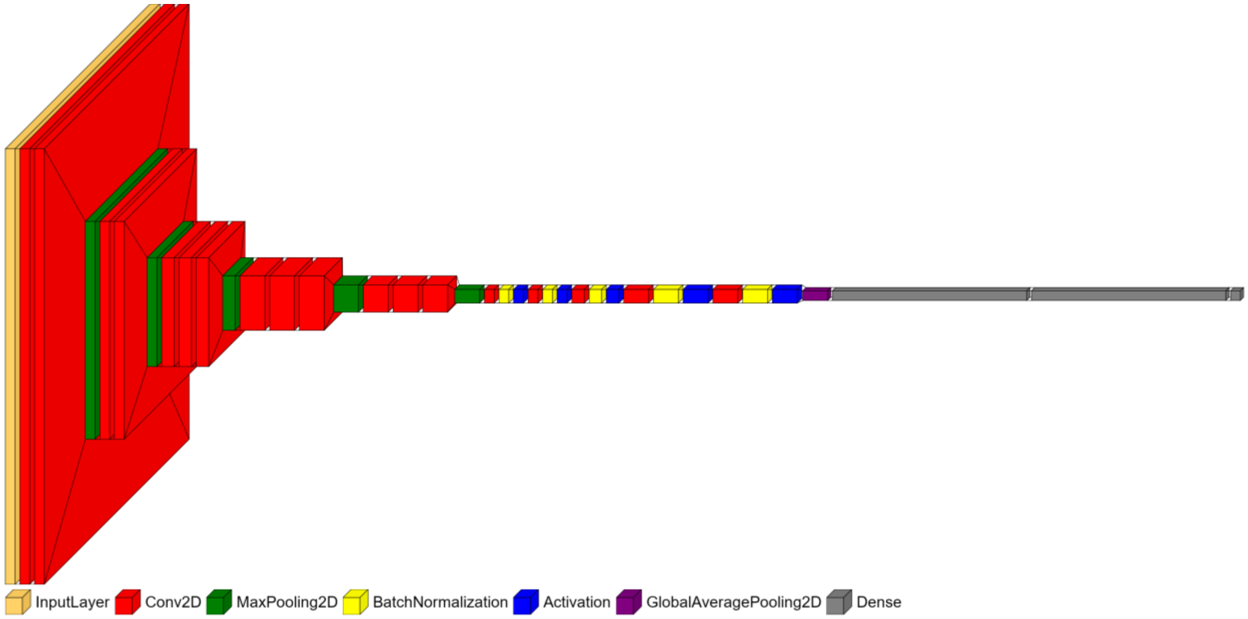

Figure S2: Schematic representation for visualization of the VGG16 model. The color map defines the different layer types in the visualization (input, convolution, max pooling, batch normalization, activation, global average pooling, and dense layers).

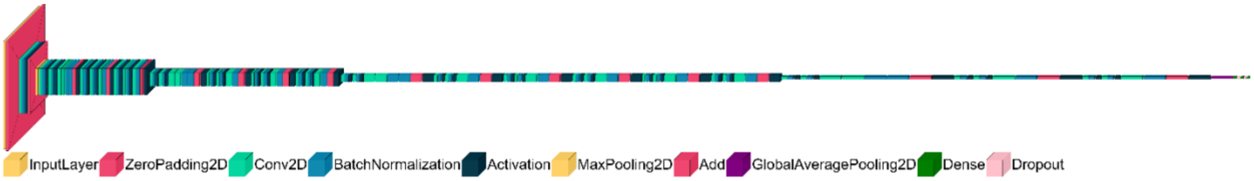

Figure S3: Schematic representation for visualization of the ResNet50 model. The color map defines the different layer types in the visualization (input, zeropadding, convolutional, batch normalization, activation, max pooling, adding, global average pooling, dense, and dropout layers).

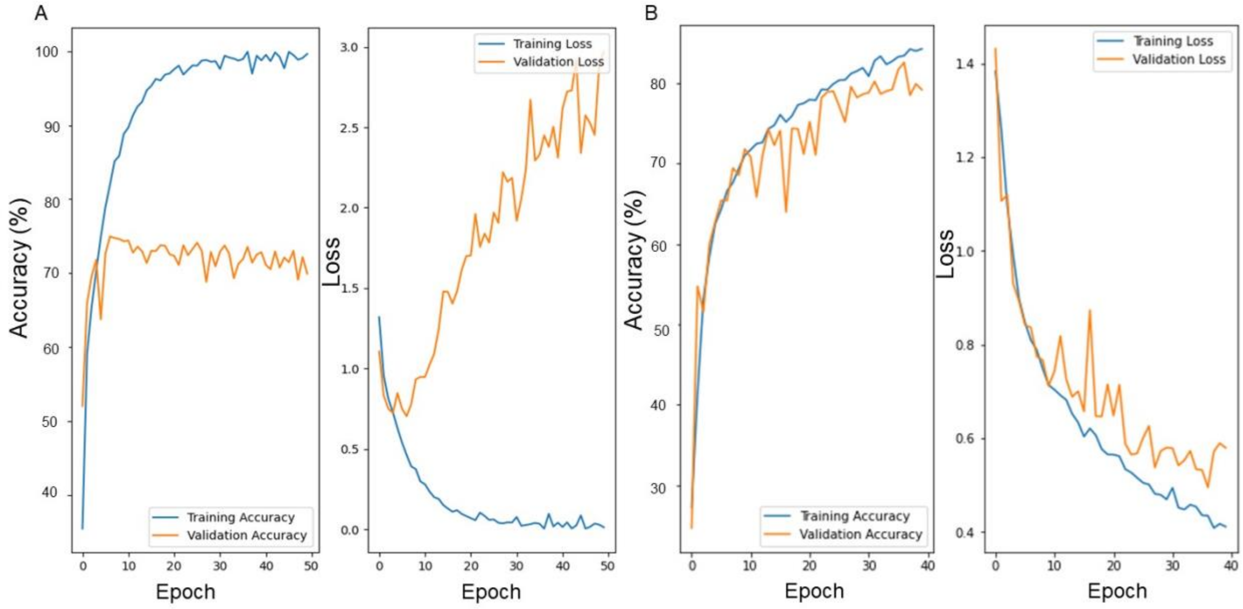

Figure S4: Training and validation accuracies of the custom CNN model before and after data augmentation. A) Training and validation accuracies and loss before data augmentation. B) Training and validation accuracies and loss after data augmentation.

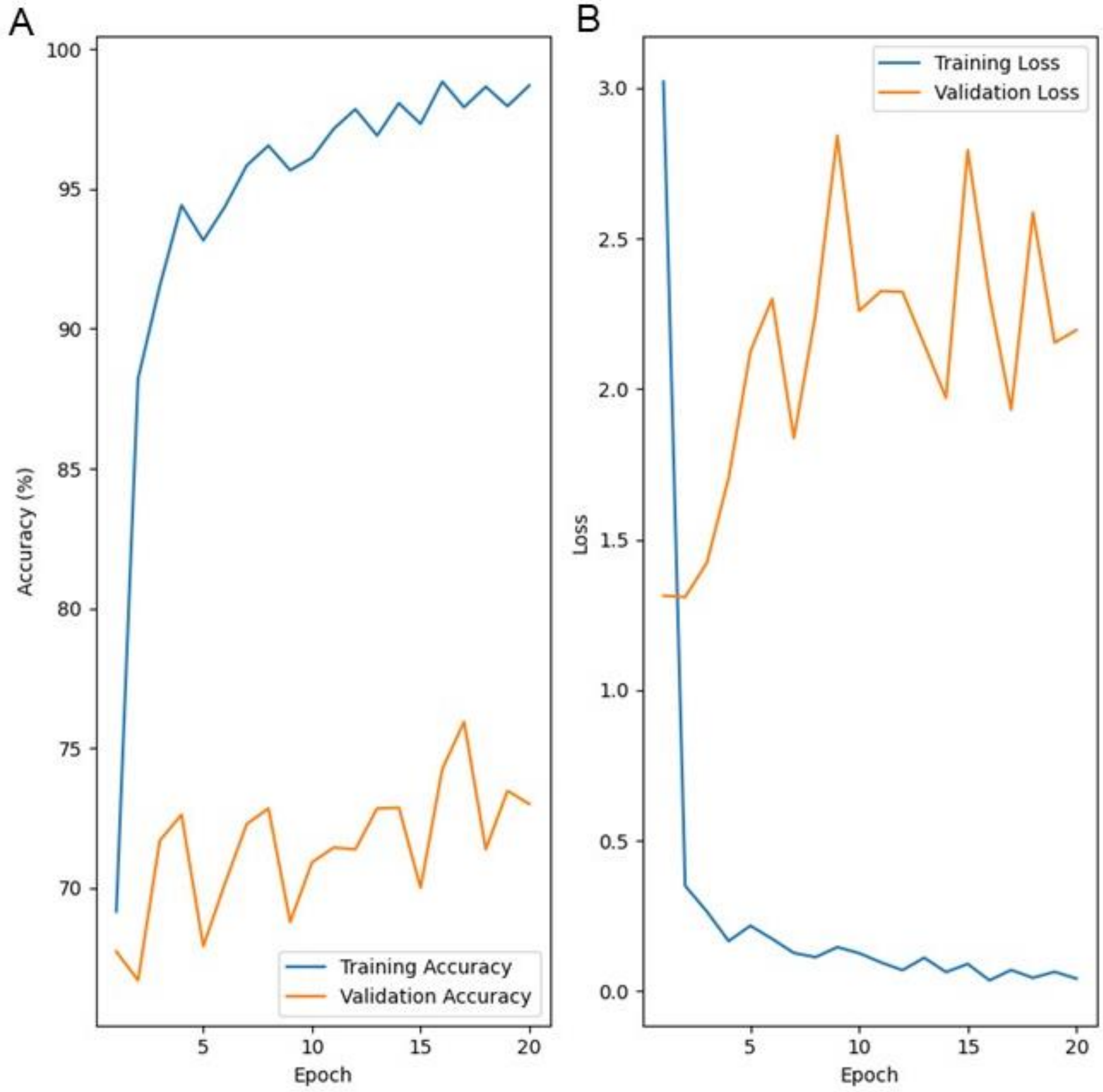

Figure S5: Training and validation accuracies and loss of the ResNet50 model. These metrics were evaluated for A) raw images, single cell, budding cell, and cell groups and for B) single cell and budding cell images.

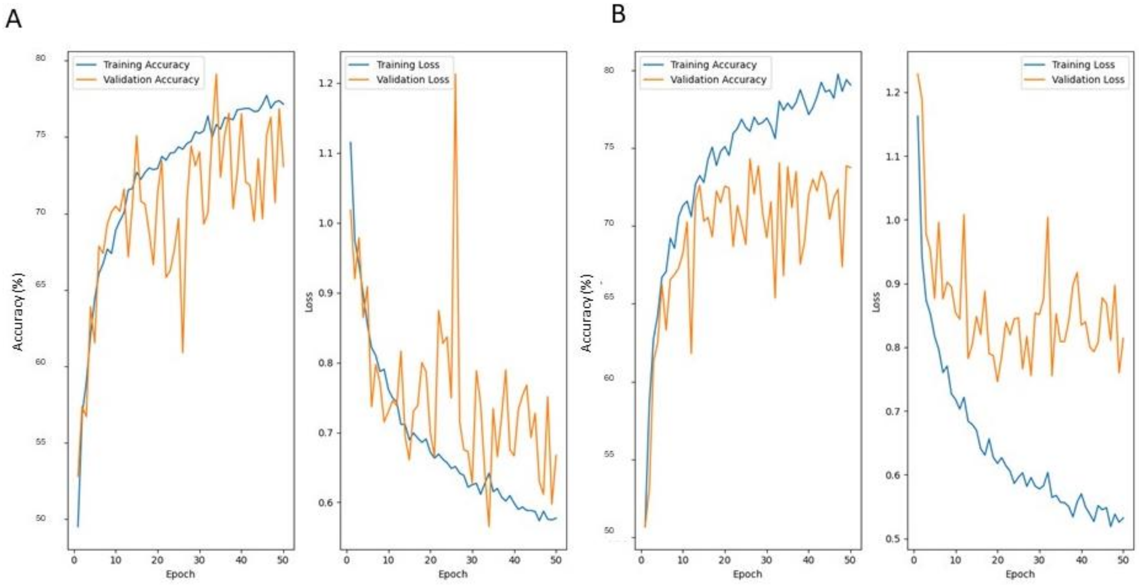

Figure S6: Training and validation accuracies and loss of the VGG16 model. These metrics were evaluated for A) raw images, single cell, budding cell, and cell groups and for B) single cell and budding cell images.

Supplementary Tables

Metrics	**Custom CNN (All images)**	***Candida albicans***	***Candida auris***	***Candida glabrata***	***Candida haemulonii***
Accuracy (%)	69.7	81.7	74.6	71.1	51.3
Sensitivity (%)	69.7	81.7	55.3	71.1	51.3
Specificity (%)	69.7	93.8	90.4	90.4	51.3
Precision (%)	72.0	78.2	55.3	72.5	82.0
Negative predictive value (%)	95.4	92.4	79.9	91.0	96.2
F_1 score (%)	69.6	79.9	63.5	71.8	63.1
AUC (%)	87.1	91.1	81.7	88.3	87.5

Table S1: Metrics of the custom CNN after data augmentation on raw cell, single cell, budding cell, and cell groups images. See “Definitions” section for quantitative definitions of each metric.

Metrics	**Custom CNN (SC and BC dataset)**	***Candida albicans***	***Candida auris***	***Candida glabrata***	***Candida haemulonii***
Accuracy (%)	69.7	81.7	74.6	71.1	51.3
Sensitivity (%)	69.7	81.7	74.6	71.1	51.3
Specificity (%)	85.5	93.8	90.4	90.1	85.5
Precision (%)	72.0	78.2	55.3	72.5	82.0
Negative predictive value (%)	94.3	92.4	79.9	91.0	96.2
F_1 score (%)	69.6	79.9	63.5	71.8	63.1
AUC (%)	87.1	81.7	81.7	88.3	87.5

Table S2: Metrics of the custom CNN after data augmentation on single cell (SC) and budding cell (BC) images. See “Definitions” section for quantitative definitions of each metric.

	Identified (%)	**Confidence (%)**
Raw images		
Candida albicans	75	93.9
Candida auris	57.5	90
Candida glabrata	72.7	82.6
Candida haemulonii	18	65.9
Single cell images		
Candida albicans	12.5	82.5
Candida auris	75.5	84.2
Candida glabrata	73	77.3
Candida haemulonii	20.5	65.67
Budding cell images		
Candida albicans	82	92.79
Candida auris	37	73.8
Candida glabrata	37.6	79.1
Candida haemulonii	21.4	68.4
Cell group images		
Candida albicans	24	56.2
Candida auris	32	76.2
Candida glabrata	51	33.1
Candida haemulonii	26	24.2

Table S3: Performance of the custom CNN model on different test image sets of four *Candida* species.

ResNet50 (All image dataset)	**Precision (%)**	**Recall (%)**
Candida albicans	25.0	26.6
Candida auris	21.5	20.1
Candida glabrata	23.6	19.6
Candida haemulonii	25.2	29.6

Table S4: Precision and recall values for the four yeast species on the validation set for the trained ResNet50 model.

	With data augmentation (single and budding cell datasets)	
	Precision (%)	**Recall (%)**
Candida albicans	68.2	88.5
Candida auris	84.8	63.0
Candida glabrata	82.8	55.7
Candida haemulonii	70.8	91.2

Table S5: Precision and recall for four different *Candida* species after training InceptionV3 for all images and training InceptionV3 on single and budding cell images together.

Metrics	**VGG16**	***Candida albicans***	***Candida auris***	***Candida glabrata***	***Candida haemulonii***
Accuracy (%)	73.0	93.3	84.2	51.8	61.0
Sensitivity (%)	72.6	93.3	84.2	51.8	61.0
Specificity (%)	72.8	80.0	72.1	77.6	61.5
Precision (%)	73.1	80.0	72.1	77.6	61.5
Negative predictive value (%)	95.1	93.3	84.2	51.8	61.0
F_1 score (%)	72.2	86.1	77.6	62.1	61.3
AUC (%)	92.7	98.0	94.6	90.0	88.0

Table S6: Results of VGG16 model and for each class of *Candida* species trained on raw cells, single cell, budding cell, and cell groups images. See “Definitions” section for quantitative definitions of each metric.

Metrics	**VGG16**	***Candida albicans***	***Candida auris***	***Candida glabrata***	***Candida haemulonii***
Accuracy (%)	73.7	76.6	75.0	71.6	71.8
Sensitivity (%)	73.7	81.0	66.7	66.2	81.0
Specificity (%)	73.5	71.3	76.0	76.2	71.3
Precision (%)	73.7	76.6	75.0	71.6	71.8
Negative predictive value (%)	73.7	76.6	75.0	71.6	71.8
F_1 score (%)	73.5	78.7	70.6	68.8	76.1
AUC (%)	91.2	94.6	87.2	89.3	93.7

Table S7: Results of VGG16 model and for each class trained on single cell and budding cell images alone. See "Definitions" section for quantitative definitions of each metric.

References

1. Bolin E, Lam W. A review of sensitivity, specificity, and likelihood ratios: evaluating the utility of the electrocardiogram as a screening tool in hypertrophic cardiomyopathy. *Congenit Heart Dis*. 2013;8(5):406-410. doi:10.1111/chd.12083

2. Classification: Precision and Recall | Machine Learning | Google for Developers. Accessed August 30, 2023. https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

3. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022;12(1):5979. doi:10.1038/s41598-022-09954-8

4. Classification: ROC Curve and AUC | Machine Learning | Google for Developers. Accessed August 30, 2023. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc