

A Design of End-to-End CNN Architecture for Colon Cancer Detection and Visualization Using ContinuousData Cleaning Method

Jie Li

Taylor's University

Weiwei Goh

weiwei.goh@taylors.edu.my

Taylor's University

Noor Zaman Jhanjhi

Taylor's University

Article

Keywords:

Posted Date: September 19th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6533575/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Design of End-to-End CNN Architecture for Colon Cancer Detection and Visualization Using Continuous Data Cleaning Method

Jie Li¹, Weiwei Goh², and Noor Zaman Jhanjhi³

¹Digital Health and Medical Advancement Impact Lab, School of Computer Science, Taylor's University, Jalan Taylor's, Subang Jaya, 47500, Selangor, Malaysia

²Digital Health and Medical Advancement Impact Lab, School of Computer Science, Taylor's University, Jalan Taylor's, Subang Jaya, 47500, Selangor, Malaysia

³Digital Health and Medical Advancement Impact Lab, School of Computer Science, Taylor's University, Jalan Taylor's, Subang Jaya, 47500, Selangor, Malaysia

*weiwei.goh@taylors.edu.my

ABSTRACT

Colon cancer threatens the quality of life and causes deaths of mankind worldwide today. Convolutional Neural Networks (CNNs) are recognized as a robust and prevalent technique in deep learning methodologies for biomedical signals, including histopathology imaging and processing. As a diagnostic tool, medical CNNs gradually facilitate the decision-making process in clinical detection. However, there is still potential improvement for an accurate model to achieve millionth-level accuracy of confirmation. Beyond that, overfitting misleads performance of a model to properly recognize pattern especially to unseen data. Further, visualized result could facilitate interpretability of a detected results from clinical end.. This research aims to introduce an end-to-end colon cancer detection using CNN with continuous parametric data cleaning method and interpretative visualization approaches.

Based on the results of the alternative settings, the raw dataset used in Experiment 1 caused overfitting during the training process, while Experiments 2 and 3 overcame the overfitting after applying Gaussian 95% and 99% rules based on RGB color distribution result. Further, there were 10 extensive experiments conducted based on settings of Experiment 3 to ensure the reliability of the results, which showed 0.99996 averaged accuracy and 0.00044 averaged loss with 1,00 precision, recall, and F1-score. Lastly, a K-Mean clustering and 3D intensity visualized sample image for interpretability improvement in clinical diagnostic process. The comparison result indicated the proposed model gained superior accuracy compared to the benchmark studies from the past 5 years. This research contributed to an end-to-end CNN model in an accurate colon cancer detection.

Introduction

According to the 2024 Cancer Statistics Report, their observation indicated there are increasing cases of colon cancer in the young generation. In the past years, many of patients were diagnosed with more serious and advanced cancers in the late stages due to delay and inaccurate detection¹. Specifically, most of the misdiagnoses were caused by an unclear boundary of objects in manual detection process. This particular problem requires an effective and reliable detection approach in histopathological diagnosis.

In the biomedical sector, artificial intelligence (AI) facilitates medical diagnosis based on advanced computability of big data in algorithms². In the field of imaging and image processing, the concept of deep learning (DL) is advanced for algorithmic feature extraction that could mathematically and statistically understand hierarchical and complex underlying patterns³. Specifically, the convolutional neural network (CNN) is widely used in processing X-rays, MRIs, CTs, and histopathological images⁴.

During the past five years, research has focused on different CNN variants to advance colon cancer detection⁵ as a support tool in diagnosis. However, in terms of performance enhancement, there is still potential for a CNN model to further achieve millionth-level accuracy to obtain a reliably supportive detection. Many studies focused only on model performance rather than imaging ability to interpret results in a diagnostic process. Therefore, an interpretable result using the imaging technique could effectively support specialists in their decision-making process. Data quality significantly determines model training quality (e.g., overfitting) in the first place so as to its performance in the future⁶. Overfitting model shows misleading performance in learning samples but shows weak identifying ability to handle unseen data. Overfitting occurs when a model is learning from noise and fluctuation rather than recognizing underlying pattern. Nevertheless, many studies ignore the application of

data cleaning as well as addressing potential overfitting in colon cancer histopathological image processing from statistical perspective. The objectives of the study are:

1. To design and develop a CNN for colon cancer detection at an accuracy level above 99.99%.
2. To apply data cleaning approach to reduce overfitting challenges in model training.
3. To visualize histopathological images after detection for improvement in interpretability in a clinical context.

In this case, this paper aims to introduce an end-to-end CNN solution to effectively detect colon cancer with a comprehensive performance measurement and visually interpretable result for histopathologists. The motivation of the research is to promote a reliable, accurate, and interpretable solution in the histopathological diagnosis.

Related Works

CNN models have recently gained popularity as an effective tool for detecting colon cancer from medical images, as they are more accurate and time-efficient than previous diagnosis techniques. The CNN models used the most are ResNet^{7–10}, Inception V3^{8,11,12}, VGGNet^{13,14}, DenseNet^{15,15,16}, and EfficientNet^{17–19}.

Specifically, Adu et al.²⁰ proposed a network based on the dual horizontal squash capsule approach to predict colon cancer tissue using LC25000. Their result showed 99.2% accuracy. Studies by Masud, et al.²¹ and Karim, et al.²², respectively, used CNN with DIP and Double Clahe showed 96.3% and 98.2% accuracy performance. In 2022, Talukder, et al.²³ proposed a CNN Model extracting hybrid ensemble features from the LC25000 dataset, and it showed model accuracy of 99.3%, which slightly improved²⁰ model performance by 0.1%. In the same year, Hage Chehade, et al.²⁴ examined model performances using XGBoost, SVM, RF, LDA, MLP, and LightGBM and gained results up to 99.0%. Also, Kumar, et al.²⁵ used DenseNet 121, which had a model accuracy of 98.6%, and Mehmood, et al.²⁶ utilized AlexNet to examine model performance with 98.4% accuracy using the same dataset—LC25000. In 2023, the accuracy of the detection model significantly improved performance metrics. Hadiyoso, et al.²⁷ adopted CNN with CLAHE and VGG16 and gained 99.0% accuracy, while Sethy, et al.²⁸ compared AlexNet, wavelet, and SVM, resulting in the highest accuracy of 99.3%. Nevertheless, in 2024, Uddin, et al.⁵ used Multi-Modal Efficient-Net improved accuracy to 0.999 with 0.001 loss after 10 experiments.

In a broader range of research, researchers have applied different colon cancer datasets over the past 3 years using various methods and evaluation metrics. In 2020, Zhou, et al.²⁹ used the 3D-Res-Net method to analyze IMRT data from 122 postoperative rectal cancer cases; they concluded there was -1.94% to 1.58% of prediction bias from experimental results. In the same year, Sarwinda, et al.⁸, adopted DenseNet-121 with KNN, achieving 98.53% accuracy in detecting colon cancer using a 10k histopathological image dataset. In 2021, Hamida, et al.³⁰ and Kuritcyn, et al.³¹ applied the WSI dataset to detect colon cancer tissue from the benign tissue.³⁰ applied DenseNet to achieve 96.9% accuracy whereas Kallipolitis, Revelos, and Maglogiannis³¹ used EfficientNet B4 but only achieved 73.0%. In the same year,³² employed EfficientNet B0-B6 and received a remarkable 99% accuracy in the model performance test. Jiang et al.³³ and Gerwert et al.³⁴ used clinical datasets, respectively, to achieve the goal of detecting colon cancer using ResNet-50 and VGG-Net, and their accuracy of the model was at the 90% level.

Methods

Data Collection

In this research, the experiments evaluated the performance in detecting colon cancer from histopathological images. Secondary data was retrieved from the Larxel Lung and Colon Cancer Histopathological Image Set (LC25000), which recorded 10,000 colon cancer histopathological images labeled benign (n=5000) and adenocarcinoma (n = 5000).

Measures

In this investigation, the raw data was measured using Gaussian distribution³⁵, which is a continuous distribution method for the identification of parametric outliers in terms of improving the quality of the data. The measures for outlier experimented both 95% and 99% normal distribution standards. Model evaluation included performance metrics (accuracy and categorical loss)³⁶, classification report (precision, recall, F1-score, and support)³⁷, and a confusion matrix³⁸ for a comprehensive understanding to model prediction. Lastly, the visualization adopted multiple-interpolation³⁹ to standardize intensity of the applied dataset.

Model Architecture

As shown in Figure.1, 768 * 768 pixels images were marked on a RGB matrix in the initial step. Continuous Parametric Gaussian 99% distribution approach was used as the data-cleaning strategy for data quality improvement by removing outliers.

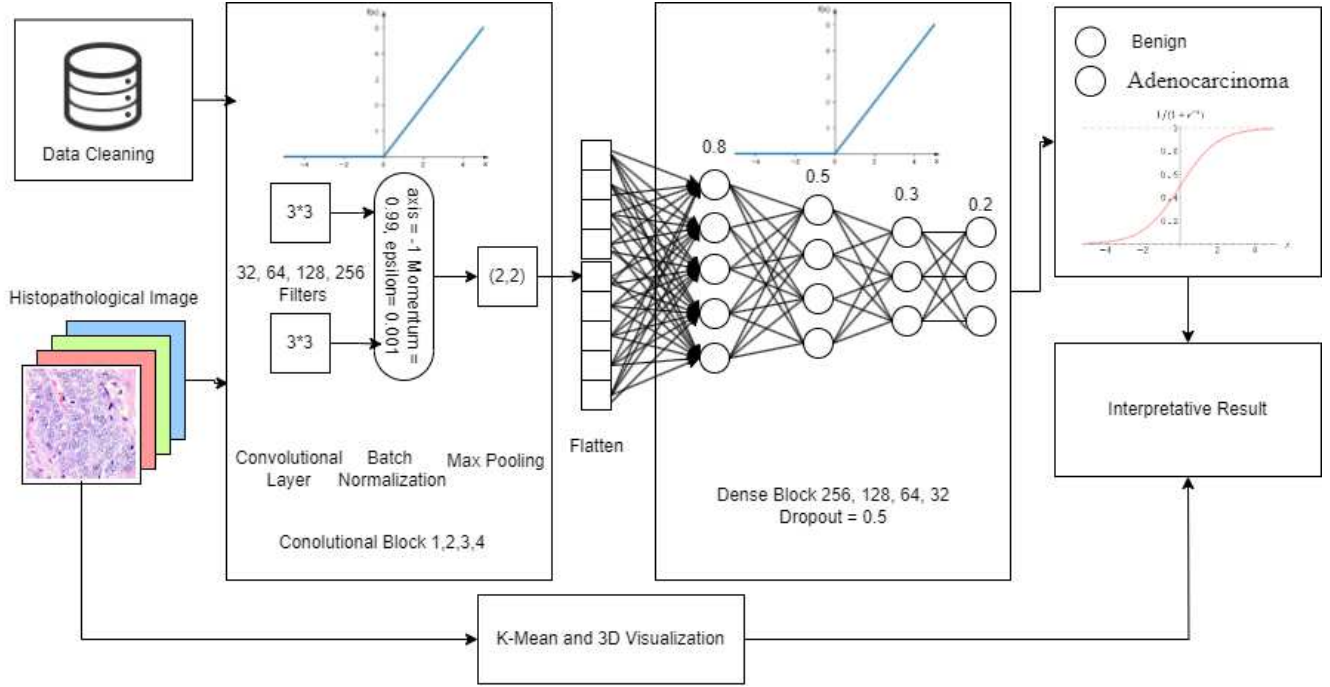


Figure 1. Model Architecture

$$y = f(x|\mu, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} E\left(-\frac{(x-\mu)^2}{(2\sigma^2)}\right) \quad x \in [\mu - 2.576\sigma, \mu + 2.576\sigma] \quad (1)$$

In Eq.1, the input variable is x , and μ stands for the mean value of distribution, squared σ is the variance, and σ , thus, is the standard deviation. Euler's Number (2.718)⁴⁰ is denoted as E . The value μ varies according to the z -score. When the 99% probability standard is applied, the value μ varies in the range of $\sigma = \pm 2.576$. While if 95% probability standard is applied, the value μ varies in the range of $\sigma = \pm 1.960$.

Then, the convolutional layer, which contains 4 individual blocks with 32, 64, 128, and 256 (3, 3) dimensioned filters, detects features in the preprocessed data using the ReLU activation (Eq.2), which explained the non-linearity of complex patterns recognition by vanishing gradient problems. If there is a positive input x , the output $\hat{y}$ remains x , but if there is a negative input x , the $\hat{y}$ changes to 0.

$$\hat{y} = f(x) = \text{MAX}(0, x) \begin{cases} y & y \geq 0 \\ 0 & y < 0 \end{cases} \quad (2)$$

Afterward, Batch Normalization standardized the inputs to each layer to improve the training speed, stability, and performance of the proposed model. As shown in Eq.3, the normalized output can be calculated from mean(μ_b) and variance (σ_b) with constant C to stabilize the model using learnable parameters γ and β for scaling and shifting the output data ($\hat{y}$) from the original RGB matrix (x_i).

$$\hat{y} = \gamma \left(\frac{x_i - \mu_b}{\sqrt{(\sigma_b^2 + C)}} \right) + \beta \quad (3)$$

To reduce spatial size in Eq.4, the max pooling layer (Max) operates a (2,2) dimensioned filter over the feature map, where $\hat{y}$ at position (i, j) in the feature map can be pooled into the (2,2) window by indices (R) at position $(m, n) \in (0, 2)$.

$$\hat{y}_{(m,n)} = \text{Max}(R_{(i,j)} \in \hat{y}_{(i,j)}) \quad (4)$$

Then, our proposed model flattened the output and compares the features in a dense 4-block layer with 0.8, 0.5, 0.3, and 0.2 learning rates again using ReLU activation in the following neural blocks. The final step of feature extraction (in Eq.5) executed the global average pooling to decide the calculated matrix in a neuron ($\hat{y}'$) for final decision making.

$$\hat{y}' = \frac{\sum_{m,n=0}^2 \hat{y}_{(m,n)}}{4} \quad (5)$$

The final output (Y) was labeled benign or adenocarcinoma using the SIGMOID function (see Eq.6). The function finalizes the output value in the proposed model where the input ($\hat{y}'$) is positive, the SIGMOID value outputs 1, $\hat{y}'$ is negative, and the SIGMOID value outputs 0. However, if $\hat{y}'$ is zero, the SIGMOID marks 0.5 instead, smoothly transforming the Y from 0 to 1 (see Fig.1) to determine a binary decision.

$$Y = f(\hat{y}') = \frac{1}{1 + e^{-\hat{y}'}} \begin{cases} \hat{y}' & \hat{y}' > 0 \\ 0.5 & \hat{y}' = 0 \\ 0 & \hat{y}' < 0 \end{cases} \quad (6)$$

To improve interpretability, K-mean clustering can be used as part of colon cancer detection through image analysis. In Eq. 7, it applies a number (k) of clusters (C) to calculate distance from each cluster to the original pixel value $x_{(i,j)}$

$$I_{k-mean}(i, j) = \sum_{l=1}^n \sum_{j=1}^k (x_{(i,j)} - C) \quad (7)$$

A 2D histopathological image can be represented as a matrix ($I(i, j)$). First, we ensure the normalization of the matrix shown in Eq.8.

$$\begin{aligned} x &= I_{norm}(i, j) \\ &= \frac{I(i, j) - \mu}{\sigma} \end{aligned} \quad (8)$$

Then, stacking or Interpolating to 3D along the z-axis (k) extends depth of the original 2D image of x indexed by the $\hat{y}'$ in Eq.9:

$$\begin{aligned} x &= I(i, j) \\ &= (1 - \alpha)I_{k_1}(i, j) + \alpha I_{k_2}(i, j) \\ &= I(i, j, k) \end{aligned} \quad (9)$$

However, when multi-slice registration happens(n), the intensity standard needs Weighted Distance-Based Interpolation $W_k(n)$ in Eq. 10 where d_k is the distance on k from the n^{th} to the rest image slices with power parameter ($p = 2$).

$$W_n(k) = \frac{d_k^{-p}}{\sum_{l=1}^n d_l^{-p}} \quad (10)$$

Hence, the multiple interpolation of intensity visualization is written in Eq.11:

$$I(i, j, k) = \sum_{l=1}^n W_n(k) I_l(i, j) \quad (11)$$

In summary, the above-proposed end-to-end model architecture with algorithms structured from data preprocessing to interpretative visualization as an accurate, rapid, and interpretable tool for comprehensive colon cancer diagnosis.

Result

Descriptive Analysis

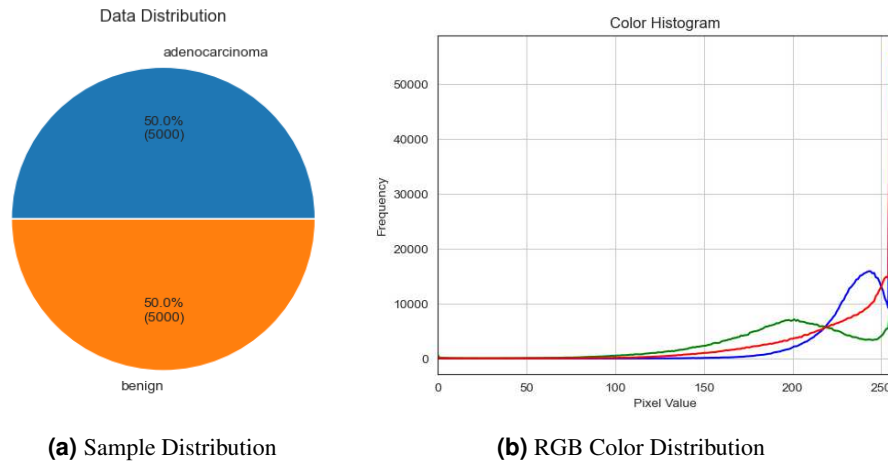


Figure 2. RGB Color and Sample Distribution

The Figure.2(a) displayed sample dataset imported. In Figure.2 (b), the color distribution test showed how the RGB colors vary in the sample dataset, providing valuable insights to statistically understand the histopathological nature of the sample data. In results, blue has the highest frequency, followed by red and green. These insights constructively guide the following stages of preprocessing, augmentation, and model training. Table.1 summarized the color distribution test result of the 10,000 images

Index	Value
Mean pixel value	206.35
Blue channel	232.90
Green channel	177.90
Red channel	208.27
The standard deviation of pixel values	43.65
Blue channel	4.67
Green channel	12.92
Red channel	10.61
Min pixel value	0
Max pixel value	255
Average width	768.00
Average height	768.00

Table 1. Color Distribution Test

from the LC25000 dataset. The mean value corresponded to the arithmetic mean of the brightness levels of total sample. A mean value approaching 255 signified brighter, whereas a mean value approaching 0 signified darker images. The average value of 206.35 indicates that the images are bright over the RGB colors. Specifically, for the blue channel, the mean value was recorded at 232.90, the green channel had its mean value of 177.90, and the red channel has 208.27.

The standard deviation quantifies the dispersion of pixel values in a standard that a high standard deviation signifies a broad spectrum of brightness levels, resulting in great contrast. Conversely, a low standard deviation signifies a limited range of brightness levels, leading to low contrast. A standard deviation of 43.65 suggested moderate contrast in the images. Specifically, the blue channel had its standard deviation of 4.67, the green channel had its value of 12.92, and the red channel had 10.61 standard deviation.

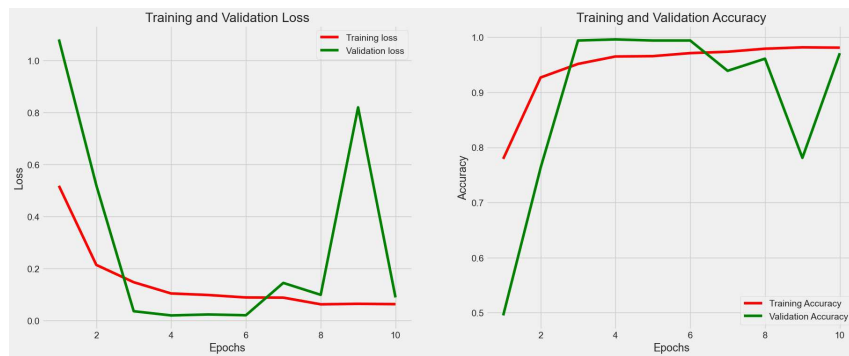
The MIN-MAX value represents the pixel with the lowest to the highest intensity in a group of samples. Specifically, a MIN value of 0 is typically used to indicate the entire black in the image. A MAX value represents the pixel with the highest brightness in dataset. A typical MAX value of 255 indicates pure white in the image. Also, the average width and height are 768 pixels, and the shape of the sample image is square. The histogram below showed the total frequency of the pixel value hit in green, red, and blue channels.

Model Optimization Experiment

The proposed model was developed. In order to monitor overfitting issue, an independent test subset was kept to validate the model's capability to unseen data. Hence, the dataset was randomly divided into training (80%), validation (10%), and test (10%) subsets. Further, the learning iteration and epoch only allows the proposed model to deal same image data once during the training process, so the training epoch was set to 10 by default. As shown in Figure.3(a) , both accuracy and loss were improved during the training and validation process until the 6th epoch, but both significantly dropped from the 8th epoch. After the 10th epoch, the model resulted 0.969 accuracy and 0.079 loss on the test dataset, which were significantly less than the its performance in training and validation processes. Hence, there was an evidence for overfitting in terms of ineffective learning.

Table 2. Model Performance Summary

Experiment	Sample Size	Splitting Rule	Epoch	Accuracy	Loss
1	10,000	80 - 10 -10	10	0.969	0.0794
2	9,542	80 - 10 -10	13	1.000	0.0007
3	9,906	80 - 10 -10	13	1.000	1.170e-05



(a) 10,000 Raw Dataset



(b) 9,542 Cleaned Dataset



(c) 9,906 Cleaned Dataset

Figure 3. Model Training Record in Full Dataset, 95%, and 99% Normal Distribution

For the overfitting issue addressed, the continuous distribution strategy was applied using parametric Gaussian's to identify and remove the outlier images for improving learning efficiency. By considering both overfitting and potential underfitting due to inadequate feature extracted, we respectively applied 95% and 99% standards in Experiments 2 (Figure.3(b)) and 3 (Figure.3(c)). From the results shown in Table 2, after applying the continuous parametric data cleaning approach, both test accuracy and loss were significantly improved. Notwithstanding, Experiment 3 had lower test loss and higher accuracy with less fluctuation in training and validation during learning process, meaning an effective noise reduction in the particular dataset using Gaussian 99%.

Classification Report and Confusion Matrix

Table 3 demonstrated classification report, explaining precision, recall, F1-score, and number of supports from Experiment 3 in the performance test of the detection model. The first metric, precision, indicated that statistically 100% of the adenocarcinoma and benign tissues in the test dataset were successfully predicted in actual observations.

Table 3. Classification Report

	Precision	Recall	F1 Score	Support
Adenocarcinoma	1.00	1.00	1.00	991
Benign	1.00	1.00	1.00	991
Accuracy			1.00	991
Macro Avg	1.00	1.00	1.00	991
Weighted Avg	1.00	1.00	1.00	991

Secondly, the recall metric explained the sensitivity of the proposed classification model. As a result, adenocarcinoma and benign have a 100% sensitivity rate, implying the true positive prediction was 100% correct in the binary decision-making process. Lastly, the F1-score explained the balance of precision and recall. From the figure above, adenocarcinoma and benign had an F1-score equal to 1.00.

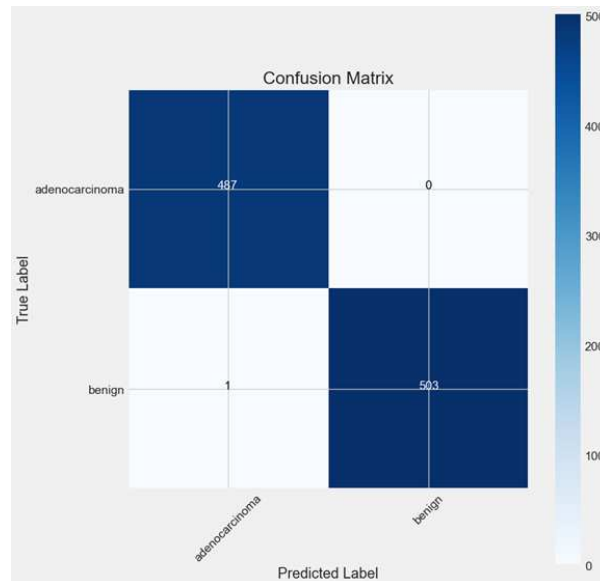


Figure 4. Confusion Matrix

The Figure. 4 demonstrated the Confusion Matrix for the Detection Model. The result showed that all of the test samples were correctly predicted as benign or adenocarcinoma. There were 991 test samples, and the result showed 487 adenocarcinoma and 504 benign tissue images were correctly identified. There were 0 False Positive and Negative identifications obtained.

Extensive Experiments

Table 4. Performance Metrics of the Extensive Experiments

Experiment	Accuracy	Loss	Category	Precision	Recall	F1	Support
3.1	1.0000	0.0007	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.2	1.0000	1.170e-05	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.3	1.0000	0.0017	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.4	1.0000	0.0003	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.5	0.9996	0.0011	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.6	1.0000	0.0002	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.7	1.0000	0.0001	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.8	1.0000	4.800e-05	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.9	1.0000	0.0002	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
3.10	1.0000	5.746e-05	adenocarcinoma	1.00	1.00	1.00	487
			benign	1.00	1.00	1.00	504
Avg	0.99996	0.00044	Macro Avg	1.00	1.00	1.00	991
Std DEV	0.00013	0.00056					

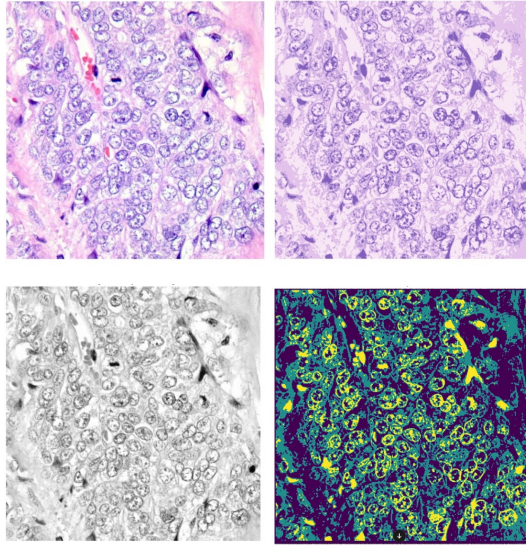
Table 4 showed comprehensive performance evaluation from the 10 extensive experiments from Experiment 3 with same model parameter setting to understand the reliability of results. The Figure. 6 in appendix recorded loss and accuracy in individual training process.

The highest accuracy was recorded as 1.000, and the lowest was 0.9996 from the 5th experiment. The lowest loss was 1.170e-05 from the 2nd experiment. In the classification report, precision, recall, and, F1-score of the experiments were all 1.00, and all the 991 testing samples were supported by the model prediction. The accuracy in average was 0.99996 with 0.00013 standard deviation and the loss in average was 0.00044 with 0.00056 standard deviation.

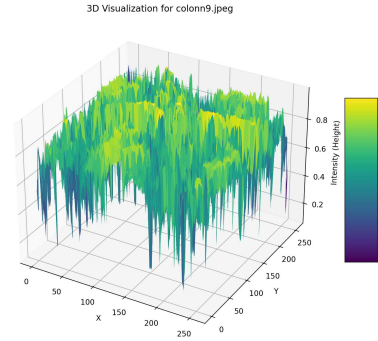
Thus, extensive experimental result indicated that the proposed lightweight model gained 0.9996 to 1.000 accuracy, 0.0067 to 4.800e-05 loss, and 1.00 precision, recall, and F1-score with 991 supported testing samples. The result also statistically implies there will be 95% probability the proposed model could have 0.99970 to 1.00000 accuracy and 0.00156 to 0.00000 loss within a normally distributed installations.

Result Visualization

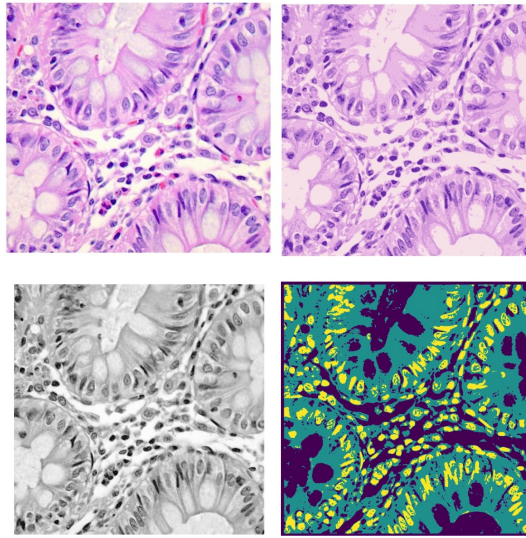
In Figure. 5 (a) and (c), both Adenocarcinoma and Benign tissue image were segmented using K-Mean Clustering based on Eq.7. The authors proceeded with K-Mean Clustering segmentation, and the result on the top right shows that the segmented image isolates distinct structures using clustering. Nuclear segmentation separates cell components (i.e., nuclei and cytoplasm) from the background in histopathological images ($x_{(i,j)}$) with $k = 3$, making it possible to clearly identify nuclear boundaries and measure their properties. This process helps clinicians assess cellular morphology, which is crucial for explaining abnormalities associated with colon cancer. In Figure.5 (b) and (d), the model functioned imaging on the output transformation to a 3-D imaging based on the multiple interpolation of intensity coordination (in Eq. 8 - 10). The intensity distance d_k was calculated in the matrix of $W_n(k)$ where $n = 9,906$ after the Gaussian's 99% distribution rule applied.



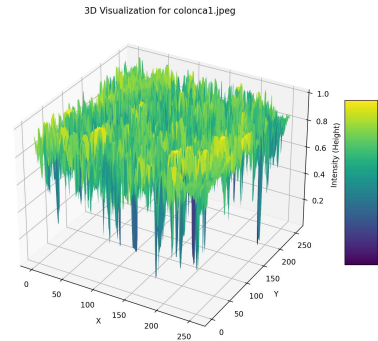
(a) Adenocarcinoma K-Mean Clustering Segmentation



(b) Adenocarcinoma 3D Intensity



(c) Benign K-Mean Clustering Segmentation



(d) Benign 3D Intensity

Figure 5. Adenocarcinoma and Benign Tissue Visualization

Discussion

Benchmark Comparison

Table 5 showed the accuracy results from previous benchmarks over the past 5 years using different datasets^{29–34} including WSI datasets, and specific patient samples. As listed in the summary table, the state-of-the-art benchmarks in the last 5 years showed overall accuracy of the model over 90%. However, in this comparison, the proposed model has the superior performance among all the models presented. The proposed model, after 10 extensive experiments, reached 0.99985 accuracy with 0.00111 loss in model performance metrics. Table 5 also compared current study with previous studies^{5,20–28,41–44} using LC25000 dataset over the past 5 years. As listed in the summary table above, the benchmarks in the last 5 years showed overall accuracy of the model over 94.0%. However, in this comparison, the proposed model had the superior performance among all the models presented. The current model outperforms all other listed models in terms of accuracy. Most models achieve a precision greater than 0.98 (98%), indicating generally high performance between different approaches. For example, Res-Net (99.87%)⁴², Inception (99.95%)²⁵, VGG-Net (99.00%)²⁷, Dense-Net (99.99%)⁴¹, and Efficient-Net (99.98%)⁵. Thus, we can indicate that the proposed model has the promising performance compared with the above benchmarks.

Table 5. Summary of Related Works on Cancer Detection Models

Ref.	Dataset	Method	Model Performance
33	590 Histopathology samples (China)	ResNet-50	$p < 0.05$, Acc > 0.9
34	629 Patient Samples	VGG-Net	AUROC: 0.9, AUPRC: 0.74
30	WSI Dataset	Dense-Net	Acc: 0.969
31	WSI Dataset	EfficientNet-B4	Acc: 0.73
32	2480 Benign / 5429 Malignant	EfficientNet B0-B6	Acc: 0.99
29	IMRT, 122 Rectal Cancer cases	3D-ResNet	Bias: -1.94% to 1.58%
5	LC25000	Multi-Modal EfficientNet	Acc: 0.999
41	LC25000	DenseNet	Acc: 0.999
42	LC25000	ResNet	Acc: 0.999
28	LC25000	AlexNet + SVM	Acc: 0.993
27	LC25000	CNN + CLAHE + VGG16	Acc: 0.990
43	LC25000	Inception V3	Acc: 0.945
24	LC25000	XGBoost, SVM, etc.	Acc: 0.990
25	LC25000	DenseNet-121	Acc: 0.986
23	LC25000	Hybrid Ensemble	Acc: 0.993
26	LC25000	AlexNet	Acc: 0.984
44	LC25000	VGG-16	Acc: 0.940
20	LC25000	DHS Capsules	Acc: 0.992
21	LC25000	CNN + DIP	Acc: 0.963
22	LC25000	Double CLAHE	Acc: 0.982
Current	LC25000	Custermized CNN	Acc: 0.99985

Normalized Color Distribution

The use of RGB color distribution analysis in data cleaning is particularly significant in the context of histopathological image analysis for colon cancer detection⁴⁵. RGB (Red, Green, Blue) color channels capture the full spectrum of color information in digital images, which is essential for distinguishing between different tissue types, staining patterns, and cellular structures⁴⁶. In our research, descriptive analysis on RGB color distribution provided a useful statistical overview on the input dataset for the further data cleaning process. By examining the distribution of pixel intensities across the RGB channels, researchers can identify and address inconsistencies or anomalies in the dataset. For example, images with unusual color distributions may indicate issues such as improper staining, scanner artifacts, or lighting inconsistencies. Detecting these outliers early allows for their correction or removal, ensuring that the dataset remains representative and free from technical biases.

Furthermore, analyzing RGB color distribution helps in standardizing the appearance of images across the dataset. This is crucial because variations in color can confound machine learning models, leading them to learn color-based artifacts rather than biologically relevant features. By normalizing color distributions, researchers can minimize the impact of such confounding factors, thereby improving the robustness and generalizability of the models.

Overcoming Overfitting

Overfitting occurs when a machine learning model learns not only the underlying patterns in the training data but also the noise and random fluctuations⁴⁷, which could lead to poor generalization on new, unseen data. In Experiment 1, the accuracy was significantly lower, and loss was significantly greater than training and validation. This implies overfitting during the model training process, since the model performed well in training, but underperformed to unseen test data. In this case, In Experiment 2, Gaussian 95% rule was applied to remove outliers to reduce the noise in original dataset. However, by applying data cleaning strategy, underfitting issue may occur due to insufficient generalization on underlying patterns from cleaned dataset. Thus, Experiment 3 further tested Gaussian 99% and the result showed a smooth, effective learning curves of both loss and accuracy metrics with less differences in understanding training, validation, and test samples. By systematically identifying and eliminating these sources of noise, researchers ensure that models are trained on data that accurately reflects the biological phenomena of interest. This process enhances the model's ability to generalize to new, unseen data and reduces the likelihood of learning spurious features that do not contribute to accurate diagnosis or prediction.

In summary, these data cleaning practices are essential for minimizing overfitting and enhancing the generalizability of machine learning models in medical image analysis. They ensure that models are trained with high-quality representative data, thus supporting reliable performance in real-world clinical applications.

Conclusion

In this research, applying the continuous parametric data cleaning and interpretative visualization approach, the proposed CNN model contributed to biomedical imaging and image processing in the discipline of colon cancer detection. In experiment

design, a random and independent test subset of data measured model performance to the unseen data to address if overfitting happened.

Regarding our research objectives, this study achieved the goal of state-of-the-art solution to detect colon cancer in averaged accuracy of 0.99996 and averaged loss of 0.00044 (See Table. 3 and Figure.6). Second, responding to the overfitting issue while training a model, a continuous parametric data cleaning method was introduced based on Gaussian 95% and 99% standards. In results, Gaussian 99% standard reduced noise effectively and sustained an efficient training process avoiding overfitting in feature generalization (See Table. 2). Lastly, the model presented K-Mean clustering and 3D intensity visualizations (See Figure. 5) to improve the interpretability of the detection result. Such a visualized result could facilitate diagnostic decision making process for the histopathologist in workplace.

Our limitations includes single source of input data (LC25000) and continuous distribution applied in data cleaning. In future, we would like to involve alternative data cleaning methods including non-parametric and discrete distribution standards across different datasets. Second, early detection is always critical for saving lives by the earliest. Hence, we are exploring the possibilities of our model and strategy in colonoscopy imaging and processing.

References

1. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA: a cancer journal for clinicians* **74**, 12–49 (2024).
2. Karalis, V. D. The integration of artificial intelligence into clinical practice. *Appl. Biosci.* **3**, 14–44 (2024).
3. Archana, R. & Jeevaraj, P. E. Deep learning models for digital image processing: a review. *Artif. Intell. Rev.* **57**, 11 (2024).
4. Abhisheka, B., Biswas, S. K., Purkayastha, B., Das, D. & Escargueil, A. Recent trend in medical imaging modalities and their applications in disease diagnosis: a review. *Multimed. Tools Appl.* **83**, 43035–43070 (2024).
5. Uddin, A. H. *et al.* Colon and lung cancer classification from multi-modal images using resilient and efficient neural network architectures. *Heliyon* **10** (2024).
6. Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D. & Khomh, F. Data cleaning and machine learning: a systematic literature review. *Autom. Softw. Eng.* **31**, 54 (2024).
7. Yang, Y. J. *et al.* Automated classification of colorectal neoplasms in white-light colonoscopy images via deep learning. *J. clinical medicine* **9**, 1593 (2020).
8. Sarwinda, D., Paradisa, R. H., Bustamam, A. & Anggia, P. Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer. *Procedia Comput. Sci.* **179**, 423–431 (2021).
9. Bodavarapu, P. N. R., Srinivas, P., Mishra, P., Mandhala, V. N. & Kim, H.-j. Optimized deep neural model for cancer detection and classification over resnet. In *Smart Technologies in Data Science and Communication: Proceedings of SMART-DSC 2021*, 267–280 (Springer, 2021).
10. Ansari, N., Faizabadi, A., Motakabber, S. & Ibrahimy, M. Effective pneumonia detection using res net based transfer learning. *Test Eng. Manag.* **82**, 15146–15153 (2020).
11. Belciug, S. Learning deep neural networks' architectures using differential evolution. case study: medical imaging processing. *Comput. biology medicine* **146**, 105623 (2022).
12. Iizuka, O. *et al.* Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci. reports* **10**, 1504 (2020).
13. Zhi, J., Sun, J., Wang, Z. & Ding, W. Support vector machine classifier for prediction of the metastasis of colorectal cancer. *Int. journal molecular medicine* **41**, 1419–1426 (2018).
14. Godkhindi, A. M. & Gowda, R. M. Automated detection of polyps in ct colonography images using deep learning algorithms in colon cancer diagnosis. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 1722–1728 (IEEE, 2017).
15. Utke, M., Zadtootaghaj, S., Schmidt, S., Bosse, S. & Möller, S. Nynetgaming-development of a no-reference deep cnn for gaming video quality prediction. *Multimed. Tools Appl.* 1–23 (2022).
16. Gao, F. *et al.* Deepcc: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **8**, 44 (2019).
17. Sarker, M. M. K. *et al.* A means of assessing deep learning-based detection of icos protein expression in colon cancer. *Cancers* **13**, 3825 (2021).

18. Faruk, M. F., Islam, M. R. & Hashi, E. K. Screening pathological abnormalities in gastrointestinal images using deep ensemble transfer learning. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, 230–235 (IEEE, 2022).
19. Izzaty, A. M. K., Cenggoro, T. W., Elwirehardja, G. N., Pardamean, B. *et al.* Multiclass classification of histology on colorectal cancer using deep learning. *Commun. Math. Biol. Neurosci.* **2022**, Article-ID (2022).
20. Adu, K. *et al.* Dhs-capsnet: Dual horizontal squash capsule networks for lung and colon cancer classification from whole slide histopathological images. *Int. J. Imaging Syst. Technol.* **31**, 2075–2092 (2021).
21. Masud, M., Sikder, N., Nahid, A.-A., Bairagi, A. K. & AlZain, M. A. A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors* **21**, 748 (2021).
22. Karim, D. Z. & Bushra, T. A. Detecting lung cancer from histopathological images using convolution neural network. In *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*, 626–631 (IEEE, 2021).
23. Talukder, M. A. *et al.* Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. *Expert. Syst. with Appl.* **205**, 117695 (2022).
24. Hage Chehade, A., Abdallah, N., Marion, J.-M., Oueidat, M. & Chauvet, P. Lung and colon cancer classification using medical imaging: A feature engineering approach. *Phys. Eng. Sci. Medicine* **45**, 729–746 (2022).
25. Kumar, N., Sharma, M., Singh, V. P., Madan, C. & Mehandia, S. An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images. *Biomed. Signal Process. Control.* **75**, 103596 (2022).
26. Mehmood, S. *et al.* Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing. *IEEE Access* **10**, 25657–25668 (2022).
27. Hadiyoso, S., Aulia, S., Irawati, I. D. *et al.* Diagnosis of lung and colon cancer based on clinical pathology images using convolutional neural network and clahe framework. *Int. J. Appl. Sci. Eng.* **20**, 1–7 (2023).
28. Sethy, P. K. *et al.* Lung cancer histopathological image classification using wavelets and alexnet. *J. X-Ray Sci. Technol.* **31**, 211–221 (2023).
29. Zhou, J. *et al.* A method of using deep learning to predict three-dimensional dose distributions for intensity-modulated radiotherapy of rectal cancer. *J. applied clinical medical physics* **21**, 26–37 (2020).
30. Hamida, A. B. *et al.* Deep learning for colon cancer histopathological images analysis. *Comput. Biol. Medicine* **136**, 104730 (2021).
31. Kuritcyn, P. *et al.* Comparison of cnn models on a multi-scanner database in colon cancer histology. In *Medical Imaging with Deep Learning* (2021).
32. Kallipolitis, A., Revelos, K. & Maglogiannis, I. Ensembling efficientnets for the classification and interpretation of histopathology images. *Algorithms* **14**, 278 (2021).
33. Jiang, L. *et al.* An improved multi-scale gradient generative adversarial network for enhancing classification of colorectal cancer histological images. *Front. Oncol.* **13**, 1240645 (2023).
34. Gerwert, K. *et al.* Fast and label-free automated detection of microsatellite status in early colon cancer using artificial intelligence integrated infrared imaging. *Eur. J. Cancer* **182**, 122–131 (2023).
35. Ilyas, I. F. & Rekatsinas, T. Machine learning and data cleaning: Which serves the other? *ACM J. Data Inf. Qual. (JDIQ)* **14**, 1–11 (2022).
36. Farhadpour, S., Warner, T. A. & Maxwell, A. E. Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. *Remote. Sens.* **16**, 533 (2024).
37. Lad, S., Chafekar, B. & Bide, P. Lung cancer classification using deep learning: A comprehensive approach with modified convolutional neural networks. In *2024 International Conference on Computational Intelligence and Network Systems (CINS)*, 1–6 (IEEE, 2024).
38. Liang, J. Confusion matrix: Machine learning. *POGIL Activity Clgh.* **3** (2022).
39. Suero Molina, E., Azemi, G., Russo, C., Liu, S. & Di Ieva, A. Artificial intelligence in brain tumors. *Comput. Neurosurg.* 201–220 (2024).
40. Yao, B., He, H., Kang, S., Chao, Y. & He, L. A review for the euler number computing problem. *Electronics* **12**, 4406 (2023).

41. Eliwa, E. H. I., El Koshiry, A. M., Abd El-Hafeez, T. & Omar, A. Secure and transparent lung and colon cancer classification using blockchain and microsoft azure. *Adv. Respir. Medicine* **92**, 395 (2024).
42. Wahid, R. R., Nisa, C., Amaliyah, R. P. & Puspaningrum, E. Y. Lung and colon cancer detection with convolutional neural networks on histopathological images. In *AIP Conference Proceedings*, vol. 2654 (AIP Publishing, 2023).
43. Swarna, I. J. & Hashi, E. K. Detection of colon cancer using inception v3 and ensembled cnn model. In *2023 International conference on electrical, computer and communication engineering (ECCE)*, 1–6 (IEEE, 2023).
44. Hossain, M., Haque, S. S., Ahmed, H., Mahdi, H. A. & Aich, A. *Early stage detection and classification of colon cancer using deep learning and explainable AI on histopathological images*. Ph.D. thesis, Brac University (2022).
45. Zhang, Q. *et al.* Unsupervised color-based nuclei segmentation in histopathology images with various color spaces and k values selection. *Int. J. Image Graph.* 2550061 (2024).
46. Hoque, M. Z., Keskinarkaus, A., Nyberg, P. & Seppänen, T. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Inf. Fusion* **102**, 101997 (2024).
47. Aliferis, C. & Simon, G. Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and ai. *Artif. intelligence machine learning health care medical sciences: Best practices pitfalls* 477–524 (2024).

Acknowledgements

This research is funded by Taylor’s Internal Research Grant Scheme - Impact Lab Grant (TIRGS-ILG/1/2023/SCS/001). A special thanks to the director and deputy director of Digital Health and Medical Advancement Impact Lab of Taylor’s University for advice and support throughout the duration of project.

Author contributions statement

Jie Li conducted experiments and wrote the content. Weiwei Goh and Noor Zaman Jhanjhi supervised experiment and result validation, as well as proofread the content of the paper.

Additional information

The authors declare no competing interests.

Data Availability

The secondary data used in this research were from LC25000, which is free and available source for global researchers at: <https://doi.org/10.48550/arXiv.1912.12142>.

Appendix

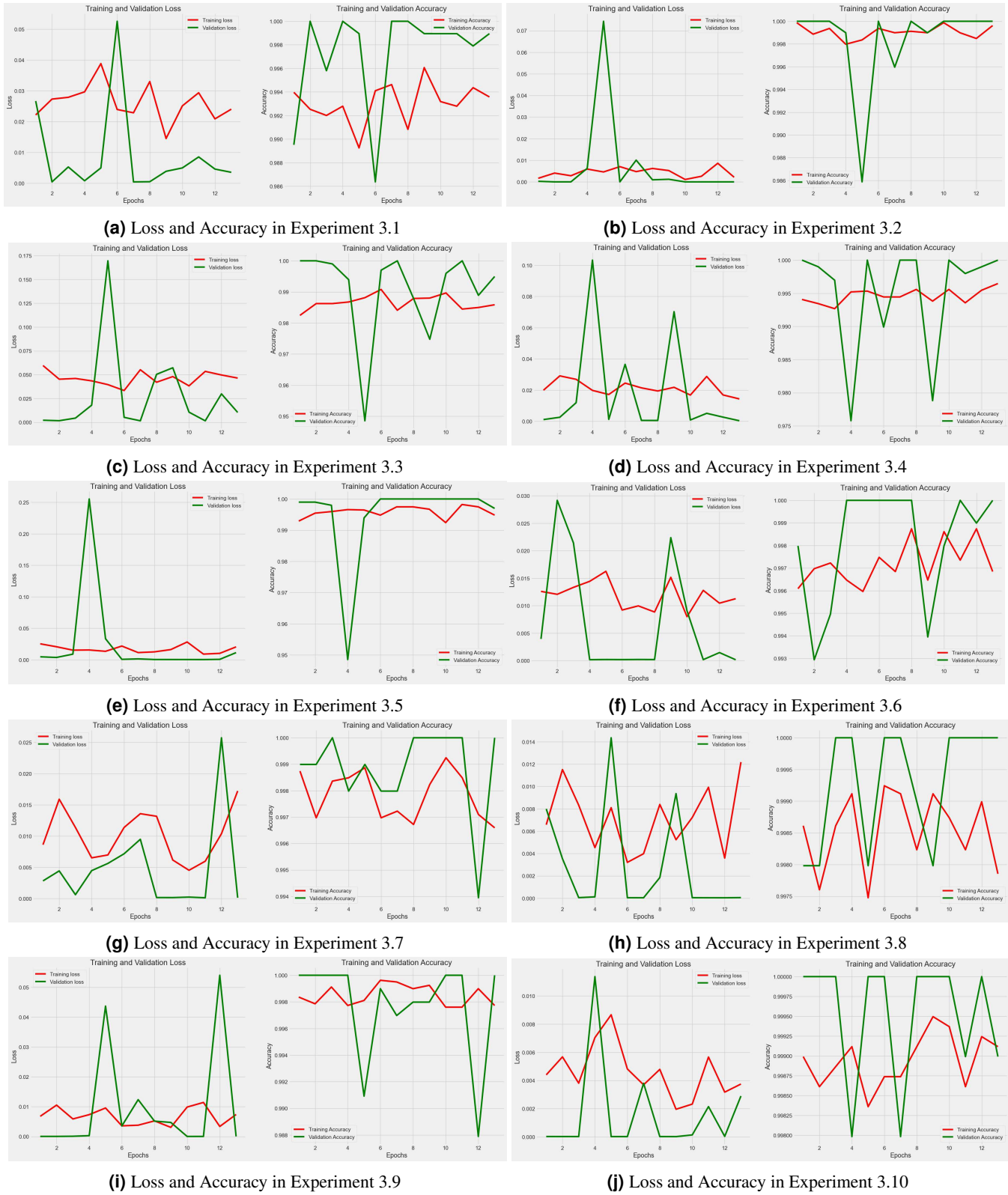


Figure 6. Training Process from Experiment 3.1 to 3.5