ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol.12 (2022) No.1, pp. 13-26
DOI:10.33168/JSMS.2022.0102

A Deepfake-Based Deep Learning Algorithm for Medical Data Manipulation Detection

YongSuk Kim, HoJung Song, JuHyuck Han

Konyang University, 158 Gwanjeodong-ro, Seo-gu, Daejeon, South Korea

yongsuk@konyang.ac.kr

Abstract. In the medical field, the abuse of manipulation data through image processing technology of deep learning is fatal. Therefore, research on detection of modulation on medical images is essential. The data set for fundus data manipulation used 356 right fundus images of 4 lesions (normal, diabetic retinopathy, glaucoma, macular degeneration) out of about 6,000 data collected by Shangong Medical Technology Co., Ltd. The training and verification dataset of the manipulation detection model used original data and U-Net manipulation data. In addition, data manipulated in the Cycle General Adversarial Network (GAN) model were used for the diversity of verification. In this paper, three ophthalmologists and two general doctors were asked to verify the above modulation data. Verification was requested for each lesion, and the verification results were shown through the Receiver operating characteristic (ROC) curve and the Area Under the Curve (AUC). The verification of this study evaluated a total of 100 randomly extracted manipulation data and original data as Observer Performance Test (OPT) for each group. When the evaluation results were digitized as average scores, the scores of ophthalmologists group: 0.72 and general doctors' group: 0.67 were recorded. The manipulated images were so similar that both ophthalmologists and general doctors could not find about 30%. However, the manipulation detection model studied in this paper was excellent in about 20% of the group OPT score with a lesion average of 0.913 in the same data group. Therefore, the manipulation detection model of this study finds the manipulated image and the original image well. The plan is to expand the scope of manipulation detection data to conduct research on various medical data. After that, it will verify its availability at the actual site.

Keywords: Deep Learning, Fundus Image, U-Net, Manipulation, Medical Image.

1. Introduction

With the recent development of image processing technology using deep learning, many studies have been conducted in various fields (Dhamo et al., 2020). Many studies are also being conducted in the medical field. In particular, cases of medical data regeneration using medical images are being announced one after another (Shen et al., 2017). The size of the precision medical market is $47.47 billion as of 2017, growing 13.3% annually. As the medical market grows, the number of clinical trials approved to verify the clinical efficacy of developed medical products also shows a steady increase (Korea Clinical Trials Information Center., 2021). However, image regeneration technology is causing many problems under the name Deep-Fake, and the possibility of abuse in the medical field is increasing. Deep-Fake was abused to avoid transparent experiments and verification by manipulating the conditions of patients' participation in clinical trials in an evaluation to verify clinical efficacy (Kim et al., 2019). In addition, medical data manipulated with Deep-Fake from the IRB (Institutional Review Board), which is essential for medical field research, can be approved in a negative manner. These problems can appear in all industries where testing and verification are conducted based on image data and can lead to critical social problems. Although many studies have been conducted on the image regeneration technology, it has not been confirmed whether such a manipulated image can be detected. Since problems caused by manipulated images in the medical field can be fatal, research on manipulated image detection is essential.

2. Materials and Methods

In this study, Materials and Methods are described in four configurations. The first data analysis describes the research data analysis and filtering process. The second describes the preprocessing process of data. Third, manipulation model and manipulation data verification, discusses the manipulation model based on U-Net and the manipulation data verification by ophthalmologists. In the last fourth, the manipulation detection model will be described.

2.1. Data Analysis

The dataset of this study utilized the dataset Ocular Disease Recreation' collected by Shangong Medical Technology Co., Ltd. of Kaggle (Larxel., 2020). This data is a dataset collected from 5,000 patients and consists of a image of the fundus of both eyes and a doctor's diagnostic keyword including the patient's age and gender. This data provides various image resolutions by capturing fundus images with various equipment in hospitals and medical centers in China. This study used four of the eight items classified in the dataset (normal, glaucoma, diabetic retinopathy, macular degeneration). Fig. 1 is the data configuration process used for model input.

Image quality for the performance of the artificial intelligence model was considered in the data selection of this study. Only image data of excellent quality with the same shape and resolution were used for the improvement of the quality of the manipulated image and the performance of the artificial intelligence model and consistency of learning data. Images with blurred retina or microvascular and images filled with black spaces of more than 30% were excluded from the data selection. In this study, only the fundus image of the right eye was selected out of about 6,000 data. Among them, three major ophthalmic diseases that cause blindness (glaucoma, diabetic retinopathy, macular degeneration) and normal fundus images were selected.

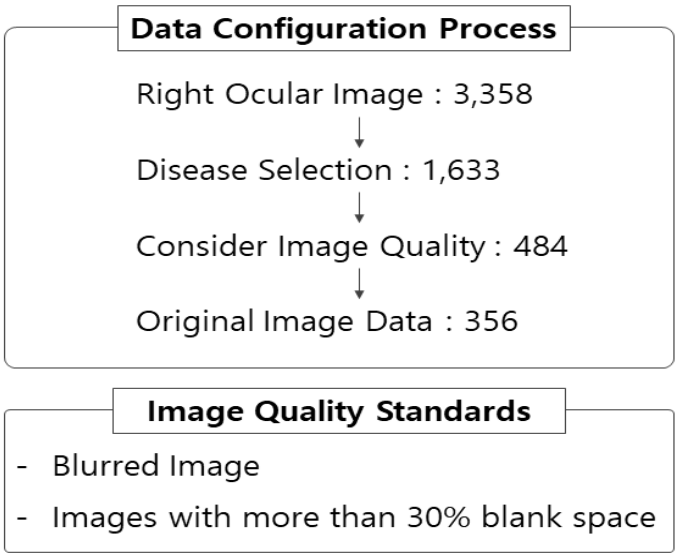

Fig. 1: Data Configuration Process

Table 1 is a table that configures the data used for model input. Of the about 6,000 data, about 356 data were selected in the manner of Fig. 1, and were classified for each lesion. As shown in Table 1, the classification results are 114 Normal, 67 Glaucoma, 78 Diabetic Retinopathies, and 97 Macular Degeneration. Normal data is the result of randomly selecting data of excellent quality among about 2,000 data. For lesion data, all data with excellent quality within the dataset were used, and a smaller number of data were used compared to Normal.

Table 1: Data Configuration

Disease	Normal	Glaucoma	Diabetic Retinopathy	Macular Degeneration
Number of Data	114	67	78	97

2.2. Data Preprocessing

In this study, two data preprocessing processes were conducted. For the efficiency of deep learning model learning, image resizing was applied to convert original data with a size of 512 x 512 into a size of 256 x 256. Through the image resizing process, the learning time of the deep learning model was reduced by more than three times to minimize time and economic loss. In addition, in order to increase the visual similarity between the original data and the manipulation data, sharp filter was applied to the original fundus image to perform preprocessing. Data preprocessing through sharp filter is a process of making the difference value between pixels and pixels larger. This proceeds by multiplying the value of the center pixel by the filtering coefficient to negatively digitize the value of the surrounding pixels so that the sum becomes 1. In this study, the middle value was filled with 5 and the edge was filled with 0, and the remaining values were filled with -1.

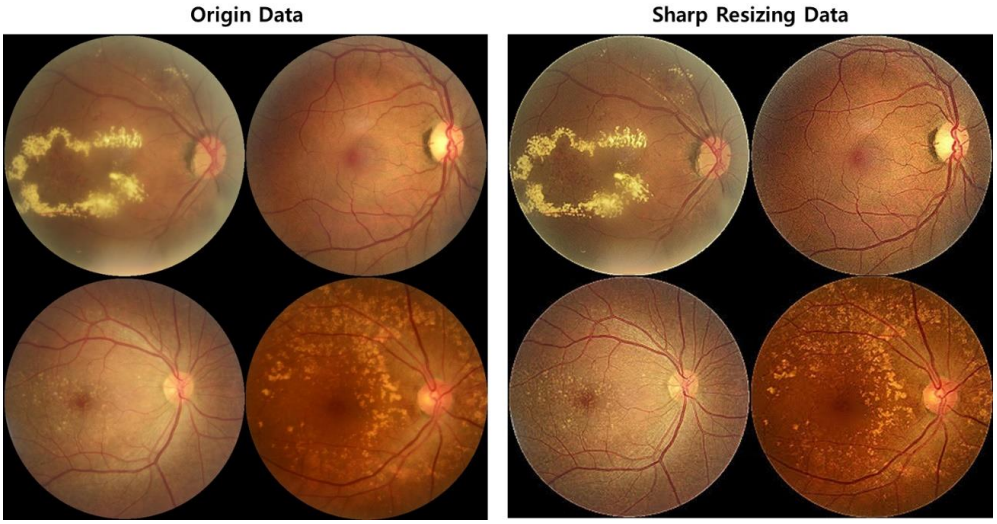

Fig. 2: Data Preprocessing(Sharp Filter, resizing)

Fig. 2 is the result of sharp filter and image resizing on the original Diabetic Retinopathy No. 15 and Macular Degeneration No. 2. For the readability of this paper, the sizes of the two data with different sizes were set to be the same and attached. As a result of preprocessing the data using a sharp filter, the overall color of the blurred data was adjusted. In addition, the contours of Microvascular or lesions have become more pronounced.

2.3. Manipulation Model and Manipulation Data Verification

The manipulation model uses the final screening data that has gone through the preprocessing process as an input. In this study, a deep learning model for manipulating fundus images is composed of U-Net. U-Net has segmentation capabilities specialized in image data and was selected to manipulate image data based on extracted features (Ronneberger et al., 2015). In this study, the use of the GAN (Generative Adversarial Network) model specialized in image manipulation was excluded to minimize time and economic loss. The results were confirmed by learning the Cycle GAN model and the U-Net model in parallel at the beginning of the study. When learning was conducted with the U-Net model, time was saved more than four times that of the Cycle GAN model. In addition, the quality of manipulated data has improved a lot.

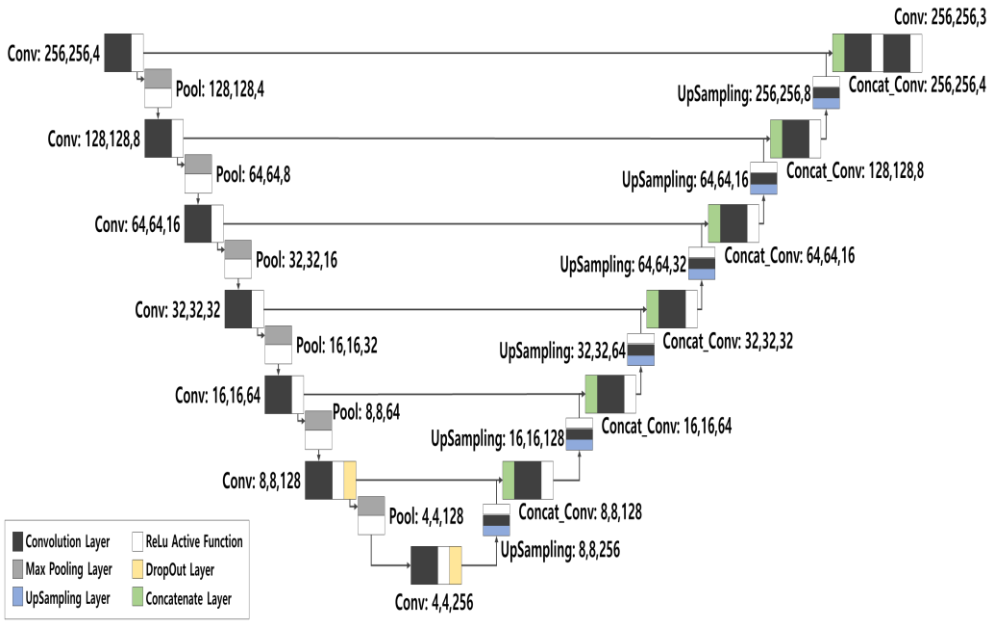

Fig. 3: U-Net Model Configuration

As can be seen in Fig. 3, the input of the model consists of four channels, not three channels of the image. This is in the form of adding a gray scale channel that is easy to extract features of blood vessels, and was constructed to include features of the area excluding the central disk from the fundus image. The feature of the U-Net model is that the network is symmetrical to identify the characteristics of the data. The U-Net model used in this study consists of a total of 7 layers. Each layer consists of 2D CNN (convolutional neural networks) and consists of nodes capable of storing about 2,000,000 weights. In addition, Adam (Adaptive moment

estimation) Optimizer was used as the optimization algorithm. Convolution operations were executed twice for each layer, and ReLu (Rectified Linear Unit) was used as an activation function for the temporal efficiency of model learning. In the left contracting path of Fig. 3, the size of the feature map is halved while performing the max-pooling operation at each layer. And each time down-sampling is performed, the number of channels doubles. In the 6th and 7th layers, dropout was used to prevent overfitting. In the right expanding path of Fig. 3, as opposed to the contracting path, the size of the Feature Map doubles in the convolution operation at each layer. And through up-sampling, the number of channels is reduced by half. In the last layer, data having the same size and channel as the input data is output. Thereafter, in order to remove noise from the manipulated data, the Blur effect was applied and post-processing was performed. The blur effect applied a Gaussian filter that flexibly changes the filter value according to distance using the standard deviation.

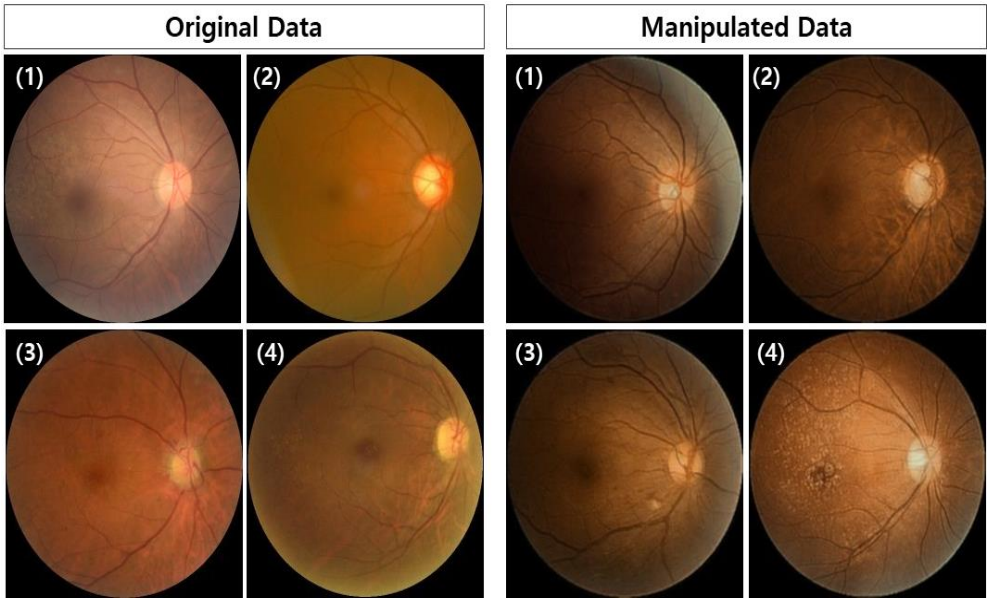

Fig. 4: Image comparison result

Fig. 4 shows the original data and manipulation data of (1) normal, (2) glaucoma, (3) diabetic retinopathy, and (4) macular degeneration. The model learned the characteristics of each lesion by itself without separate segmentation. Visually checked, there seems to be no significant difference between the original data and the manipulated data. In this study, an image data quality assessment technique was used to quantify the difference between the original data and the manipulation data. Data were verified using RMSE (Root Mean Square Error),

SSIM (Structural Similarity), and FID (Fréchet Inception distance). RMSE is a measure used to indicate a difference between original data and manipulation data. The smaller the difference between the data, the closer to zero is output (Mason et al., 2019). SSIM is a method of measuring similarity with original data for distortion caused by image data conversion. The more similar the data is, the closer the value is to 1.0 (Sara et al., 2019). FID calculates the distance between the feature vector of the original image and the feature vector of the manipulated image using Inception V3. The more similar the data is, the closer the value is to zero is output (Obukhov et al., 2020).

Table 2: Results of Image Quality Evaluation Indicators

Values / Disease	RMSE	SSIM	FID
Normal	38.58	0.65	254.32
Glaucoma	39.93	0.65	310.90
Diabetic Retinopathy	36.76	0.61	284.28
Macular Degeneration	37.78	0.64	253.49

Table 2 is the result of image quality evaluation indicators for manipulation data. Although it was difficult to visually distinguish between the original data and the manipulated data, there were many differences in image evaluation indicators. This is a phenomenon that occurs because when a filter is applied to data during preprocessing and post-processing, it looks visually similar, but there is a difference in pixel values (Wang et al., 2020).

In this study, data was verified by ophthalmologists and general doctors to evaluate the clinical effectiveness of manipulated data. The requested data consisted of 45 original data and 5 manipulated data, and verification was conducted in the form of randomly mixing the data to find the manipulated data. The number of medical personnel who participated was 3 ophthalmologists (more than 10 years of experience) and 2 general doctors (more than 5 years of experience). Verification was requested for each lesion, and the verification results were quantified into groups of ophthalmologists and general doctors using ROC (Receiver Operating Characteristic) Curve and AUC (Area Under the ROC Curve). ROC Curve and AUC are widely used as performance evaluation indicators for models that distinguish classes. The higher the AUC, the better the performance of the class-separating model (Huang et al., 2005).

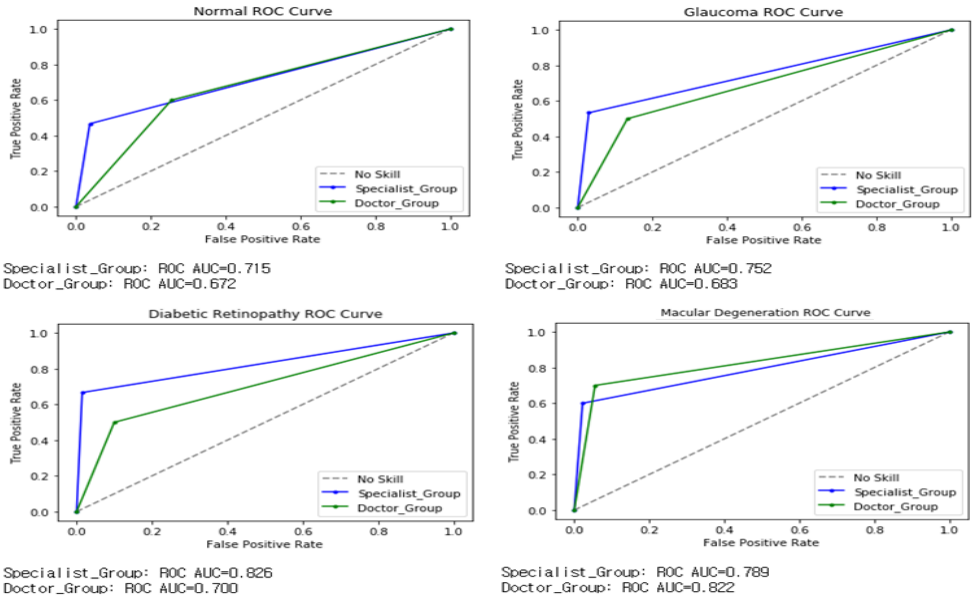

Fig. 5: Results of doctor group`s evaluation

Fig. 5 visualizes and shows the ROC Curve and AUC of the ophthalmologist and general doctors group for each lesion. The horizontal axis is a case in which data that is not manipulated by False Positive is determined as manipulated data. The vertical axis is a case in which data manipulated with True Positive is determined as manipulated data. The part of the graph where the line is bent can be seen as an indicator of classification. The area below the line of the graph is AUC, and the higher the detection performance, the closer to 1 is output. Data verification did not provide patient information of the manipulated image, so the original was conducted in the same manner. In order to minimize the effect of omission of patient information on image classification, doctors in each group classified only images excluding patient information. The AUC results of all lesions except macular degeneration in Fig. 5 were more predominant in the ophthalmologist group. In the normal case, the AUC of the ophthalmologist group was 0.715, and the AUC of the general doctor group was 0.672. In the case of glaucoma, the AUC of the ophthalmologist group was 0.752 and the AUC of the general doctor group was 0.683. In the case of normal and glaucoma, similar detection performance was shown. However, in the case of diabetic retinopathy, the AUC of the ophthalmologist group was 0.826, which best found the manipulated data. In addition, in the case of macular degeneration, the AUC of the general doctor group was 0.822, which was about 3% higher than that of the ophthalmologist group.

The overall statistics of this data verification can be seen as the average of the AUC scores. The average of the AUC scores was 0.77 for ophthalmologists and 0.71 for general doctors. This means that the ability to detect the manipulation of the fundus image varies depending on the doctor's proficiency. In addition, the longer the doctor's career, the higher the overall detection ability. Furthermore, the data manipulated with the U-Net artificial intelligence model of this study can be seen as having no visual difference from the original data so that ophthalmologists and general doctors cannot find 20-30%.

2.4. Manipulation Detection Model

The manipulation detection model of this study uses the structure of Sparse CNN, a deep learning network (Liu et al., 2015). The input of the model is 256 x 256, using the original fundus data used for learning the manipulation model and data manipulated through the manipulation model. In addition, for the development of a model with the performance of detecting all manipulated data in various ways, manipulated data in the Cycle GAN model is included in the model learning and testing. The dataset used for model learning at Cycle GAN utilized Keggle's 'MESSIDOR-2 DR Grades' and 'Glaucoma Detection' (Webster., 2018, Zhang., 2021). Each dataset includes image data of diabetic retinopathy and glaucoma and a doctor's diagnosis. The input data of the Cycle GAN model progressed to a size of 256 x 256, the same as the input data of the U-Net model.

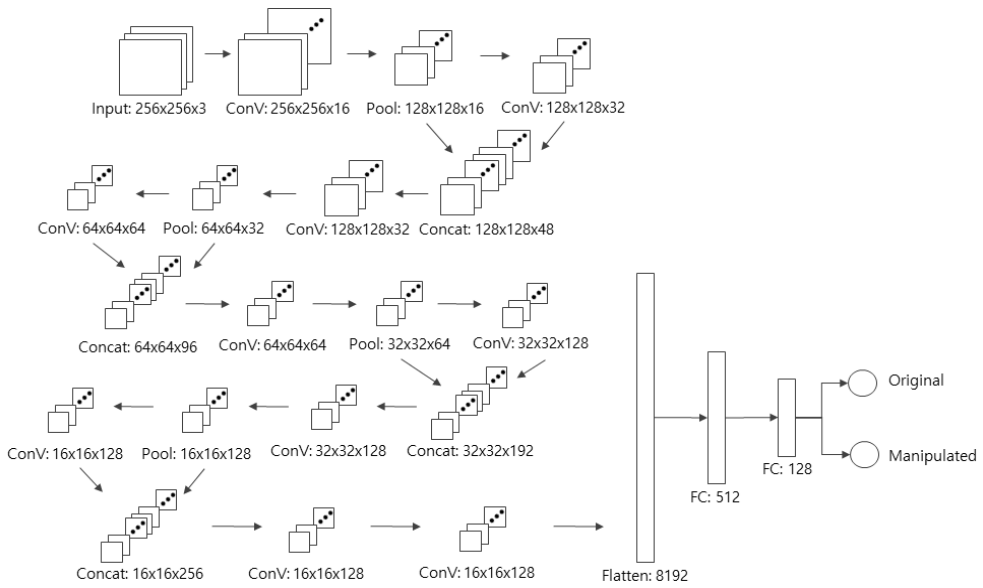

Fig. 6: Sparse CNN Model`s Configuration

As can be seen in Fig. 6, the data input to the model captures the characteristics of the data input through pooling layer and convolution layer. In addition, the size of Feature Map gradually decreases and the number of channels increases. The concatenate function was used to minimize the problem of losing the initial weight value as the layer progresses. In the final layer, it shows the form of 16 x 16 x 128, and images are classified by configuring a flatten matrix in the form of a one-dimensional array through the flatten process in a fully connected layer. In the final process, the model performs binary classification on the input data as original data or manipulation data. In addition, as can be seen from the characteristics of Sparse CNN, this model tried to minimize feature loss that occurs in the process of image convolution. Detection of the manipulated image is determined from the value of the pixel itself and the difference value between the pixels. Therefore, in the convolution process of extracting the pattern of the image, it is important to maintain the feature so that the value between pixels is not omitted. This is also the reason why the dropout used to organize weights in the last layer of this model was not used.

The manipulation detection model is configured to replace different manipulation detection capabilities with artificial intelligence models according to the doctor's experience. This is designed to prevent damage caused by manipulation of medical data that may occur based on fundus images, and is an artificial intelligence model that can be applied to clinical trials, medical diagnosis, and medical insurance.

3. Results and Discussion

The manipulation detection test of this study was based on about 1,000 data, including about 350 original data, about 200 manipulated data, about 400 Cycle GAN original data and manipulation data, through a data screening process based on about 6,000 data. The training and validation data set for learning the manipulation detection model used 900 data, which is 90% of the total data. The test set used a total of 100 data as 50 randomly selected data and 50 data for group verification.

This study shows a single critical point as shown in Fig. 7 as a binary classification that judges the original and manipulation data of each lesion. In addition, the horizontal axis of Fig. 7 is the case where the original data is determined as manipulation data, and the vertical axis is the case where the manipulation data is determined as manipulation data.

In addition, the test in Fig. 7 included the data used for group-specific verification in Fig. 5, and in this paper, the manipulation detection capability of

ophthalmologist and general doctor group and the performance of the manipulation detection model were compared. According to the doctor group's evaluation of Fig. 5, in the case of normal, the group of ophthalmologists showed 0.715 AUC.

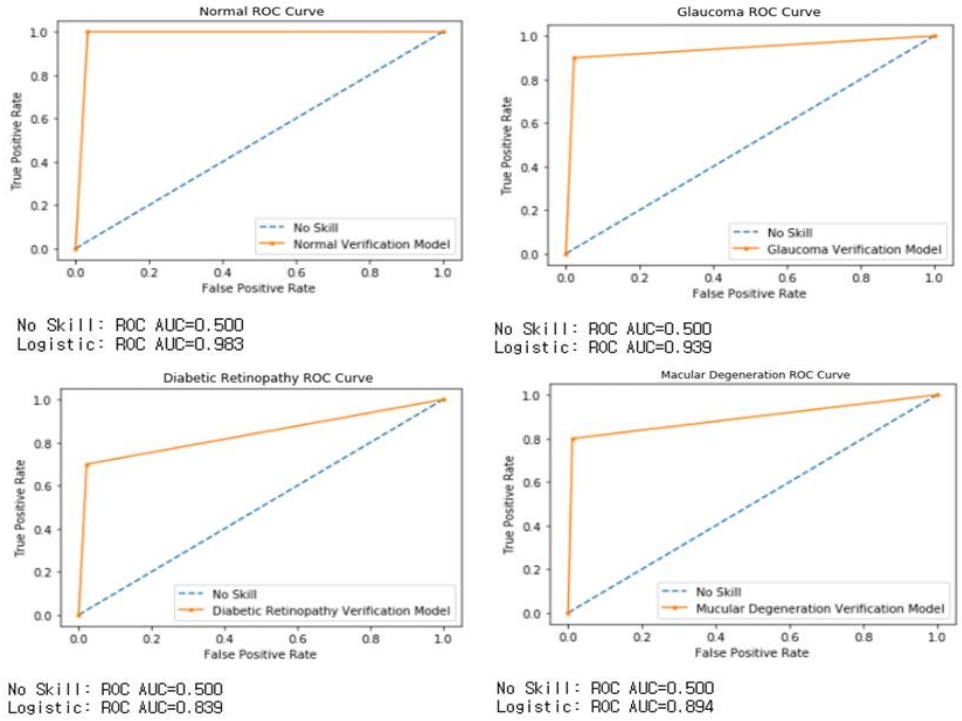

Fig. 7: Result of Model`s evaluation

This is about 25% lower detection rate than the performance of the manipulation detection model. In addition, in the case of glaucoma, the manipulation detection model showed excellent performance by about 18% with AUC 0.752 of the ophthalmologist group and AUC 0.939 of the manipulation detection model. In the case of diabetic retinopathy, AUC 0.826 in the ophthalmologist group and AUC 0.839 in the manipulation detection model showed no significant difference as in other lesions. Finally, in the case of macular degeneration, where the performance of the general doctor group was evaluated higher, the performance of the manipulation detection model was approximately 7% better with AUC 0.822 of the general doctor group and AUC 0.894 of the manipulation detection model.

As can be seen in Table 3, the detection result of the original data showed that the normal was precision 1.00, recall was 0.97 and F1-Score was 0.98. In the case of glaucoma, precision was 0.99 and recall was 0.98 and F1-Score was 0.98. In the

case of diabetic retinopathy, precision 0.97 was recall 0.98 and F1-Score 0.97. In addition, in the case of macular degeneration, precision 0.98 was recall 0.99 and F1-Score 0.98. The detection ability of the model for the original data showed a high score of 0.97 or more for each indicator, and the performance of the detection model for the original data was excellent.

Table 3: Test Results by Detection Model

		Precision	Recall	F1-Score
Origin	Normal	1.00	0.97	0.98
	Glaucoma	0.99	0.98	0.98
	Diabetic Retinopathy	0.97	0.98	0.97
	Macular Degeneration	0.98	0.99	0.98
Manipulated	Normal	0.97	0.97	0.97
	Glaucoma	0.82	0.90	0.86
	Diabetic Retinopathy	0.78	0.70	0.74
	Macular Degeneration	0.89	0.80	0.84

As a result of detection of the manipulation data, normal was 0.97 and recall w as 0.97 and F1-Score was 0.97. In the case of glaucoma, precision was 0.82 and rec all was 0.90 and F1-Score was 0.86. In the case of diabetic retinopathy, precision 0. 78, recall 0.70 and F1-Score 0.74. In the case of macular degeneration, precision 0.8 9, recall 0.80 and F1-Score 0.84. The detection ability of the manipulation detection model for manipulation data showed a score of 0.7 or higher for each indicator, and a high score on average of 0.97 for normal data. In addition, the ROC Curves of eac h lesion are shown in Fig. 7, and AUC showed normal: 0.98, glaucoma: 0.93, diabet ic retinopathy: 0.83, macular degeneration: 0.89. Each lesion showed a high score o f 0.8 or higher and an average AUC of 0.91 or higher. As a result of the experiment in this study, the average detection ability was 91%. On the result index, the score w as 0.98 on average for original data and 0.85 on average for manipulation data. Ther efore, the manipulation detection model of this study is not affected by the data man ipulation method of Cycle GAN and U-Net. In addition, it is judged that the manipu lation detection performance is excellent.

4. Conclusion

In this study, in order to solve problems that may arise from image regeneration in

the medical field, manipulation data was regenerated using fundus image data and verified through medical personnel, and a manipulation data detection model was studied based on this. In addition, data selection and preprocessing processes were performed to improve the performance of the manipulation model. After that, the fundus image data was manipulated using a U-Net-based manipulation model. The image data quality evaluation of the manipulation data differed from the original data. However, the detection results of the two doctor groups showed a detection rate of 77% in the ophthalmologist group and 71% in the general doctor group. This seems to influence the ability to detect images according to the doctor's experience.

The manipulation detection model used an artificial intelligence model based on Sparse CNN. An artificial intelligence model was constructed to detect manipulated data in various ways using manipulated data of the U-Net model and manipulated data of the Cycle GAN model. The experimental results of this study show an average detection ability of 91%. This was 14% higher than the detection results of ophthalmologists who performed detection only with U-Net data and 20% higher than the detection results of general doctors. In addition, as a result of the detection, the score was 0.98 on average for the original data and 0.85 on average. Therefore, the manipulation detection model of this study showed about 14% higher detection performance based on the group of ophthalmologists regardless of various manipulation methods for fundus data. Future research plans to study manipulation models and manipulation detection models using CXR or 3D medical image data. After that, it plans to verify that this model can be used in the actual field.

Acknowledgements

This paper was supported by the Konyang University Research Fund in the second half of 2020.

References

Dhamo, H., Farshad, A., Laina, I., Navab, N., Hager, G.D., Tombari, F., & Rupprecht, C. (2020). Semantic image manipulation using scene graphs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5213-5222.

Shen, D., Wu, G., & Suk, H. I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, 19, 221–248.

Korea Clinical Trials Information Center. (2021). Korea clinical trials: Global Clinical Trial Status. WEB: https://www.koreaclinicaltrials.org/kr/contents/datainfo_data_01_tab03/view.do/. Korea/

Kim, H., Jung, D.C., & Choi, B.W. (2019). Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: adversarial attacks, *Journal of the Korean Society of Radiology*, 80, 259-273.

Larxel. (2020). Ocular Disease Recognition: Right and left eye fundus photographs of 5000 patients. WEB: https://www.kaggle.com/andrewmvd/ocular-disease-recognition-odir5k/

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, 234-241. DOI: https://arxiv.org/abs/1505.04597

Mason, A., Rioux, J., Clarke, S.E., Costa, A., Schmidt, M., Keough, V., Huynh, T., & Beyea, S. (2019). Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images. *IEEE transactions on medical imaging*, 39, 1064-1072.

Sara, U., Akter, M., & Uddin, M.S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications*, 7, 8-18.

Obukhov, A., & Krasnyanskiy, M. (2020). Quality Assessment Method for GAN Based on Modified Metrics Inception Score and Fréchet Inception Distance. *Proceedings of the Computational Methods in Systems and Software*, 102-114.

Wang, S.Y., Wang, O., Zhang, R., Owens, Andrew., & Efros, A.A. (2020). CNN-generated images are surprisingly easy to spot... for now. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8695-8704.

Huang, J., & Ling, C.X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17, 299-310.

Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Penksy (2015). Sparse convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 806-814. DOI: https://doi.org/10.1109/CVPR.2015.7298681

Webster, D. (2018). MESSIDOR-2 DR Grades: Adjudicated DR Severity, DME, and Gradability for the MESSIDOR-2 fundus dataset. WEB: https://www.kaggle.com/google-brain/messidor2-dr-grades/

Zhang, E. (2021). Glaucoma Detection. OCT Scans, Retinal Imaging. WEB: https://www.kaggle.com/sshikamaru/glaucoma-detection/