Homology-Based Image Processing for Automatic Classification of Histopathological Images of Lung Tissue

Nishio, Mizuho

Nishio, Mari

Jimbo, Naoe

Nakane, Kazuaki

(Citation)
Cancers,13(6):1192

(Issue Date)
2021-03

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/90008152

cancers

MDPI

Article

Homology-Based Image Processing for Automatic Classification of Histopathological Images of Lung Tissue

Mizuho Nishio [1,*], Mari Nishio [2], Naoe Jimbo [3] and Kazuaki Nakane [4]

1 Department of Radiology, Kobe University Graduate School of Medicine, 7-5-2 Kusunoki-cho, Chuo-ku, Kobe 650-0017, Japan
2 Division of Pathology, Department of Pathology, Kobe University Graduate School of Medicine, 7-5-1 Kusunoki-cho, Chuo-ku, Kobe 650-0017, Japan; marin@med.kobe-u.ac.jp
3 Department of Diagnostic Pathology, Kobe University Graduate School of Medicine, 7-5-2 Kusunoki-cho, Chuo-ku, Kobe 650-0017, Japan; naoe1123@med.kobe-u.ac.jp
4 Department of Molecular Pathology, Osaka University Graduate School of Medicine and Health Science, Osaka 565-0871, Japan; k-nakane@sahs.med.osaka-u.ac.jp
* Correspondence: nmizuho@med.kobe-u.ac.jp; Tel.: +81-78-382-6104; Fax: +81-78-382-6129

Simple Summary: The purpose of this study was to develop a computer-aided diagnosis (CAD) system for automatic classification of histopathological images of lung tissues. Homology-based image processing (HI) was proposed for CAD. For developing and validating CAD with HI, two datasets of histopathological images of lung tissues were used. The private dataset consists of 94 histopathological images that were obtained for the following five categories: normal, emphysema, atypical adenomatous hyperplasia, lepidic pattern of adenocarcinoma, and invasive adenocarcinoma. The public dataset consists of 15,000 histopathological images that were obtained for the following three categories: lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue. For the two datasets, our results show that HI was more useful than conventional texture analysis for the CAD system.

Abstract: The purpose of this study was to develop a computer-aided diagnosis (CAD) system for automatic classification of histopathological images of lung tissues. Two datasets (private and public datasets) were obtained and used for developing and validating CAD. The private dataset consists of 94 histopathological images that were obtained for the following five categories: normal, emphysema, atypical adenomatous hyperplasia, lepidic pattern of adenocarcinoma, and invasive adenocarcinoma. The public dataset consists of 15,000 histopathological images that were obtained for the following three categories: lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue. These images were automatically classified using machine learning and two types of image feature extraction: conventional texture analysis (TA) and homology-based image processing (HI). Multiscale analysis was used in the image feature extraction, after which automatic classification was performed using the image features and eight machine learning algorithms. The multicategory accuracy of our CAD system was evaluated in the two datasets. In both the public and private datasets, the CAD system with HI was better than that with TA. It was possible to build an accurate CAD system for lung tissues. HI was more useful for the CAD systems than TA.

Keywords: pathology image; lung cancer; homology; Betti number; texture analysis; machine learning

Citation: Nishio, M.; Nishio, M.; Jimbo, N.; Nakane, K. Homology-Based Image Processing for Automatic Classification of Histopathological Images of Lung Tissue. *Cancers* **2021**, *13*, 1192. https://doi.org/10.3390/cancers13061192

Academic Editors: Ognjen Arandjelović and Stefan Delorme

Received: 14 January 2021
Accepted: 8 March 2021
Published: 10 March 2021

1. Introduction

In 2020, 228,820 new lung cancer cases are projected to occur in the United States [1]; lung cancer is the leading cause of cancer-related deaths in the United States, with almost one-quarter of all cancer deaths being caused by lung cancer. An estimated 606,520 Americans will die from cancer in 2020, with 72,500 male and 63,220 female Americans dying from lung cancer [1].

Currently, histopathological and molecular subtypes are important in lung cancer diagnoses to determine a treatment strategy, and accurate histopathological diagnoses allow clinicians to select targeted treatment options that are specific to each patient. For example, erlotinib (Tarceva; Genentech, South San Francisco, CA, USA) is a tyrosine kinase inhibitor effective in lung cancer patients with mutated epidermal growth factor receptor [2]. Clinicians determine the use of tyrosine kinase inhibitor based on histopathological diagnoses of the mutated epidermal growth factor receptor. Generally, immunohistochemistry is used for the diagnosis of the mutated epidermal growth factor receptor.

Digital pathology systems [3,4] have improved over the years and are now capable of producing high-resolution histopathological images. Using digital pathology systems, histopathological assessments can be performed using a computer display rather than a light microscope. In addition, this system has enabled computer-aided diagnosis (CAD) for histopathological diagnosis. Currently, CAD is being used for detection and diagnosis in several medical fields [5–7], and CAD has the potential to improve the speed and accuracy of histopathological diagnoses of lung cancer [3].

CAD frequently utilizes machine learning to improve its diagnostic accuracy. In order to use medical images in CAD, image feature extraction is required for machine learning. For evaluation of tumor aggressiveness, tumor heterogeneity is an important factor [8,9]. In CAD of cancers, texture analysis is frequently used for image feature extraction to assess tumor heterogeneity [8,9].

In recent years, homology-based image processing has been increasingly used [10–17]. For example, Nishio et al. showed that homology-based image processing was useful for estimating the risk of lung cancer [15], and Nakane et al. showed that colon cancer could be accurately segmented on histopathological images using homology-based methods [14]. In homology-based methods, Betti numbers are import metrics for image feature extraction. These numbers are calculated from binarized images obtained from medical images (please refer to Figure 2 of [13] and Figure S1 of [17] for the calculation of Betti numbers). In the current study, it was assumed that Betti numbers obtained with homology-based image processing were useful for evaluation of tumor heterogeneity in image feature extraction.

The purpose of this study was to develop a CAD system for the automatic classification of histopathological images. To the best of our knowledge, image feature extraction of histopathological images of lung tissue has not been performed using homology-based image processing, and the performance of CAD has not been appraised when homology-based image processing has been used. For the purpose of this study, private and public datasets were used. In the proposed method, the histopathological images were automatically classified using image features extracted based on the homology method and several machine learning algorithms. For comparison with the proposed method, conventional texture analysis was used for image feature extraction.

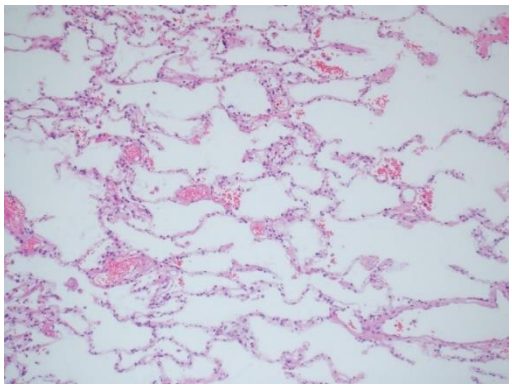
2. Materials and Methods

This retrospective study was approved by the institutional review board of our institution (permission number: B200033); the requirement for acquiring informed consent was waived.

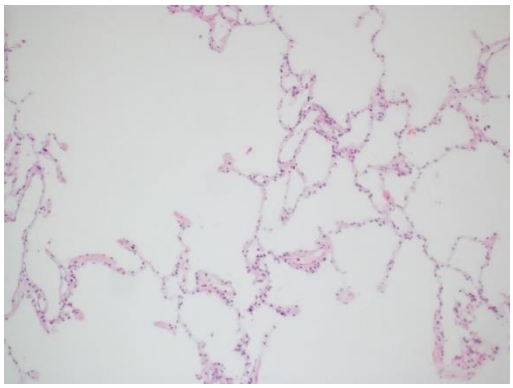
2.1. Private Dataset

In the private dataset, ninety-four histopathological images of lung tissue were obtained from lung surgery specimens. They belonged to five categories of lung tissue (normal, emphysema, atypical adenomatous hyperplasia (AAH), lepidic pattern of adenocarcinoma (LP), and invasive adenocarcinoma (AC)), consisting of 20 normal, 20 emphysema, 23 AAH, 19 LP, and 12 AC images. The histopathological diagnosis of the 94 images was confirmed by two board-certified pathologists (M.N. and N.J.). These histopathological images were obtained by means of hematoxylin and eosin staining. The image resolution of the 94 images was 1600 × 1200 pixels with RGB channels at 100× magnification

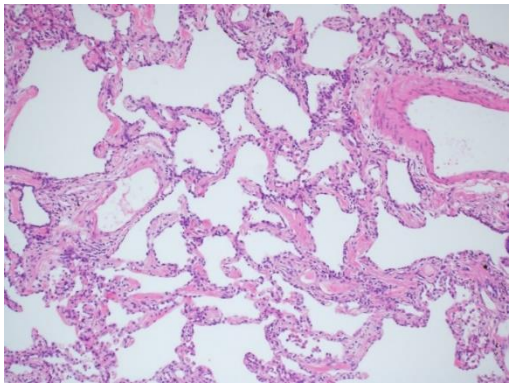
(the magnification of the objective lens being 10×). Figure 1A–E show representative histopathological images of the five categories, respectively.



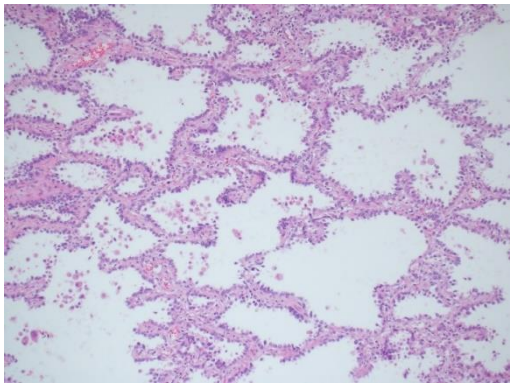
(**A**) normal



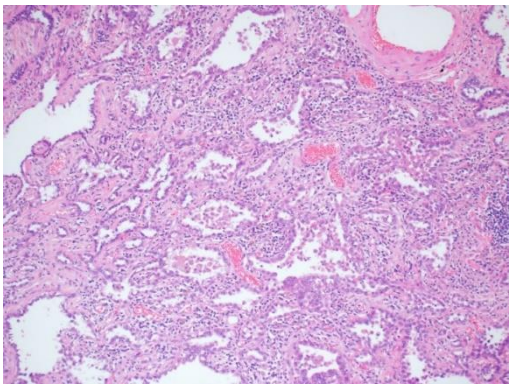
(**B**) emphysema



(**C**) AAH



(**D**) LP



(**E**) AC

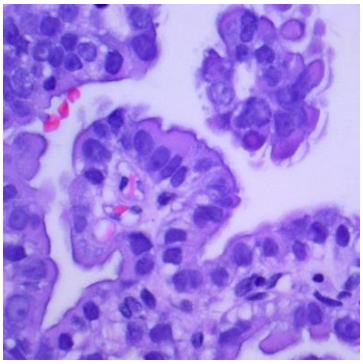
Figure 1. Representative histopathological images of (**A**) normal, (**B**) emphysema, (**C**) AAH, (**D**) LP and (**E**) AC. The magnification is 100× (the magnification of the objective lens being 10×). AAH, atypical adenomatous hyperplasia; LP, lepidic pattern of adenocarcinoma; AC, invasive adenocarcinoma.

For developing and evaluating the CAD system, the 94 histopathological images of the private dataset were randomly divided into a training set with 50 images, a validation set with 20 images, and a testing set with 24 images. Because the number of images in the private dataset was small, image patches were extracted from the images for each of the three sets. Ten image patches with image resolution 1024 × 1024 pixels were randomly extracted from one histopathological image. In addition, vertical and horizontal flipping were randomly applied to the image patches as in data augmentation of deep learning [6].

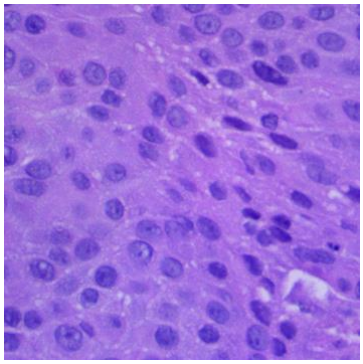
Finally, a training set with 500 image patches, a validation set with 200 image patches, and a testing set with 240 image patches were used for the CAD system.

2.2. Public Dataset

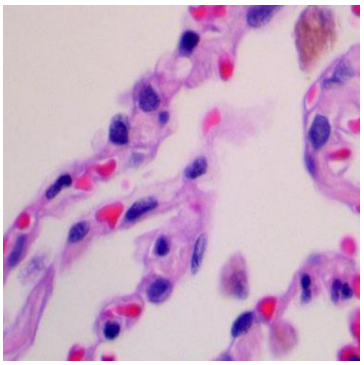
The public dataset (LC25000) contains 25,000 color images with five classes of 5000 images each [18]. All images are 768 × 768 pixels in size. From LC25000, 15,000 histopathological images of three classes of lung tissue (lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue) were selected. Figure 2A–C show representative histopathological images of the three categories, respectively.



(**A**) Lung adenocarcinoma



(**B**) Lung squamous cell carcinoma



(**C**) Benign lung tissue

Figure 2. Representative histopathological images of (**A**) lung adenocarcinoma, (**B**) lung squamous cell carcinoma and (**C**) benign lung tissue.

As in the private dataset, the 15,000 histopathological images of LC25000 were divided into a training set with 9000 images, a validation set with 3000 images, and a test set with 3000 images. The image patch extraction was not used for the public dataset.

2.3. Outline of CAD System

Figure 3 shows an outline of the CAD system for the private dataset. Except for the output, the same processing was performed for the public dataset. The RGB images were fed into the CAD system and then the image features were extracted. A machine learning algorithm classified the image based on the extracted features. To train the machine learning algorithm of the CAD system and optimize parameters of the CAD system, image features of the training and validation sets were used, respectively. Finally, image features of the testing set were used for assessing the performance of the CAD system. The programming language used for the development of the CAD system was Python (version 3.7, http://www.python.org/ (accessed on 13 November 2020)).

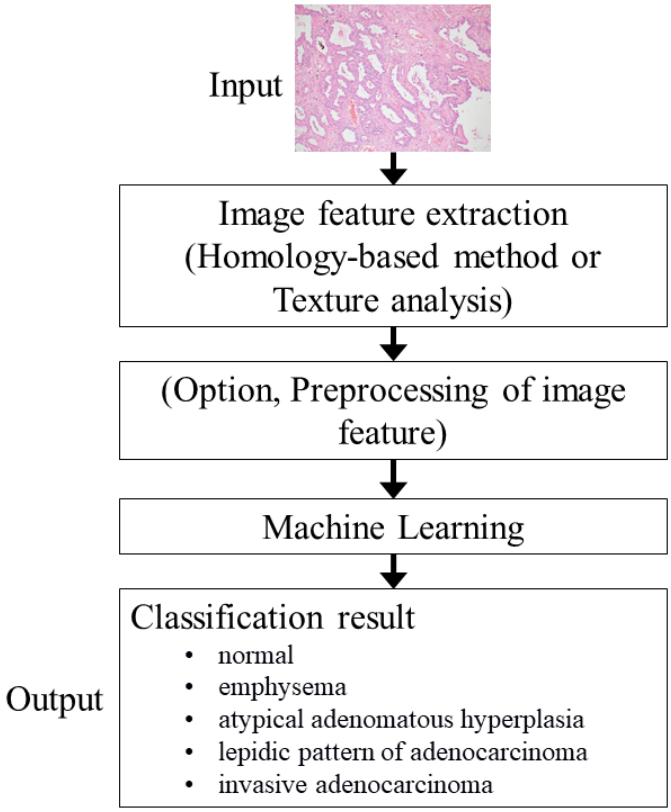


Figure 3. Outline of the CAD system in private dataset. Note: Except the output, the same processing was performed for the public dataset. CAD, computer-aided diagnosis.

2.4. Image Feature Extraction

To perform homology-based image processing, Betti numbers (b0 and b1) were calculated for the histopathological images with RGB channels. Figure 4 shows an outline of the Betti number calculation process for histopathological images. To calculate the Betti numbers, a grayscale image converted from the RGB images was prepared. Because Betti numbers are calculated using a binarized image in which each pixel can have two values (0 and 1), the grayscale image was binarized before calculating the Betti numbers. For binarization, thresholding was performed using predefined pixel values. The binarized images obtained via thresholding were processed using our in-house homology software to calculate the Betti numbers. The process of calculating the Betti numbers from binarized images has been described elsewhere [13–17]. Briefly, in a two-dimensional binarized image, b0 (the zero-dimensional Betti number) is the number of connected components in the image, and b1 (the one-dimensional Betti number) is the number of one-dimensional or "circular" holes in the image. For the predefined pixel value of thresholding, 0, 5, 10, . . . , 245, 250, and 255 were used. For multiscale analysis, the image resolution was changed in calculating the Betti numbers. In the private dataset, the image resolutions of 1024 × 1024, 512 × 512, 256 × 256, and 128 × 128 pixels were used for multiscale homology-based image processing. In the public dataset, the image resolutions of 768 × 768, 384 × 384, 192 × 192, and 96 × 96 pixels were used. Image features of the Betti numbers at different image resolutions were concatenated, based on our multiscale homology-based image processing. A schematic illustration of the multiscale homology-based image processing is shown in Figure 5.

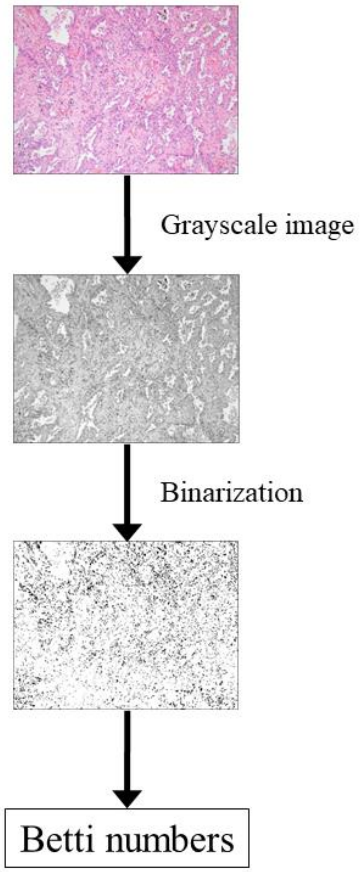


Figure 4. Outline of the process of calculating Betti numbers.

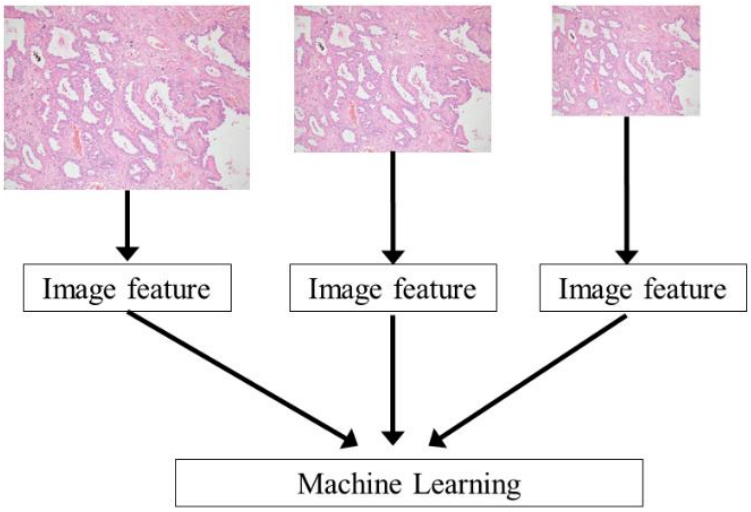


Figure 5. Schematic illustration of the multiscale homology-based image processing.

For the conventional method, image feature extraction was performed using texture analysis via PyRadiomics (version 3.0, https://pyradiomics.readthedocs.io/en/latest/ (accessed on 14 November 2020)) [19]. Texture analysis was performed on a grayscale image converted from the original RGB image. The target of texture analysis was the entire image. The image feature names of the texture analysis are listed in the Supplementary Materials (Table S1). Briefly, 18, 23, 16, 16, 14, and 5 image features were calculated for First Order, Gray Level Co-occurrence Matrix, Gray Level Run Length Matrix, Gray Level

Size Zone Matrix, Gray Level Dependence Matrix, and Neighboring Gray Tone Difference Matrix, respectively. The multiscale analysis was also performed for texture analysis.

2.5. Preprocessing of Image Features and Machine Learning

In the current study, scikit-learn (version 0.23.2, https://scikit-learn.org/stable/ (accessed on 14 November 2020)) was used for both preprocessing of the image features and the machine learning algorithms [20]. After the image feature extraction, preprocessing of the image features was performed. For the preprocessing, the standardization of image features and feature selection were utilized. The mean and standard deviation of each feature was used for the standardization of image features using the sklearn.preprocessing.StandardScaler class. After the standardization, the feature selection was performed using the sklearn.feature_selection.SelectKBest class and sklearn.feature_selection.f_classif function, where the number of selected features was set to 20% of the original image features. Both the feature standardization and the feature selection were optional. The image features with or without the preprocessing were fed into the machine learning algorithms. The machine learning algorithms included (0) perceptron, (1) logistic regression, (2) kNN, (3) support vector machine with linear kernel, (4) support vector machine with radial basis function kernel, (5) decision tree, (6) random forest, and (7) gradient tree boosting. For gradient tree boosting, xgboost (version 1.2.0, https://xgboost.readthedocs.io/en/latest/ (accessed on 14 November 2020)) was used [21]. These machine learning algorithms were trained with the image features of training set and their default hyperparameters provided by the implementation of scikit-learn and xgboost.

2.6. Performance Evaluation

Performance evaluation was performed using the multicategory classification accuracy obtained in the testing set. For determining the optimal CAD, the validation accuracy was calculated for all possible combinations of normalization, feature selection, image resolution, and machine learning algorithms. For both homology-based image processing and texture analysis, single-scale and multiscale analyses were performed. All combinations of image resolutions (1024 × 1024, 512 × 512, 256 × 256, and 128 × 128 pixels for the private dataset, and 768 × 768, 384 × 384, 192 × 192, and 96 × 96 pixels for the public dataset) were used for multiscale analysis.

3. Results

Tables 1–4 and Tables S2–S5 show prediction results of the CAD systems for the private and public datasets. In each entry of Tables S2–S5, the most accurate result and its corresponding algorithm were selected among the eight machine learning algorithms. In the optimal machine learning algorithm of Tables 1–4 and Tables S2–S5, 0–7 represents perceptron, logistic regression, kNN, support vector machine with linear kernel, support vector machine with radial basis function kernel, decision tree, random forest, and gradient tree boosting, respectively. 0 and 1 in the normalization and feature selection of these tables represent "without preprocessing" and "with preprocessing", respectively.

Table 1. Validation and testing five-category accuracies of the optimal CAD with homology-based image processing in the private dataset. Note: The optimal CAD was selected based on the validation accuracies of Table S2.

Normalization	Feature Selection	Image Resolutions (Pixels)	Validation Accuracy	Testing Accuracy	Optimal Machine Learning Algorithm
0	1	256 × 256	0.9000	0.7833	6

Table 2. Validation and testing five-category accuracies of the optimal CAD with texture analysis in the private dataset. Note: The optimal CAD was selected based on the validation accuracies of Table S3.

Normalization	Feature Selection	Image Resolutions (Pixels)	Validation Accuracy	Testing Accuracy	Optimal Machine Learning Algorithm
0	1	1024 × 1024	0.8650	0.7083	1

Table 3. Validation and testing three-category accuracies of the optimal CAD with texture analysis in the public dataset. Note: The optimal CAD was selected based on the validation accuracies of Table S4.

Normalization	Feature Selection	Image Resolutions (Pixels)				Validation Accuracy	Testing Accuracy	Optimal Machine Learning Algorithm
1	0	1024 × 1024	512 × 512	256 × 256		0.9927	0.9940	2
0	0	512 × 512	256 × 256	128 × 128		0.9927	0.9923	7
0	0	1024 × 1024	512 × 512			0.9927	0.9920	7
0	0	1024 × 1024	512 × 512	256 × 256	128 × 128	0.9927	0.9943	7
1	0	512 × 512	256 × 256	128 × 128		0.9927	0.9923	7
1	0	1024 × 1024	512 × 512			0.9927	0.9920	7
1	0	1024 × 1024	512 × 512	256 × 256	128 × 128	0.9927	0.9943	7

Table 4. Validation and testing three-category accuracies of the optimal CAD with texture analysis in the public dataset. Note: The optimal CAD was selected based on the validation accuracies of Table S5.

Normalization	Feature Selection	Image Resolutions (Pixels)				Validation Accuracy	Testing Accuracy	Optimal Machine Learning Algorithm
0	0	1024 × 1024	512 × 512	256× 256	128 × 128	0.9923	0.9933	7
1	0	1024 × 1024	512 × 512	256× 256	128 × 128	0.9923	0.9933	7

Tables S2 and S3 show validation accuracies of the CAD systems with homology-based image processing and texture analysis for all possible combinations in the private dataset, respectively. Tables 1 and 2 show the validation and testing accuracies of the optimal CAD systems with homology-based image processing and texture analysis selected from Tables S2 and S3, respectively. According to Tables 1 and 2, the testing accuracy of the optimal CAD with the homology-based image processing (78.33%) was better than that with the texture analysis (70.83%). Random forest and logistic regression were used in Tables 1 and 2, respectively. Because single-scale analysis was used in the entries of Tables 1 and 2, the usefulness of multiscale analysis was limited for the private dataset.

Tables S4 and S5 show validation accuracies of the CAD systems with homology-based image processing and texture analysis for all possible combinations in the public dataset, respectively. Tables 3 and 4 show validation and testing accuracies of the optimal CAD systems with homology-based image processing and texture analysis selected from Tables S4 and S5, respectively. According to Tables 3 and 4, the best testing accuracy of the optimal CAD with the homology-based image processing (99.43%) was better than that with the texture analysis (99.33%). Gradient tree boosting was frequently used in Tables 3 and 4. Because no entry of single-scale analysis was found in Tables 3 and 4, multiscale analysis was useful in the public dataset.

Figures 6 and 7 show the confusion matrices between the ground truth and prediction, which were obtained with the optimal CAD systems for the private and public datasets, respectively.

		Prediction				
		emphysema	normal	AAH	LP	AC
Ground Truth	emphysema	40	0	0	0	0
	normal	10	50	0	0	0
	AAH	0	0	43	7	0
	LP	0	0	35	25	0
	AC	0	0	0	0	30

Figure 6. Confusion matrix between the ground truth and prediction obtained with the optimal CAD system for the private dataset.

		Prediction		
		Benign lung tissue	Lung adenocarcinoma	Lung squamous cell carcinoma
Ground Truth	Benign lung tissue	1016	2	0
	Lung adenocarcinoma	4	993	7
	Lung squamous cell carcinoma	0	4	974

Figure 7. Confusion matrix between the ground truth and prediction obtained with the optimal CAD system in the public dataset.

4. Discussion

The results of this study indicate that it is possible to construct an accurate CAD system by using homology-based image processing for the multicategory classification of lung tissue (normal, emphysema, AAH, LP, and AC in the private dataset, and lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue in the public dataset). Our results show that the accuracy of the multicategory classification with homology-based image processing was better than that with texture analysis in the private and public datasets.

Classification of AAH, LP, and AC is important because it affects patient prognosis and survival [22]. For instance, the identification of pure LP has been shown to have excellent prognoses for patients with stage I lung cancer [23]. However, accurate classification of such patterns can be challenging [24]. Because the classification accompanies the

subjective nature of pathologists, interobserver variability of the pathologists' diagnosis can be problematic. Our CAD system might be helpful in solving this problem.

To our knowledge, few studies have used machine learning or deep learning to predict the histological subtype classification of lung tissue [3]. One study performed a six-category classification of histologic patterns in lung adenocarcinoma and benign tissue (lepidic, acinar, papillary, micropapillary, solid, and benign) [25]. Another study performed the five-category classification (solid, micropapillary, acinar, cribriform, and nontumor) in lung adenocarcinoma and nontumor tissue [26]. Compared with these two studies, our novelty is that our CAD system distinguished AAH from the other four categories. In addition, while these two studies used deep learning, our study used machine learning.

To evaluate the efficacy of homology-based image processing in the large dataset, the public dataset obtained from LC25000 was used in this study. The results for the public dataset show that homology-based image processing was more useful than conventional texture analysis in the classification between lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue.

In this study, it was assumed that homology-based image processing was useful for evaluating tumor heterogeneity in the CAD system of lung cancer. Because our results show that CAD with homology-based image processing was more accurate than that with texture analysis, our assumption was validated. One major advantage of homology-based image processing over texture analysis is topological invariance [14]. Because of this property, Betti numbers are not changed by continuous transformation. It is speculated that in the CAD system with homology-based image processing, topological invariance makes image features more robust, compared with texture analysis.

The multiscale analysis improved the accuracy of both homology-based image processing and texture analysis for the public dataset. It is speculated that because the image resolution is essential information for image classification, multiscale analysis was useful for the two methods of image feature extraction. On the other hand, the usefulness of multiscale analysis was not clear for the private dataset. This might be caused by an imbalance between dataset size and number of image features in the multiscale analysis. Further study is needed to establish the usefulness of the multiscale analysis in homology-based image processing.

According to Figure 6, classification between AAH and LP was difficult in our optimal CAD system. One major reason for this result is the size of the private dataset. Generally, machine learning and deep learning yield relatively poor performance for small datasets. Although we used patch-level accuracy for mitigating the effect of the small dataset, we could not avoid deterioration in the classification between AAH and LP. To overcome this problem, a larger dataset should be used.

Our study has several limitations. First, the private dataset was small. For mitigating the effect of the small dataset, patch-level accuracy was evaluated in the private dataset. In addition, a public dataset was also used in this study. Second, no external validation was performed. Overfitting of our CAD system may have occurred in the external validation. For future studies, we need to investigate the effectiveness of our CAD systems using datasets obtained from other affiliations. Third, the subtype classification of adenocarcinoma was not fully investigated. Of adenocarcinoma, minimally invasive adenocarcinoma and adenocarcinoma in situ [22] were not considered for the private dataset. Classification between AAH, minimally invasive adenocarcinoma, adenocarcinoma in situ, and invasive adenocarcinoma should be performed in future development of our CAD system. Fourth, classification of adenocarcinoma between lepidic predominant, acinar predominant, papillary predominant, micropapillary predominant, etc. was not performed. This classification should be investigated in future study. Fifth, we did not compare our CAD system with that with deep learning. Because the private dataset was small, it was speculated that the performance of the CAD system with deep learning might be low in the private dataset. Therefore, we did not use deep learning in this study. Sixth, although several studies investigated CAD systems for the prognosis, survival, and genetic

features of lung cancers [27–29], we did not predict them in the current study. Because the classification between LP and AC is directly related to prognosis and survival of lung cancer [23], we believe that our CAD system is useful for evaluating the prognosis and survival of lung cancer.

5. Conclusions

It was possible to build an accurate CAD system for the automatic classification of lung tissue. Homology-based image processing was more useful for CAD systems than conventional texture analysis.

Supplementary Materials: The following are available online at https://www.mdpi.com/2072-6694/13/6/1192/s1, Figure S1: Binarized image of handwritten character and its Betti numbers, Table S1: Feature names of texture analysis, Tables S2–S5: Prediction results of CAD.

Author Contributions: Conceptualization, K.N. and M.N. (Mizuho Nishio); methodology, M.N. (Mizuho Nishio); software, M.N. (Mizuho Nishio); validation, M.N. (Mari Nishio) and N.J.; formal analysis, M.N. (Mizuho Nishio); investigation, M.N. (Mizuho Nishio); resources, M.N. (Mizuho Nishio), M.N. (Mari Nishio), and N.J.; data curation, M.N. (Mari Nishio) and N.J.; writing—original draft preparation, M.N. (Mizuho Nishio); writing—review and editing, all authors; visualization, M.N. (Mizuho Nishio); supervision, K.N.; project administration, M.N. (Mizuho Nishio); funding acquisition, M.N. (Mizuho Nishio). All authors have read and agreed to the published version of the manuscript.

Funding: The present study was partly supported by JSPS KAKENHI (grant numbers JP19K17232 and JP19H03599).

Institutional Review Board Statement: This retrospective study was approved by the institutional review board of our institution (permission number: B200033).

Informed Consent Statement: Patient informed consent was waived by the review board.

Data Availability Statement: The public dataset is available in a publicly accessible repository. Please see [18]. The private dataset is not available because of regulation of privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [CrossRef] [PubMed]
2. Shepherd, F.A.; Rodrigues Pereira, J.; Ciuleanu, T.; Tan, E.H.; Hirsh, V.; Thongprasert, S.; Campos, D.; Maoleekoonpiroj, S.; Smylie, M.; Martins, R.; et al. Erlotinib in previously treated non-small-cell lung cancer. *N. Engl. J. Med.* **2005**, *353*, 123–132. [CrossRef]
3. Wang, S.; Yang, D.M.; Rong, R.; Zhan, X.; Fujimoto, J.; Liu, H.; Minna, J.; Wistuba, I.I.; Xie, Y.; Xiao, G.; et al. Artificial intelligence in lung cancer pathology image analysis. *Cancers* **2019**, *11*, 1673. [CrossRef] [PubMed]
4. Jara-Lazaro, A.R.; Thamboo, T.P.; Teh, M.; Tan, P.H. Digital pathology: Exploring its applications in diagnostic surgical pathology practice. *Pathology* **2010**, *42*, 512–518. [CrossRef]
5. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
6. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **2018**, *9*, 611–629. [CrossRef]
7. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA J. Am. Med. Assoc.* **2016**, *316*, 2402–2410. [CrossRef] [PubMed]
8. Davnall, F.; Yip, C.S.; Ljungqvist, G.; Selmi, M.; Ng, F.; Sanghera, B.; Ganeshan, B.; Miles, K.A.; Cook, G.J.; Goh, V. Assessment of tumor heterogeneity: An emerging imaging tool for clinical practice? *Insights Imaging* **2012**, *3*, 573–589. [CrossRef]
9. Cook, G.J.; O'Brien, M.E.; Siddique, M.; Chicklore, S.; Loi, H.Y.; Sharma, B.; Punwani, R.; Bassett, P.; Goh, V.; Chua, S. Non-Small Cell Lung Cancer Treated with Erlotinib: Heterogeneity of (18)F-FDG Uptake at PET-Association with Treatment Response and Prognosis. *Radiology* **2015**, *276*, 883–893. [CrossRef] [PubMed]
10. Yan, C.; Nakane, K.; Wang, X.; Fu, Y.; Lu, H.; Fan, X.; Feldman, M.D.; Madabhushi, A.; Xu, J. Automated gleason grading on prostate biopsy slides by statistical representations of homology profile. *Comput. Methods Programs Biomed.* **2020**, *194*, 105528. [CrossRef]
11. Nakane, K.; Tsuchihashi, Y.; Matsuura, N. A simple mathematical model utilizing topological invariants for automatic detection of tumor areas in digital tissue images. *Diagn. Pathol.* **2013**, *8*, S27. [CrossRef]

12. Qaiser, T.; Sirinukunwattana, K.; Nakane, K.; Tsang, Y.-W.; Epstein, D.; Rajpoot, N. Persistent Homology for Fast Tumor Segmentation in Whole Slide Histology Images. *Procedia Comput Sci.* **2016**, *90*, 119–124. [CrossRef]
13. Qaiser, T.; Tsang, Y.-W.; Taniyama, D.; Sakamoto, N.; Nakane, K.; Epstein, D.; Rajpoot, N. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *Med. Image Anal.* **2019**, *55*, 1–14. [CrossRef]
14. Nakane, K.; Takiyama, A.; Mori, S.; Matsuura, N. Homology-based method for detecting regions of interest in colonic digital images. *Diagn. Pathol.* **2015**, *10*, 36. [CrossRef]
15. Nishio, M.; Kubo, T.; Togashi, K. Estimation of lung cancer risk using homology-based emphysema quantification in patients with lung nodules. *PLoS ONE* **2019**, *14*, e0210720. [CrossRef] [PubMed]
16. Nishio, M.; Nakane, K.; Tanaka, Y. Application of the homology method for quantification of low-attenuation lung region in patients with and without COPD. *Int. J. COPD* **2016**, *11*. [CrossRef]
17. Nishio, M.; Nakane, K.; Kubo, T.; Yakami, M.; Emoto, Y.; Nishio, M.; Togashi, K. Automated prediction of emphysema visual score using homology-based quantification of low-attenuation lung region. *PLoS ONE* **2017**, *12*, e0178217. [CrossRef]
18. Borkowski, A.A.; Bui, M.M.; Brannon Thomas, L.; Wilson, C.P.; DeLand, L.A.; Mastorides, S.M. Lung and colon cancer histopathological image dataset (LC25000). *arXiv* **2019**, arXiv:1912.12142.
19. Van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
21. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference Knowl Discov Data Min KDD '16, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
22. Travis, W.D.; Brambilla, E.; Noguchi, M.; Nicholson, A.G.; Geisinger, K.; Yatabe, Y.; Beer, D.G.; Powell, C.A.; Riely, G.J.; Van Schil, P.E.; et al. International Association for the Study of Lung Cancer/ American Thoracic Society/European Respiratory Society: International multidisciplinary classification of lung adenocarcinoma-An executive summary. *Proc. Am. Thorac. Soc.* **2011**, *8*, 381–385. [CrossRef] [PubMed]
23. Kadota, K.; Villena-Vargas, J.; Yoshizawa, A.; Motoi, N.; Sima, C.S.; Riely, G.J.; Rusch, V.W.; Adusumilli, P.S.; Travis, W.D. Prognostic significance of adenocarcinoma in situ, minimally invasive adenocarcinoma, and nonmucinous lepidic predominant invasive adenocarcinoma of the lung in patients with stage I disease. *Am. J. Surg. Pathol.* **2014**, *38*, 448–460. [CrossRef]
24. Warth, A.; Stenzinger, A.; Von Brünneck, A.C.; Goeppert, B.; Cortis, J.; Petersen, I.; Hoffmann, H.; Schnabel, P.A.; Weichert, W. Interobserver variability in the application of the novel IASLC/ATS/ERS classification for pulmonary adenocarcinomas. *Eur. Respir. J.* **2012**, *40*, 1221–1227. [CrossRef] [PubMed]
25. Wei, J.W.; Tafe, L.J.; Linnik, Y.A.; Vaickus, L.J.; Tomita, N.; Hassanpour, S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci. Rep.* **2019**, *9*, 1–8. [CrossRef]
26. Gertych, A.; Swiderska-Chadaj, Z.; Ma, Z.; Ing, N.; Markiewicz, T.; Cierniak, S. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci. Rep.* **2019**, *9*, 1–12. [CrossRef] [PubMed]
27. Koyasu, S.; Nishio, M.; Isoda, H.; Nakamoto, Y.; Togashi, K. Usefulness of gradient tree boosting for predicting histological subtype and EGFR mutation status of non-small cell lung cancer on 18F FDG-PET/CT. *Ann. Nucl. Med.* **2020**, *34*, 49–57. [CrossRef]
28. Pinheiro, G.; Pereira, T.; Dias, C.; Freitas, C.; Hespanhol, V.; Costa, J.L.; Cunha, A.; Oliveira, H.P. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Sci. Rep.* **2020**, *10*, 3625. [CrossRef]
29. Li, J.; Wang, H.; Li, Z.; Zhang, C.; Zhang, C.; Li, C.; Yu, H.; Wang, H. A 5-Gene signature is closely related to tumor immune microenvironment and predicts the prognosis of patients with non-small cell lung cancer. *Biomed. Res. Int.* **2020**, *2020*, 2147397. [CrossRef] [PubMed]