Classification of Cervical Cancer Using Deep Learning and Machine Learning Approach

Yerang Park
Gachon University
Young Jae Kim
Gachon University
Woong Ju
Ewha Womans University Medical Center
Kyehyun Nam
Soonchunhyang University Hospital
Soonyung Kim
NTL Medical Institute
Kwang Gi Kim (kimkg@gachon.ac.kr)
Gachon University Gil Hospital

Research Article

Keywords: Human and material resources, cervical cancer, Resnet-50, XGB, SVM

Posted Date: March 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-254234/v1

Version of Record: A version of this preprint was published at Scientific Reports on August 9th, 2021. See the published version at https://doi.org/10.1038/s41598-021-95748-3.

Classification of cervical cancer using deep learning and machine learning approach

Ye Rang Park[1], Young Jae Kim[2], Woong Ju[3], Kyehyun Nam[4], Soonyung Kim[5], Kwang Gi Kim[2]

[1]Dept. of Health Sciences and Technology, Gachon Advanced Institute for Health Sciences and Technology (GAIHST), Gachon University

[2]Department of Biomedical Engineering, Gachon University College of Medicine, Incheon, Republic of Korea

[3]Department of Obstetrics & Gynecology Ewha Womans University Seoul Hospital

[4]Department of Obstetrics & Gynecology, Soonchunhyang University, Bucheon Hospital

[5]R&D Center, NTL Medical Institute

ABSTRACT

Human and material resources are scarce in countries such as developing countries with a high rate of cervical cancer. In such an environment, the introduction of automatic diagnostic technology that can replace specialists is urgent. Finding best method of the known methods can accelerate the adoption of computer-aided diagnostic tools for cervical cancer. In this paper, we

would like to investigate which method, machine learning or deep learning, has higher classification performance in diagnosing cervical cancer.

Using 4,119 sheets, cervical cancer was classified to positive or negative class using Resnet-50 for deep learning, XGB, SVM and RF for mechine learning. In both experiments, square images which of vaginal wall regions are cut were used. In the machine learning, 10 major features were extracted from a total of 300 features.

All tests were validated by 5-fold cross-validation, and receiver operating characteristics(ROC) analysis yielded the following AUC: Resnet-50 0.97(CI 95% 0.949-0.976), XGB 0.82(CI 95% 0.797-0.851), SVM 0.84(CI 95% 0.801-0.854), RF 0.79(CI 95% 0.804-0.856). Deep learning was 0.15 point higher ($p < 0.05$) than the average (0.82) of three machine learning methods.

We propose an better algorithm among the previously known or newly proposed algorithms for diagnosis of cervical cancer using cervicography images.

Introduction

Cervical cancer is the second most common cancer in women worldwide, with a death rate of 60%. In particular, about 85% of the deaths are women in developing countries every year [1-3]. Cervical cancer of the uterus is characterized by a long period of pre-invasive stages. It is sufficiently preventable as screening tests enable effective treatment of precancer stage lesions [4,5]. Nevertheless, it is analyzed that the mortality rate in developing countries is particularly

high because they are not receiving the benefits of preventive policies such as free vaccination programs and national examination programs provided by the state. [6,7].

There is a typical method of determining cervical cancer [8]. Cervicography is a test in which morphological abnormalities of the cervix are read by external professional readers after applying 5% acetic acid to the cervix and then magnifying the cervix up to 50 times the maximum with camera [9]. However, this method has a limitation in that it needs sufficient human and material resources given that accurate reading of cervical dilatation tests requires a professional reader licensed for international reading[10]. In addition, it is essential to increase objectivity through systematic and regular reader quality control as there may exist inter-intra observer errors. In addition, results may vary depending on the subjective view of the reader in cervical reading and researcher's health condition [11,12].

To compensate for these shortcomings, computer-aided diagnostic tools such as classic machine learning(ML) and deep learning(DL) recently been used to recognize patterns using computers are useful for medical diagnosis [13,14]. ML is originally a high-level concept of DL, and it refers to a series of processes that analyze and learn data and make decisions based on learning information [15]. Artificial neural networks modeled after neurons as human brain structures are included in ML. Such artificial neural network has limitations in that it cannot learn or handle new data due to the problem of gradient converging to zero and falling to the local maximum value. And DL has now been pushed beyond limits of the existing artificial neural network using pre-learning and dropout methods called DL [16]. In this paper, ML and DL concepts were used separately.

In the 2000s, ML-based cervical lesion screening techniques began to be actively studied [17]. In 2009, an artificial intelligence (AI) research team in Mexico conducted a study to classify negative and positive cervicography images using k-nearest neighbor algorithm(K-NN). With images of about 50 people, k-NN classified negative and positive images with a sensitivity of 71% and a specificity of 59% [18].

In another case of ML classification, in 2020 Indonesia, image processing was applied to cervicography images and classification of negative and positive images was conducted using support vector machine (SVM), resulting in an accuracy of 90% [19].

Since 2016, when the Fourth Industrial Revolution drew attention, numerous studies have focused on image classification by DL. In the field of cervical research, many research teams around the world are paying attention to detection and classification using DL[20]. In 2019, Utah State University in the United States used a faster region convolution neural network(F-RCNN) to automatically detect the cervical region in cervicography images and classify dysplasia and cancer, with an AUC of 0.91 [21]. In 2017, Japan conducted an experiment in which 500 images of cervical cancer were classified into three grades [severe dysplasia, carcinoma in situ (CIS), and invasive factor (IC)] using the research team's self-developed neural network, showing an accuracy of about 50% as an early stage experiment [22]. ML and DL are still actively studied in the medical field, especially in cervical lesion screening, with various techniques.

As mentioned earlier, DL is known to produce higher performance as a way of supplementing limitations of ML. In this study, we classify Cervicography images as negative and positive using ML and DL techniques which are computer-aided diagnostic tools under the same

environment and evaluate the performance of them. The reason why it is important to compare the two methods and find a method that performs better is that it is closely related to the faster introduction of automatic diagnostic technology in countries such as developing countries. Through this verification process, algorithms that are more suitable for classification of negative and positive cervical cancer are identified for clinical application. Such algorithms can be used to assist the diagnosis of cervical cancer.

Methods

Development environments. The neural network models were developed with Ubuntu 18.04 OS. It was used for learning using four of NVIDIA's GeForce RTX 2080 TI. As for the Graphic Driver, a 440.33 version of linux 64-bit released in November 2019, a CUDA 10.2 version, and a CUDNN 7.6.5 version were used. Python 3.7.7 was used as the language with Keras 2.3.1 library based on tensorflow 1.15.0. For feature extraction, pyradiomics 3.0 developed by a research group called radiomics at Harvard Medical University and scikit-learn 0.23.1 developed by a code project organized by Google were used.

Data description. The Institutional Review Board of Ewha Womans University Mokdong hospital approved (IRB No. EUMC 2015-10-004) this retrospective study and waived the requirement for informed consent for both study populations. All methods were performed in accordance with the relevant guidelines and regulations. A total of 4,119 collected image were filmed with three models for Cervicam series. Sizes of images obtained from each equipment were different: 1280x960 pixels with Dr.Cervicam equipment, 1504x1000 pixels with Dr.Cervicam WiFi equipment, and 2048x1536 pixels with Dr.Cervicam C20 equipment. This

equipment provides a cervical magnification examination video that magnifies the cervix by about 50 times. Grades by cervical magnification test are generally classified into negative, atypical, and positive. Only negative and positive were used in this study. These images consisted of 1,984 normal negative images and 2,135 abnormal positive images. Data were accurately verified by tissue biopsy. For the number of images used for learning, a total of 2,636 images (1,270 negative and 1,366 positive) were used for the training set and a total of 659 images (317 negative and 342 positive) were used for the validation. Test set used a total of 824 images (397 negative and 427 positive).

Data pre-processing. Cervicography images were generally obtained with width longer than height. The cervical area was located in the center of the image and the vaginal wall was often photographed on the left and right sides. In the stage of ML feature analysis, the entire input area is screened and the features are extracted, so it is recommended to remove areas other than the target area. The left and right ends were cropped to the same size and made into squares so that the width was equal to the height, provided that the cervical region was centered in the image. Likewise, in DL, the same pre-processed image was used as input to meet the same conditions.

Study design for ML analysis. The overall process of ML is shown in Figure 1. Train sets were pre-processed and converted into grayscale images as described earlier. After extracting more than 300 features from the pre-processed and converted image through the feature extraction stage, only major variables affecting the classification were selected through the Lasso model. ML model used in this study included Extreme Gradient Bosst (XGB), Support Vector Machine (SVM), and Random Forest (RF) for conducting classification learning process with selected

variables. Through the trained model, 5-fold cross-validation was performed with test set to evaluate model performance.

Eighteen first order features, 24 Grey Level Co-occurrence Matrix (GLCM), 16 Grey Level Run Length Matrix (GLRLM), 16 Grey Level Size Zone Matrix (GLSZM), and 226 Laplacian of Gaussian (LoG)-filtered-first-order features were included as second order features. A total of 300 features from five categories were extracted from train set images [23].

First-order feature is a value that relies only on each pixel value of the image for analyzing one-dimensional characteristics such as mean, maximum, and minimum not expressed in the image. Second, the GLCM of second-order-feature is a matrix that takes into account the spatial relationship between the reference pixel and the adjacent pixel. Adjacent pixels are east (0), north-east (45), north (90), and north-west (135) of reference pixels. Third, the second-order-feature GLRLM is a matrix that calculates how continuous pixels have the same value within a given direction and length. GLSZM identifies nine adjacent pixel zones, a matrix that calculates how continuous pixels with the same value are. Finally, LoG-filtered-first-order is a method of applying the Laplacian of Gaussian filter and then selecting the first order features. LoG filter is the application of Laplaceian filter after smoothing the image with Gaussian filter, a technique commonly used to find contours, the point of rapid change in the image.

ML generally adopt only key features to create easy-to-understand models, better-performing models, and fast-learning models. The lasso feature selection method using L1 regularization was adopted, with only a few important variables selected and coeffients of other variables

reduced to zero. This method is known to be simpler and more accurate than other models. Thus, it is often used to select variables [24].

The selected 10 features included variance included in First order feature, RunLengthNonUniformity, LongRunHighGrayLevelEmphasis, LongRunEmphasis included in GLRLM feature. GrayLevelNonUniformity, LargeAreaEmphasis, LargeAreaLowGrayLevelEmphasis, ZoneVariance, SizeZoneNonUniformity included in GLSZM feature, and log-sigma-10-0-mm-3D-Energy included in LoG filtered first order.

GrayLevelNonUniformity measures the variability of gray-level intensity values in the image, with a lower value indicating more homogeneity in intensity value. Variance is the the mean of squared distances of each intensity value from the mean value. This is a measure of the spread of the distribution about the mean. RunLengthNonUniformity measures the similarity of run lengths throughout the image, with a lower value indicating more homogeneity among run lengths in the image. LongRunHighGrayLevelEmphasis measures the joint distribution of long run lengths with higher gray-level values. LargeAreaEmphasis is a measure of the distribution of large area size zones, with a greater value indicating larger size zone and more coarse texture. LargeAreaLowGrayLevelEmphasis measures the proportion in the image of the joint distribution of larger size zones with lower gray-level values. Log-sigma-10-0-mm-3D-Energy is a calculation of energy in three dimensions by applying a LoG filter. ZoneVariance measures the variance in zone size volumes for zones. LongRunEmphasis is a measure of the distribution of long run lengths, with a greater value indicating longer run lengths and more coarse structural textures. SizeZoneNonUniformity measures the variability of size zone volume in the image, with a lower value indicating more homogeneity in size zone volume [25].

ML classification architectures. We used architecture XGB, RF, and SVM for ML classification. XGB is one of boosting methods that combines weak predictive models to create strong predictive models [26]. As shown in Figure 2-A, a pre-pruning method is used to compensate for the error of the previous tree and create the next tree. The RF in Figure 2-B is one of methods of bagging. After random selection of variables, multiple decision trees are created. Results are integrated into an ensemble technique to classify the data [27]. SVM is one of linear regression methods. When classifying two classes as shown in Figure 2-C, it finds data next to the line closest to the line called vector and then selects the point where the margin of the line and support vector maximizes [28]. A default value was designated for each parameter. Default values for main parameters of XGB were: loss function = linear stress; evaluation metric = rmse; learning rate eta = 0.3; and max default = 6. The SVM's main parameter default values were: normalization parameter C =1, the kernel used by the algorithm = RBF, and the kernel factor gama = 1. Default values for key parameters of RF were: number of crystals n_estimators = 100, the classification performance evaluation indicator = 'gini', max depth of the crystal tree = None, and the minimum number of samples required for internal node division = 2.

Study design for DL analysis. The entire DL process is shown in Figure 3. After preprocessing the same way as ML, the model was created with the Resnet-50 architecture. The generated model was applied to test set and model performance was evaluated through 5-fold cross-validation.

DL classification architecture. We used Resnet-50 architecture, one of convolution neural networks (CNN) (Figure 4-A). As shown in Figure 4-B, the traditional CNN method was used to find the optimal value of input x through the learning layer, while Resnet was used to find the

optimal F(x)+x by adding input x after the learning layer. This approach has the advantage of optimizing input values for the next layer [29].

In this study, we used ImageNet to generate weight, the most common used one in the field of image recognition, consisting of 1.25 million real-life images and 1,000 classes. This weight was applied to transfer learning [30]. Parameters for learning were set to batch size of 40 and epoch 300 suitable for computing power. The learning rate was set to be 0.0001 to prevent significant changes in transition learning weights. For proper learning speed, the image was resized to 256x256.

Evaluation process. Cross validation is one of the evaluation methods to prevent overcompatibility and improve accuracy in evaluating model performance. In this paper, we validated classification performances of two algorithoms with 5-fold cross validation, a method in which all datasets were tested once each with a total of five verifications. For implementation under the same condition, the same five training sets and test sets were used in each method.

Results

Visualization. The bar graph in Figure 5 shows 10 selected features and the importance of each feature of ML method. Features with values greater than zero had positive linear relationships while features smaller than zero meant negative linear relationships.

To determine which area the Resnet recognized as negative or positive, results of the test set were visualized using Class Activation Map (CAM) to show which areas were given more weights (Figure 6).

Evaluation. To evaluate performances of the XGB, SVM, RF and the Resnet-50, results were validated by a 5-fold cross validation and evaluated as precision, recall, f1-score, and acurracy indicator as shown in Figure 7.

Resnet-50 had an AUC of 0.97 (95% confidence interval [CI]: 94.9%-97.6%). XGB had an AUC of 0.82 (95% CI: 79.7%-85.1%). SVM had an AUC of 0.84 (95% CI: 80.1%-85.4%). RF had an AUC of 0.79%. Resnet-50 was 0.15 points higher than the average (0.82) of three ML algorithms ($p < 0.05$) (Figure 8).

Discussion

Principal findings. In this study, we compared performance by automatically classifying negative and positive cervical images using DL and ML among previously known artificial intelligence techniques. Resnet-50 architecture were 15% higher than the average value of ML methods XGB, RF, and SVM architecture performance.

Results. In this study, we compared performances of ML algorithms XGB, SVM, and RF and DL algorithm Resnet-50 among automatic classification techniques for cervical images to determine which algorithm would be more suitable to help with accurate diagnosis by clinicians. Using 1,984 negative scenes and 2,135 positive scenes, a total of 4,119 cervicography images

were used to select 10 features out of 300 features after pre-processing of images in a linear regression. They were then trained with three algorithms (XGB, SVM, and RF) to create a ML classification model. DL classification model with Resnet-50 architecture was also generated with pre-processing images. With both techniques, the assessment achieved more reliable results when all datasets were tested once using 5-fold cross validation. AUC values for XGB, SVM, and RF were 0.82, 0.84, and 0.79, respectively. Resnet-50 showed an AUC value of 0.97. ML algorithms did not exceed 0.80 for accuracy, while Resnet-50 showed an accuracy of 0.9065 with a relatively better performance.

Clinical implications. Generally, lesions are diagnosed by compiling thickness of aceto-white area, presence of transformation zone, and tumor identification when diagnosing cervical cancer in clinical practice. Given this complexity in the diagnostic process, it is judged that the end-to-end method of DL, which divides the problem into multiple parts and then obtains answers for each and combines results, might have contributed to its improved performance for cervical cancer classification with a single step in the learning system, compared to step-by-step ML method that splits the problem into multiple parts, obtain the answers for each, and then add the results together.

Research implications. In terms of algorithms, DL selects and trains meaningful features among all features by itself, while in ML, features that are deemed unnecessary are removed by human judgment with certain techniques. This might have led to the decrease in learning performance. Since DL learns low-level features in the initial layer and high-level features as layer deepens, the weight of high-level features not trained in ML is added up. Thus, DL is found to have performed better.

Additionally, we will add a DL-based cervical detection model to the process in the future. This study has used arbitrary crop as a pre-processing method, but if a cervical area detection model is used, a more accurate comparative study will be possible because only the exact desired area can be analyzed.

Strengths and limitations. This study is the first to compare the performance of DL and ML in the field of automatic cervical cancer classification. Compared to other studies that have produced results using only one method of DL or ML, this work has the advantage of enabling cervical clinicians to objectively evaluate which automation algorithms are better as a computer-aided diagnostic tool.

In pre-processing, the same width is cut from both ends to remove vaginal wall areas taken at both ends of the image, assuming that the cervix was exactly in the middle during pre-processing. However, not all images have the cervix in the center. In addition, not all images have the exact form of a circle. In other words, the cut image may still contain vaginal walls which is unnecessary or contain the cervix that should be analyzed is cut. This may lead to poor accuracy for the comparison. In addition, when selecting ML features, the lasso technique was adopted and 10 features were selected. However, adopting a different feature selection method or selecting features more than or less than 10 might result in completely different results. The fact that human intervention is involved in the process itself has the disadvantage of not being able to accurately compare it with DL by making results inaccurate.

Conclusion

In this study, the performance of the existing ML and DL techniques was objectively evaluated and compared in the classification of negative and positive cervical cancer under the same environment.

The results of this study can serve as a criterion for objective evaluation of which technique clinicians will choose as a computer-assisted diagnostic tool in the future. In addition, when diagnosing cervical cancer, it can help to consider diagnostic factors in various ways by outputting both the automatically selected features (DL) and the randomly selected features (ML).

In future studies, a more accurate comparison of cervical cancer classification performance will be conducted by adding a detection model that accurately detects and analyzes only the cervix, and by minimizing human intervention in ML through finding and adopting the optimal feature selection technique.

The results of these additional studies are convinced that the automatic diagnosis of cervical cancer will be able to accelerate the introduction of computer-assisted diagnostic tools that will produce more accurate and reliable results in countries or regions where there is an urgent need.

References

1. Steven, E. W. Cervical cancer. Lancet. 361, 2217–2225 (2003).

2. Rebecca, L. S., et al. Cancer Statistics 2019. A Cancer Journal for Clinicians. 69, 7–34 (2019).

3. K. Canfell et al. Mortality impact of achieving WHO cervical cancer elimination targets: a comparative modelling analysis in 78 low-income and lower-middle-income countries. Lancet, 395, 591–603 (2020).

4. Adolf, S. Cervicography: A new method for cervical cancer detection. Am. J. Obstet. Gynecol. 139, 815–821 (1981).

5. M. F. Janicek et al. Cervical Cancer : Prevention , Diagnosis , and Therapeutics. A Cancer Journal for Clinicians. 51, 92–114 (2008).

6. J. S. Mandelblatt et al. Costs and Benefits of Different Strategies to Screen for Cervical Cancer in Less-Developed Countries. JNCI J. Natl. Cancer Inst. 94, 1469–1483 (2002).

7. T. C. J. R. WRIGHT. Cervical Cancer Screening in the 21st Century: Is it Time to Retire the PAP Smear? Clin. Obstet. Gynecol. 50, (2007).

8. M. Ottaviano, et al. Examination of the cervix with the naked eye using acetic acid test. Am. J. Obstet. Gynecol., 143, 139–142 (1982).

9. W. Small, et al. Cervical Cancer : A Global Health Crisis. A Cancer Journal for Clinicians. 123, 2404-2412 (2017).

10. M. Schiffman, et al. The Promise of Global Cervical-Cancer Prevention. The New England Journal of Medicine. 353, 2101–2104 (2005).

11. Sezgin, M. Ismail et al. Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. Br. Med. J. 298, 707–710 (1989).

12. D. C. Sigler, et al. Inter- and intra-examiner reliability of the upper cervical X-ray marking system. J. Manipulative Physiol. Ther., 8, 75–80 (1985).

13. Richard, L. Comparison of computer-assisted and manual screening of cervical cytology. Gynecol. Oncol., 104, 134–138 (2007).

14. Afzal, H. S., et al. A deep learning approach for prediction of Parkinson's disease progression. Biomed. Eng. Lett., 10, 227–239, (2020).

15. Alexzandru, K., et al. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Mol. Pharm. 14, 4462–4475 (2017).

16. K. G. Kim. Deep Learning. 22, 351–354 (2016).

17. Xiaoyu, D., et al. Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods. in proceedings of the IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), 631–635.(IEEE, 2018)

18. Héctor-Gabriel, A., et al. Aceto-white temporal pattern classification using k -NN to identify precancerous cervical lesion in colposcopic images. Comput. Biol. Med. 39, 778–784 (2009).

19. Muhammad, T., et al. Classification of Colposcopy Data Using GLCM- SVM on Cervical Cancer. in proceedings of 2020 International Conference on Artificial Intelligence Information and Communication. 373–378 (ICAIIC, 2020).

20. H. Y. Kim, et al. A Study on Nucleus Segmentation of Uterine Cervical Pap-Smears using Multi Stage Segmentation Technique. J Korean Soc Med Inf. 5, 89–97 (1999).

21. Limming, H., et al. An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening. JNCI. 111, 923–932, (2019).

22. Masakazu, S., et al. Application of deep learning to the classification of images from colposcopy. Oncology Letters. 3518–3523 (2018).

23. Avrim, L., et al. Selection of relevant features and examples in machine learning. Artif. Intell., 97, 245–271 (1997).

24. Valeria, F., et al. Feature selection using lasso. VU Amsterdam Res. Pap. Bus. Anal. 30, 1–25, (2017).

25. Robert, M., et al. Textural Features for Image Classification. IEEE Trans. Syst. Man. Cybern. SMC-3, 610–621 (1973).

26. Tianqi, C., et al. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785–794 (KDD, 2016).

27. Andy, L., et al. Classification and regression by randomForest. R news, 2, 18–22 (2002).

28. William, S. N. What is a support vector machine? Nat. Biotechnol. 24, 1565–1567 (2006).

29. Sasha, T., et al. Resnet in resnet: Generalizing residual architectures. arXiv Prepr. arXiv1603.08029, (2016).

30. Raja,. R., et al. Self-taught learning: transfer learning from unlabeled data. in Proceedings of the 24th international conference on Machine learning. 759–766 (ICML, 2017).

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2017-0-01630) supervised by the IITP(Institute for Information & communications Technology Promotion), and GRRC program of Gyeonggi province. [GRRC-Gachon2020(B01), AI-based Medical Image Analysis], Intelligent SW Technology Development for Medical Data Analysis), and the Technology development Program(S2797147) funded by the Ministry of SMEs and Startups(MSS, korea).

Author contributions

Y.R.P. contributed the deep learning and machine learning analysis, drafted the manuscript, and generated the figures. Y.J.K and K.G.K contributed to design the study, draft and revise the manuscript. W.J, K.N, and S.K contributed recruiting participants and data anonymizing. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Figure legends

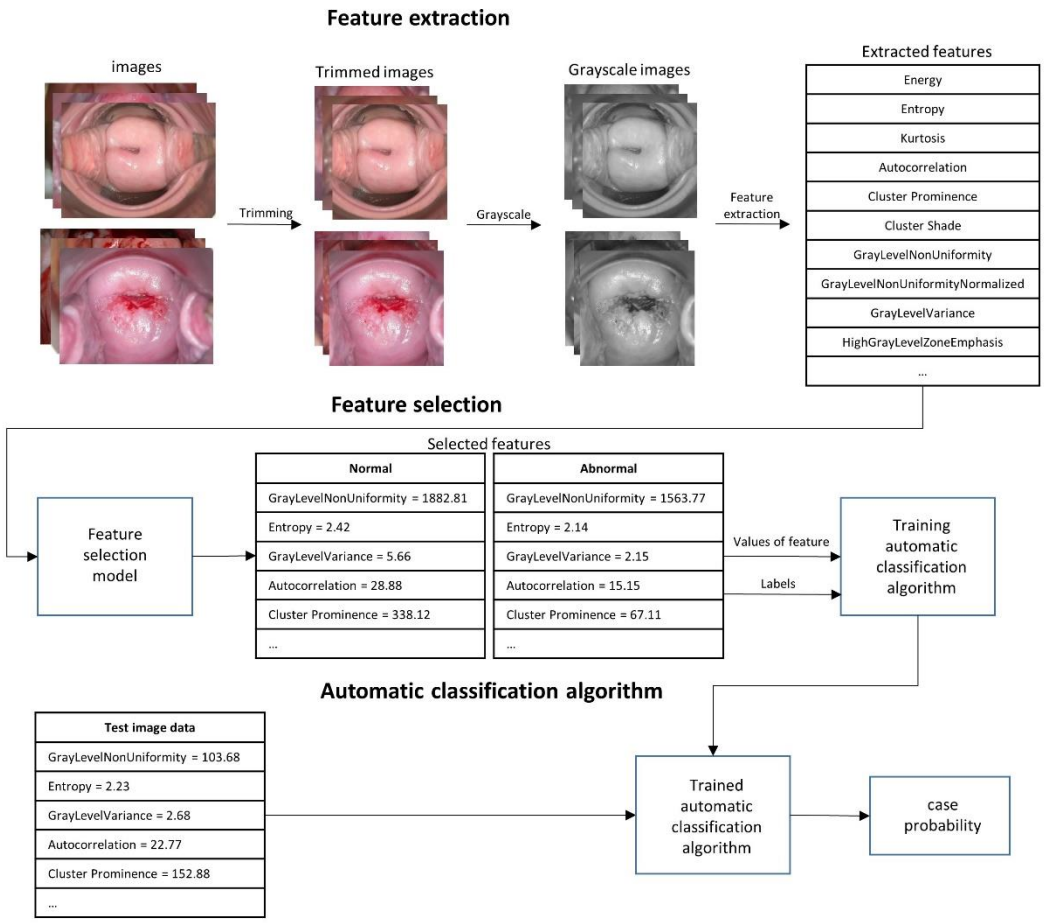


Figure 1. ML model training process for cervical cancer classification. The trimmed and grayscaled input data, negative and positive groups, are used to extract radiomics features. After feature selection process, 10 features are trained for cervical cancer classification. The trained model classify the test image and produces results.

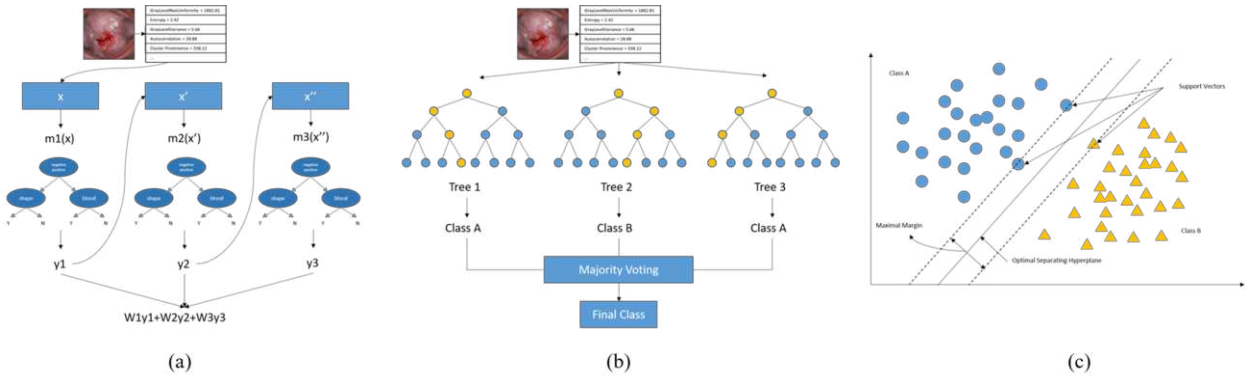


Figure 2. The diagrams of ML model architectures for training cervical cancer classification. (**a**) In XGB, The y value extracted through one feature is used as an input value to predict the next feature, and the final y values after this process is repeated are combined with the weights. (**b**) RF creates several decision trees and synthesize the results of each tree to determine the final class. (**c**) After linearly classifying the data, SVM finds the boundary with the largest width through an algorithm.

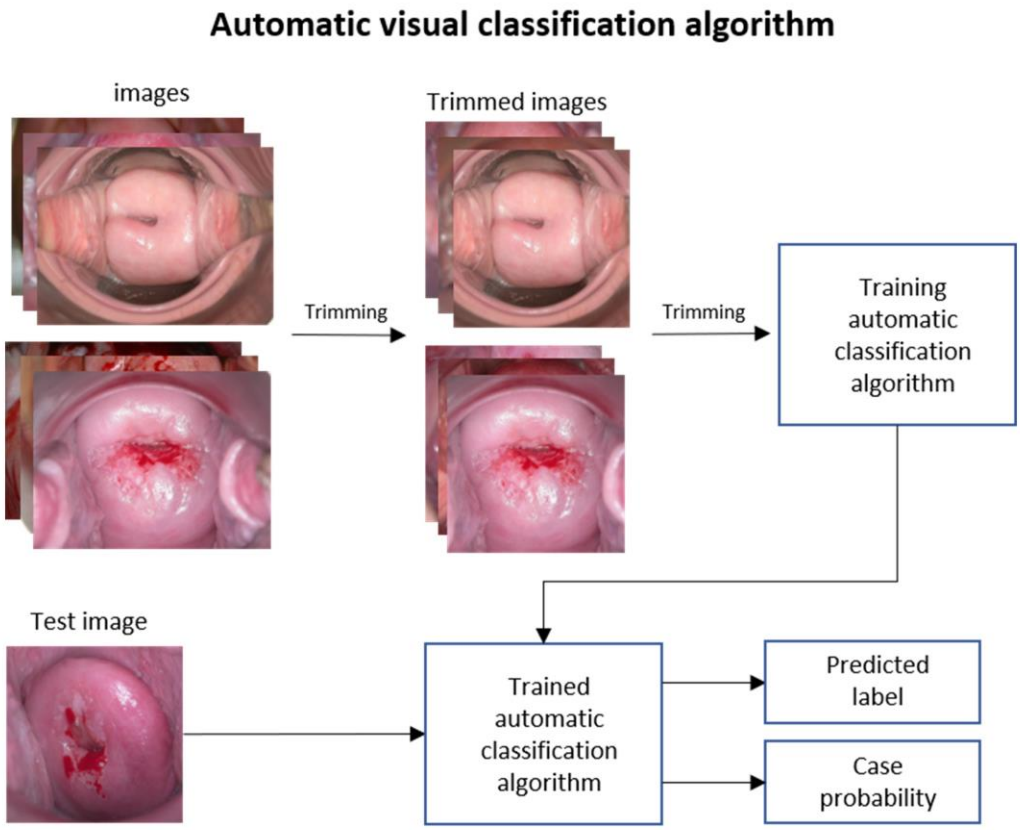
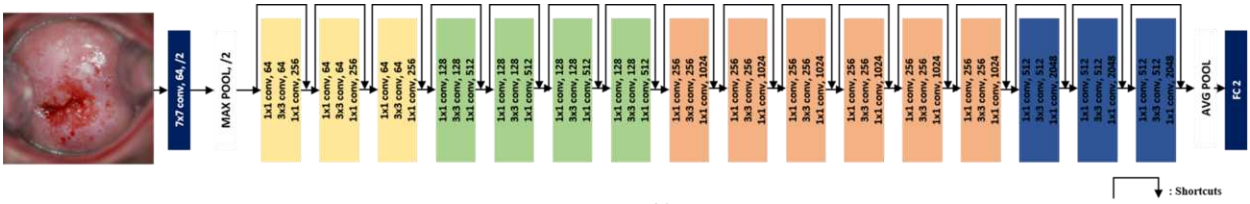
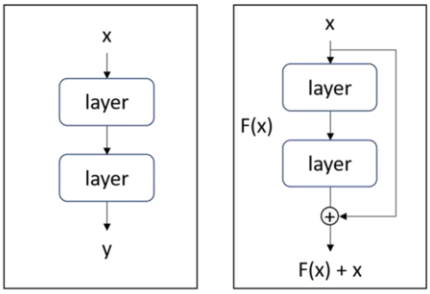


Figure 3. DL model training process for cervical cancer classification. The input images are trimmed and used to training data. The trained model predicts test data.



(a)



(b)

Figure 4. (**a**) ResNet50 Architecture of DL, by reducing the dimension by adding 1x1 conv to each layer, the amount of computation is reduced and the training speed is accelerated. (**b**) learning method of existing CNN(left) and ResNet (right). By adding shortcuts that adds input value to output value every two layers, errors is reduced faster than existing CNNs.

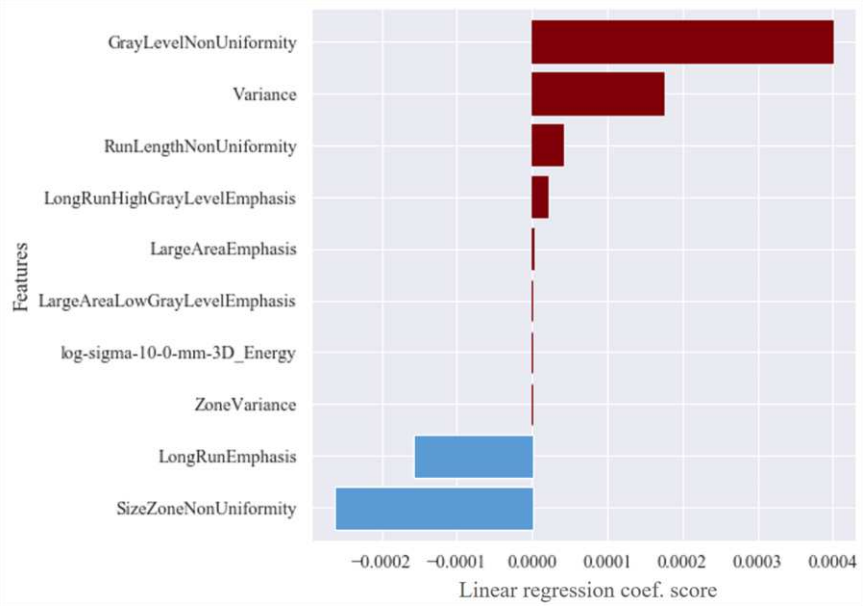


Figure 5. Selected 10 features by Lasso regression efficient score. 8 features showed positive coefficient scores(red) while 2 features showed negitve coefficient scores(blue).

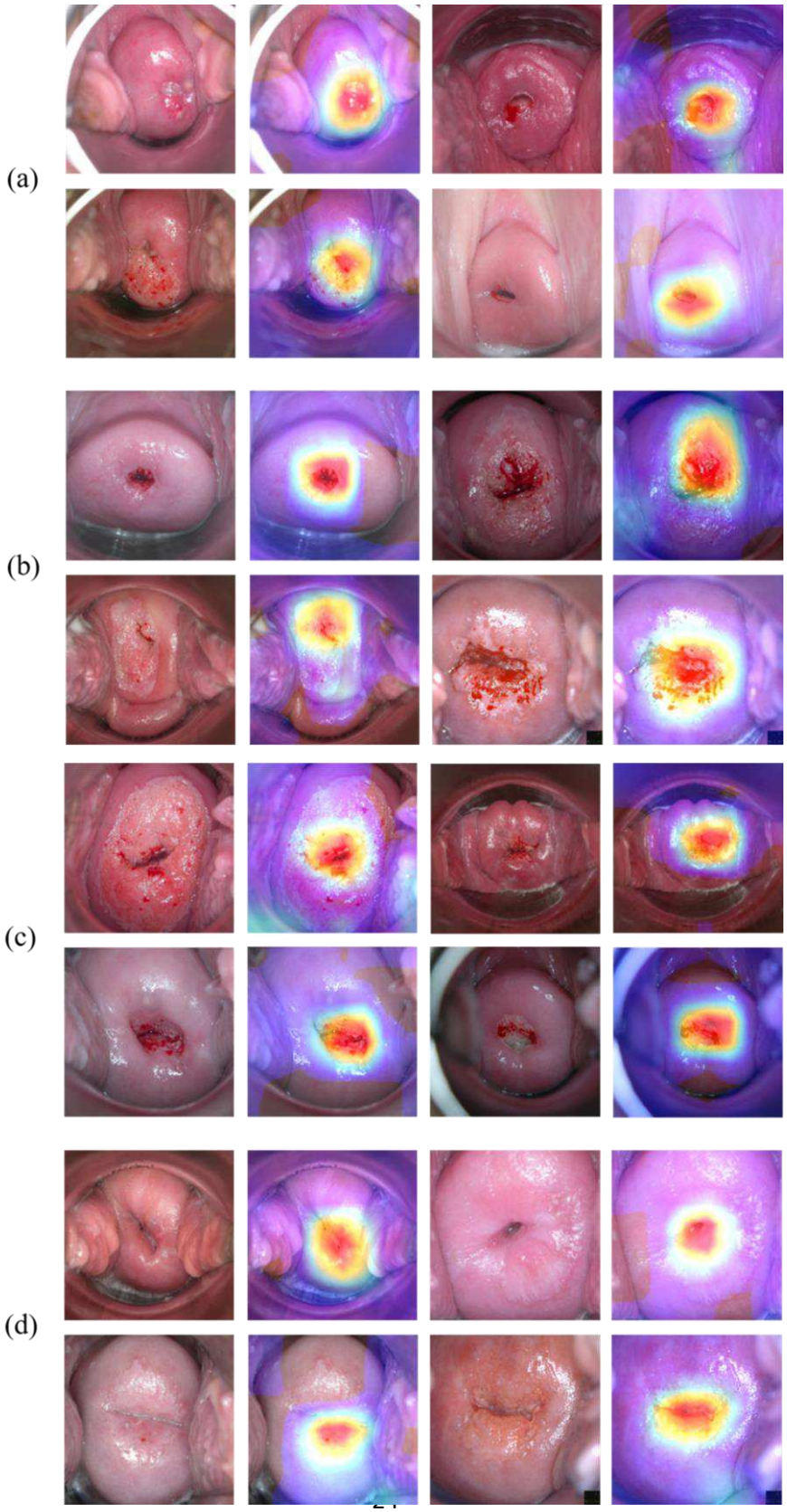

(a)
(b)
(c)
(d)

Figure 6. Examples of CAM images of test sets. (**a**)True Negative(ground truth: negative, predict: negative) (**b**)True Positive(ground truth: positive, predict: positive) (**c**)False Positive(ground truth: negative, predict: positive) (**d**)False Negative(ground truth: positive, predict: negative).

Metrics	XGB	SVM	RF	ResNet50
Precision	0.75 ± 0.14	0.79 ± 0.07	0.80 ± 0.06	0.94 ± 0.02
Recall	0.73 ± 0.03	0.73 ± 0.08	0.67 ± 0.08	0.86 ± 0.01
F1-score	0.74 ± 0.03	0.76 ± 0.08	0.73 ± 0.12	0.90 ± 0.10
Accuracy	0.74 ± 0.03	0.76 ± 0.01	0.71 ± 0.10	0.90 ± 0.09
AUC	0.82 ± 0.09	0.84 ± 0.06	0.79 ± 0.04	0.97 ± 0.07

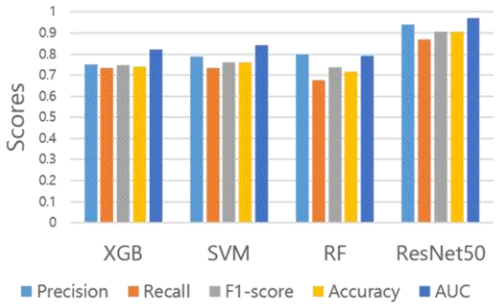


Figure 7. Internal cross validation results for test data to predict cervical cancer. ResNet-50 showed the highest performance in all metrics, and SVM showed the highest performance when compared with AUC among ML models excluding ResNet-50.

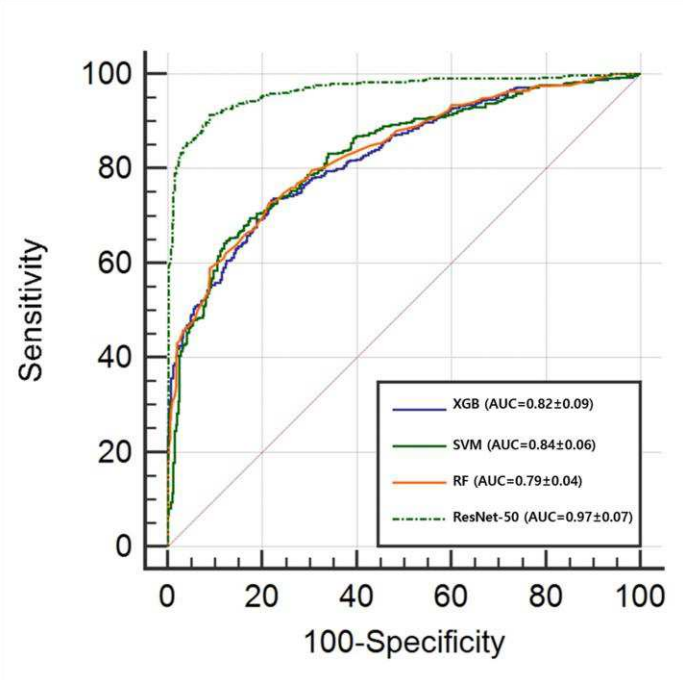


Figure 8. Mean ROC comparison graph of 5-fold cross validation of each method to predict cervical cancer.

Figures

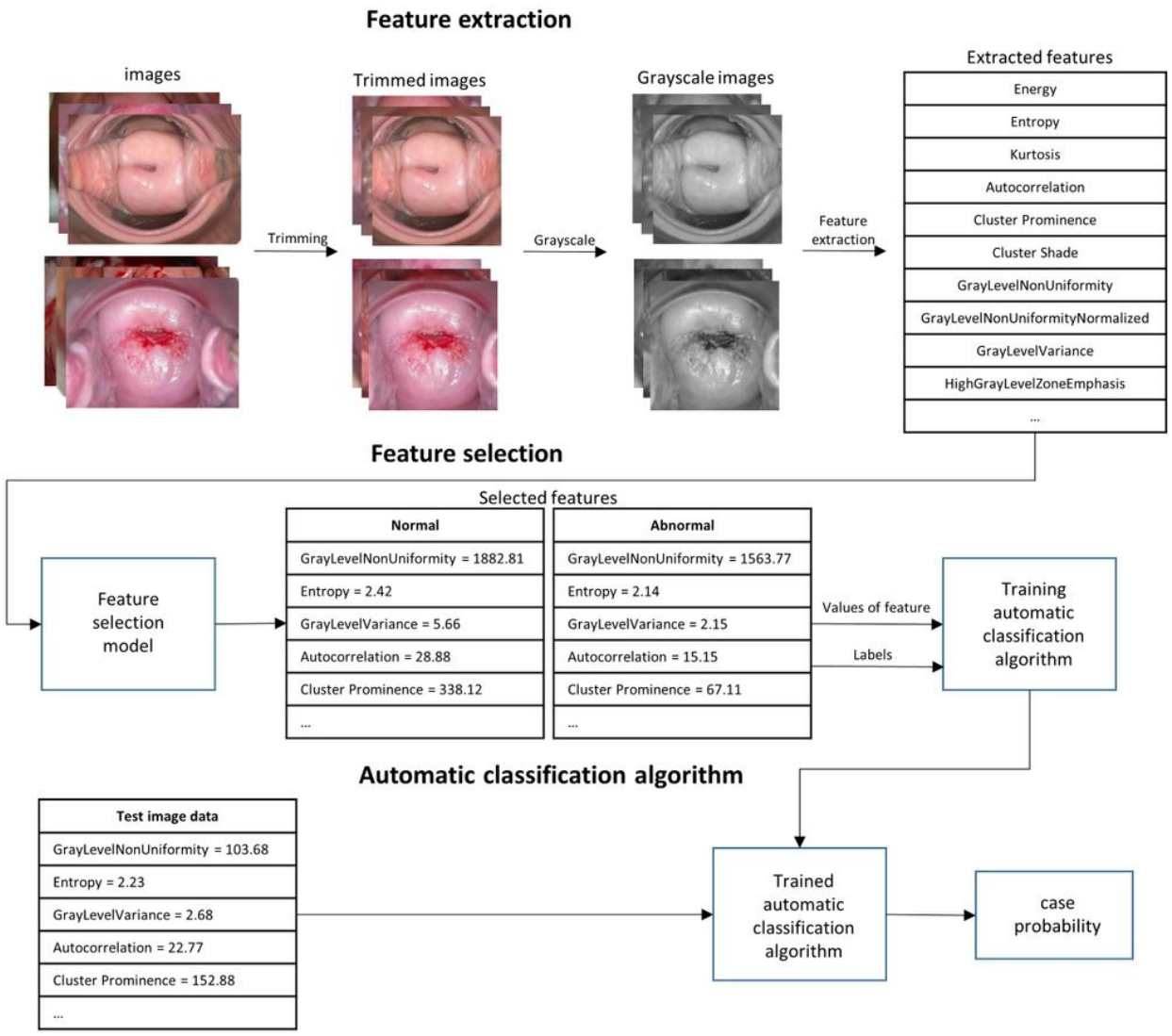


Figure 1

ML model training process for cervical cancer classification. The trimmed and grayscaled input data, negative and positive groups, are used to extract radiomics features. After feature selection process, 10 features are trained for cervical cancer classification. The trained model classify the test image and produces results.

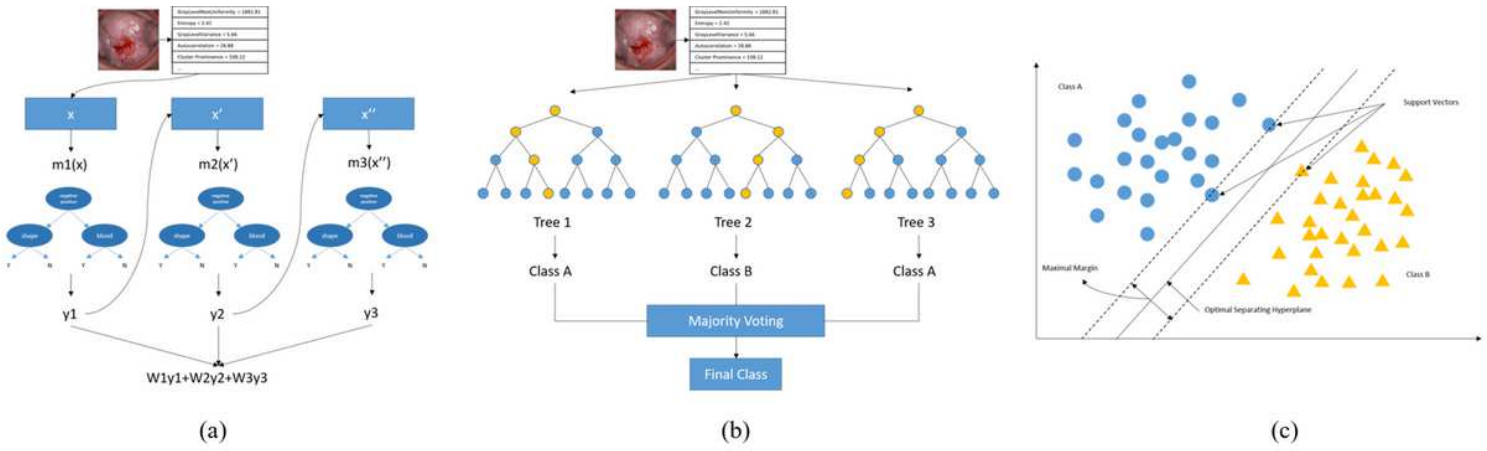


Figure 2

The diagrams of ML model architectures for training cervical cancer classification. (a) In XGB, The y value extracted through one feature is used as an input value to predict the next feature, and the final y values after this process is repeated are combined with the weights. (b) RF creates several decision trees and synthesize the results of each tree to determine the final class. (c) After linearly classifying the data, SVM finds the boundary with the largest width through an algorithm.

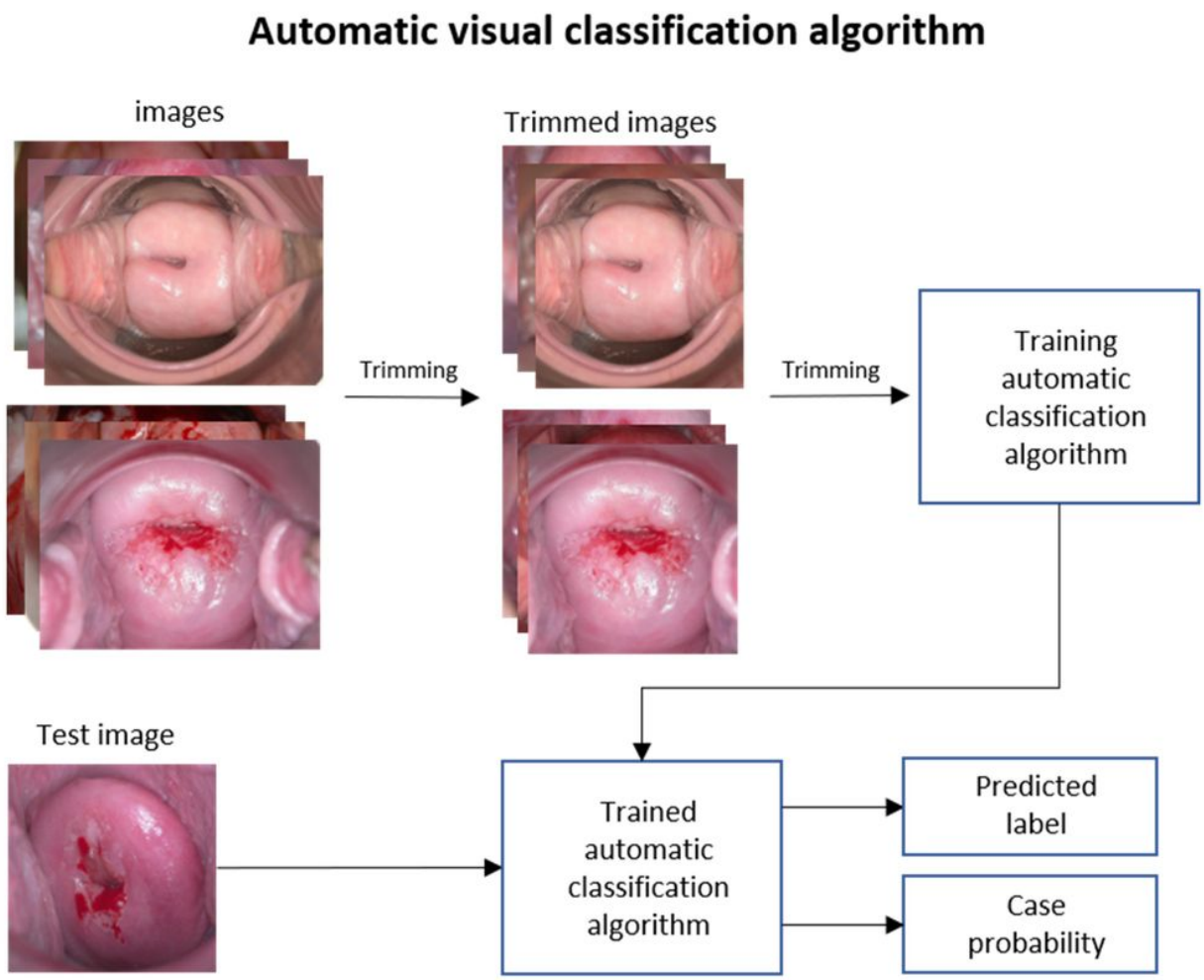


Figure 3

DL model training process for cervical cancer classification. The input images are trimmed and used to training data. The trained model predicts test data.

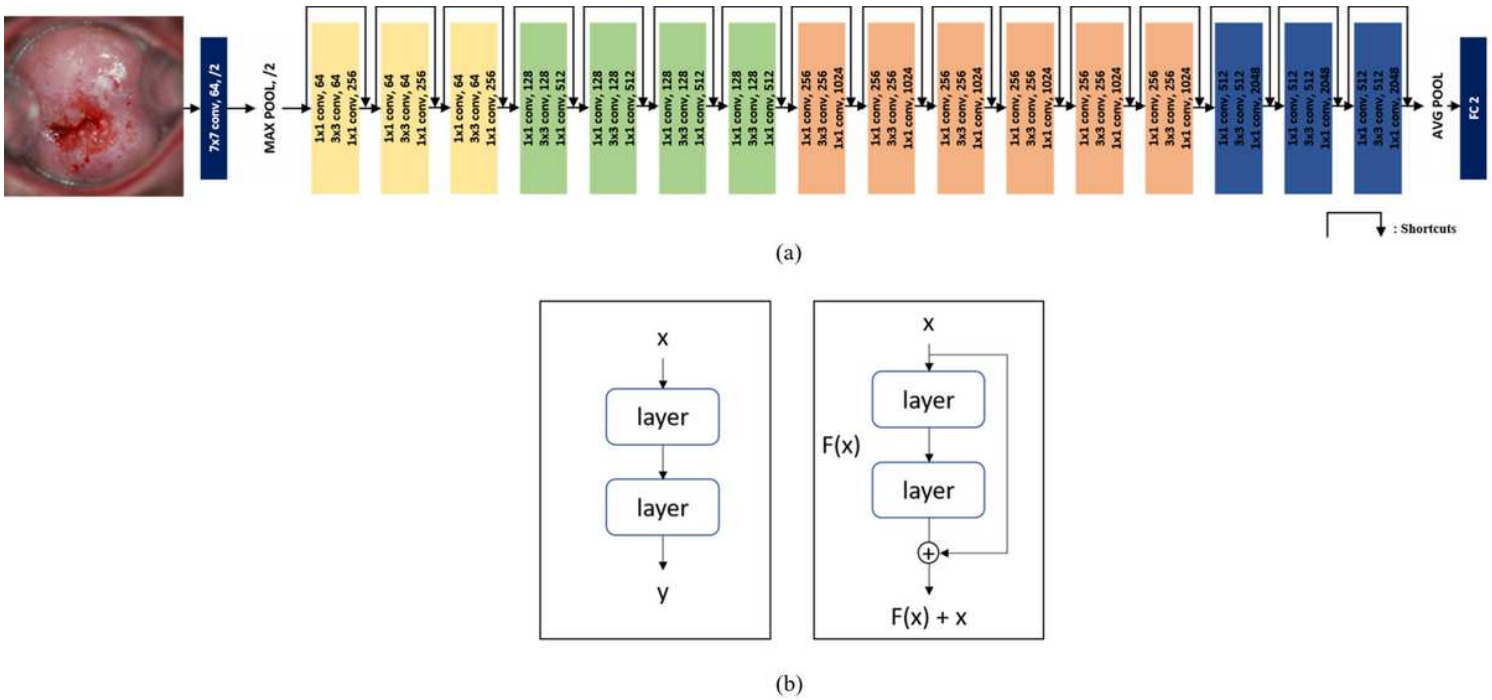


Figure 4

(a) ResNet50 Architecture of DL, by reducing the dimension by adding 1x1 conv to each layer, the amount of computation is reduced and the training speed is accelerated. (b) learning method of existing CNN(left) and ResNet (right). By adding shortcuts that adds input value to output value every two layers, errors is reduced faster than existing CNNs.

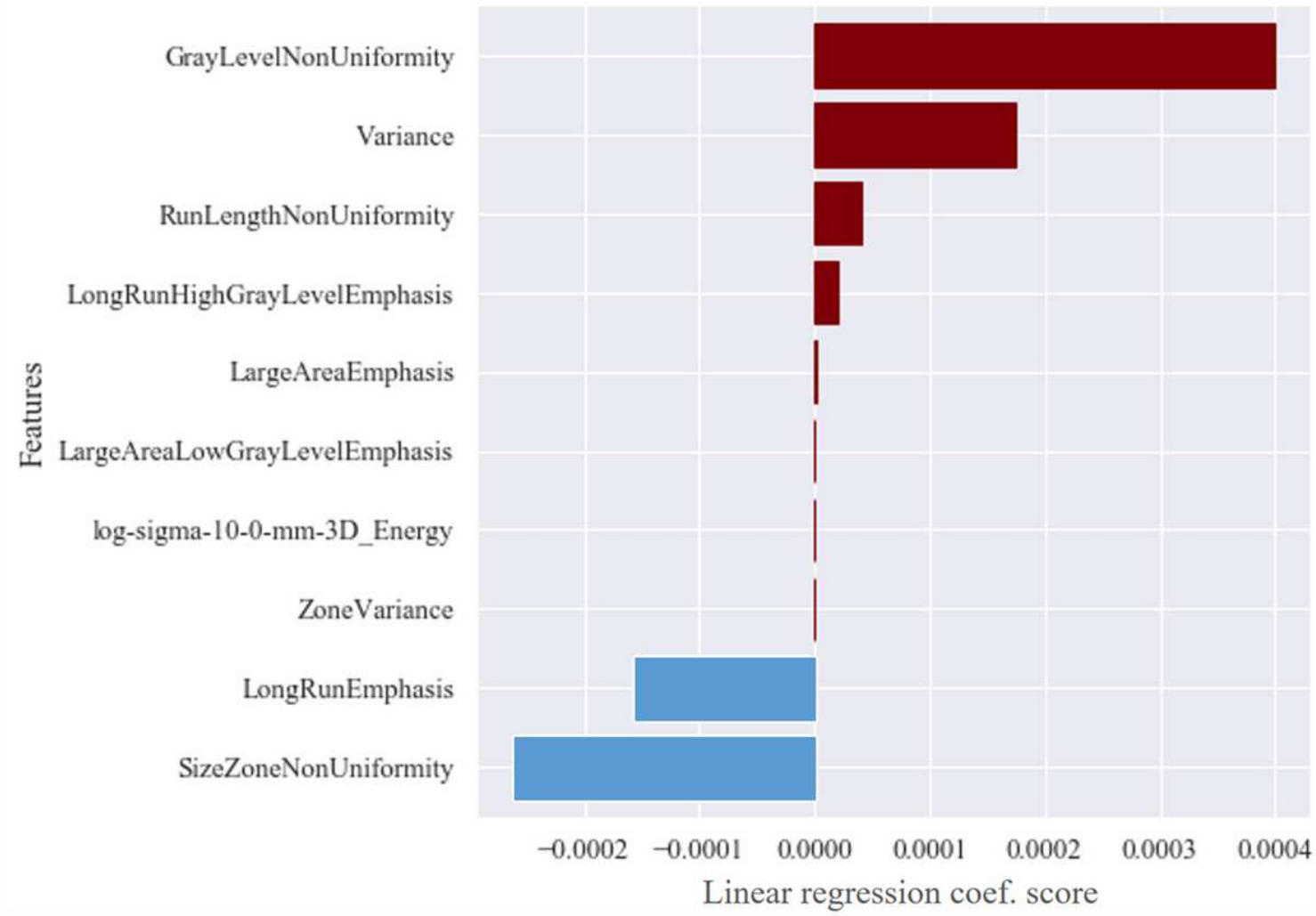


Figure 5

Selected 10 features by Lasso regression efficient score. 8 features showed positive coefficient scores(red) while 2 features showed negitve coefficient scores(blue).

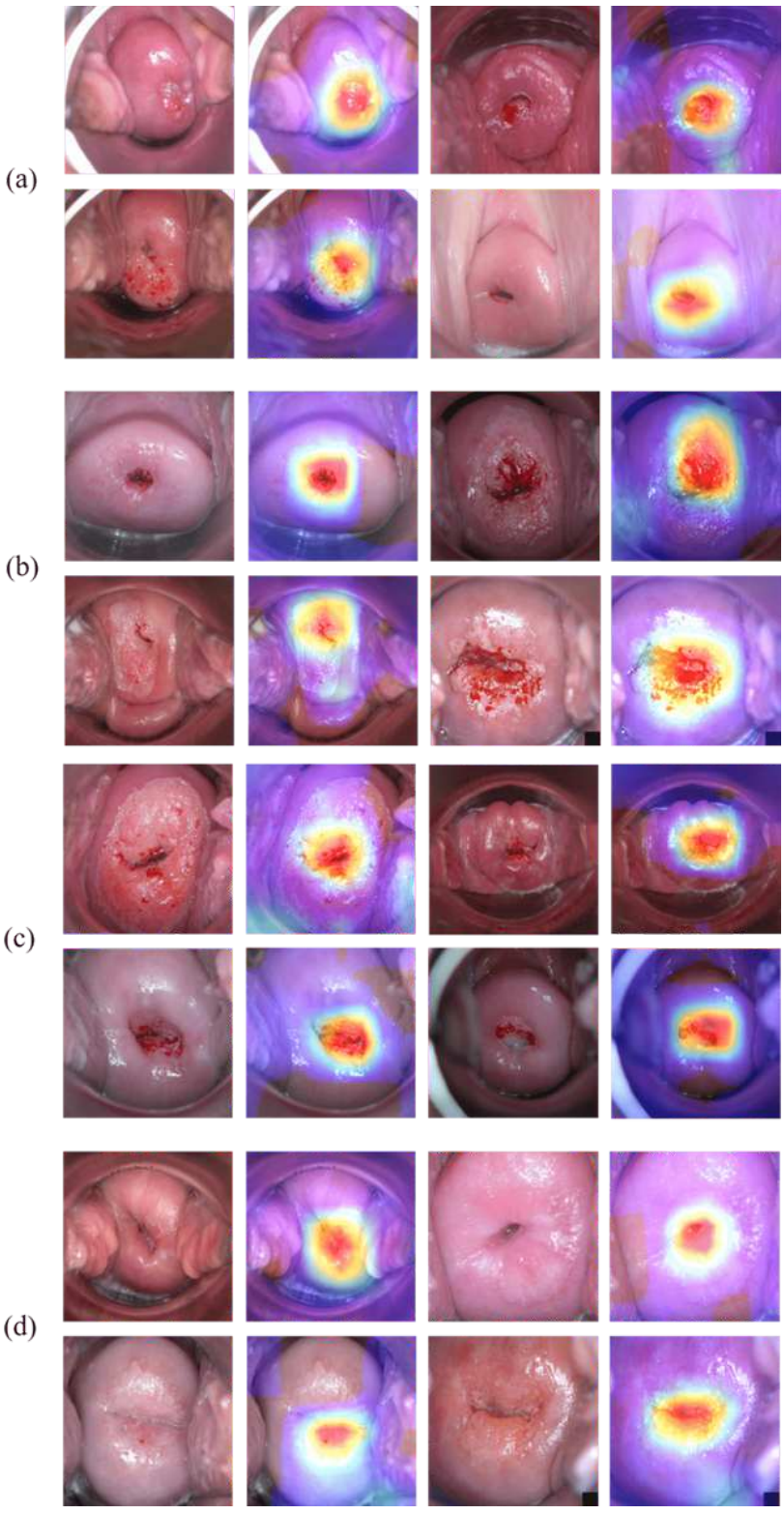


Figure 6

Examples of CAM images of test sets. (a)True Negative(ground truth: negative, predict: negative) (b)True Positive(ground truth: positive, predict: positive) (c)False Positive(ground truth: negative, predict: positive) (d)False Negative(ground truth: positive, predict: negative).

Metrics	XGB	SVM	RF	ResNet50
Precision	0.75 ± 0.14	0.79 ± 0.07	0.80 ± 0.06	0.94 ± 0.02
Recall	0.73 ± 0.03	0.73 ± 0.08	0.67 ± 0.08	0.86 ± 0.01
F1-score	0.74 ± 0.03	0.76 ± 0.08	0.73 ± 0.12	0.90 ± 0.10
Accuracy	0.74 ± 0.03	0.76 ± 0.01	0.71 ± 0.10	0.90 ± 0.09
AUC	0.82 ± 0.09	0.84 ± 0.06	0.79 ± 0.04	0.97 ± 0.07

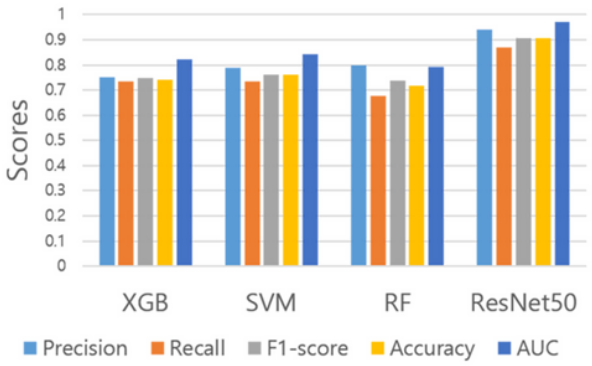


Figure 7

Internal cross validation results for test data to predict cervical cancer. ResNet-50 showed the highest performance in all metrics, and SVM showed the highest performance when compared with AUC among ML models excluding ResNet-50.

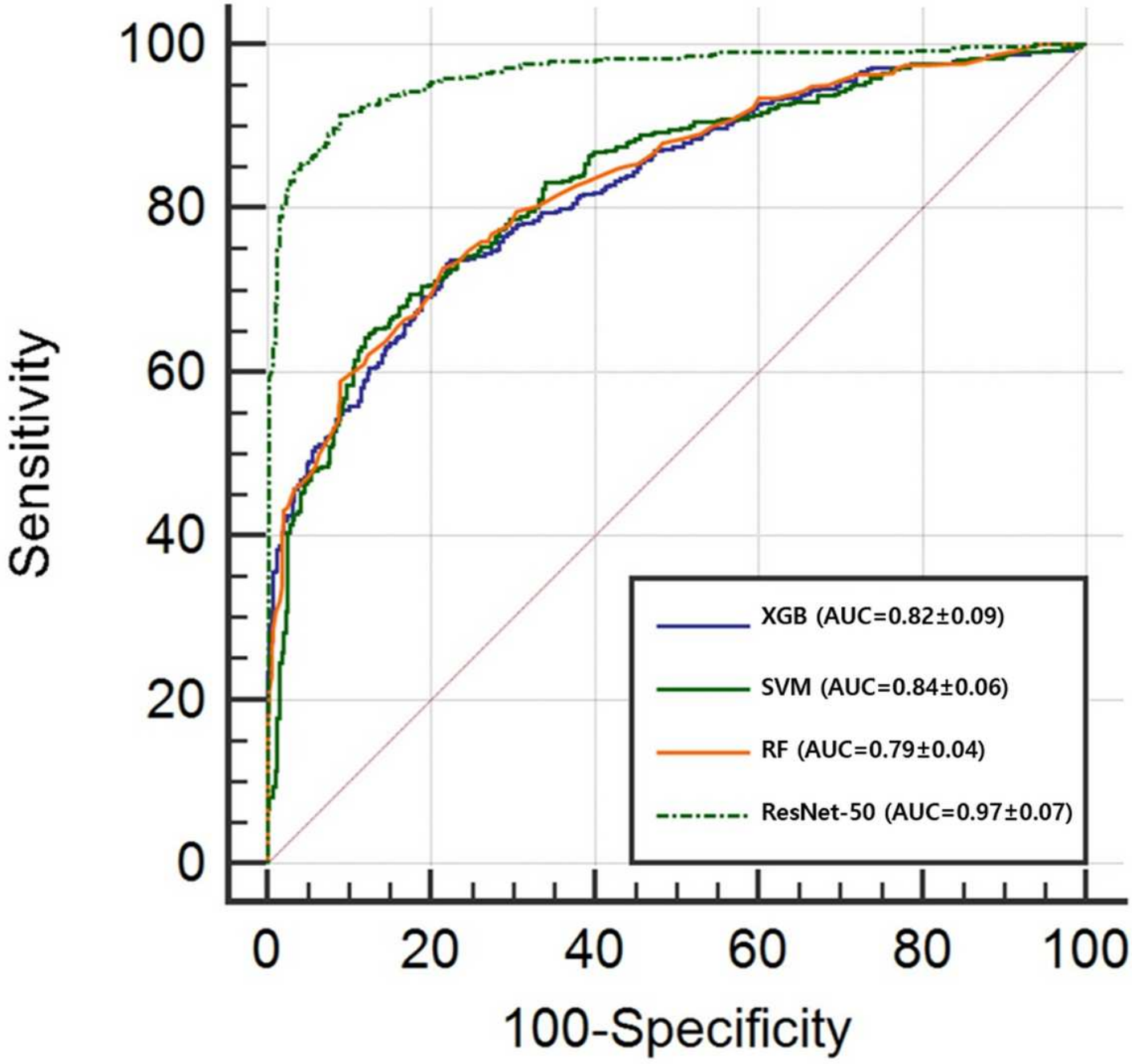


Figure 8

Mean ROC comparison graph of 5-fold cross validation of each method to predict cervical cancer.