An Efficient Machine Learning based Cervical Cancer Detection and Classification

Ahmed N. Al Masri[1], Hamam Mokayed [2]

[1] American University in the Emirates, Dubai, United Arab Emirates
[2] LTU University of Technology, Sweden
Emails: ahmed.almasri@aue.ae ; Hamam.mokayed@ltu.se

.

Abstract

Cervical cancer (CC) is the fourth commonly occurring cancer among females over the globe. It accounts for 7.9% of woman cancer as identified by world health organization (WHO). The most important reason for increased mortality due to cervical cancer is the deficiency of effective initial treatment. The asymptomatic nature is a main problem faced in the analysis of CC from initial stage. Recently, computer aided diagnosis (CAD) model has gained significant attention in the disease diagnostic process. At the same time, machine learning (ML) finds its use in several medical applications and is utilized as classifier for the initial detection of cancerous cells occurs from cervix area of uterus. With this motivation, this study introduces an intelligent ML based CAD (IML-CAD) technique to classify cervix cancer. The IML-CAD technique involves different stages of operations to detect and classify the cancerous cervix cells. In addition, the IML-CAD technique involves histogram based segmentation to determine the affected regions. Moreover, local binary patterns (LBP) based feature extractor and least squares support vector machine (LS-SVM) based classifier is designed for CC classification. To showcase the better performance of the IML-CAD technique, a series of simulations is performed and the experimental results highlighted the superior performance of the IML-CAD technique over the other techniques.

Keywords: Machine learning, Cervical cancer, Pap Smear images, CAD model, Image classification.

1. Introduction

Cervical cancer (CC) is the main medical conditions for ladies around the world, and particularly in non-industrial nations [1]. According to the insights given by WHO In 2012, roughly 270, 000 ladies kicked the bucket from CC over 90% of this passing's happening in lower and central pay nations. It has 445,000 novel cases in 2012. With no serious consideration, the pace of this pass was anticipated to ascend by twenty-five percentage [2-4]. The earlier manifestations of CC incorporate unusual period, hefty feminine cycle, sporadic feminine cycle, or spotting [5]. A Pap smear is basic, fast, as well as basically easy evaluating system for CC. The cycle incorporates the assortment of cell from lady's cervix at the time of pelvic test is spreading over a magnifying lens slide to assessment [6-8].

The disadvantages of such test's territory deficient medical part, lacking person consistency, helpless reproducibility's of judgment, insufficient support, as well as carelessness by experts directing the test

DOI: 10.5281/zenodo.5504344

because of profoundly monotonous nature of test [9-10]. Though afterward the approach of pap smear test and HPV inoculation to forestall diseases in more youthful ladies, which is lesser than equivalent to eighteen years old, it was lessening in quantity of mortalities because of cancer around there, the death rate in non-industrial country is yet high [11-14]. The CC creates in a lady's cervix. The cervix is low, restricted piece of ladies' uterus or belly. Cancer happens because of an infection named human papillomavirus, or HPV and this infection could spread via sexual contact. All grown-ups were tainted with HPV sooner or later and this disease may disappear all alone and subsequently end up being innocuous. Yet, Productive contamination by higher-hazard HPV kinds is shown as cervical level moles or condyloma that shed irresistible virions from their surface [15].

In [16], a new ensemble approach is introduced to foresee the danger of CC. By embracing a voting system, this technique tends to the difficulties related to past investigations on CC. A data correction system is developed to enhance the exhibition of the expectation. A quality help module is additionally included as a discretionary procedure to upgrade the vigor of the forecast. Various estimations are performed to assess the proposed technique. The outcomes demonstrate that the probability of creating CC can be viably anticipated utilizing the voting methodology. Ijaz et al. [17] proposes a CCPM provides earlier expectations of CC utilizing hazard features as information sources. The CCPM first eliminates anomalies by utilizing exception identification strategies, for example, DBSCAN and iForest and by expanding the quantity of cases in the dataset in a fair way, for instance, through SMOTE and with SMOTE Tomek. At long last, it utilizes RF as classification.

Khamparia et al. [18] proposed a new IoHT driven profound learning system for identification and arrangement of CC in Pap smear pictures utilizing idea of transfer learning. Following exchange learning, CNN was joined with various regular ML strategies like KNN, NB, LR, RF, and SVM. In the proposed structure, highlight extraction from cervical images is performed utilizing pre-prepared CNN models are taken care of thick and leveled layers for ordinary and unusual cervical cell classifier.

Rehman et al. [20] presents a CC cell identification and characterization framework dependent on CNNs. The cell pictures are taken care of into a CNNs model to extricate profound learned highlights. At that point, an ELM based classification characterizes the information images. CNNs model is utilized through move learning and tweaking. Options in contrast to the ELM, MLP, and AE based classifications are likewise researched. This article gives an audit of the science, counteraction, and therapy of CC, and talks about the worldwide CC emergency and endeavors to improve the avoidance and therapy of the infection in immature nations. It is utilized cross country Swedish segment and wellbeing registers to follow an open populace of 1,672,983 young ladies and ladies who were 10 to 30 years old from 2006 through 2017. The authors surveyed the relationship between HPV inoculation and the danger of intrusive CC, controlling for age at follow-up, schedule year, district of home, and parental attributes, including instruction, family pay, homelands of birth, and maternal sickness history.

This paper presents an intelligent ML based CAD (IML-CAD) model for CC classification. The IML-CAD technique involves different stages of operations to detect and classify the cancerous cervix areas. In addition, the IML-CAD technique involves histogram-based segmentation to determine the affected regions. Moreover, local binary patterns (LBP) based feature extractor and least squares-support vector machine (LSSVM) based classifier is designed for CC classification. In order to showcase the better performance of the IML-CAD model, a series of simulations take place and the experimental results highlighted the superior performance of the IML-CAD technique over the other techniques.

2. The Proposed IML-CAD Technique

Fig. 1 depicts the working process of IML-CAD technique. The IML-CAD technique involves histogram based segmentation, LBP based feature extractor, and LS-SVM based classifier.

DOI: 10.5281/zenodo.5504344

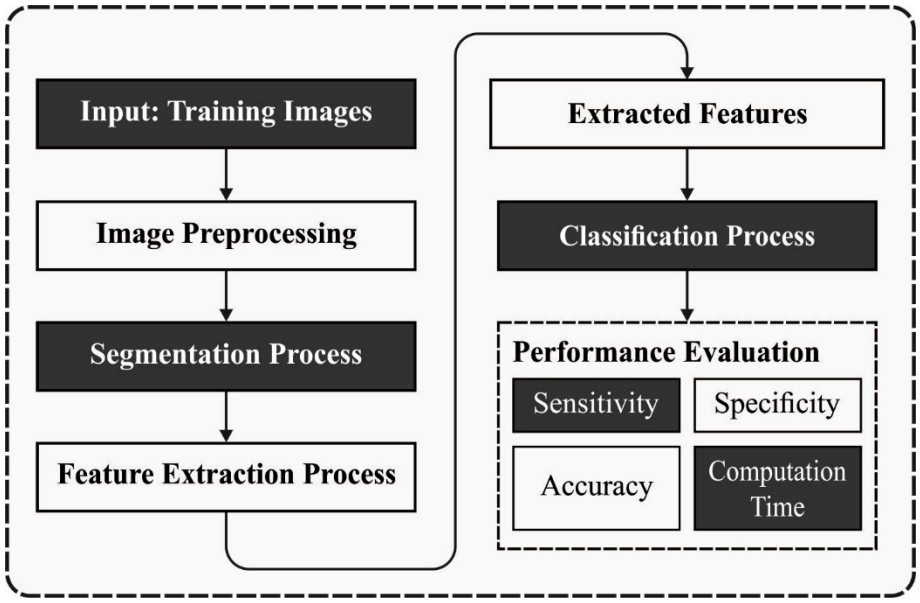

Figure 1 Process involved in the CC diagnosis

2.1 Image Segmentation

The pap smear images are segmented for determine the existence of infected areas in the image. It involves the task of dividing an image into several non-overlapping regions through optimum threshold values. A new technique called P-tile utilized this kind of data as object that is darker compared to alternative background and uses a determined fraction such as percentile (1/p) of whole image portions. The thresholds could be created through examining the intensity from the fraction of image pixel is minimal than predetermined value. Thus, intensity is recognized by:

$$c(g_k) = \sum_{i=0}^{k} g_i \qquad (1)$$

The threshold value T can be defined by $sc(T) = 1/p$.

2.2 Feature Extraction

Primarily, LBP [19, 21] is presented for feature extraction of the segmented images. Additionally, LBP is utilized for other applications such as image retrieval, palm print and face recognitions attained accomplishment because of its speed (not essential for tuning the variables) and efficiency. The LBP is determined according to the connection among center pixel and its nearby neighbors in the image. The connection that is gathered among center pixel and its neighbors is depending upon edge amongst the center and neighbor pixels. when the neighboring pixel value is greater than/equivalent to center pixel, that LBP bit is encoded by 'one'; or else it is coded as 'zero' Eqs. (2)-(3).

$$LBP_{P,R} = \sum_{i=1}^{P} 2^{(i\text{-}1)} \times f_1\big(I(g_i) - I\,(gc)\big) \qquad (2)$$

$$f_1(x) = \begin{cases} 1\ x \geq 0 \\ 0\ \ else \end{cases} \qquad (3)$$

DOI: 10.5281/zenodo.5504344

Where $\mathrm{I}(g_c)$ represents gray value of intermediate pixel, $\mathrm{I}(g_p)$ denotes gray value of its neighbors, P indicates amount of neighbors and R represents radius of neighbourhood. Afterward calculating the LBP pattern for every pixel (j, k) , the entire image is denoted by creating a histogram as presented in Eq. (3).

$$\mathrm{H}_{LBP}(l) = \sum_{j=1}^{N_1}\sum_{\mathrm{k}=1}^{N_2} f_2\left(LBP\ (j,k), l\right); l \in [0, (2^P - 1)] \quad (4)$$

$$f_2(x,y) = \begin{cases} 1\ x = y \\ 0\ else \end{cases} \quad (5)$$

Where the size of input image denotes $N_1 \times N_2$. The histogram of this pattern maintains the data to distribute the edges in the image.

2.3 Image Classification

LS-SVM is a binary classification model used for CC diagnosis. SVM [22] is real-world use of statistical learning concept in multi-dimensional functions. Here (x_i, y_i), $1 < \mathrm{i} < N$, indicates group of data comprising N trained instance. Every sample should confirm to the condition $x_i \in R^d$. y_i determines the class of equivalent instance, x_i. Hence $y_i \in \{-1, 1\}$ and d denotes amount of dimension of input data. The separate hyperplane is determined by

$$w \cdot x_i + b = 0, \qquad 1 \le \mathrm{i} \le N. \quad (6)$$

When this hyperplane occurs, then linear separation is attained. The instance which is closest one to the separate hyperplane is named as support vector. In boundary (support vectors), Eq. (6) is changed to

$$w \cdot x_i + b = \pm 1. \quad (7)$$

$$y_i \cdot (w \cdot x_i + b) \ge 1. \quad (8)$$

Thus, the problem is detecting w and b. It has different hyperplanes that could partitioned the 2-class data however SVM creates an optimal hyperplane. These hyperplanes have a highest distance to SV. The boundary of a separate hyperplane is $2/\|w\|$. Hence, they need to detect an optimum hyperplane, it must minimalize $\|w\|$. For easiness they could substitute $(1/2)\|w\|^2$ using $\|w\|$. Consequently, they are handling an optimization problem. It implies that they should minimalize $(1/2)\|w\|^2$ subject to Eq. (8).

For nonlinear problems positive slack parameters ζ_i are presented. Hence the problem reformed as

$$\mathrm{Min}\frac{1}{2}\|w\|^2 + \mathrm{C} \cdot \sum_{i=1}^{n} \zeta_i \ \ s.t\ y_i \cdot (w \cdot x_i + b) \ge 1 - \zeta_i \quad (9)$$

$$\zeta_\mathrm{i} \ge 0, 1 \le \mathrm{i} \le N.$$

In (9), C denotes penalty factor. It is presented for controlling the tradeoff among boundary maximalization and error minimalization. Such problems are resolved using Lagrange multipliers. So, the classifier decision function turns into

$$F(\mathrm{x}) = sign\left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot K\left(x_i, x_j\right) + b\right), \quad (10)$$

Whereas α_i represents Lagrange multiplier. $K\left(x_i, x_j\right) = \phi(x_i) \cdot \phi\left(x_j\right)$ indicates kernel function via few other mapping functions, $\phi(\mathrm{x})$. QP resolver is utilized for finding α_i. Afterward w and b is attained as follows

DOI: 10.5281/zenodo.5504344

$$w = \sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \phi(x_i), \qquad (11)$$

$$b = \frac{1}{N_{\mathrm{SV}}} \sum_{i} \left(y_i - \sum_{j} \alpha_j \cdot y_j \cdot K(x_j, x) \right). \qquad (12)$$

where N_{SV} indicates amount of SV and x denotes input unknown instance. Some common kernel functions are given below:

linear: $k(x, y) = x \cdot y + 1$,

polynomial: $k(x, y) = (x \cdot y + 1)^{\sigma}$,

RBF: $k(x, y) = exp\left(-\|x - y\|/(2 \cdot \sigma^2)\right)$,

quadratic: $1 - \|x - y\|^2/(\|x - y\| + \sigma)$,

In traditional SVM, the limitations are inequality that require quadratic programming solution whereas it raises the computational complexity for high sized datasets, where in LSSVM [23, 24], every inequality constraint are altered to equality constraint types. Hence, LSSVM resolves scheme of linear equation rather than quadratic program. Finally, its result provides support values and outcomes. The support values are relevant to the error rather than support vector in traditional SVM. In regression case, LSSVM is given as:

$$\min \frac{1}{2} \|\mathrm{w}\|^2 + \gamma \frac{1}{2} \sum_{i=1}^{N} \xi_i^2 \qquad (13)$$

Where ξ_i represents slack parameters and $\gamma \geq 0$ denotes regularization variable. High γ doesn't allow other slack parameters and subsequently increase the method complicity however lower γ implies a method with highly trained error. Thus, it is crucial for finding appropriate value for γ and it is one of LSSVM tuning variable that must be adapted accurately. Assuming subsequent equality constraints in Eq. (14) Lagrangian form of Eq. (13) is determined as follows:

$$y_i = w\varphi(x_i) + b + \xi_i, i = 1, \ldots, N \qquad (14)$$

$$L_{LSSVM} = \frac{1}{2} \|w\|^2 + \gamma \frac{1}{2} \sum_{i=1}^{N} \xi_i^2 - \sum_{i=1}^{N} \alpha_i (w\varphi(x_i) + b + \xi_i - y_i), \qquad (15)$$

Where $\alpha_i s(\mathrm{i} = 1, \ldots, \mathrm{N})$ denotes Lagrange multiplier. By employing optimum condition for Eq. (16), the subsequent formulation are given by:

$$\begin{cases} I) \ \dfrac{\partial L_{LSSVM}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i) \\ II) \dfrac{\partial L_{LSSVM}}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \\ III) \dfrac{\partial L_{LSSVM}}{\partial \xi_i} = 0 \rightarrow \alpha_i = \gamma \xi_i, i = 1, \ldots, N \\ IV) \ \dfrac{\partial L_{LSSVM}}{\partial \alpha_i} = 0 \rightarrow w\varphi(x_i) + b + \xi_i - y_i = 0, \quad i = 1, \ldots, N \end{cases} \qquad (16)$$

DOI: 10.5281/zenodo.5504344

Finally, the reformed optimization problem is given by:

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & k(x_1,x_1)+\frac{1}{c} & \cdots & k(x_1,x_N) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & k(x_N,x_1) & \cdots & k(x_N,x_N)+\frac{1}{c} \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_N \end{bmatrix} \quad (17)$$

Lastly, exclusive form of LSSVM predictive function is given as:

$$f(x) = \sum_{i=1}^{N} \alpha_i \, k(x, x_i) + b. \quad (18)$$

3. Performance Validation

For experimental validation, the Herlev database [25] is employed to ensure the overall classification performance. It includes 918 images and 7 class labels, comprising 3 class labels into normal and 4 class labels into abnormal. Fig. 2 shows the sample test images.

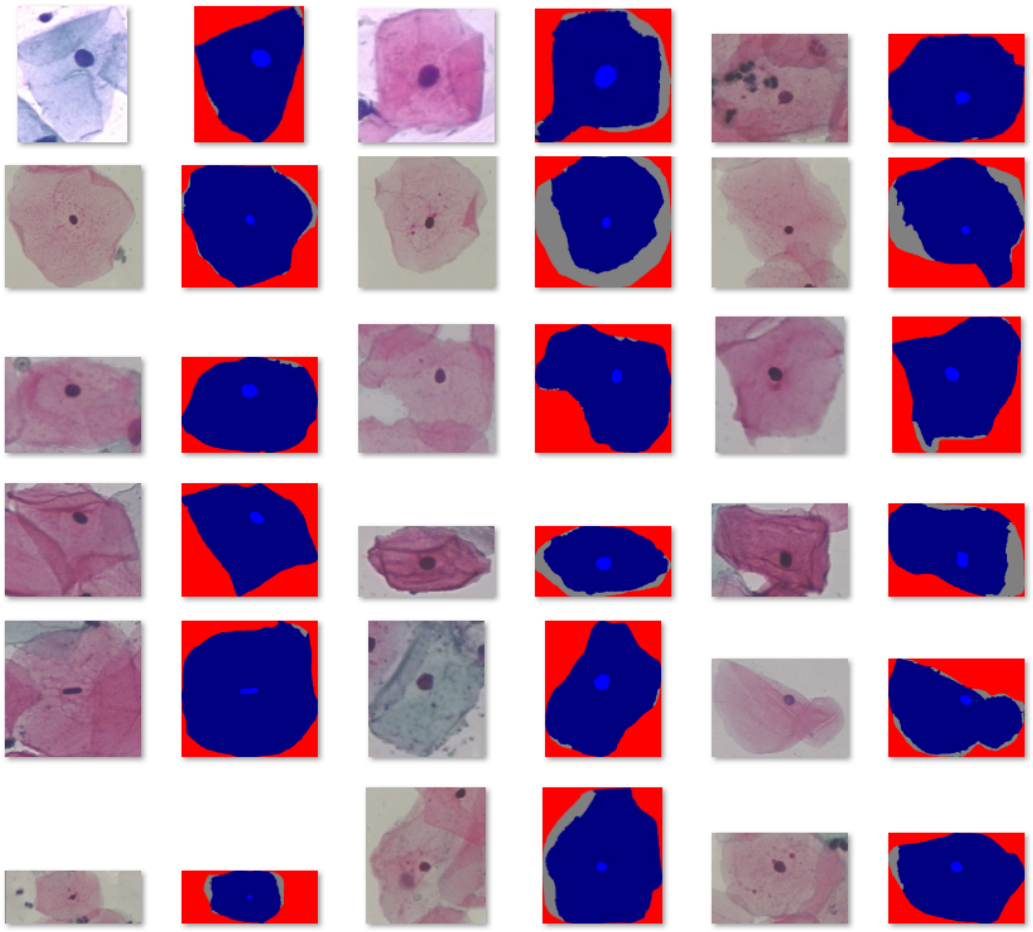

Figure 2 Sample Test Images

DOI: 10.5281/zenodo.5504344

Table 1 and Fig. 3 investigates the performance analysis of the IML-CAD technique with existing techniques. On examining the results interms of accuracy, the DenseNet-169 and Inception-Resnet-v2 techniques have attained poor results with the accuracy of 0.6979 and 0.6930 respectively. In line with, the DenseNet-121and CYENET techniques have obtained moderate accuracy of 0.7242 and 0.9230 respectively. However, the IML-CAD technique has resulted to an enhanced accuracy of 0.9356.

Table 1 Results analysis of IML-CAD technique on CC classification

Methods	Accuracy	Sensitivity	Specificity
IML-CAD	0.9356	0.9389	0.9690
DenseNet-121	0.7242	0.5986	0.7683
DenseNet-169	0.6979	0.6500	0.7148
Inception-Resnet-v2	0.6930	0.6670	0.7060
CYENET	0.9230	0.9240	0.9620

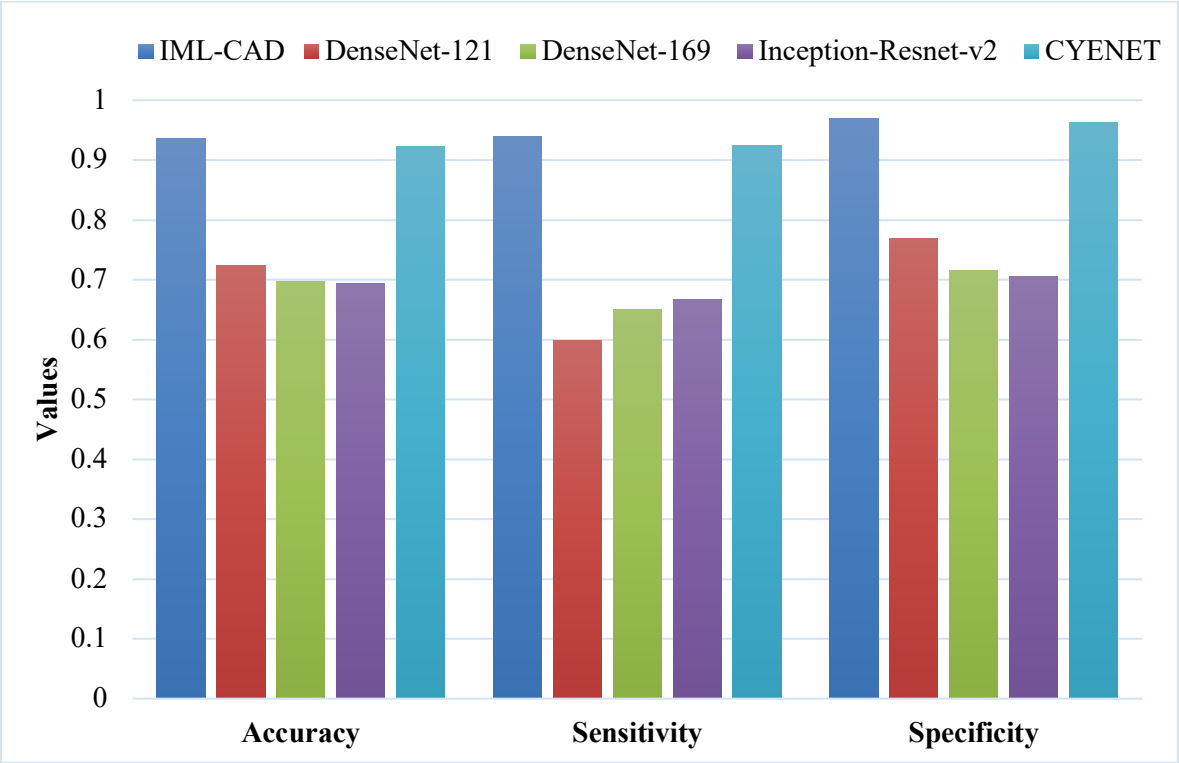

Figure 3 Optimal results on the validation set

On examining the results interms of sensitivity, the DenseNet-121 and DenseNet-169 techniques have attained poor results with the sensitivity of 0.5986 and 0.6500 respectively. In line with, the Inception-Resnet-v2 and CYENET techniques have obtained moderate sensitivity of 0.6670 and 0.9240 respectively. However, the IML-CAD technique has resulted to an enhanced sensitivity of 0.9389. Finally, on examining the results interms of specificity, the DenseNet-121 and DenseNet-169 techniques have attained poor results with the specificity of 0.7683 and 0.7148 respectively. In line with, the Inception-Resnet-v2 and CYENET techniques have obtained moderate specificity of 0.7060 and 0.9620 respectively. However, the IML-CAD technique has resulted to an enhanced specificity of 0.9690. The above-mentioned results analysis ensured the betterment of the IML-CAD technique over the other techniques.

DOI: 10.5281/zenodo.5504344

Table 2 Computational Time (CT) analysis of IML-CAD Technique

Methods	Time (s)
IML-CAD	56.89
DenseNet-121	80.46
DenseNet-169	82.19
Inception-Resnet-v2	72.91
CYENET	68.64

Table 2 and Fig. 4 depicts the CT analysis of the IML-CAD technique with existing techniques. The table values denoted that the DenseNet-121 and DenseNet-169 techniques have shown ineffectual outcome with the higher CT of 80.46s and 82.19s respectively. Moroever, the Incpetion-Resnet-v2 and CYENET techniques have accomplished moderate CT of 72.91s abd 68.64s respectively. However, the IML-CAD technique has resulted to supreme outcome with the lower CT of 56.89s. Therefore, the IML-CAD technique can be utilized as an effective tool to diagnose CC using pap smear images.

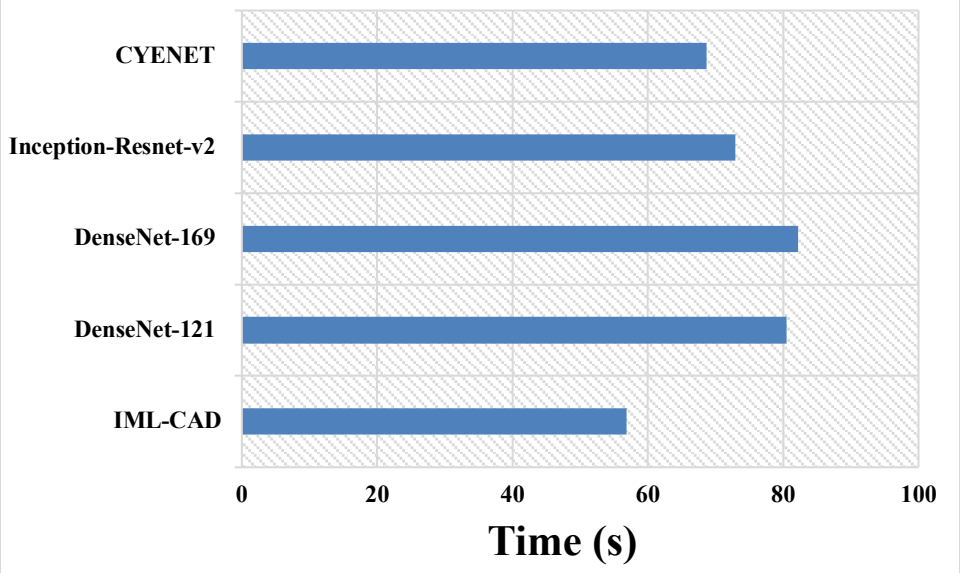

Figure 4 CT analysis of IML-CAD technique with other techniques

4. Conclusion

This paper has presented an effective IML-CAD model for CC classification. The IML-CAD technique involves different stages of operations to categorize the cancerous cervix cells. In addition, the IML-CAD technique involves histogram-based segmentation to determine the affected regions. Moreover, LBP based feature extractor and LSSVM based classifier is designed for CC classification. To showcase the better performance of the IML-CAD model, a series of simulations take place and the experimental results highlighted the superior performance of the IML-CAD technique over the other techniques. Therefore, the IML-CAD model is found to be an effective tool for CC diagnosis. As a part of future scope, the efficiency of the IML-CAD model can be extended by the design of deep learning models.

DOI: 10.5281/zenodo.5504344

References

[1] Dong, N., Zhao, L., Wu, C. H., & Chang, J. F. (2020). Inception v3 based cervical cell classification combined with artificially extracted features. *Applied Soft Computing, 93,* 106311.

[2] Yusufaly, T. I., Kallis, K., Simon, A., Mayadev, J., Yashar, C. M., Einck, J. P.,... & Meyers, S. M. (2020). A knowledge-based organ dose prediction tool for brachytherapy treatment planning of patients with cervical cancer. *Brachytherapy, 19* (5), 624-634.

[3] Shao, J., Zhang, Z., Liu, H., Song, Y., Yan, Z., Wang, X., & Hou, Z. (2020). DCE-MRI pharmacokinetic parameter maps for cervical carcinoma prediction. *Computers in biology and medicine, 118,* 103634.

[4] Zhang, T., Luo, Y. M., Li, P., Liu, P. Z., Du, Y. Z., Sun, P.,... & Xue, H. (2020). Cervical precancerous lesions classification using pre-trained densely connected convolutional networks with colposcopy images. *Biomedical Signal Processing and Control, 55,* 101566.

[5] Hua, W., Xiao, T., Jiang, X., Liu, Z., Wang, M., Zheng, H., & Wang, S. (2020). Lymph-vascular space invasion prediction in cervical cancer: exploring radiomics and deep learning multilevel features of tumor and peritumor tissue on multiparametric MRI. *Biomedical Signal Processing and Control, 58,* 101869.

[6] Lu, J., Song, E., Ghoneim, A., & Alrashoud, M. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Generation Computer Systems, 106,* 199-205.

[7] Ghoneim, A., Muhammad, G., & Hossain, M. S. (2020). Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems, 102,* 643-649.

[8] Nayak, M., Das, S., Bhanja, U., & Senapati, M. R. (2020). Elephant herding optimization technique based neural network for cancer prediction. *Informatics in Medicine Unlocked, 21,* 100445.

[9] Kim, S. I., Lee, S., Choi, C. H., Lee, M., Kim, J. W., & Kim, Y. B. (2020). Prediction of disease recurrence according to surgical approach of primary radical hysterectomy in patients with early-stage cervical cancer using machine learning methods. *Gynecologic Oncology, 159*, 185-186.

[10] Agus Pratondo, Chee-Kong Chui, Sim-Heng Ong (2017). Integrating machine learning with region-based active contour models in medical image segmentation, *Journal of Visual Communication and Image Representation*, 43, 1-9, 2017

[11] Deepa, B., & Sumithra, M. G. (2019). An intensity factorized thresholding based segmentation technique with gradient discrete wavelet fusion for diagnosing stroke and tumor in brain MRI. *Multidimensional Systems and Signal Processing, 30* (4), 2081-2112.

[12] William, W., Ware, A., Basaza-Ejiri, A. H., & Obungoloch, J. (2018). A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Computer methods and programs in biomedicine, 164,* 15-22.

[13] Zhang, C., Leng, W., Sun, C., Lu, T., Chen, Z., Men, X.,... & Qin, J. (2018). Urine proteome profiling predicts lung cancer from control cases and other tumors. *EBioMedicine, 30,* 120-128.

[14] Matsuo, K., Purushotham, S., Moeini, A., Li, G., Machida, H., Liu, Y., & Roman, L. D. (2017). A pilot study in using deep learning to predict limited life expectancy in women with recurrent cervical cancer. *American journal of obstetrics and gynecology, 217* (6), 703.

[15] Iliyasu, A. M., & Fatichah, C. (2017). A quantum hybrid PSO combined with fuzzy k-NN approach to feature selection and cell classification in cervical cancer detection. *Sensors, 17* (12), 2935.

[16] Lu, J., Song, E., Ghoneim, A., & Alrashoud, M. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Generation Computer Systems, 106,* 199-205.

[17] Ijaz, M. F., Attique, M., & Son, Y. (2020). Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors, 20* (10), 2809.

[18] Khamparia, A., Gupta, D., de Albuquerque, V. H. C., Sangaiah, A. K., & Jhaveri, R. H. (2020). Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. *The Journal of Supercomputing,* 1-19.

[19] Zhang, C. W., Jia, D. Y., Wu, N. K., Guo, Z. G., & Ge, H. R. (2021). Quantitative detection of cervical cancer based on time series information from smear images. Applied Soft Computing, 107791.

[20] Rehman, A. U., Ali, N., Taj, I., Sajid, M., & Karimov, K. S. (2020). An Automatic Mass Screening System for Cervical Cancer Detection Based on Convolutional Neural Network. *Mathematical Problems in Engineering, 2020.*

[21] Khan, K. A., Shanir, P. P., Khan, Y. U., & Farooq, O. (2020). A hybrid Local Binary Pattern and wavelets based approach for EEG classification for diagnosing epilepsy. *Expert Systems with Applications, 140,* 112895.

[22] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics, 15* (1), 41-51.

DOI: 10.5281/zenodo.5504344

[23] Tian, Z. (2020). Short-term wind speed prediction based on LMD and improved FA optimized combined kernel function LSSVM. *Engineering Applications of Artificial Intelligence, 91,* 103573.
[24] Razavi, R., Bemani, A., Baghban, A., Mohammadi, A. H., & Habibzadeh, S. (2019). An insight into the estimation of fatty acid methyl ester based biodiesel properties using a LSSVM model. *Fuel, 243,* 133-141.
[25] DTU/Herlev Pap Smear Database. (2008). http://mde-lab.aegean.gr/index.php/downloads

DOI: 10.5281/zenodo.5504344