Intelligence-Based Medicine 5 (2021) 100031

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

Classification of Cervical Intraepithelial Neoplasia (CIN) using fine-tuned Convolutional Neural Networks

Oluwatomisin E. Aina [a,*], Steve A. Adeshina [a], Adeyinka P. Adedigba [b], Abiodun M. Aibinu [b]

[a] Department of Electrical & Electronics Engineering, Nile University of Nigeria, Abuja, Nigeria
[b] Department of Mechatronics Engineering, Federal University of Technology, Minna, Nigeria

ARTICLE INFO

Keywords:
Cervical cancer
Computer-aided diagnosis
Cervical intraepithelial
Neoplasia (CIN)
Convolutional Neural Network (CNN)
Discriminative
Fine-tuning
Transfer learning
Visual inspection with acetic acid (VIA)

ABSTRACT

Convolutional Neural Network (CNN) is considered one of the most successful deep learning techniques used in classification or diagnosis of medical images. However, CNN requires a high computational resource and time; and a large dataset which most medical images (cervix) do not possess. In order to compensate for these shortcomings, we propose an optimized fine-tuned CNN model to classify cervix images into Cervical Intraepithelial Neoplasia grades (CIN 1,2,3) normal and cancerous cervix images. This classification ensures that patients are diagnosed correctly, and appropriate treatments are administered. Deep learning techniques such as Data Augmentation, 1 cycle policy for optimal learning rates selection, Discriminative Fine-Tuning, Mixed Precision Training were used to optimize the fine-tuned DenseNet CNN model. The model achieved 96.3% accuracy, the specificity of 98.86%, and sensitivity of 94.97% on the datasets.

1. Introduction

In recent times, CNNs have been used in detecting various diseases accurately due to its inherent ability to automatically learn useful representations and features of images [1]. A notable success is seen in Refs. [2,3], where CNN classified skin and breast cancer at an expert-level. Hence, it has been explored by various researchers in detecting cervical cancer, a developmental disease caused by the Human Papilloma Virus (HPV). It was shown in literature that scaling up cervical cancer screening tests can eliminate the disease within 30 years [4]. Human Papilloma Virus test (HPV), Cytology (Pap Smear), Visual Inspection with Acetic Acid (VIA), Colposcopy and Cervigram are cervical cancer screening methods [5]. These methods are heavily subjected to the level of expertise and experience of the health practitioner carrying out the test. Furthermore, some of these methods are capital intensive and requires sophisticated infrastructure and equipment. VIA is a cheaper alternative to the other tests as it involves staining the cervix with Acetic Acid. It involves inspecting the cervix with the naked eyes for the presence of abnormalities (Aceto-white lesion) on the cervix. The aceto-white lesion is shown in Fig. 1. Cervicograph and colposcopy is an upgrade to VIA as images of the cervix are captured with a camera (Cervicograph) or viewed at higher magnification (Colposcopy). Images observed from the colposcopy could also be captured.

These captured images are useful in developing screening systems used to diagnose lesions/cancer automatically. The World Health Organization (WHO) defined the abnormal growth observed during Visual Inspection with Acetic acid (VIA) or any screening tests as the Cervical Intraepithelial Neoplasia (CIN) [6,7]. The CIN classes or grades define the severity of the abnormalities with CIN 1,2,3 representing mild, moderate and severe, respectively.

In literature, Computer-Aided Detection (CAD) systems based on CNN have been proposed in classifying the images of the cervix based on the CIN grades. Medical (cervix) datasets are small in size, have imbalance classes and have high dimensionality [8]. Therefore, transfer learning is usually used to train the models to classify images. Transfer learning is a technique that involves leveraging "knowledge" used to train a model with millions of images on another model with fewer images (medical images). It is also useful in speeding up the training process since model "B" reuses parameters obtained during the training of model "A" to make predictions or classifications.

It has been observed from literature that most classifications are binary, where the CIN grades are categorized into Negative (Normal/CIN 1) and Positive CIN (CIN 2+) [9]. However, having specific information on the CIN grade is essential to the health practitioner in administering a

* Corresponding author.
E-mail addresses: oluwatomisin.aina@nileuniversity.edu.ng (O.E. Aina), steve.adeshina@nileuniversity.edu.ng (S.A. Adeshina), adeyinka.adedigba@futminna.edu.ng (A.P. Adedigba), abiodun.aibinu@futminna.edu.ng (A.M. Aibinu).

https://doi.org/10.1016/j.ibmed.2021.100031
Received 29 September 2020; Received in revised form 24 March 2021; Accepted 8 April 2021

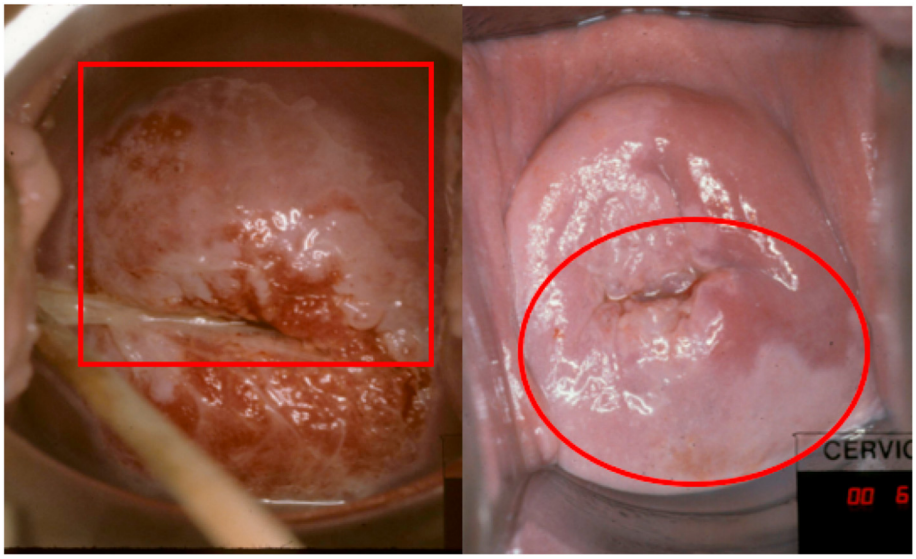

Fig. 1. Image showing aceto-white lesion mapping on the cervix.

follow-up test and treatment. In addition, authors in Ref. [10] suggested that in young women, CIN 2 could be potentially treated differently than CIN 3. It was suggested that less invasive treatments like cervix excision could be avoided if a woman is diagnosed with CIN 2. Hence, it is crucial to develop a system that can detect CIN grades separately.

It was also observed that proposed methods in literature usually require many epochs (e.g., 3000 in Ref. [11]) before the model converges. Too many epochs used in training usually results in more computation requirements (resources and time). It stems from the fact that these proposed methods do not have a defined method of selecting the hyperparameters used in training the model. In Ref. [12], the authors stated that they arrived at the appropriate hyperparameters "After many training and parameter adjustments …". The hyperparameter is a variable of the training model that must be pre-defined before the deployment of the training model [13]. These hyperparameters such as Learning Rate (LR), momentum, weight decay, batch size contributes to the overall performance and computation time of the CNN model.

In this work, we propose a fine-tuned CNN model (DenseNet) that to the best of our knowledge, has one of the best performances reported in literature. The model:

1. Classifies the CIN grades into Normal, CIN 1,2,3, and Cancer and obtained an Accuracy of 96.3%, 94.97% Sensitivity and 98.86% Specificity.
2. Optimizes and schedules the hyperparameters (learning rate) using 1 cycle policy and discriminative learning.
3. Reduce memory requirements while maintaining the state-of-art accuracy in the classification of cervix lesions by implementing a Mixed Precision Training that trains the CNN model using a single-precision format to speed up the computations and reduce the memory requirements
4. Visualizes the areas of the cervix the algorithm analyses to make predictions.

2. Literature review

We reviewed research work reported by authors to classify cervical cancer lesions using the Convolution Neural Network. This review focused on images obtained via the cervigram or colposcopy. Emphasis is made on CNN architecture adopted, hyperparameter selection, run-time, and evaluation metrics used. Also, we reviewed past methods in selecting an optimum learning rate.

2.1. Convolution Neural Network and detection of cervical cancer lesions

Cervical Cancer is a developmental disease that starts as an abnormal lesion in the transformation zone of the cervix. It could take between 10 and 20 years for lesions to progress to cancer [14,15]. However, due to weakened immunity and other factors, cancer progression could happen even between 1 and 2 years. The abnormal lesions known as Cervical Intraepithelial Neoplasia (CIN) is categorized into CIN 1,2 and 3 based on the level of dysplasia with CIN 1 being mild and CIN 2 and 3 moderate and severe respectively. These grades can either regress, progress or persist after numerous tests. Regression is the ability of the CIN grade to return to its less developed stage (Cervix with CIN 1 becoming Normal). Persistence means that the grade remains unchanged while progression occurs if the CIN advances to a more advanced stage. It is therefore essential for health practitioners to know the CIN grades of the patients to make informed decisions on the appropriate treatments and follow-up treatment plans. One of the significant challenges for health practitioners is determining the location of the transformation zone. The transformation zone is a vital part of the cervix as it is the location where cervical cancer originates. The location of the transformation zone differs in women based on age and other factors. Health practitioners need to differentiate these classes to provide the best treatment based on transformation zone types. Researchers in Refs. [5,16,17] have proven that CNN can be used in classifying the cervix into three types based on the location of the transformation zone.

Authors in Ref. [12] used a fine-tuned DenseNet 121/169 model to diagnose the presence of a moderate or high lesion in a colposcopy image (binary classification). 1709 colposcopy images were obtained from a local hospital and experts annotated the Region of Interest of the Images (ROI). The number of images was small; hence, data augmentation techniques such as random blur, cropping was used. The authors stated that many adjustments were made to obtain an optimized selection of the hyperparameters. The learning rate was set to $1e^{-5}$. The DenseNet121 model has the best accuracy of 73.08%, specificity of 78.55% and sensitivity of 57.6%.

Authors in Ref. [9], showed that CNN could be used to learn features from highly non-linear correlation across different modalities. The modalities include a low input cervigram image and non-image inputs (pap smear results, HPV test, cervigram result, PH value, and patient age). Also, a joint fully connected layer to model all modalities to classify the cervix lesions into positive CIN (CIN 2+) and negative CIN (Normal/CIN 1). Sensitivity of 87.83%, specificity of 90% and accuracy of 88.9%, were

achieved which outperformed most algorithms that use multi-modal inputs to classify cervical dysplasia. Another similar multi-feature model is illustrated in Ref. [18] which obtained an accuracy of 77.39%, 80.87% of sensitivity and 73.91% of specificity. Multi-modal inputs provide complementary information; however, other input information may not be accessible in low-resource region. For example, carrying out an HPV or Pap smear is expensive for a woman in such regions. Also, using multi-modal inputs assumes equal priority for all the features (image and non-image), which is not realistic in real-life scenarios.

The authors in Ref. [19] took an approach to diagnose cervical cancer by considering the preacetic and post acetic cervix images. It was proposed that this approach could utilize multimodal features and reduce requirements for data acquisition. These images were registered using the cross-correlation and projection transformation. The cervix region was extracted using the K-means clustering algorithm and trained using CNN (VGG 16, ResNet50 and DenseNet121). In addition, the learning rate used in training the model was 0.001, and there were no details on how this was optimized and scheduled to improve the model's performance. Classification accuracy of 86.3%, a sensitivity of 84.1%, and specificity of 89.8% were obtained.

To prevent overfitting and compensate for the small dataset sets used in the automatic screening of cervical cancer, the authors in Ref. [11] used data augmentation techniques to increase the number of datasets. The authors obtained datasets from the National Cancer Institute to determine if a cervix image is normal/CIN 1 or CIN2+. A ColpoNet CNN algorithm was designed based on the concept of concatenation of layers. An accuracy of 81.35% was obtained when compared with other CNN architectures. The authors stated that the number of epochs was increased to 3000 to obtain an accuracy of 83.95%. The authors did not disclose any information on selecting the learning rate specificity and sensitivity of the model, which are key metrics in medical diagnosis. Authors in Ref. [20] achieved an accuracy of 83.33% and 91.66% for the binary and multiclass CIN classification using an ensemble of MobileNetV2 networks. However, ensemble networks are challenging to implement in real-life cases as such networks are usually very complex.

It can be deduced from the reviewed papers that little information was given on selecting the hyperparameters. According to Ref. [13], the hyperparameter is a variable that must be pre-defined before the deployment of the training model. These hyperparameters such as Learning Rate (LR), momentum, weight decay, batch size contributes to the overall performance and computation time of the CNN model. However, the learning rate is an essential hyperparameter. It has the highest impact on the training process as well and the overall performance of the model [21].

2.2. Selecting the optimum learning rate

Conventional selection of learning rate involves taking a guess (grid search) by manually selecting a LR. It also involves setting a schedule that decreases the value (decay) as the model approaches convergence (Learning Rate Annealing) [22]. The conventional method was time-consuming and inaccurate because the LR did not adjust to the model's parameters. As a result of this, the introduction of learning rate optimization algorithms that can adjust the learning rate to the parameters of the model automatically was considered. Examples of these algorithms are Adam, RMSprop, Adagrad, and AdaDelta [23,24]. Unfortunately, adaptive learning rates are computationally expensive and do not always result in global minima due to saddle points. Saddle points give a false representation of the local minima, which slows down the learning process [25].

The Cyclical Learning Rate (CLR) presented in Ref. [21] solved the saddle point problem by rapidly increasing the LR for some iterations during a training epoch and gradually reducing it for the rest of the iteration in the epoch. The rapid increment overcomes plateaus, while gradual decrements ensure that the algorithm converges to the global optimum. CLR used in selecting the optimum learning rate reduced the number of iterations by 65% in classifying the CIFAR-10 dataset and achieved the same accuracy with conventional methods of selecting LR.

The super-convergence (1 cycle policy) was introduced in Refs. [26,27] to reduce the model's training time by using a large LR. The difference between super-convergence and CLR is that it has one cycle as opposed to CLR with multiple cycles. Super-convergence also improves the performance of the training model when the data is small.

In this work, we decided to use 1 cycle policy as suggested in Ref. [26]. We discovered that when used in other domains, it achieved an excellent performance. Such domains include: classification of skin lesions [27], prediction of drug functions from chemical structure [28] and automatic detection of plant disease [29]. In all these, an excellent performance was achieved. In addition to selecting the optimum learning rate, other hyperparameters such as batch size, momentum, weight decay are tuned to improve the performance of the model. The DenseNet CNN model was used in this work because of its excellent performance when evaluated on top object recognition tasks (CIFAR-10, ImageNet, CIFAR-100). In addition, DenseNet reduces the vanishing gradient problem, improves feature propagation and promotes feature reuse [28].

3. Preliminary studies

The goal of this section is to provide background information on the methodology adopted in this work. The areas considered are Supervised Learning and Hyperparameters of CNNs.

3.1. Supervised Learning for classification of CNNs

The goal of this section is to provide background information on the methodology adopted in this work. The areas considered are Supervised Learning and Hyperparameters of CNNs.

3.2. Supervised Learning for classification of CNNs

In deep learning, a given training dataset is made up of input x and corresponding values denoted by y. The algorithm aims to learn parameters (weight w and bias b) to predict the value of y for an input x. The predicted value is labeled as $\hat{y}$, and the algorithm's task is to ensure that the difference between $\hat{y}$ and y (loss) is small. The goal of this learning algorithm is to minimize the total cost incurred during training. It is represented mathematically as:

$$J(w,b) = \frac{1}{2n}\sum L(y,\hat{y}) \tag{1}$$

From the equation, n is the total number of training samples, L is the loss function for each of the data-point. Also, $J(w,b)$ is the cost function that measures how well the algorithm predicts the output.

Equation (1) shows the cost function as the average loss function over the training samples n. It is important to note that methods such as the Mean Square Error (MSE), Mean Absolute Errors (MAE), Cross-Entropy Loss etc. are loss functions. They are used to calculate the loss based on the type of classification or regression problem. Thus, the deep learning problem becomes an optimization problem as it learns parameters that globally minimize the cost $J(w,b)$.

In practice, Gradient Descent (GD), a popular optimization algorithm used in obtaining a point (value) that minimizes $J(w,b)$. It is obtained by calculating the partial derivative of the function with respect to the weights (backpropagation). The obtained function is used to update the weights. The updated weight directs or guides the algorithm to converge at the minimum point. Mathematically,

$$w_{j+1} : w_j - \alpha\frac{\partial}{\partial w_j}J(w,b) \tag{2}$$

From equation (2), $\frac{\partial}{\partial w_j}J(w,b)$ is the partial derivative of the cost function obtained via the backpropagation that computes the gradients

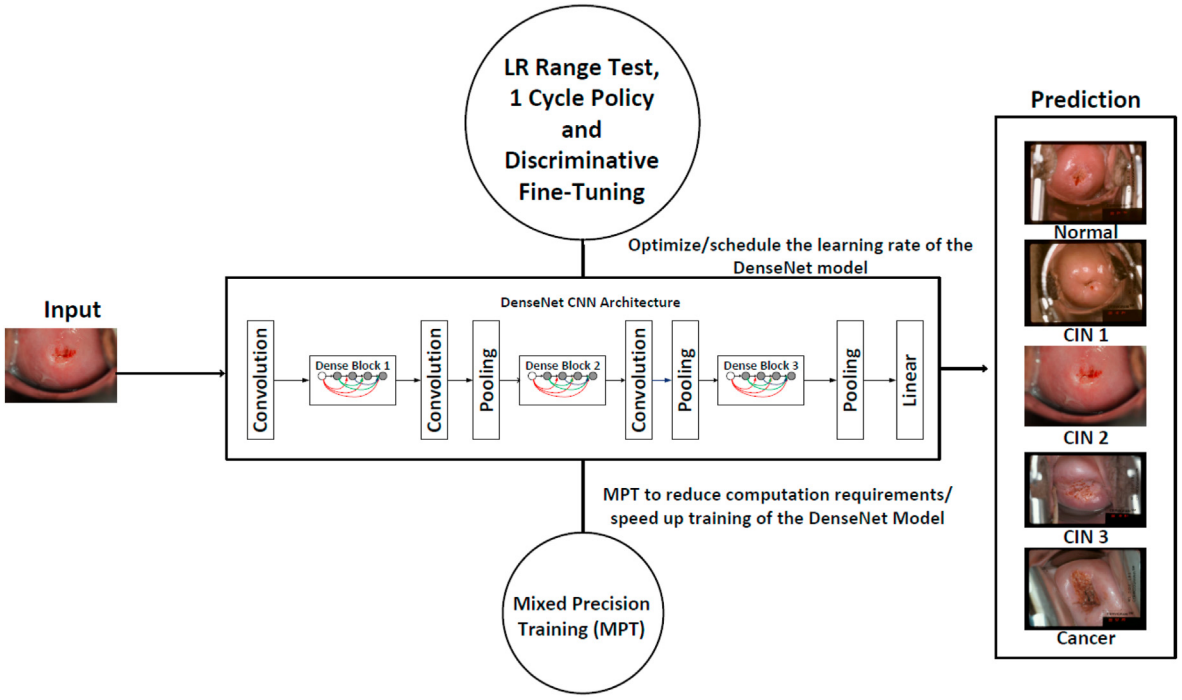

Fig. 2. Proposed CNN architecture (Densenet) Methodology for the classification of CIN.

used to update the weights of the model, α is the learning rate. The learning rate defines steps required for the algorithm to converge to its minimum value. If the LR is too small, the algorithm converges after training with many epochs. In contrast, with a large LR, the algorithm may not converge to the minimum value.

It is important to note that the Stochastic Gradient Descent (SGD), as shown in (1), is a simplified variation where the gradient is estimated based on a random sampling of training examples [29]. SGD is simple to implement on datasets with many training examples, and it has a better generalization performance. However, its learning rate needs to be tuned manually. The learning rate is an essential hyperparameter as it has the highest impact on the training process and the overall performance of the model [21].

3.3. Hyperparameters of CNNs

Hyperparameters play an essential role in the model's performance and training time. It is usually pre-set before training. It is important to note that hyperparameters and model's parameters are inherently different. The model's parameters are learnt from the data to make predictions [30] while hyperparameters such as learning rate, weight decay, batch size determines the quality of the training process and how fast a model will converge. Furthermore, hyperparameters are independent of the training dataset as they are pre-set. The generalization and performance of the model is dependent on the model's parameters obtained during training.

Table 1
Distribution of the dataset from the different studies.

	Normal	CIN 1	CIN 2	CIN 3	Cancer	Total number of images
ALTS	304	116	70	93	4	587
Biopsy	5	14	15	15	2	51
CVT	20	10	19	35	5	148
NHS	2	44	40	88	8	182
TOTAL NUMBER/ CLASS	331	184	144	231	19	909

There is a relationship between the batch size and the "*n*" in equation (1); the term is the number of training samples. The number of training samples is usually divided into section or batches, a best practice in training. Weight updates of the model are done after each batch. It is therefore paramount to select the optimum batch size that speeds up training without trading off the generability of the model.

In order to accelerate the training, momentum can be tuned in the model. Also, momentum can reduce steepness that is common with SGD, making the model converge at its global minima faster [24]. Weight decay is a hyperparameter and common regularization technique. Regularization is an important technique that adds a penalty term (weight decay) to reduce the complexity of a model during training [31]. A model is said to be complex if it tries to learn all the different patterns (variations) in the dataset, thereby overfitting. Such a model makes poor predictions on unseen data (test data) reducing the generability of the model. Therefore in this work, an optimum hyperparameter selection technique is proposed that guarantees quick convergence of the objective function and better generalization of learned parameters to unseen dataset.

4. Methodology

4.1. Adopted methodology

In this paper, we adopted different techniques in classifying the cervix image based on the CIN. The overview of the methodology is summarized in Fig. 2. We augmented the cervix to increase the number and variations in the dataset, which serves as an input to the pre-trained CNN architecture. The pre-trained CNN architecture was fine-tuned to suit the cervix classification task. The LR was selected by using the LR range test as described in the previous section. For faster convergence, we introduce Discriminative finetuning, which takes advantage of 1 cycle policy to select the optimum learning rate for each layer of the model. Mixed Precision Training is implemented to reduce memory requirements and speed up training. The intuitions behind these techniques are discussed in this section.

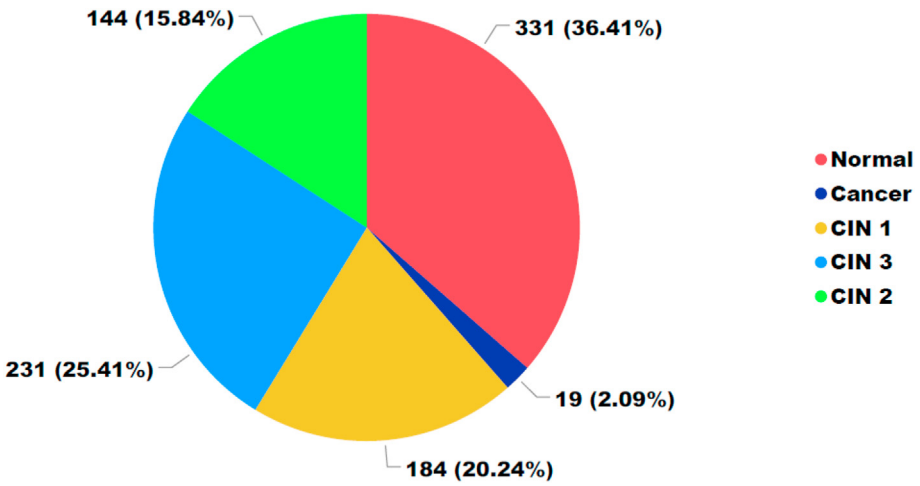

Fig. 3. Dataset distribution based on the CIN grade.

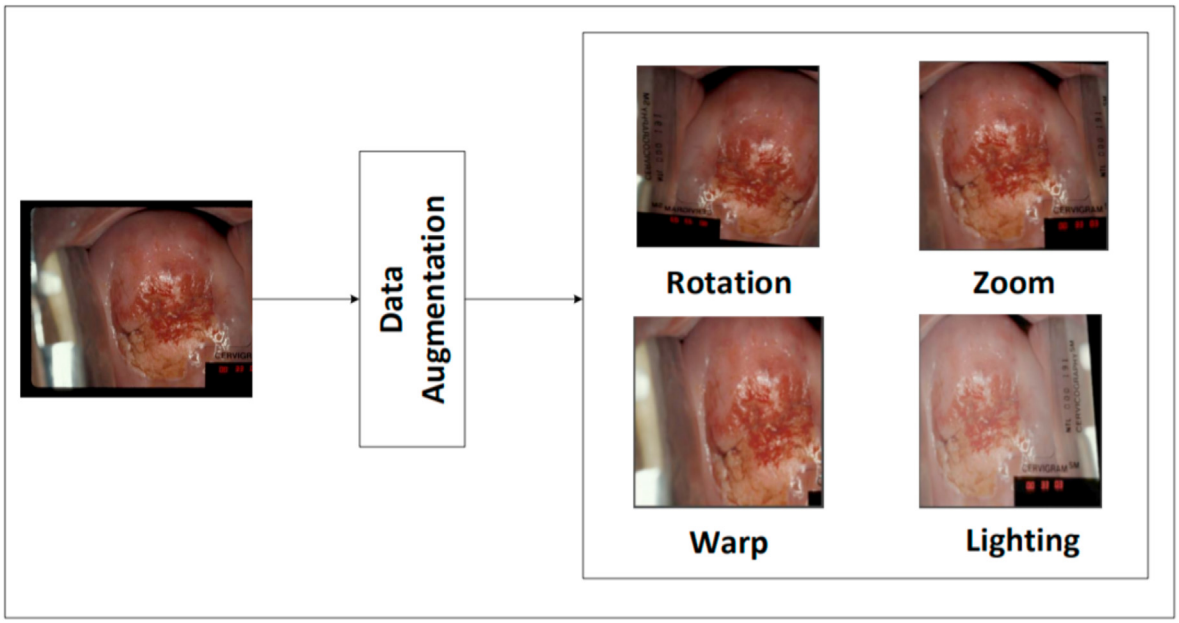

Fig. 4. Data Augmentation of sample cervix images.

4.2. Dataset description

The datasets used were obtained from the National Cancer Institute, Bethesda, USA. The dataset contains four different studies: Costa Rica Natural History (NHS), ASCUS-LSIL Triage Study (ALTS), Biopsy, and Costa Rica Vaccine Trial (CVT). We were granted access to the pilot dataset, which has 2120 images. These images were classified based on the presence of precancerous lesions after histology diagnosis. It is important to note that images without the histology results were removed from the training dataset. Thus, the total number of images obtained from NCI used in this work 909.

The distribution of the dataset from the different studies are summarized in Table 1.

As a result of this, it is observed that the classes were unbalanced. Class imbalance can result in misclassification of the training algorithm. The model will be biased towards the class with the majority data samples ("Normal" as seen in Fig. 3) and has an impact on the effectiveness of CNN.

As a result of this, the CNN overfits and does not generalize well on unseen data.

4.2.1. Dataset splitting and augmentation

In order to evaluate the model, we split the dataset using the Leave-P-out cross-validation technique. The "P" refers to the number of points (images) removed from the dataset; this validation set was separated before training the model. It is observed in Table 1 that the number of images in each class was imbalanced. Therefore, we implemented the data augmentation techniques to achieve a balanced dataset before splitting the data into training and validation set.

Data augmentation technique alters the original image's pixel to form new images and increase variations of the dataset. It synthesizes new images from the original image by flipping, rotation, Gaussian blurring, warping, lighting translation etc., to prevent bias and overfitting the CNN model (see Fig. 4). The model's performance and generalization were increased by training and validating the model on these additional data points, which could be similar to real-life test images [32–34].

It is important to note that generalization of a model is that ability of a make accurate predictions on unseen data points. Further details on the data augmentation techniques used are summarized in Table 2.

4.3. DenseNet CNN architecture

DenseNet is made up of a composite layer for batch normalization, rectified linear units, pooling and convolution operations. It has several Densenet blocks with transitional layers between them [28] as shown in Fig. 2. It has different variations (DenseNet-121,169,201 and 264) based on the number of layers in the network.

One of DenseNet architecture's many advantages is that it reduces the vanishing gradient problem, common to gradient-based learning algorithms like CNN. In DenseNet, the input to a layer l is the concatenation of the feature maps of the preceding layer $x_0, x_1; \ldots; x_{l-1}$. Hence, each layer is fed from all preceding features in the dense block [35]. Therefore, the output to the layer is:

$$x_l = H_l([x_0, x_1; \ldots; x_{l-1}]) \tag{3}$$

However, a drawback to this model is the computational requirements required since it has numerous hidden layers. Mixed Precision Training was therefore used to speed up the training time and computational requirements of this model. In addition, the hyperparameters (learning rate) of the network were optimized by selecting an optimal learning rate for each layer of the densenet model.

4.3.1. Hyperparameter selection and optimization

In transfer learning, parameters of the pre-trained network are transferred to the new task. However, it is crucial to tune the hyperparameters to suit the new task. The learning rate is considered the most vital hyperparameter to tune when compared to the batch size, momentum, weight decay [31]. In the literature review section of this work, we reviewed various methods of selecting and optimizing the learning rate. We adopted the LR range test and 1 cycle policy because it reduced the computation time without hindering the performance of the model.

LR range test was useful in selecting an optimal rate (LR), the most critical hyperparameter during training. It adjusts the weight of the model with respect to the loss function (see Algorithm 1 for details). It was adopted because it required only a single one-epoch trial experiment to determine the optimal LR, reducing the computational cost. Also, it reduced the number of iterations required for a CNN model to reach its optimal performance.

In CLR, an initial LR α_{min} pre-set increases linearly by a predetermined step size to α_{max} (upper limit) in the first cycle. It later drops from α_{max} to α_{min} in the second cycle. As a result of this, it can also be referred to as the triangular learning rate policy. This cycle continues throughout the training of the model. However, in super-convergence, the large LR ($\alpha_{max\ large}$) is obtained from the LR range test while α_{min} is one-tenth of the $\alpha_{max\ large}$. In the first cycle, the α_{min} increases to $\alpha_{max\ large}$ and further decreases to a LR lower than α_{min} in the second cycle. Other hyperparameters such as batch size, momentum and weight decay must be tuned for optimum performance of the model. It is based on pre-setting a boundary (maximum and minimum LR) in which the LR changes at every layer in response to the loss function.

In the initial training, we set the LR to $1e^{-3}$ (default pre-trained LR) with a weight decay of $1e^{-4}$. We observed from the graph that the loss at that learning rate was considerably low. We further unfreeze other layers and trained the models for a few more epochs to observe the relationship between the LR and the loss of the model. A graph showing this relationship is observed in Fig. 5. The part of the graph with the steepest LR is selected. Hence, the maximum and minimum boundary to be selected is $1e^{-3}$ and.

Algorithm 1 Optimal Hyperparameter Selection PseudoCode

procedure LR_{test}
  Input: (α_{min} : minimum learning rate,
    α_{max} : maximum learning rate,
    $\mathcal{D}$: size of dataset, bs: batch size)
  Output: $(\alpha_{min}, \alpha_{max})$

  $j \leftarrow \frac{\mathcal{D}}{bs}$
  $pct \leftarrow$ random number between 0.5 and 1
  // pct determines how rapidly the learning rate increases or reduces

while $j \leq pct \times j$ **do**:
  $\alpha_j \leftarrow \alpha_{min} + \frac{\alpha_{max} - \alpha_{min}}{j}$ *// Increase the learning rate*
  $w_j \leftarrow w_{j-1} - \alpha_j \frac{\partial J(w_{j-1}, b)}{\partial w_{j-1}}$ *// perform parameter update*
  $J(w_j, b) \leftarrow \mathcal{L}(F(x_i|w_j, b))$ *// estimate the local loss* $j \leftarrow j + 1$
end **while**

while $j > pct \times j$ **do**:
  $\alpha_j \leftarrow \alpha_{max} - \frac{\alpha_{max} + \alpha_{min}}{j}$ *// Increase the learning rate*
  $w_j \leftarrow w_{j-1} - \alpha_j \frac{\partial J(w_{j-1}, b)}{\partial w_{j-1}}$ *// perform parameter update*
  $J(w_j, b) \leftarrow \mathcal{L}(F(x_i|w_j, b))$ *// estimate the local loss* $j \leftarrow j + 1$
end **while**

  plot α_j vs $J(w_j, b)$
  $(\alpha_{min}, \alpha_{max}) \leftarrow$ steepest slop of the graph where loss reduces the most.

Table 2
Augmentation parameters.

Data Augmentation	Parameters
Rotation	Rotation angle: 30 & 60
Gaussian blurring	Kernel Size: 3
Zoom	Scale: 1.3
Lightning	Intensity: 2
Warp	Magnitude: 0.4

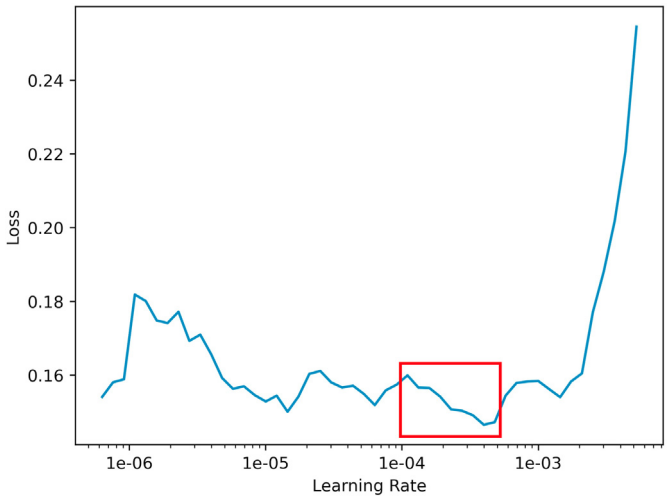

Fig. 5. Graph showing the optimal LR for the model.

In addition, in order to complement the optimum learning rate selected, we selected a cyclic momentum as suggested by the author in Ref. [27]. In cyclic momentum, the momentum (m) reduces from the m_{max} to m_{min} in the first cycle of 1 cycle policy and increases from m_{min} to m_{max} in the second half cycle. The m_{max} is kept constant when the LR decreases further lower than the α_{min}. We selected the boundaries of the momentum between $0.85-0.95$ as suggested.

4.4. Discriminative fine-tuning

Fine-tuning of a pre-trained CNN network involves tuning or adjusting the pre-trained network to suit the required task. In this paper, the required task is to classify the cervix lesions into the five categories discussed previously.

The knowledge was transferred from CNN model pre-trained on ImageNet, a dataset with over 15 million images classified into 22,000 images [36]. In fine-tuning, it is important to remove the classification (SoftMax) layer with the 1000 and to replace it with the number of classes in the required task (5 classes in for this paper).

CNN is made up of layers that learn different features of an image. Low-level features such as edges, lines and corners are learnt in earlier layers. In comparison, later layers learn more complex features specific to the required task (cervix lesions). Features learnt in earlier layers of deep CNN models are common to most images; therefore, those layers and the weights are frozen. The weights of these layers from the pre-trained networks are used directly in the new task. However, the last layer (fully connected) is trained to ensure that the weights of the network are specific to the task. If CNN's performance is not satisfactory, further layers (middle) is unfrozen and trained again until the CNN model classifies the images correctly. This technique, known as gradual unfreezing, was first introduced in Ref. [37].

It has been established that different layers in the CNN learn different features; therefore, different layers should be tuned with different learning rates [38]. Therefore, each layer will have specific parameters to be learnt and updated using layer-specific learning rates.

Therefore, from the SGD equation to update the parameters of the model

$$w_{j+1} : w_j - \alpha \frac{\partial}{\partial w_j} J(w, b) \tag{4}$$

In this work, this was modified to reflect layer-specific learning rates for weight updates and represented mathematically as:

$$w^l_{j+1} : w^l_j - \alpha^l \frac{\partial}{\partial w^l_j} J(w, b) \tag{5}$$

From the equation, $\frac{\partial}{\partial w_j} J(w, b)$ is the partial derivative of the cost function obtained via the backpropagation that computes the gradients used to update the weights of the model. Also, α is the learning rate, w^l_j is the parameter in layer l, and α^l is the learning rate of the lth layer, ranging from $(1, \ldots, L)$. L is the number of layers in the model. The α^l of the last layers is obtained by using the LR range test.

4.5. Mixed Precision Training

Numerous computations occur during the training of a CNN. For example, at the convolution layer, where the images are convolved with a filter to form the feature map. Conventionally, these computations are usually done in FP-32 (Single precision), which requires high memory consumption and bandwidth. FP-16 (half precision) is ideal in reducing the computational requirements; however, its dynamic range is narrower than FP-32, which may result in loss of model accuracy [39]. If the gradients of the model are smaller than the FP16 dynamic range 2^{-24} to 2^{15}, the gradient is truncated and result in inaccurate weight updates. From (4), if the weight updates multiplied by the learning rate are too small or too large, and the magnitude is outside FP-16's range, no updates are done. Hence the equation becomes:

$$w^l_{j+1} : w^l_j \tag{6}$$

Therefore, Mixed Precision Training (MPT) combines the strengths of FP-16/32 by using FP16 where appropriate and adding loss scaling to compensate for the small gradient values [39]. For example, FP32 is used for layers that output probabilities. The process of MPT for a layer is shown in Fig. 6.

As a result of the combined strength of FP16/32, MPT reduces memory and computational cost of running the model without loss of accuracy. It means larger batch-sizes can be used for training deep learning models compared to traditional methods. Consequently, there is a significant reduction in training time and improved performance in some cases. Due to the efficient use of memory and computational resources, MPT is beneficial for researchers in low-resources countries.

5. Results

Experiments were carried out to explore the effectiveness of the techniques mentioned above in optimizing the DenseNet model. It is important to note that the evaluation of the model was done on the out-sample dataset. A comparison was made between a pre-trained DenseNet model and a DenseNet model with Discriminative fine-tuning and Optimal learning rate. Also, the role of Mixed Precision Training in increasing the batch size and overall performance of the model was considered.

5.1. Learning rate selection and scheduling

The DenseNet model was trained using learning rate annealing, where a single learning rate was selected for all the layers but was scheduled to decay after every 10 epochs. The performance of the model was compared with a DenseNet model, whose learning rate was selected

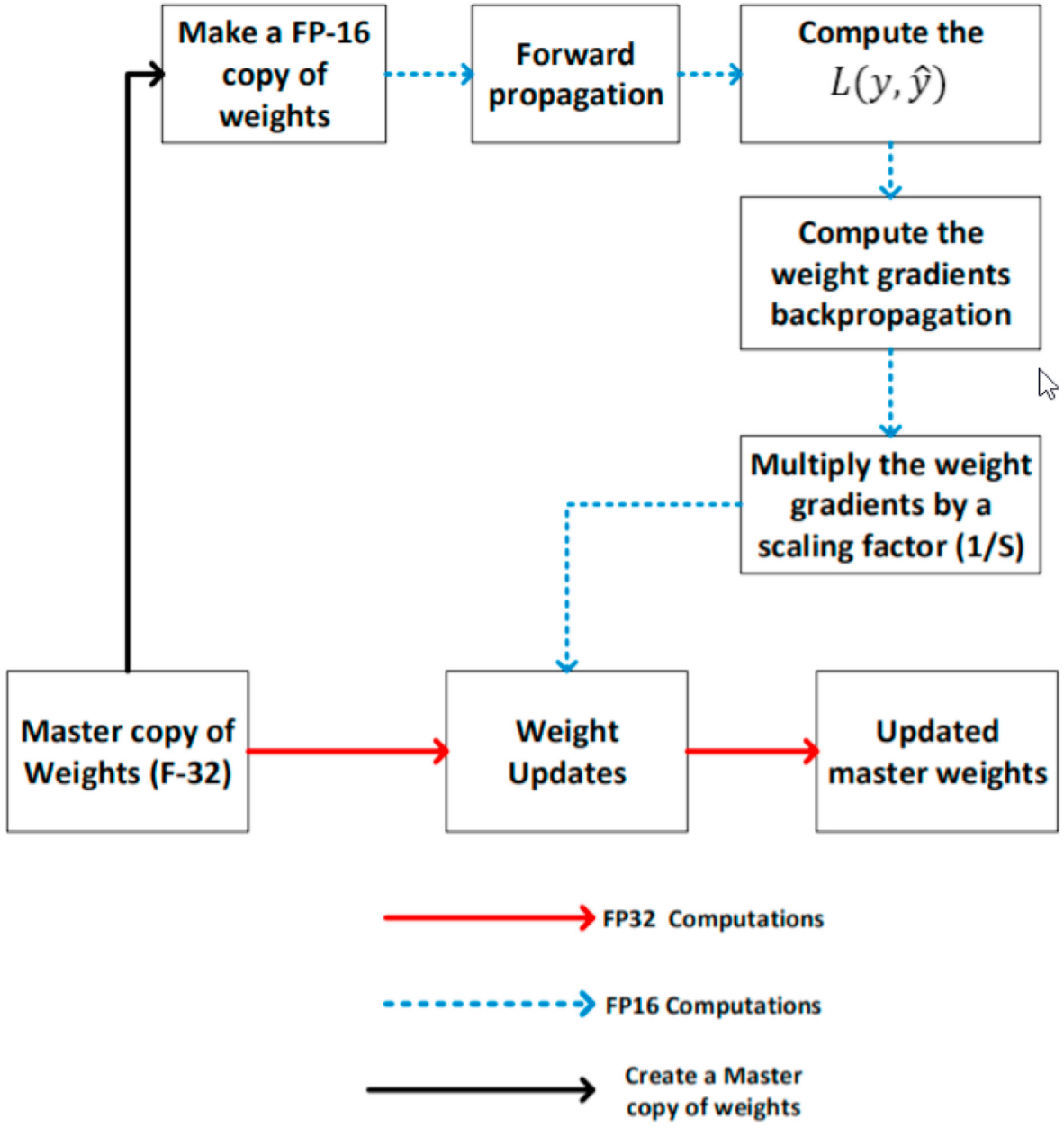

Fig. 6. Block diagram of mixed precision training.

using the LR range Test and 1cycle policy. Also, different learning rates were used to train different layers of the model using discriminative fine-tuning. The run-time and the accuracy of the model is shown in Table 3.

5.2. Effect of Mixed Precision Training

The DenseNet was trained on an ASUSTeK Computer with a processor Intel(R) Core (TM) i7-8750H CPU and NVIDIA GTX 2070 8G GPU. However, this limits the number of batch size the model could accommodate. It was observed that using MPT in training reduces the computation requirements. In the DenseNet model without DFT, the maximum batch size used was 20. However, with the model that uses MPT, a maximum batch size of 42 can be used.

Table 3
Performance of learning rate scheduling.

Model	Number of Epochs	Best Accuracy
Densenet + Learning rate annealing	100	51.2%
Proposed Model	30	96.3%

5.3. Evaluation metrics

In comparing and evaluating the performance of the model, the Multiclass Sensitivity and Specificity of the model are required. Multiclass Sensitivity (M.Sens) is the ability of a screening system to identify the CIN class correctly. In contrast, Multiclass Specificity (M.Spec) can determine that the image does not belong to a particular class [40]. Mathematically,

$$M.Sens = \frac{TP}{TP + FN} \tag{7}$$

$$M.spec = \frac{TN}{TN + FP} \tag{8}$$

In understanding the definitions for multiclass classification, it is

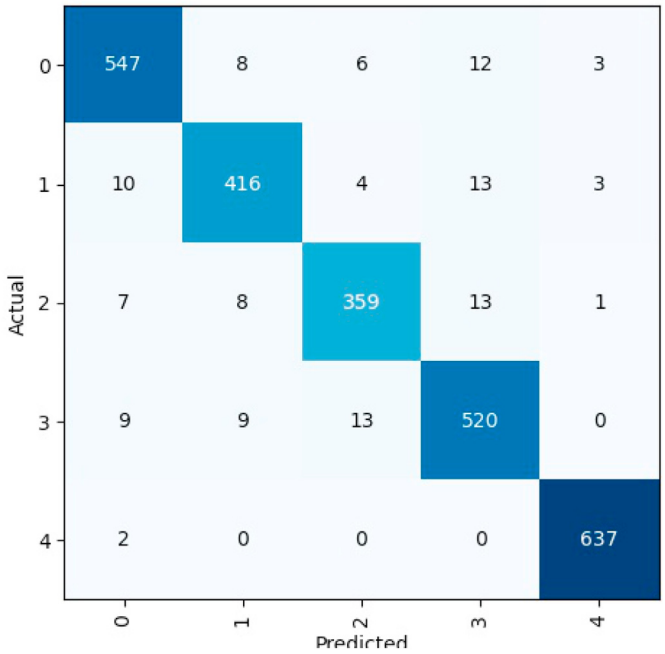

Fig. 7. Confusion Matrix of the proposed Model.

Table 4
Performance Evaluation of each class based on out-sample data.

Class	M.Sens (%)	M.space. (%)	PPV(%)	NPV (%)	F1 Score (%)	Acc. (%)
Normal	94.97	98.62	95.13	98.57	95.04	97.75
CIN 1	93.27	98.84	94.33	98.61	93.97	96.68
CIN 2	92.53	98.96	93.98	98.37	93.24	97.95
CIN 3	94.37	98.15	93.19	98.46	93.77	96.16
Cancer	99.69	99.72	98.91	99.92	99.29	99.64

important to note that if a class is misclassified, it is negative. At the same time, if it correctly classified, it is positive. TP (True Positive) indicates that the cervix image belonged to a particular class and was predicted correctly. TN (True Negative) indicates that the cervix image did not belong to a particular class and was classified correctly. FP (False Positive) indicates that the image does not belong to the same class, but it predicts it as belonging to that class. False Negative (FN) shows that the image belongs to a particular class, but it predicts it does not. TP, FN, TN and FP values can be obtained from the Confusion Matrix. The confusion matrix gives the predictability ability of the model, as shown in Fig. 7.

The multiclass Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and F1 score of each class obtained from the confusion matrix is summarized in Table 4. The average accuracy of the model across all the classes is 96.3%.

It can be observed from Table 4 that the model was sensitive to each category. However, it is observed that it could predict the cancer category the best when compared to other categories. It can be attributed to the model's capacity to detect cancerous features in the image without misclassification.

Table 5
Performance Comparison with other proposed methods.

Reference	Sensitivity (%)	Specificity (%)	Accuracy (%)
[9]	57.56	78.55	73.08
[11]	–	–	81.35
[12]	87.83	90	88.9
[15]	80.87	73.91	77.39
Proposed DenseNet	94.97	98.86	96.3

6. Discussion

6.1. Performance comparison

The overall accuracy of the model obtained is 96.3%. However, for diseases like cervical cancer that can be treated if detected early, sensitivity and specificity are preferred. It is seen that the model has high sensitivity and specificity, as shown in Table 4; thus, it can classify cervix images into the correct class. The implication of this is the appropriate treatment and follow-up treatment based on the CIN grade can be administered correctly by the health practitioners.

In comparing the performance of the model with other proposed methods in literature, it is observed that our proposed methodology had one of the best performances. The performance comparison is summarized in Table 5.

In addition, the inference and pre-processing times or computational was also computed. However, it is essential to note that these parameters are dependent on the hardware and system capability of the computer where the training and evaluation of the model are carried out. Hence, based on the system's capability as described in section 5.2, it was observed that the model could make inference on an image in about 0.26 ms.

6.2. Visualization of CNN using Grad-Cam

The visualization of the region of interest where the CNN network analyzes to make predictions could provide health practitioners further information on how the model works. Recall that the abnormal lesions that lead to cervical cancer start from the transformation zone of the cervix. Detection of the transformation zone in the cervix is a daunting task for health practitioners, therefore hindering administering appropriate treatments. Such treatments include Conization, Loop Electrical

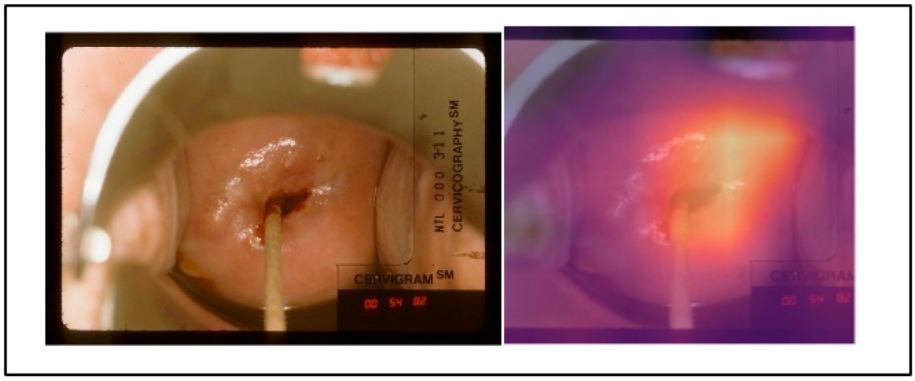

Fig. 8. Visualizing the section of the cervix CNN model analyses to make a prediction.

Excision Procedure (LEEP), where treatments are made on the transformation zone. Gradient-weighted Class Activation Mapping (Grad-Cam) highlight these important regions of the cervix (the transformation zone). It highlights the import regions (features) that CNN uses in prediction by using the gradients of the target to produce a localization map [41]. It can be observed in Fig. 8 that there is a sharp colour contrast in the cervix image on the right indicating the most prominent gradient used in prediction. The use of Grad-Cam in cervical cancer screening will give insight into the reasons why the model made predictions. It could foster trust in health practitioners when using the deployed model. Furthermore, it gives insight to researchers on ways of improving screening models.

predictions. It could foster trust in health practitioners when using the deployed model. Furthermore, it gives insight to researchers on ways of improving screening models.

7. Conclusion

The classification of cervix images into the CIN grades is critical to health practitioners in providing follow-up screening tests that may be required for possible treatments options. We proposed an algorithm that can classify the cervix images into five grades by optimizing a fine-tuned CNN model. In extending this work, we hope to explore the options of deploying the algorithm to a hardware device like a mobile phone for cervical cancer screening. It will be an alternative screening method for women in low-resource regions, where the ratio of experienced health practitioners to women's total population is small. It could also provide the screening personnel with the opportunity to liaise and share opinions with doctors on the appropriate treatments that can be administered and subsequent follow-up tests needed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We will like to thank the National Cancer Institute, Bethesda, USA, for providing us with the cervix images used in training the CNN model. We will also like to thank our clinician, Dr. Stephen O. Ohize, for reviewing our manuscript.

References

[1] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on mri. Zeitschrift für Medizinische Physik 2019;29(2):102–27.
[2] Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. nature 2017;542(7639):115–8.
[3] Bejnordi BE, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 2017;318(22):2199–210.
[4] Gultekin M, Ramirez PT, Broutet N, Hutubessy R. World health organization call for action to eliminate cervical cancer globally. 2020.
[5] Aina OE, Adeshina SA, Aibinu A. Classification of cervix types using convolution neural network (cnn). In: 2019 15th international conference on electronics, computer and computation (ICECCO). IEEE; 2019. p. 1–4.
[6] Santesso N, et al. World health organization guidelines for treatment of cervical intraepithelial neoplasia 2–3 and screen-and-treat strategies to prevent cervical cancer. Int J Gynecol Obstet 2016;132(3):252–8.
[7] Papillomavirus H. Related diseases report world. 2014. Version posted on, https://www.hpvcenter.net. in March 17.
[8] Al-Stouhi S, Reddy CK. Transfer learning for class imbalance problems with inadequate data. Knowl Inf Syst 2016;48(1):201–28.
[9] Xu T, Zhang H, Huang X, Zhang S, Metaxas DN. Multimodal deep learning for cervical dysplasia diagnosis. In: International conference on medical image computing and computer-assisted intervention. Springer; 2016. p. 115–23.
[10] McAllum B, Sykes PH, Sadler L, Macnab H, Simcock BJ, Mekhail AK. Is the treatment of cin 2 always necessary in women under 25 years old? Am J Obstet Gynecol 2011;205(5). 478–e1.
[11] Saini SK, Bansal V, Kaur R, Juneja M. Colponet for automated cervical cancer screening using colposcopy images. Mach Vis Appl 2020;31(3):1–15.
[12] Zhang T, et al. Cervical precancerous lesions classification using pretrained densely connected convolutional networks with colposcopy images. Biomed Signal Process Contr 2020;55:101566.
[13] Bengio Y. Practical recommendations for gradient-based training of deep architectures. In: Neural networks: tricks of the trade. Springer; 2012. p. 437–78.
[14] Burd EM. Human papillomavirus and cervical cancer. Clin Microbiol Rev 2003;16(1):1–17.
[15] Ostör A. Natural history of cervical intraepithelial neoplasia: a critical review. Int J Gynecol Pathol: official journal of the International Society of Gynecological Pathologists 1993;12(2):186–92.
[16] Yang X, Zeng Z, Teo SG, Wang L, Chandrasekhar V, Hoi S. Deep learning for practical image recognition: case study on kaggle competitions. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining; 2018. p. 923–31.
[17] Payette J, Rachleff J, Van de Graaf C. Intel and mobileodt cervical cancer screening kaggle competition: cervix type classification using deep learning and image classification. 2017.
[18] Xu T, et al. Multi-feature based benchmark for cervical dysplasia classification evaluation. Pattern Recogn 2017;63:468–75.
[19] Peng G, Dong H, Liang T, Li L, Liu J. Diagnosis of cervical precancerous lesions based on multimodal feature changes. Comput Biol Med 2021;130:104209.
[20] Buiu C, Dănăilă V-R, Rădută CN. Mobilenetv2 ensemble for cervical precancerous lesions classification. Processes 2020;8(5):595.
[21] Smith LN. Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV). IEEE; 2017. p. 464–72.
[22] Zeiler MD. Adadelta: an adaptive learning rate method. 2012. arXiv preprint arXiv:1212.5701.
[23] LeCun Y, Bengio Y, Hinton G. Deep learning, nature 2015;521(7553):436–44.
[24] Ruder S. An overview of gradient descent optimization algorithms. 2016. arXiv preprint arXiv:1609.04747.
[25] Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. Advances in neural information processing systems. 2014. p. 2933–41.
[26] Smith LN, Topin N. Super-convergence: very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications, vol. 11006. International Society for Optics and Photonics; 2019. p. 1100612.
[27] Smith LN. A disciplined approach to neural network hyperparameters: Part 1–learning rate, batch size, momentum, and weight decay. 2018. arXiv preprint arXiv:1803.09820.
[28] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4700–8.
[29] Bottou L. Stochastic gradient descent tricks. In: Neural networks: tricks of the trade. Springer; 2012. p. 421–36.
[30] Wang B, Gong NZ. Stealing hyperparameters in machine learning. In: 2018 IEEE symposium on security and privacy (SP). IEEE; 2018. p. 36–52.
[31] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.
[32] Chen C, Bai W, Davies RH, Bhuva AN, Manisty CH, Augusto JB, Moon JC, Aung N, Lee AM, Sanghvi MM, et al. Improving the generalizability of convolutional neural network-based segmentation on cmr images. Frontiers in cardiovascular medicine 2020;7:105.
[33] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of Big Data 2019;6(1):60.
[34] ADEDIGBA AP, ADESHINAT SA, AIBINU AM. Deep learningbased mammogram classification using small dataset. In: 2019 15th international conference on electronics, computer and computation (ICECCO). IEEE; 2019. p. 1–6.
[35] G. Pleiss, D. Chen, G. Huang, T. Li, L. van der Maaten, K. Weinberger, Memory-efficient implementation of densenets. arxiv 2017, arXiv preprint arXiv:1707.06990.
[36] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–105.
[37] Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018. arXiv preprint arXiv:1801.06146.
[38] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?. In: Advances in neural information processing systems; 2014. p. 3320–8.
[39] Micikevicius P, et al. Mixed precision training. 2017. arXiv preprint arXiv:1710.03740.
[40] Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Cont Educ Anaesth Crit Care Pain 2008;8(6):221–3.
[41] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 618–26.