Cervical Cancer Classification Using Deep Learning Techniques

Pierjos Francis COLERE MBOUKOU[1], Ali IDRI[1,2], Ferdaous IDLAHCEN[1] and Hasnae ZEROUAOUI[1]

[1] *Modeling, Simulation and Data Analysis, Mohammed VI Polytechnic University, Benguerir, Morocco*
[2] *Software Project Management Research Team, ENSIAS, Mohammed V University, Rabat, Morocco*

francis.mboukou@um6p.ma, ali.idri@um5.ac.ma, ferdaous.idlahcen@um6p.ma, hasnae.zerouaoui@um6p.ma

Abstract— Cervical cancer is one of the most common cancers in women. In 2020, it accounted for an estimated 604 000 new cases and about 342 000 deaths worldwide. It occurs when there are presences of abnormal cells within the cervix, which still grow uncontrollably. These cancer cells can also spread to other organs like lungs, liver and bladder which complicates the problem. If detected early, cervical cancer is one of the most successfully treatable cancers. In this paper, we propose and compare Deep Convolutional Neural Network (CNN) architectures for an automatic binary classification of cervix histological and cytological images based on fined tuned versions of seven deep learning techniques (VGG16, VGG19, DenseNet201, InceptionResNetV2, InceptionV3, ResNet50 and MobileNetV2). As empirical evaluations, four classification performance criteria (accuracy, recall or sensitivity, precision and F1-score), Scott Knott (SK) statistical test to select the best cluster of the outperforming architectures, and Borda Count voting method to rank the best performing models were used. The proposed architectures were evaluated, by 5-fold cross-validation, using two datasets: Liquid-based Cytology (LBC) Pap Smear and The Cancer Genome Atlas cervical histopathology slides (CESC) which contain 963 cytological and 600 histological images respectively. By classifying cervical images into normal and abnormal or into Cervical Squamous Cell Carcinoma(SCC) and Endocervical Adenocarcinoma(AC) types, the seven deep learning architectures gave higher accuracy values. For the LBC Pap Smear dataset, InceptionV3 was the best performing technique with an accuracy of 99.02%, a sensitivity of 99.67% , precision of 98.4% and f1-score of 99.03%. As for the CESC dataset, VGG16 had the most optimum accuracy, sensitivity, precision and f1-score of 99.33%, 99.83%, 98.85% and 99.34% respectively.

Keywords— Cervical Cancer, Classification, Deep Convolutional Neural Networks, Image Processing, Pap smear, whole-slide imaging.

1. Introduction

Cervical cancer is the fourth most prevalent female malignancy in terms of incidence and mortality. According to GLOBOCAN 2020, an estimated 604,000 new cases and 342,000 related deaths occurred overall. These statistics conceal a massive global inequality since 91% of deaths are reported in low- and middle-income countries (LMICs) that lack effective screening programs. [1][2]

The current screening strategies of cervical cancer, conventional Pap smear and Liquid-based cytology Pap smear, are used to assess for precancerous lesions within the cervix. These lesions are called Cervical Intraepithelial Neoplasia (CIN) and are divided into low-grade intraepithelial lesions (LSIL), known as CIN1, or high-grade intraepithelial lesions (HSIL), known as CIN2 and CIN3. Pap smear screening outcomes are crucial for follow-up decisions, generally, low grade tend to regress to normal, while high grade need further examinations and biopsy to confirm diagnosis and appropriate treatment subsequently. However, interpretations of cytology-based Pap smears and biopsy-based histology slides are subjective or biased experiences due to high workloads, laborious preparation process, and features complexity, which leads to the use of intelligent systems. [3][4]

Much recent progress has been made to improve disease detection through medical imaging, which provides facilities for diagnosis and and decision-making of several diseases such as cervical cancer [4][5]. To enhance and help oncologist's early detection and diagnosis, several algorithms are developed and evaluated. Several studies [6-9] reported, mostly, that Deep Learning (DL) techniques performed better in cervical cancer detection, and provided high accurate classifications than classical Machine Learning (ML) techniques. For instance, the study [6] showed that the use of deep CNN architecture ResNet50 gave better results (97% of accuracy) compared to existing ML techniques such as SVM (84%) and Random Forest (79%). This revealed the power and the efficiency of the use of deep CNN architectures in cancer detection and classification tasks.

In this study , we develop and evaluate the performances measured in terms of accuracy, sensitivity, specificity, precision, and f1-score of seven of the recent Deep Learning

techniques for Cervical Cancer (CC) pathology classification over two datasets: LBC Pap Smear and CESC datasets. To the most effective of our knowledge, this study is the first to evaluate and compare seven DL techniques (VGG16, VGG19, DenseNet201, InceptionResNetV2, InceptionV3, ResNet50 and MobileNetV2) using the Scott Knott (SK) statistical test and also the Borda Count voting method in Cervical Cancer image processing and classification. It is also the first to use both pathology-related procedures images, cytology and histopathology. The SK test has hugely been applied to compare cluster, and rank multiple machine learning models for parameters tuning in several fields such as software engineering [29-31] and cancer classification such as breast cancer[40]. Therefore, we use the SK test since: (1) it has high performance compared to other statistical tests as as presented by Jollife et al. [32] and Calinski et al. [33], and (2) it selects the best non-overlapping groups of machine learning techniques. In addition, to rank the most effective SK selected techniques, we use the Borda Count voting method [34–35]. This study considers two main research questions (RQs):

- (RQ1): What is the overall performance of DL techniques in Cervical Cancer classification?
- (RQ2): Is there any DL techniques which distinctly outperform the others?

The main contributions of this empirical study are the following:

(1) Designing seven DL architectures: VGG16, VGG19, DenseNet201, InceptionResNetV2, InceptionV3, ResNet50 and MobileNetV2 in CC classification.

(2) Avoiding overfitting by using weight decay and L2 regularizers.

(3) Evaluating the seven DL architectures over two datasets: LBC Pap Smear and CESC.

(4) Comparing the performances of the seven architectures using SK clustering test and Borda Count voting method.

The rest of the paper is organized as follows: Section 2 presents an overview of the seven DL techniques used in this paper. Section 3 explores related work associated with cervical cancer. In Section 4, we present the configuration and parametrization of the seven DL techniques. Section 5 describes data preparation which includes data acquisition and image processing. The empirical methodology followed throughout the research is reported in Section 6. Section 7 presents and discusses the empirical results. Section 8 reports the threats of validity of the study. Section 9 concludes and draws the line towards future works.

2. Deep Learning Architectures Background

This study uses seven DL architectures (VGG16, VGG19, InceptionV3, DenseNet201, ResNet50, InceptionResNetV2 and MobileNetV) as they were the most frequently used for diagnosing cancer and they provided high accuracy classification values. Then they were compared to a CNN which was used as the baseline model.

2.1. Convolutional Neural Network

A convolutional neural network (CNN, ConvNet) is a type of deep neural networks that are frequently used for computer vision tasks [11]. CNN model consists generally of 5 layers which are: input layer, convolutional layers, pooling layers, fully connected layer, and output layer. It is, as well, known that CNN model is often trained to allow feature extraction, feature selection, and classification or regression. Interpretation of how the network understands the image is difficult because of the black boxes modeling of CNN. Nevertheless, it is known that feature extraction obtained in network layers works better than human built features [34].

2.2. VGG16 and VGG19

VGG16 (Visual Geometry Group) is a convolution neural network architecture for visualization and image classification proposed by K. Simonyan and A. Zisserman[10]. It won the ILSVR (ImageNet) challenge in 2014. VGG16 accepts as an input 224 x 224 with 3 channel RGB image; the architecture of the model consists of convolution layers blocks of 3x3 filters with a stride 1, the same padding and maxpool layer of 2x2 filters of stride 2. It also has 3 fully connected layers (FC) and a softmax for the output. The Rectified Linear Unit (ReLU) non-linearity is present in all hidden levels. The 16 in VGG16 refers to 16 layers of the model. This is a large network, and it has about 138 million parameters. The big difference between VGG16 and VGG19 is the number of layers as shown in Figure 2*. VGG19 has 19 layers. The architecture of VGG16 and VGG19 are represented in Figure 12 in Appendix.

2.3. InceptionV3

There are three versions of Inception CNN architecture. InceptionV3 is the third version of Google's Inception Convolutional Neural Network, introduced during the ImageNet Recognition Challenge with a specialization in image analysis and object detection [12]. Starting as a GoogleNet module, InceptionV3 has 42 deep layers with a default input size fixed to 299x299 [36]. Blocks of parallel convolutional layers with 3 different sizes of filters (1x1, 3x3, 5x5) constitue this architecture. Additionally, 3×3 max pooling is also performed. The outputs are combined and forwarded to the next inception module as shown in Figure 13 in Appendix.

2.4. ResNet50

ResNet is a deep network architecture proposed by He Kaiming et al. [13]. ResNet50, one of the variants of ResNet, is a convolutional neural network that has 50 layers deep and can classify images into 1000 object categories. The network has an image input size of 224 x 224. The architecture of ResNet is inspired by VGG (Figure 14 in Appendix) and follows two design rules: (1) for the same output feature map size, the layers have the same number of filters and (2) if the feature map size is halved, the number of filters is doubled so as to preserve the time complexity per layer.

2.5. InceptionResNetV2

InceptionResNetV2 is a convolutional neural network that contains 164 layers deep and can classify images into 1000 object categories as shown in Figure 15. The network has a 299x299 as input size [37].

2.6. DenseNet 201

DenseNet201 is a variant of DenseNet (Dense Convolutional Network). Its architecture is similar to ResNet. The difference is about dense blocks which are densely connected together [38]. A Dense block is composed of a Batch Normalization, ReLU activation and 3x3 convolution. Each layer receives in input all previous layers output feature maps (Figure 16 in Appendix).

2.7. MobileNetV2

MobileNet series V2 introduces inverted residuals and contains 53 layers [39]. It is a lightweight architecture that performs a single convolution on each color channel rather than combining all three and flattening it (Figure 17 in Appendix).

3. Related work

To carry out this study, a literature review on the use of machine learning in cervical cancer has been performed. The main findings of this review are:

- Classification is the investigated objective in their studies. And binary classifiers are more accurate than that of multiclass classifiers which can detect the type of abnormalities in the cells.
- First of all, we can assume that most of the successful deep learning-based cervical cancer detection methods were based on a dataset using Pap smear or histology whole-slide slides. Among all datasets in Table 1, the Herlev dataset is the most commonly used dataset for cervical cell classification.

Table 1: Image datasets for cervical cancer.

Dataset	Size	Classes	Type	Author
MobileODT	1448	3	Colposcopy	MobileODT [18]
Zenodo	962	4	Cytology	Franco et al. [19]
ALTS	938	2	Cytology	Alts Group [16]
Herlev	917	187	Cytology	Dr J. Jantzen [14]
DANS-KNAW	963	4	Histology	Hussien [15]
CRIC	400	6	Cytology	M.T. Rezende et al.[17]

- Most of the studies preprocessed their input images by normalizing images and using data augmentation.
- Deep learning methods with pre-trained networks and those with transfer learning mechanisms are more accurate than networks trained from scratch.
- Among criteria (Accuracy, F1, Precision, Recall, Zijdenbos similarity index) to evaluate the performance of the DL techniques, accuracy is frequently used as shown in Table 2.

Table 2: Summary of selected papers on cervical cell classification.

Authors	Techniques	Classification type	Performance measures	Findings and results
Kurnian-ingsih et al. [20]	VGG-Like Net	Binary and Multi-classification	Accuracy, Sensitivity, Specificity, AUC	The classifier yields a sensitivity score of more than 96% for the binary classification problem and yields a higher result of more than 95% for the 7-class problem on Herlev dataset.
Hyeon et al. [9]	Logistic Regression, Random Forests, AdaBoost and SVM	Binary	F1, Precision, Recall	Authors use different classifiers namely: logistic regression, random forests, AdaBoost and SVM for classification of the pap-test images into normal and abnormal. From these classifiers, the highest scoring one is the SVM classifier with an F1-score of 78% on a dataset collected locally.
Lin et al. [21]	AlexNet, GoogLeNet, ResNet and DenseNet	Binary and Multi-classification	Accuracy	The pre-trained models were fine-tuned on the Herlev cervical dataset. They achieved classification accuracies of 94.5%, 71.3% and 64.5%, for two-class (abnormal versus normal), four-class (normal, low-grade squamous intraepithelial lesion (LSIL), high-grade squamous intraepithelial lesion (HSIL) and carcinoma-in-situ (CIS) [22]) and seven-class classification tasks respectively.
Promworn et al. [23]	resnet101, densenet161, AlexNet, vgg19-bn and squeeznet1-1	Binary and Multi-classification	Accuracy, Sensitivity, Specificity	The models are retrained on the Herlev dataset. Based on accuracy, densenet161 was the best performer on both binary classification (94.38%) and multiclass classification (68.54%). AlexNet and resnet have achieved 100% of sensitivity on binary classification. Whereas densenet161 was the best performer on multiclass classification with 68.18%. Again, based on specificity, densenet161 was superior with values 82.61% for binary and 69.57% for multiclass classification.
Dong et al. [24]	Inception-V3	Binary	Accuracy, Sensitivity, Specificity	Features extraction such as color, texture and morphology along with the Inception-V3 model for the classification of cervical cells.The proposed algorithm achieved an overall accuracy of 98.2%, the sensitivity of 99.4% and specificity of 96.73% for normal abnormal classification on the Herlev dataset.

4. Deep Learning Architectures Configuration

In order to classify cervical cells, we are going to use, as seen in Section 3, pre-trained networks and binary classification since they perform well (accuracy, recall, F1, precision,etc.). Therefore this section exposes the parameter tuning of the DL models. The parameter values of these DL architectures are too long and can be available by emailing the authors of this study.

4.1. DL Architectures Configuration

To build an automatic binary BC classification based on LBC Pap Smear [25] and CESC [26] datasets, different DL architectures have been implemented using several parameters tuning experiments. All images from these two datasets were resized to 224x224 pixels except those of InceptionV3 and InceptionResNetV2 models that were resized to 299x299 since it is the default input size in their architectures. For the training, we used the transfer learning technique where we downloaded the seven DL techniques pre-trained in the ImageNet dataset which contains 1.2 million images for training, 50 000 for validation and 100 000 for testing and with 1 000 classes [27]. Then, for the parameter tuning, we set the batch size to 32 and the number of epochs to 200. As for the optimization, we used Adam (adaptive moment estimation) [28] with an initial learning rate set to 0. 000001. In addition, L2 regularizer were used to reduce the overfitting for different models. A fully connected layer was trained with the ReLU, followed by a dropout layer with a probability of 0.5. We changed the last dense layer to two classes in output corresponding to normal and abnormal instead of 1000 classes as was used for ImageNet.

4.2. CNN Baseline model

The proposed method has the following parameters (Table 3):

- Input layer in which inputs are images of 244x244 dimension and the number of channels (3 stands for RGB).
- Convolutional layer which computes the convolutional operation in the input images using kernel filters (3x3 in our CNN) to extract fundamental features. It specializes in reducing the size of the input matrix but keeping the important features. This means it detects only the features that contribute the most to the image.
- Pooling Layer (Maxpooling in our case) reduces the number of parameters and computation by downsampling the representation. The max pooling is set to 2x2 with a stride of 2.
- Fully connected layer (Dense) which treats the input data as a vector and produce an output as a single vector. The last one,fully connected output layer with sigmoid activation with 2 output filters for our binary classification problem.

Table 3 recaps the tuning of the CNN layers used in this study. Note that every output shape has 'None' instead of the batch size in order to facilitate changing of batch size at runtime.

5. Data Preparation

This section presents the data preparation process we followed for the two datasets which consists of:

(1) Data acquisition.

(2) Data augmentation.

Type of layer	Output shape	No Parameters
Convolutional Layer (Conv2D)	(None, 222, 222, 64)	1792
Max pooling (MaxPooling2)	(None, 111, 111, 64)	0
Flatten	(None, 788544)	0
Dense	(None, 128)	100933760
Dropout	(None, 128)	0
Dense	(None, 2)	258
Total of parameters	Trainable parameters	Non trainable parameters
100,935,810	100,935,810	0

Table 3: CNN Baseline layers configuration.

(3) Rescaling and Pickling images

Data preparation process is summarized in Figure 1 below.

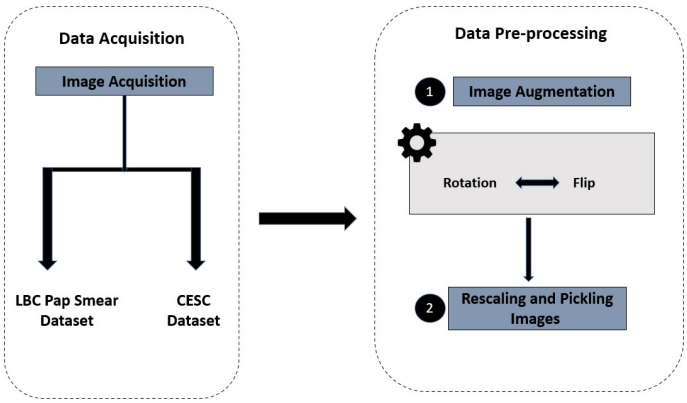

Figure 1: Data preparation process.

5.1. Data acquisition

In this study, to build the binary classification DL models, we used two datasets (LBC Pap Smear and CESC) which contain images.

5.1.1. LBC Pap Smear

LBC Pap Smear dataset [25] was collected using liquid-based cytology (LBC) technique at 400x magnification in the Obstetric and Gynecology department of Gauhati Medical College and Hospital. It contains 963 cytological images as follows: 613,163, 113, and74 images belong respectively for NILM, LSIL, HSIL, and SCC as shown in Table 4. In order to have a binary classification problem, we consider NILM as a "normal" category and LSIL, HSIL, and SCC into one category "abnormal".

5.1.2. CESC Dataset

CESC dataset was collected from the cancer genome atlas (TCGA) and pre-processed by Ferdaous Idlahcen et al [26]. It consists of 600 histological images: 300 images for ACC and 300 SCC. Images of CESC dataset are 1024x1024 pixels images at 20x magnification.

Table 4: LBC Pap Smear Dataset description.

Class	Category	Quantity
Normal	NILM	613
Abnormal	LISL	163
	HSIL	113
	SCC	74
Total		963

5.2. Data processing

Data preprocessing is converting data from one format to another more user-friendly, desired, and relevant.

5.2.1. Data Augmentation

Data augmentation is mostly used to avoid the risk of overfitting [41]. As we can notice in Section 5.1 for the LBC Pap Smear dataset the number of images in each class normal and abnormal is imbalanced. Indeed 63% of the dataset images belong to normal class. One of the issues of the imbalanced data is misleading classification accuracy. So classes need to be balanced. On the other hand, as DL models require a lot of data to perform correctly, CESC dataset does not have many images (600 only). Both datasets were resampling by using data augmentation in order to overcome this limitation. Thus, we generated new images by following the data augmentation process hereafter and as shown in Figure 2:

- For each image to augment, we randomly choose one of these operations : rotation of 90 degree, flip.
- If rotation is chosen. We exclusively use clockwise or counterclockwise rotation.
- Else, we randomly use again one of these transformations : vertical flip, horizontal flip.

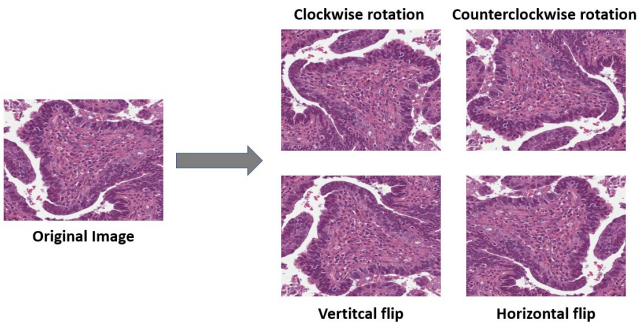

Figure 2: Data augmentation process.

So, by performing the process, we obtained results are summarized in Table 5 and Table 6.

Table 5: CESC Dataset description after data augmentation.

Class	Quantity
Adenocarcinoma (ACC)	600
Squamous Cell Carcinoma (SCC)	600
Total	1200

Table 6: LBC Pap Smear Dataset description after data augmentation.

Class	Quantity
Normal	613
Abnormal	613
Total	1226

5.2.2. Rescaling and Pickling images

After having augmented datasets, we prepared data to be used for the models training. First, we rescaled images by changing their range. That means we scaled images from integers 0-255 to floats 0-1.

Most DL networks are designed in a way so that they can only accept images of a fixed size. Then, we resized the input images to 224x224 pixels except those of InceptionV3 and InceptionResNetV2 models that required 299x299 images as the default input size in their architectures. Finally, as the datasets now contain a large number of images, we saved the resized data (images) into pickle files in order to be used for later works (training, evaluating, etc.), without having to repeat all the data processing again.

6. Empirical design

In this part we present the empirical design of this study. We present therefore:

(1) cross-validation used to evaluate the models,

(2) performance criteria used to evaluate the seven DL architectures,

(3) Scott Knott statistical test we used to cluster the DL techniques according to their accuracy values,

(4) Borda Count method we used to rank the DL techniques of the best SK cluster according to accuracy, recall, precision, and F1-core,

(5) experimental process we followed to carry out all the empirical evaluations,

(6) the abbreviations we used to shorten the names of DL techniques.

6.1. Cross-validation

We evaluate the models using a cross-validation method named Stratified k-Fold (k=5 in our study). It splits the dataset into 5 folds such that each fold contains approximately a percentage of samples from each target class equal to that of the entire set. Then, it chooses 4 folds which is the training set. The remaining fold is the test set. We trained the model on the training set. On each iteration, a new model is trained. Next, we validated on the test set and saved the result of the validation. We repeated the process 5 times. In the end, we have validated the model on every fold. Finally, we got the final score: the average of results that we got.

6.2. Statistical tests

6.2.1. Scott Knott

Scott-Knott, proposed by Scott and Knott in 1974 [42*], is a hierarchical clustering algorithm used in the application of the analysis of variance (ANOVA). Due to its simplicity and robustness [29, 31, 32, 33], it is mostly used to find distinct homogeneous overlapping groups distinction based on the multiple comparisons of treatment means.

6.3. Borda count

The Borda Count Method is a simple tool that is used in elections and decision-making in various situations. In this technique, points are given to candidates based on their ranking; 1 point for last choice, 2 points for second-to-last choice, and so on until we are at the top. After all of the votes are tallied, the most points option is the best, and thus the winner [34][35].

In this study, we used Borda count technique to find the best performing DL technique from the four performance measures with equal weights. It is intended to be able to choose different options and candidates, rather than the option that is preferred by the majority. As a consensus-based voting system, the opposite of this is a majority system. This strategy was adopted to make sure that we do not favor a particular performance criterion than another.

6.4. Performance measures

As mentioned in the previous sections, this present experiment uses four metrics to evaluate the performance of the seven DL classifiers: accuracy, recall, precision, and F1 score. These popular parameters are defined as follows:

$$Accuracy = \frac{TP+TN}{TN+TP+FP+FN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1-score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4)$$

Where:

- TP (True Positive) means malignant case is identified as malignant.
- FP (False Positive) : benign case is identified as malignant.
- TN (True Negative) : benign case is identified as benign.
- FN (False Negative) : malignant case is identified as benign.

Here, accuracy is the measure of all the correctly identified cases. Precision is a metric that quantifies the number of correct malignant predictions made. That means minimizing benign cases which are identified as malignant. Recall is a metric that measures the amount of right malignant predictions made out of all possible ones; it reduces the number of benign cases labeled as malignant. And F1-score is the weighted average (harmonic mean) of Precision and Recall. As a result, it considers both false positives and false negatives.

6.5. Experiment process

The methodology we followed to carry out all the empirical evaluations is illustrated in Figure 3. It consists of five steps that are well described below. Note that similar methodologies were used in [30, 31, 42, 43, 44].

The evaluation process involves five steps:

(1) Assess the accuracy of each variant of the deep learning architectures (VGG16, VGG19, DensNet201, MobileNetV2, ResNet50, InceptionV3, InceptionResNetV2) and of the baseline CNN for both datasets LBC Pap Smear and CESC.

(2) Select the deep learning architectures outperforming the CNN baseline (accuracy higher than 5% of the CNN).

(3) Transform the accuracy values of the selected DL models using Box-cox method since the Scott Knott test required its inputs to be roughly normally distributed.

(4) Cluster the selected DL using Scott Knott test and select the DL techniques of the best SK cluster (have the best accuracy and statistically indifferent).

(5) Rank the DL techniques of the best SK cluster using Borda count voting system based on the four performance measures (accuracy, recall, precision and F1-score) and select the top deep learning architectures.

6.6. Abbreviation

In order to facilitate the reading of the names of the DL techniques, we abbreviate the name of each variant of DL techniques as demonstrated in Table 7.

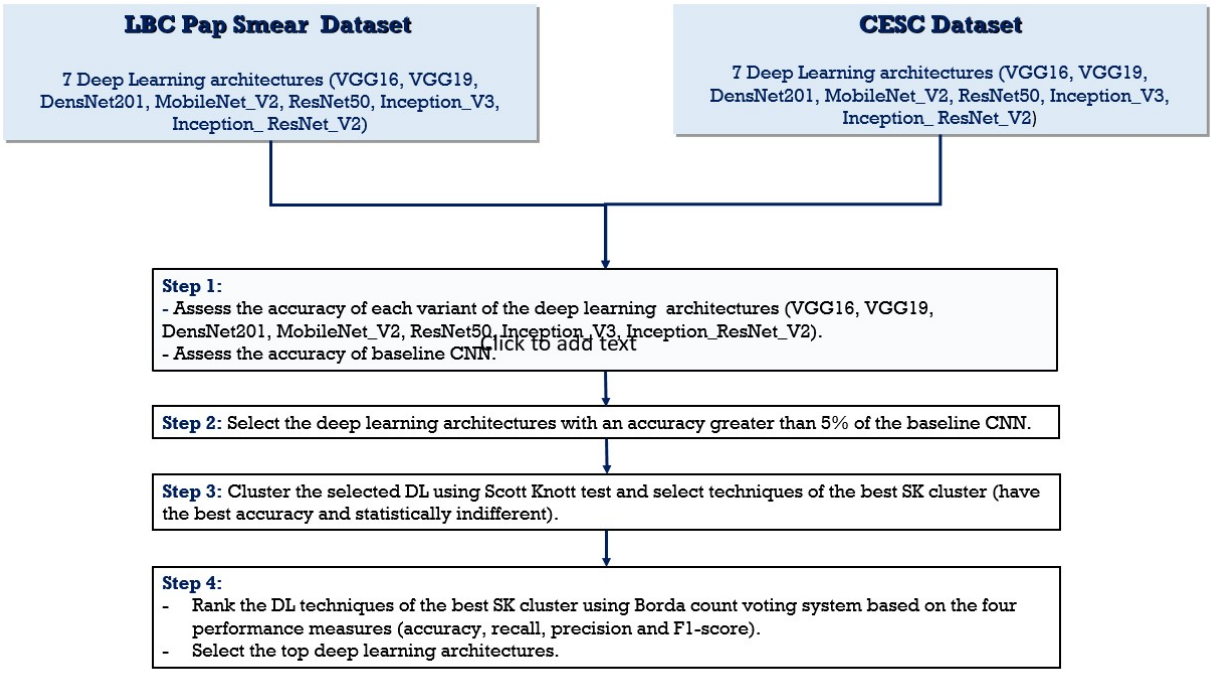

Figure 3: Steps of the Experimental Process.

DL Techniques	Abbreviation
CNN	CNN
VGG16	VGG16
VGG19	VGG19
ResNet50	RES50
InceptionV3	INV3
InceptionResNetV2	INRES
DensNet201	DENS
MobilNetV2	MOB

Table 7: Abbreviations used for the DL techniques.

7. Results and Analysis

This part shows and discusses the results of the empirical evaluations of seven DL techniques stated since the beginning of this study, over two datasets (LBC Pap Smear and CESC). Remember that the performances of the DL techniques were evaluated using four criteria (accuracy, recall, precision and F1-score). For each dataset, we first compared the performance in terms of accuracy of each DL technique with that one of the CNN baseline model, and we only kept DL techniques with an accuracy greater than 5% of the baseline. Then, we used the SK statistical test to cluster the selected DL techniques, and Borda count to rank the DL techniques belonging to the best SK cluster.

7.1. Accuracy evaluation and comparison of the seven DL techniques

This section compares the accuracy values of the seven DL techniques to the CNN baseline model and their accuracy values to each other. Note that the empirical evaluations of the DL techniques were performed using a computer with Processor: Intel® Core™ i5-7200U CPU @ 2.50GHz × 4 and 4 Go in RAM running on a Ubuntu 18.04.5 LTS. Python 3 was used with the two DL frameworks Keras and Tensorflow as deep learning backend. R was used with ScottKnott package for performing the Scott-Knott clustering algorithm.

7.1.1. LBC Pap Smear dataset

First of all, we present and discuss the accuracy of each DL technique depending on the number of epochs over LBC Pap Smear dataset as shown in Table 8 and Figure 4 shows their evolution as a function of epochs. Thereafter, we compare the accuracy values obtained by the seven DL techniques to the baseline CNN accuracy.

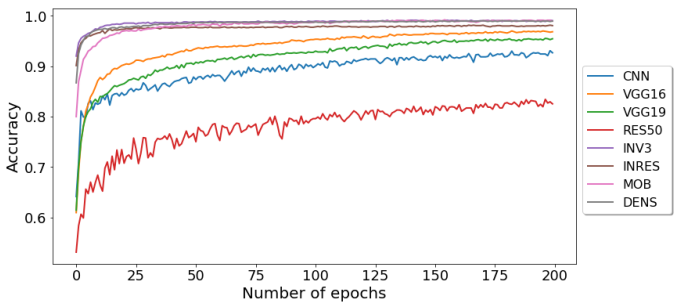

Figure 4: Accuracy evolution of the seven deep learning architectures and the baseline CNN over the LBC Pap Smear dataset.

We can see that InceptionV3, InceptionResNetV2, Den-

DL Techniques	Accuracy (%)
CNN	92.65
VGG16	96.81
VGG19	95.43
ResNet50	82.51
InceptionV3	99.02
InceptionResNetV2	98.04
DensNet201	98.94
MobilNetV2	98.94

Table 8: Accuracy values over the LBC Pap Smear dataset.

DL Techniques	Accuracy (%)
CNN	83.75
VGG16	99.33
VGG19	98.67
ResNet50	91.75
InceptionV3	99.08
InceptionResNetV2	97.83
DensNet201	99.17
MobilNetV2	98.50

Table 9: Accuracy values over the CESC dataset.

sNet201 and MobileNetV2 outperformed the baseline CNN but VGG16, VGG19, InceptionV3, ResNet50 did not achieve an accuracy values greater than 5% of the baseline CNN. Worst, we can even see that ResNet50 accuracy (82.51%) is much smaller than CNN accuracy. However, the best accuracy value was achieved by InceptionV3 (99.02%) followed by DensNet201 and MobileNetV2 (98.94%). Thus, InceptionV3, InceptionResNetV2, DensNet201 and MobileNetV2 were selected for the next steps of the evaluation process.

7.1.2. CESC Dataset

Figure 5 and Table 9 show the accuracy values of the baseline CNN, VGG16, VGG19, DenseNet201, InceptionResNetV2, InceptionV3, ResNet 50 and MobileNetV2 over the CESC dataset. Note that all DL models achieved an accuracy values greater than 5% of the baseline CNN. Therefore, all the seven DL techniques were selected for the SK statistical test. VGG16 was the best technique with 99.33%. In general, they correctly classify cervical cells as shown by the high accuracies.

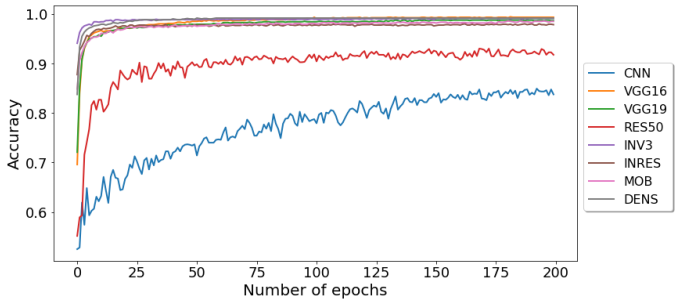

Figure 5: Accuracy evolution of the seven deep learning architectures and the baseline CNN over the CESC dataset.

7.2. Scott Knott statistical test and Borda Count

This subsection clusters the DL techniques selected in step 7.1 using SK test and ranks them using Borda count method.

7.2.1. LBC Pap Smear

InceptionV3, InceptionResNetV2, DensNet201 and MobileNetV2 are DL techniques kept for the LBC Pap Smear dataset. Figure 6 shows the results of SK test. We obtained 1 cluster which is also the best and contains these selected DL techniques. This means that the DL techniques are statistically identical or similar according based on the accuracy. So, they were kept for the Borda count method and their accuracy, recall , precision and F1-score in % are shown in Table 10.

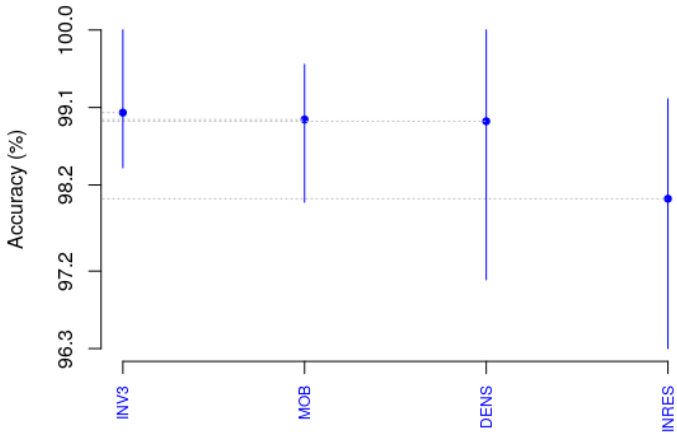

Figure 6: Results of SK test for the DL techniques over the LBC Pap Smear dataset.

Borda count applied on the criteria of the Table 10 revealed that InceptionV3 is ranked first followed by DensNet201 which was really close to MobileNetV2 in terms of score. So, Table 11 summarizes the results of Borda count method over LBC Pap Smear dataset.

DL Techniques	Accuracy	Recall	Precision	F1-score
InceptionV3	99.02	98.4	99.67	99.03
InceptionResNetV2	98.04	97.58	98.53	98.05
DensNet201	98.94	98.43	99.51	98.95
MobilNetV2	98.94	99.18	98.70	98.94

Table 10: Perfomance criteria values over the LBC Pap Smear dataset.

Rank	Deep Learning Techniques	Score
1	InceptionV3	14
2	DensNet201	12
3	MobilNetV2	11
4	InceptionResNetV2	4

Table 11: Ranks of the DL techniques belonging to the best SK cluster of LBC Pap Smear dataset.

7.2.2. CESC

Figure 7 presents the results of SK test over the CESC dataset. We obtained 2 clusters over the 7 DL techniques. The best cluster contains all techniques except ResNet50. Only DL techniques in the best cluster were used in Borda count voting method as summarized in Table 12.

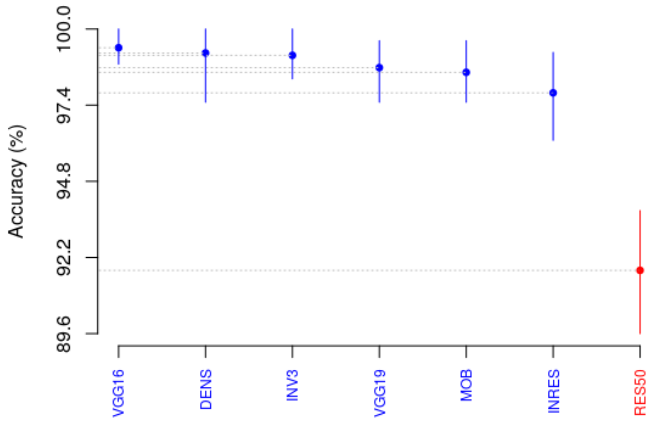

Figure 7: Results of SK test for the DL techniques over the CESC dataset.

DL Techniques	Accuracy	Recall	Precision	F1-score
VGG16	99.33	98.85	99.83	99.34
VGG19	98.67	98.03	99.33	98.67
InceptionV3	99.08	98.38	99.83	99.09
InceptionResNetV2	97.83	96.74	99.00	97.86
DensNet201	98.94	98.43	99.51	98.95
MobilNetV2	98.50	98.18	98.83	98.51

Table 12: Perfomance criteria values over the CESC dataset.

By applying Borda count voting method, VGG16 is ranked first, DensNet201 second and the third place was for InceptionV3 as in Table 13.

Rank	Deep Learning Techniques	Score
1	VGG16	23
2	DensNet201	20
3	InceptionV3	18
4	VGG19	12
5	MobileNetV2	8
6	InceptionResNetV2	5

Table 13: Ranks of the DL techniques belonging to the best SK cluster of CESC dataset.

To summarize the results of the SK test and Borda count voting method for both datasets, there are two points to emphasize:

(1) The VGG16, VGG19 and ResNet50 accuracy results were not good compared to the baseline CNN over the LBC Pap Smear dataset. They were eliminated from the SK test. Although for the CESC dataset, VGG19 and ResNet50 were kept for the SK test we observe that they did not give good results compared to other DL techniques (VGG19 4th and ResNet50 belongs to the last cluster).

(2) The DensNet201 technique gave good results for both datasets. Indeed, for both datasets, we can observe that DensNet201 belongs to the best cluster and is ranked second. For LBC Pap Smear dataset InceptionV3 is ranked first but third for CESC dataset. Even if VGG16 is the best technique for the CESC dataset but it is not even selected for SK test regarding LBC Pap Smear dataset.

We can therefore conclude that DensNet201 gave good results compared to the other techniques regardless the datasets. It is the compromise and the best technique for Cervical Classification. Then, InceptionV3 comes in second position.

8. Threats of validity

This section describes the threats to this paper's validity with respect to external and internal validity.

8.1. Internal validity

This work used the cross-validation evaluation method and the Adam optimizer. The main purpose of using the Adam optimizer over the classical Stochastic Gradient Descent (SGD) is due to its good performance and fast convergence of the learning rate [45]. However, note that its generalization performance tend to be significantly worse than that of SGD in some scenarios [46]. The learning rate of SGD is often difficult to tune, since the magnitudes of different parameters vary widely, and adjustment is required throughout the training process. Another internal threat for this study

is the tuning of DL architectures hyperparameters to avoid overfitting by using weight decay and L2 regularizers.

8.2. External validity

This paper only used two datasets that contain histological and cytological images. As seen in Section 3, others studies used different datasets to classify cervical cells. One of the most used datasets is Herlev. Thus, we cannot generalize the obtained results for all the datasets with the same type of images and the same features. However, it will be a benefit to redo this study using the same and different DL techniques with other datasets (such as Herlev and MobileODT) in order to confirm or refute the findings of this work or experiment.

9. Conclusion and Future Works

This work presented and discussed the results of an empirical comparative study of seven recent deep learning techniques (VGG16, VGG19, DenseNet201, InceptionResnetV2, InceptionV3, ResNet50 and MobileNetV2) for Cervical Cancer Imaging classification. In this paper, to evaluate these DL models, we used four performance criteria, SK statistical test and Borda Count to rank these seven DL techniques over two datasets (LBC Pap Smear and CESC). The baseline model was CNN. The results of this study are:

(RQ1): What is the overall performance of DL techniques in Cervical Cancer classification?

In terms of accuracy, we observed that InceptionV3, DensNet201, MobileNetV2 and InceptionResNetV2 gave the best results and outperformed the baseline CNN regardless of the dataset used in this experiment. However, for LBC Pap Smear dataset, VGG16, VGG19 and ResNet50 underperformed the baseline CNN.

(RQ2): Is there any DL techniques, which distinctly outperform the others?

For both datasets DensNet201 is ranked second. As InceptionV3 is ranked first on LBC Pap Smear and third on CESC dataset. Therefore, we conclude that DensNet201 is the best technique for Cervical Classification as it is the compromise of these two datasets. Then InceptionV3 technique can be considered as an option.

As said in 8.2, we will redo this study as oncoming works using the same and different Deep Learning techniques with other datasets (Herlev and MobileODT) in order to confirm or refute this work results.

10. Aknowledgement

This work was carried out as part of an end-of-studies internship for a Data Science License within the Al Khawarizmi department of the Mohammed VI Polytechnic University (UM6P). It is therefore an introductory research internship. This internship was made possible thanks to the help and collaboration of several people. Therefore, we would like to thank them wholeheartedly and express our gratitude.

To Mr. Ali IDRI
Head of the Web and Mobile Engineering department at ENSIAS,Mohammed V University and Professor in the Modeling, Simulation and Data Analysis program at Mohammed VI Polytechnic University.
Carrying out this work in the field of health is certainly governed by our interest in this subject, but also for your success, both nationally and internationally, by applying Machine Learning concepts in this area to facilitate decision-making.

To Madams Hasnae ZEROUAOUI and Ferdaous IDLAHCEN
PhD Students at Mohammed VI Polytechnic University.
We thank you for your availability and especially your advice which has contributed to our reflection. You were always listening to our many questions and interested in advancing this work. We wouldn't be where we are now without your support, sympathy and dynamism. You were a great help in reading and interpreting the results. You have exponentially played a major role in the development of this study.

And with you, it was very pleasant to work in a simple and good mood. You were the perfect internship supervisor to whom we owe a lot for this study, which will certainly have brought us a lot from a scientific point of view but also humanly through your meeting. Receive our gratitude and sincere thanks now.

We also want to thank the entire teaching team of the Al Khawarizmi department of UM6P, we are thinking in particular of Ahmed Ratnani, Hassan Machkour, Nabila Idar and all the teachers, who provided us with their experience, skills and the necessary tools for the realization of this work.

Finally, we would like to express our gratitude to the friends and colleagues who brought us, from near and far, their moral and intellectual support throughout the internship.

Appendix

Figure 8-11 represent diagrams of training and validation for the LBC Pap Smear and CESC Dataset and Figure 12 to 17 show the seven Deep Learning Architectures (InceptionV3, DensNet201, MobileNetV2 and InceptionResNetV2, VGG16, VGG19 and ResNet50).

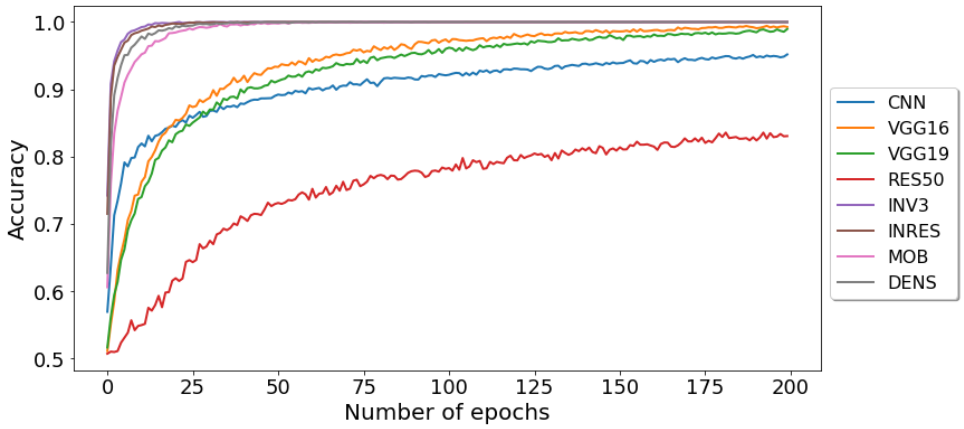

Figure 8: LBC Pap Smear : Training Accuracy.

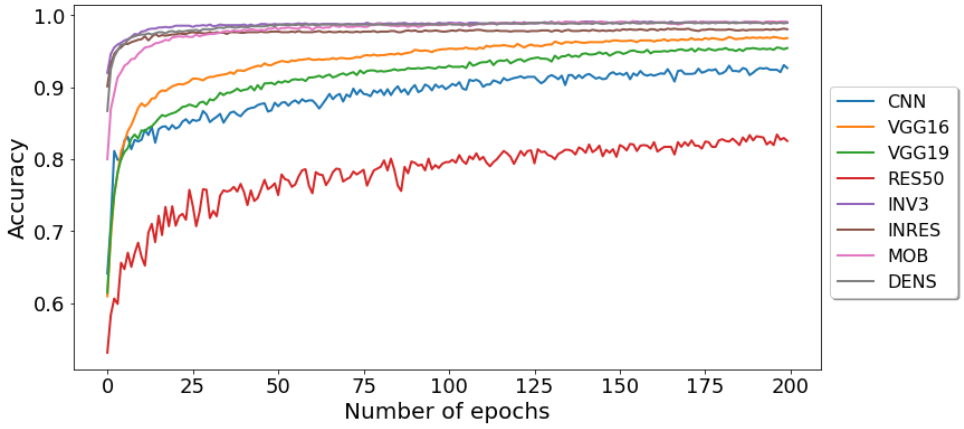

Figure 9: LBC Pap Smear : Validation Accuracy.

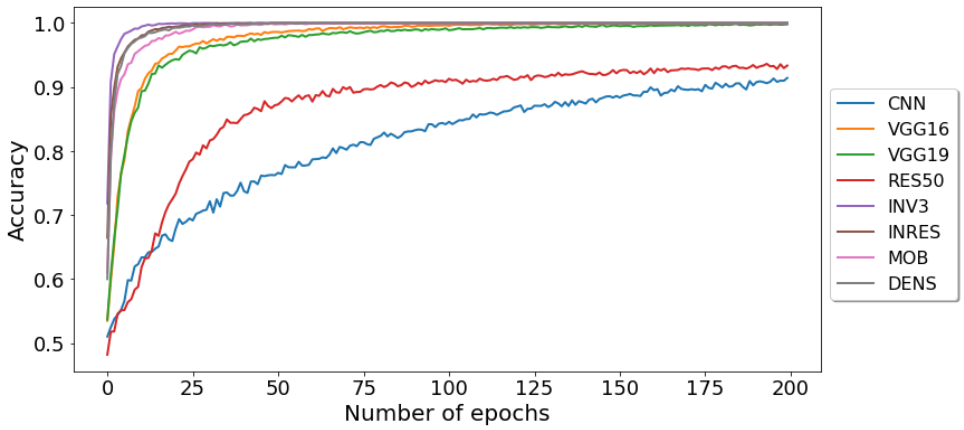

Figure 10: CESC : Training Accuracy.

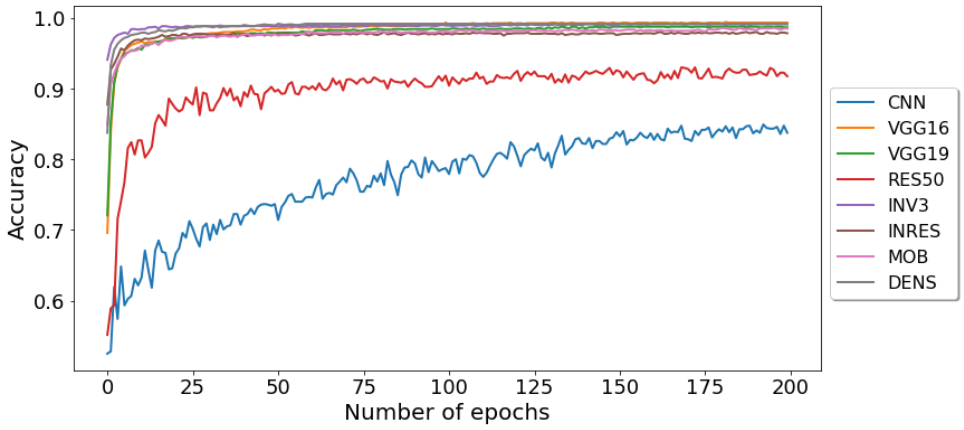

Figure 11: CESC : Validation Accuracy.

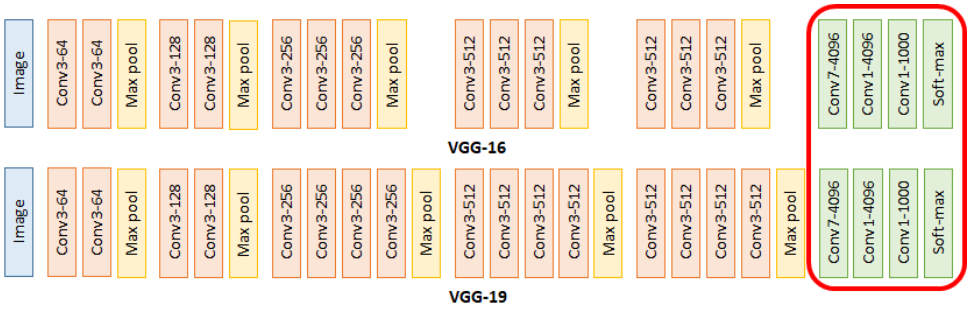

Figure 12: VGG16 and VGG19 architectures.

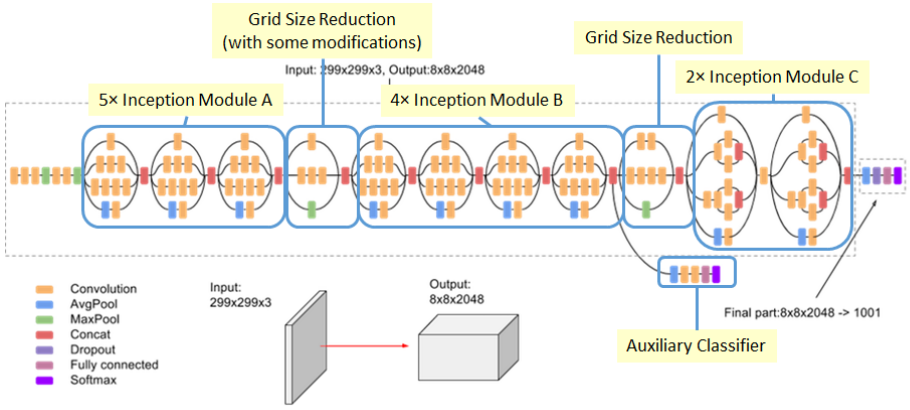

Figure 13: InceptionV3 architecture.

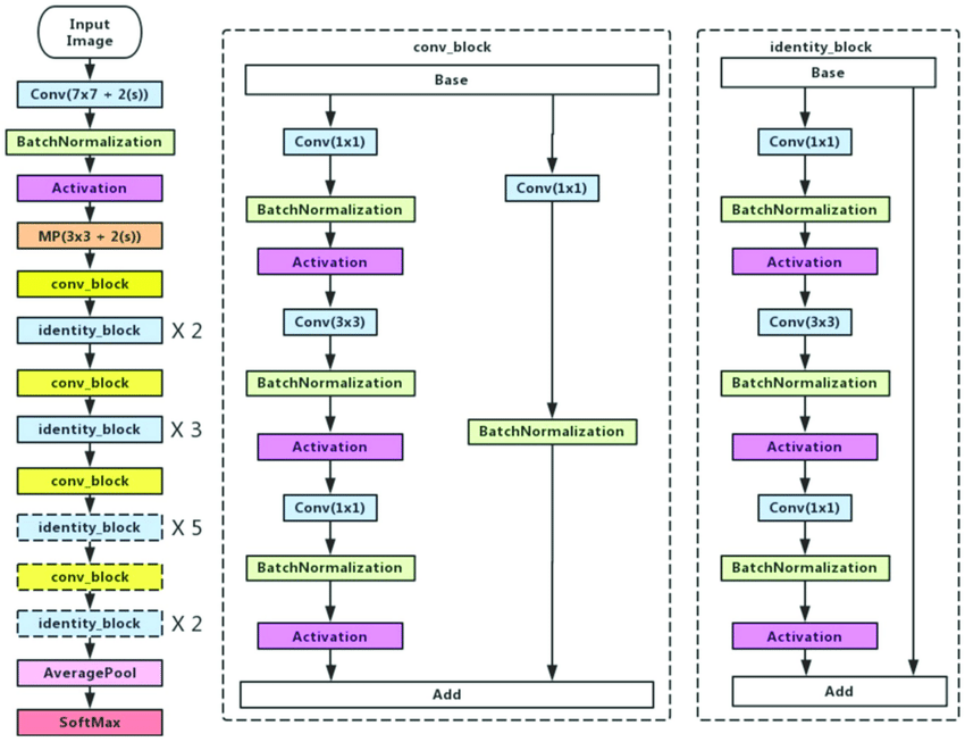

Figure 14: ResNet50 architecture.

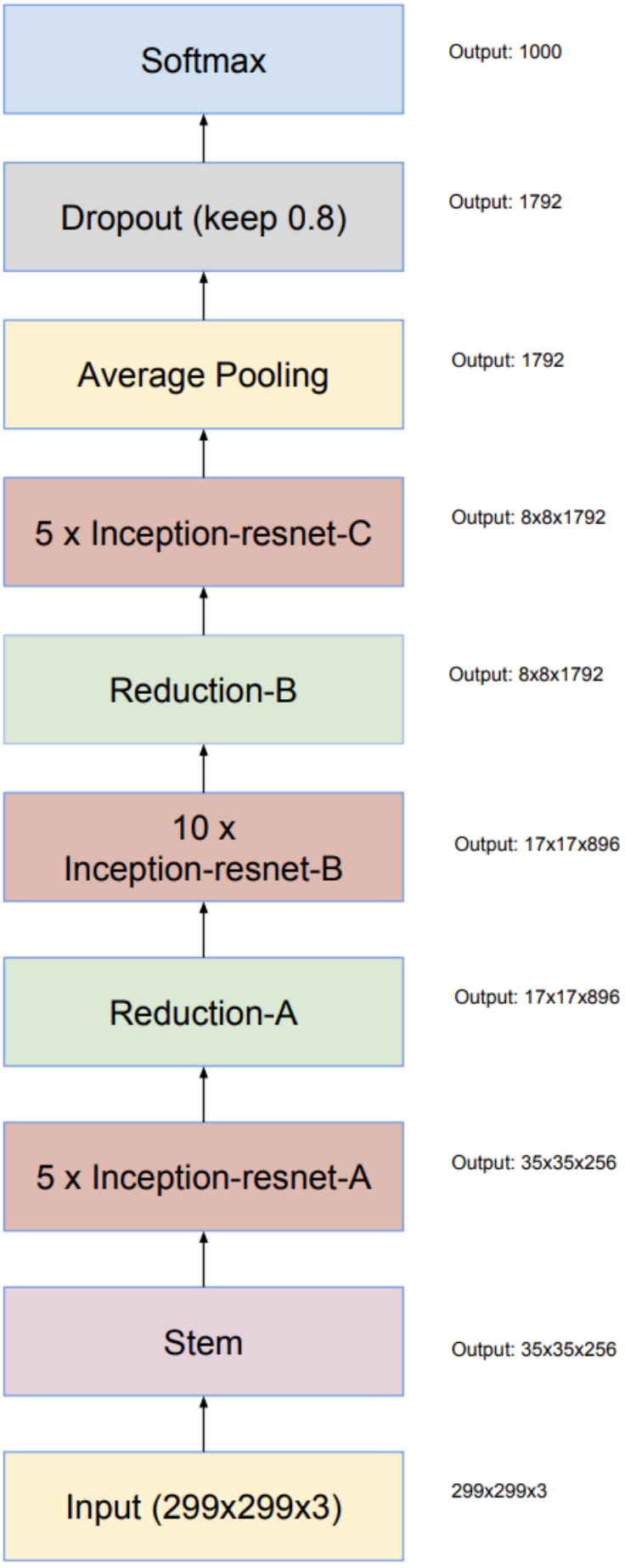

Figure 15: InceptionResnetV2 architecture.

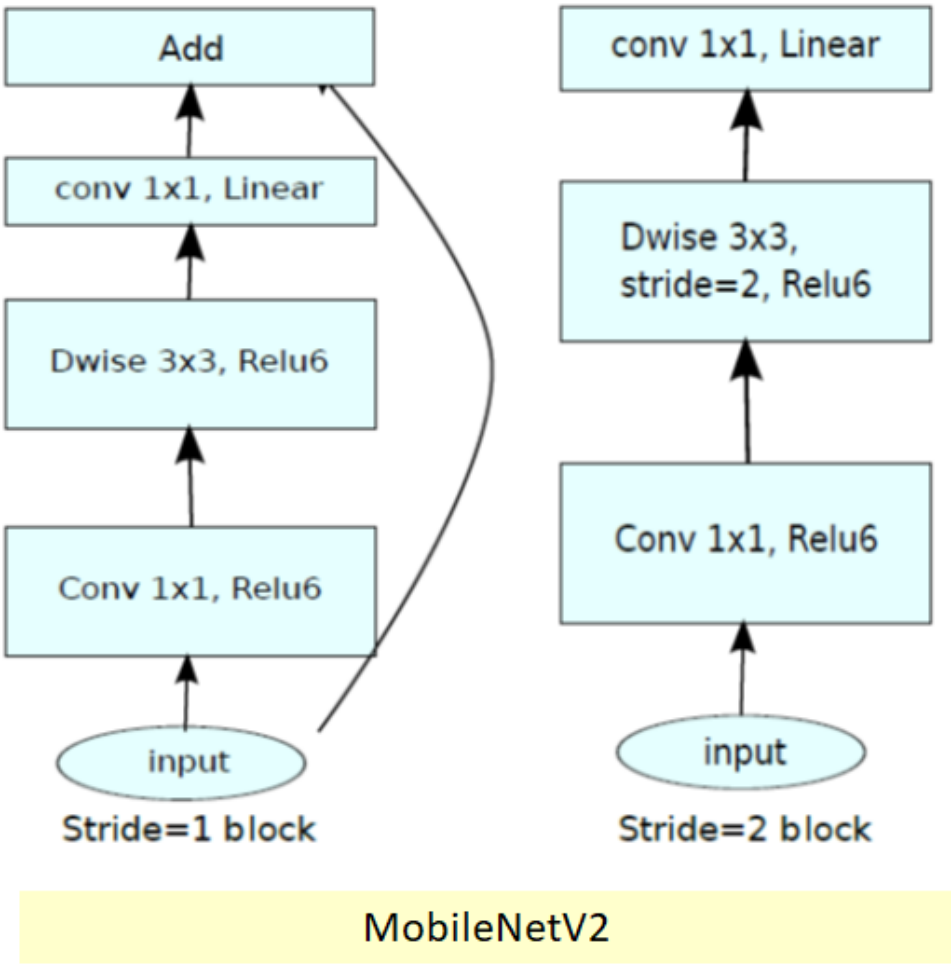

Figure 16: MobileNetV2 architecture.

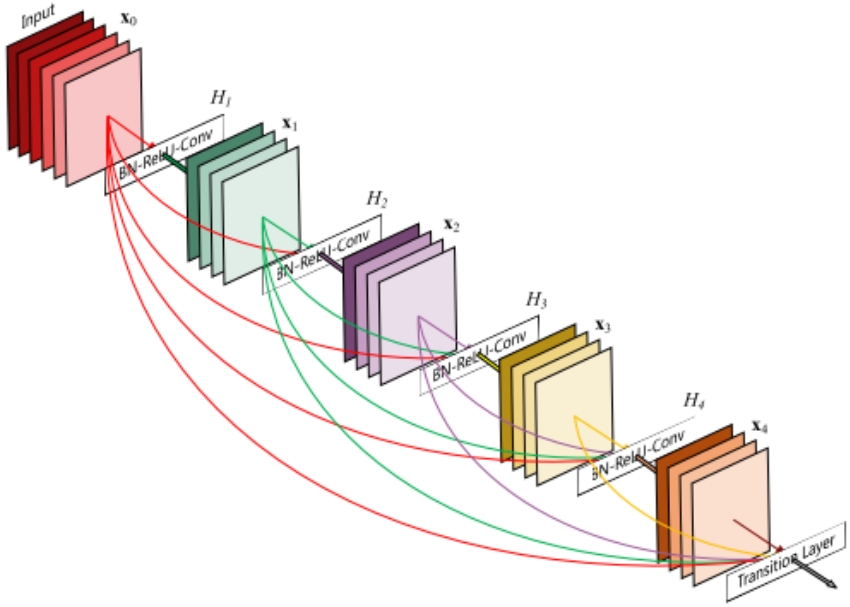

Figure 17: DensNet201 architecture.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", vol. 71, no. 3, pp. 209-215, June 2021.

[2] "Global Burden of Cancer in Women", pp. 15-24.

[3] World Health Organization (WHO), Cervical Cancer, Overview.

[4] T. G. Debelee, S. R. Kebede, F. Schwenker and Z. M. Shewarega, "Deep Learning in Selected Cancers' Image Analysis", pp. 9–10, 2020.

[5] W. A. Musrafa, A. Halim, M. A. Jamlos and S. Z. S. Idrus, "A Review: Pap Smear Analysis Based on Image Processing Approach", Journal of Physics: Conference Series, pp. 1-12, 2019.

[6] Y. R. Park, Y. J. Kim, W. Ju, K. Nam, S. Kim, and K. G. Kim, "Classification of cervical cancer using deep learning and machine learning approach", pp. 1-2, 2021.

[7] A. Tripathi, A. Arora, A. Bhan, "Classification of Cervical Cancer Detection using Machine Learning Algorithms", 2021.

[8] J. Singh, S. Sharma, "Prediction of Cervical Cancer Using Machine Learning Techniques", International Journal of Applied Engineering Research ISSN 0973-4562 vol. 14, no. 11, pp. 2576, 2019.

[9] J. Hyeon, H.J. Choi, K.N. Lee, B.D. Lee, "Automating Papanicolaou Test Using Deep Convolutional Activation Feature", Proceedings of the 2017 18th IEEE International Conference on Mobile Data Management (MDM), Daejeon, Korea, 29 May – 1 June 2017.

[10] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", 2014.

[11] M.V. Valueva, N.N. Nagornov, P.A. Lyakhov, G.V. Valuev, N.I. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation", Mathematics and Computers in Simulation, pp. 232–243, 2020.

[12] Tang, Intelligent Mobile Projects with TensorFlow, Packt Publishing, pp. Chapter 2, May 2018.

[13] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", Microsoft Research Asia, 2015.

[14] J. Jantzen, J. Norup, G. Dounias, B. Bjerregaard, "Pap-smear benchmark data for pattern classification", In Proceedings of the Nature Inspired Smart Information Systems (NiSIS 2005), Albufeira, Portugal, 3–5 October 2005.

[15] E. Hussain, "Liquid based cytology pap smear images for multi-class diagnosis of cervical cancer", Data Brief, 2019.

[16] A. Group "Human Papillomavirus Testing for Triage of Women with Cytologic Evidence of Low-Grade Squamous Intraepithelial Lesions: Baseline Data from a Randomized Trial", J. Natl, pp. 92, 397–402, Cancer Inst, 2000.

[17] M.T. Rezende, A.H.G. Tobias, R. Silva, P. Oliveira, F.S.D. Medeiros, D. Ushizima, C.M. Carneiro, A.G.C. Bianchi, "CRIC Cervix Cell Classification", 2020.

[18] MobileODT, "Intel & Mobile ODT Cervical Cancer Screening", 2017.

[19] R.A.S. Franco, M.A.G. Carvalho, G.P. Coelho, P. Martins, J.L0 Enciso, "Dataset of Cervical Cell Images for the Study of Changes Associated with Malignancy in Conventional Pap Test", ZENODO, 2018.

[20] A.K.H.S. Kurnianingsih, L.E. Nugroho, Widyawan, L. Lazuardi, A.S. Prabuwono, T. Mantoro, "Segmentation and Classification of Cervical Cells Using Deep Learning", IEEE Access, 2019.

[21] H. Lin, Y. Hu, S. Chen, J. Yao, L. Zhang, "Fine-Grained Classification of Cervical Cells Using Morphological and Appearance Based Convolutional Neural Networks", IEEE Access, 2019.

[22] R. Nayar, D.C. Wilbur, "The Pap test and Bethesda 2014", Cancer Cytopathol, 2015.

[23] Y. Promworn, S. Pattanasak, C. Pintavirooj, W. Piyawattanametha, "Comparisons of PAP-Smear Classification with Deep Learning Models", In Proceedings of the 14th annual IEEE International Conference on Nano/Micro Engineering and Molecular Systems, Bangkok, Thailand, 11–14 April 2019.

[24] N. Dong, L. Zhao, C. Wu, J. Chang, "Inception v3 based cervical cell classification combined with artificially extracted features" Appl. Soft Comput, 2020.

[25] E. Hussain, Lipi B. Mahanta, H. Borah, C. Ray Das, "Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions", 2020.

[26] Ferdaous Idlahcen, "DCNN For Uterine Cervical Neoplams Pathology: An Empirical Comparison of Medical Image Classifiers", April 26, 2020.

[27] L. Fei-Fei, J. Deng, and K. Li, "ImageNet: Constructing a large-scale image database", J. Vis., vol. 9, no. 8, pp. 1037–1037, 2010.

[28] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–15, 2015.

[29] A. L. C. Ottoni, E. G. Nepomuceno, M. S. de Oliveira, and D. C. R. de Oliveira, "Tuning of reinforcement learning parameters applied to SOP using the Scott–Knott method", Soft Comput., vol. 24, no. 6, pp. 4441–4453, 2020.

[30] A. Idri, M. Hosni, and A. Abran, "Improved estimation of software development effort using Classical and Fuzzy Analogy ensembles", Appl. Soft Comput. J., vol. 49, pp. 990–1019, 2016.

[31] B. Ghotra, S. McIntosh, A. E. Hassan, "A Large-Scale Study of the Impact of Feature Selection Techniques on Defect Classification Models", pp. 4-10.

[32] I. T. Jolliffe, O. B. Allen, and B. R. Christie, "Comparison of Variety Means Using Cluster Analysis and Dendrograms", vol. 25, pp. 259–269, 1989.

[33] T. Calinski and L. C. A. Corsten, "Clustering Means in ANOVA by Simultaneous Testing", Biometrics, vol. 41, no. 1, p. 39, 1985.

[34] P. Emerson, "The original Borda count and partial voting", Soc. Choice Welfare, vol. 40, no. 2, pp.353–358, 2013.

[35] J. L. García-Lapresta and M. Martínez-Panero, "Borda count versus approval voting: A fuzzy approach", Public Choice, vol. 112, no. 1, pp. 167–184, 2002.

[36] C. Szegedy, V. Vanhoucke, J. Shlens, and Z. Wojna,", Rethinking the Inception Architecture for Computer Vision," 2014.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "the Impact of Residual Connections on Learning", pp. 4278–4284.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks", 2017.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks", Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 4510–4520, 2018.

[40] H. Zerouaoui, A. Idri, and K. El Asnaoui, "Machine learning and image processing for breast cancer: A systematic Map", pp. 1–20.

[41] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning", 2017.

[42] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning", 2017.

[43] A. Idri, I. Abnane, and A. Abran, "Evaluating Pred(p) and standardized accuracy criteria in software development effort estimation", J. Softw. Evol. Process, vol. 30, no. 4, pp. 1–15, 2018.

[44] A. Idri and I. Abnane, "Fuzzy Analogy Based Effort Estimation: An Empirical Comparative Study", IEEE CIT 2017 - 17th IEEE Int. Conf. Comput. Inf. Technol., no. Ml, pp. 114–121, 2017.

[45] Z. Zhang, "Improved Adam Optimizer for Deep Neural Networks", 2018 IEEE/ACM 26th Int. Symp. Qual. Serv. IWQoS 2018, pp. 1–2, 2019.

[46] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," in Advances in Neural Information Processing Systems, 2017.