Classification of Cervical Cancer types

Module: Neural Computing and Deep Learning

Table of Contents

1 Introduction and Background

Cervical cancer is a variant in cancer that can be seen in the cervix part of a women. A cervix is an opening between vagina and womb and a part of their reproductive system. Infection is the cause for most of this type of cancer. This can be prevented by early cervical screening which finds the changes to these cells and treat them before it turns into cancer. The nature of this cancer is it grows very slow, but the severity depends upon the size of these cells developed inside the body. The main cause for this cancer is having weakened immune system and injecting hormonal drugs while a woman is pregnant and due to kidney, bladder cancer in the past.

The prevention and cure for this cervical cancer can be done through cervical screening and vaccination. It is not difficult to prevent this cancer if it caught in the early stages of its development. Every woman who suffers from this can be cured if the cancerous cells detected before it is further developed to result in cancer. Even though cervical screening can cure the cancer but the significant challenge here is to identify the suitable treatment method for each individual because the method of treatment is different to every person which depends on the physiological differences in patient. The patients are getting treatment that doesn't work for them because doctors not able the identify the exact position of the cervix and its type and by result they are unable to decide the correct treatment for the patient.

MobileODT is one of the companies which is bringing the AI in to the cervical cancer screening and supports the world health organisation mission. The workflow of this company will be improved majorly if the capability of identifying the patient's cervix type and mapping to correct treatment is provided. To establish a solution to this, MobileODT along with Intel as its partner challenged the Kagglers to create an algorithm that correctly classifies the cervix type given the lots of woman cervix images. The aim of this challenge is to get the accurate algorithm to classify the cervix type so that the health providers can prevent ineffective treatments and provide exact treatment to the patient.

2 Literature Survey

(Mango, 1994) discussed the applications of computer-based algorithms in detecting the cancer cells in the cervical cancer. His solution describes the usage of neural networks and conventional PAP smear test. The major cause for cervical cancer is Human papilloma virus (HPV). Lack of precise assessing of these types of HPV results in providing ineffective treatment to the patient. (Muñoz, 2003) provides a statistical survey of the woman suffering from cervical cancer. Munoz surveyed by using common protocol and questionnaire. (Punitha, 2016) experiments various architecture of artificial neural network for classification of cervical types. She also given overview of working and detection of cervical cancer for better understanding of the problem context.

(Ghouti, 2020) proposed a pipeline that includes two pretrained models one for cervical detection and later for classification. The results outperformed with best accuracy and performance. (Hossain, 2020) build a model using convolutional neural network for cervix cell detection and classification. They also investigated using autoencoders and other multilayer layer perceptron models for this classification and feature extraction. (Won, 2018) provides a various classification technique and briefs the advantages of feature selection for better classification of cervical types. He also described about sampling techniques such as over and under sampling which helps through the classification of images in the model.

(Rahmadwati and Naghdy, 2011) says texture analysis of the nuclei structure plays key part in classification of cervix types. The team also provided a two-step classification model using the Gabor filter banks localization and abnormality spread in the image. The Gabor filter technique is used to extract the vector of features from the dataset. (Yessi Jusman, 2014) gives a review of advantages and disadvantages of various cervical screening techniques with their implementation. The stages of the computer system such as feature extraction and selection, enhancement and classification are reviewed. (Cömert, 2019) used autoencoders as a deep learning model with SoftMax as a activation for classifying the dataset. The classification rate achieved is 97.8% with best performance classification metrics. And introduced the new methods for diagnosis for the cervical cancer.

3 Data Pre-processing Approach

The dataset used for this project are collection of cervical images. The different images in the dataset are not cancerous. But some patients need further tests because the transformational zones are not always clearly visible to decide the process of treatment. The decision to choose correct treatment for patient is being difficult to the health providers and critical for patient if wrong treatment is given. Based on different transformation zone locations the cervix types can be classified into three types.

The dataset is of two types. One is for training and second one is for testing. The training set contains three folders each for cervix type labelled as Type1, Type2, Type3. The test dataset contains all three types of cervix images without labels. There are total of 1481 cervical images in training set in which 250 images are type1, 781 images are type2 and 450 images are of type cervical type. The test set contains total of 512 cervical images of three cervix types. The distribution of the train dataset of all three types of cervical images are shown below. The visualization shown below is obtained by using matplotlib library on the train dataset.

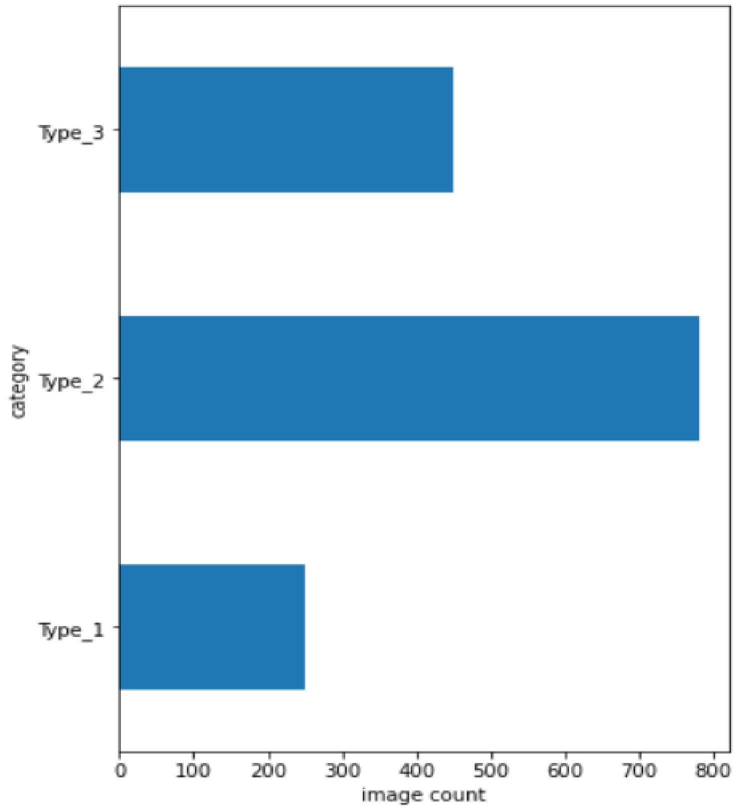


Figure 1: Distribution of Train dataset and cervix types

The sample cervix type image for each type is shown below. The cervical images given are of higher dimensions which ranging from 400 to 4300 pixels. These cervical images are three channelled images of red, green, and blue. The pixel ranges and their distributions using matplotlib are shown below.

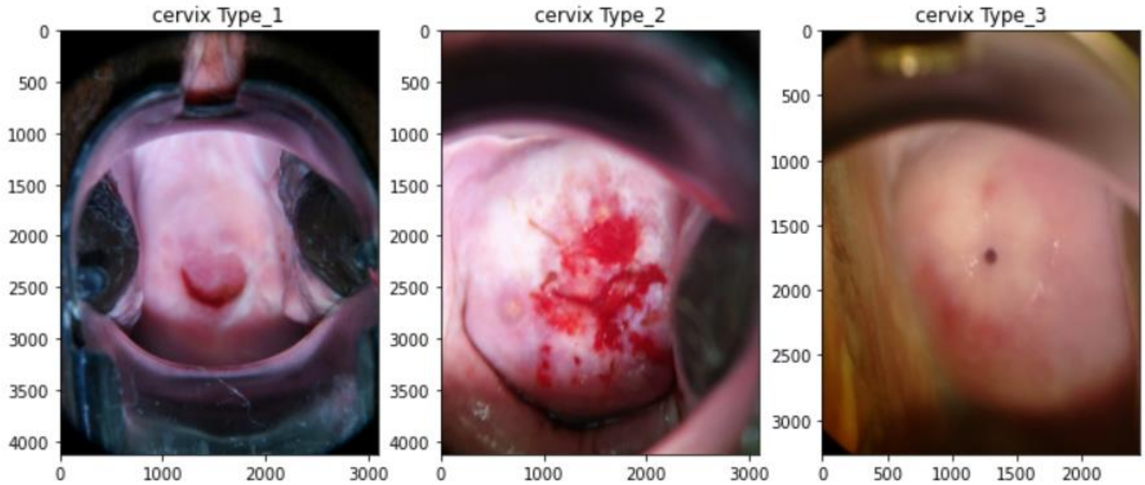


Figure 2: Cervical Images for each type.

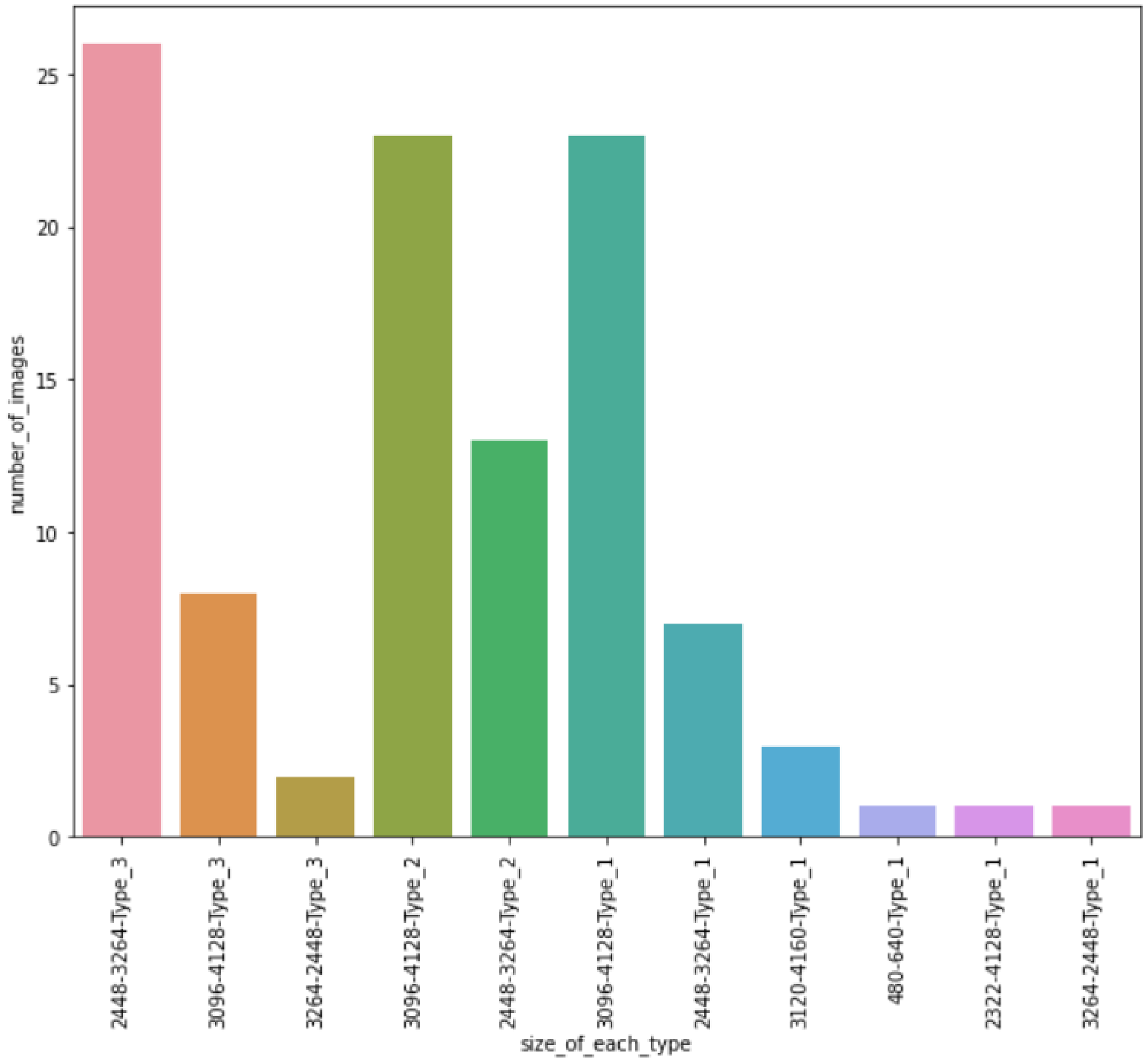


Figure 3: Distribution of cervical images with their dimensions

The higher dimensional images are impractical to input the machine learning model. To overcome this, the cervical images are scaled into smaller dimensions of size into (224,224) pixels size. The resulting cervical images of 224*224 pixels can be fed as input to the model which produced a better model performance in classifying the cervix types. The sample images of the reduced dimensional cervical images are shown below.

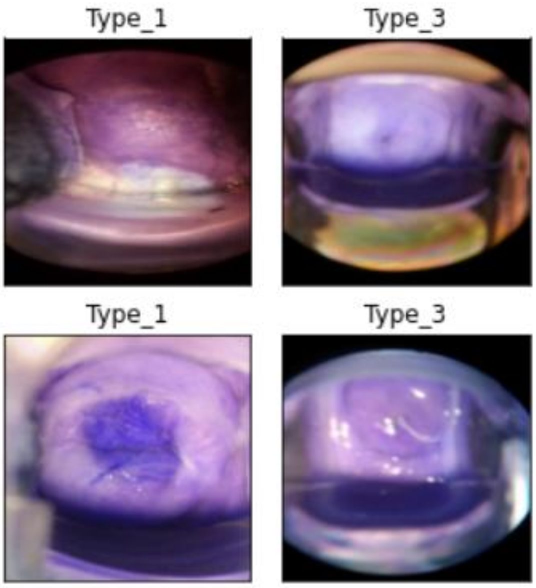


Figure 4: Reduced dimensional cervix images

The train dataset is splitted into three categories such as train-files, Validation-files, and test-files. Train-files are 80% of the entire train dataset and the remaining 20% is allocated for test-files. The train-files are again divided into Validation-files which is 10% of the train-files. Both train and validation files are used for model building and test files is used in model evaluation. The cervical images in train dataset are converted into gray scale images by using scaling from sklearn.

The train dataset images now have a value between 0 and 255 as their pixels. The shape of the train, validation and test files are changed to (1065, 224, 224, 3) (119, 224, 224, 3) (297, 224, 224, 3). The labels for train, validation and test are first converted into labels by using sklearn Label Encoder and the labels are given 0 for type1, 1 for type 2 and 2 for type3 cervix type. Then these labels are converted into one hot encoder values by using built in OneHotEncoder. The test dataset that contains 512 cervical images are transformed into the train datset form with suitable scaling of pixels so that while predicting the model there will be no obstacles to prevent the model prediction on test dataset.

4 Machine Learning Approaches

Image processing and classification is the active areas of research in the field of pattern recognition. In this task to classify the cervical cancer images to their respective types we use convolutional neural networks and transfer learning using pretrained models. Convolutional neural network is a deep learning algorithm which is well suited for Image classification. The application of CNN is well known for achieving solutions for problems such as object detection, image recognition and image classification.

The first model I used here is CNN model. The architecture of CNN model comprises of three convolutional layers and following by three dense layers and one dropout layer at the end before output layer. The three convolutional layers are Conv2D especially because the input to the model are images which are two dimensional. ReLU activation function is used in each layer with convolution size of (5*5) and two as value for strides, padding as same then random normal as kernel initializer. Convolution in CNN is the process in which we take the matrix of values and passed through the image to be transformed from the kernel. The values for feature map are calculated used the below formula where f is input, h is kernel and m, n are rows and columns of result matrix.

$$G[m,n] = (f * h)[m,n] = \sum_{j} \sum_{k} h[j,k] f[m-j, n-k]$$

The value padding width must follow the below requirement where p and f are padding and dimension of filter.

$$P = (f - 1)/2$$

The dimensions of output matrix are calculated based on the stride and the padding by using below formula.

$$Nout = [(Nin + 2p - f)/2 + 1]$$

After each convolutional layer a 2D Max pooling layers is added with pool size of (2*2). 64, 128, 512 filters are used in the three convolutional layers respectively. At the end of convolutional layers, a Flatten layer is added which flattens the values from last convolutional layers and given as input to the Dense layers following in the network. There are three dense layers in the network.

Firstly, a layer with 512 nodes and ReLU as its activation function with random normal as its kernel initializer. The next dense layers consist of 128 node depth with same activation function and kernel initializer. Drop out layer is added to prevent the model from over fitting. The threshold value given for the dropout layer is 0.2. Finally, output layer with three nodes, because the task in hand is to classify the cervical images into three types. SoftMax activation function is used in the output layer because it is a classification problem.

Transfer learning is the second method applied to solve this task. VGG19 pretrained model is a simple CNN model used for ImageNet challenge. This model is developed by the university of oxford in 2014. Visual Geometry group is the acronym for VGG. And 19 in VGG19 implies that there it is a 19 layered architecture. The network architecture for this pretrained model comprise of convolutional blocks including pooling elements.

This model is trained against the image net weights and this training is used to classify the cervical images using some additional fully connected dense layers which is normally called classification layers or network. This custom network includes a flatten layer and one two dense layers, one being fully connected layer with 512 nodes in depth and relu as activation function and the other is output layer with three nodes and SoftMax as activation function. We freeze the pretrained layers and fit the model for the custom images and evaluate the model for performance.

ResNet50 pretrained model is used to classify the cervix types. ResNet50 is 50layered deep CNN pretrained model. These pretrained models can be imported directly from the applications module of keras API. Residual networks are famously called ResNet. The prime speciality of resent is the element of skip connections. They can alleviate the problem of vanishing gradient by bringing up alternate passthrough. Also, this enables the model to learn the identity function which ensures that the front layers of the network perform best than the last layers.

The best performance for this problem to classify the cervix type is from VGG19 pretrained model with accuracy of 91.9%. And CNN best performance is recorded as 71.2% in accuracy. The other experimental analysis of these CNN and pretrained models are shown in below table figure.

5 Results

SNO	OPTIMIZER	Learning Rate	Epochs	Accuracy
1	RMSPROP	0.008	40	91.9%
2	RMSPROP	0.02	20	74.23%
3	SGD	0.01	40	79.45%
4	Adam	0.02	40	71.54%
5	Adam	0.009	50	73.26%

Figure: Performance of VGG19 model

The best performance is from VGG19 model with 91% accuracy and while performing hyperparameter tuning I have observed that increase in learning rate is leading to high convergence and increase in number of epochs is giving better model performance.

SNO	OPTIMIZER	Learning Rate	Epochs	Accuracy
1	SGD	0.01	50	71.23%
2	SGD	0.02	50	67.98%
3	Adam	0.01	50	68.34%
4	Adam	0.02	60	65.34%

Figure: performance of CNN model

The accuracy of CNN model drastically decreasing with increase in epochs and batch size. The model trained with Adam as optimizer and learning rate 0.02 got the least accuracy of 65%. Even though the batch size variation and increase in depth of convolution provided better performance but the model is overfitted and leading to poor generalization of model.

SNO	OPTIMIZER	Learning Rate	Epochs	Accuracy
1	Adam	0.01	40	67.23%
2	Adam	0.02	55	62.36%
3	SGD	0.01	40	60.56%
4	SGD	0.02	40	61.54%

Figure: Performance of ResNet50

Resnet50 model best performance in this experiment is 67%. The problem encountered while training resnet50 is that it is time consuming and the loss the converging frequently which is resulting in lowering the accuracy of the model.

The best performance of the VGG19 pretrained model for classifying the cervix type is shown in below graph. The graph depicts the model loss and accuracy for both training data and validation data.

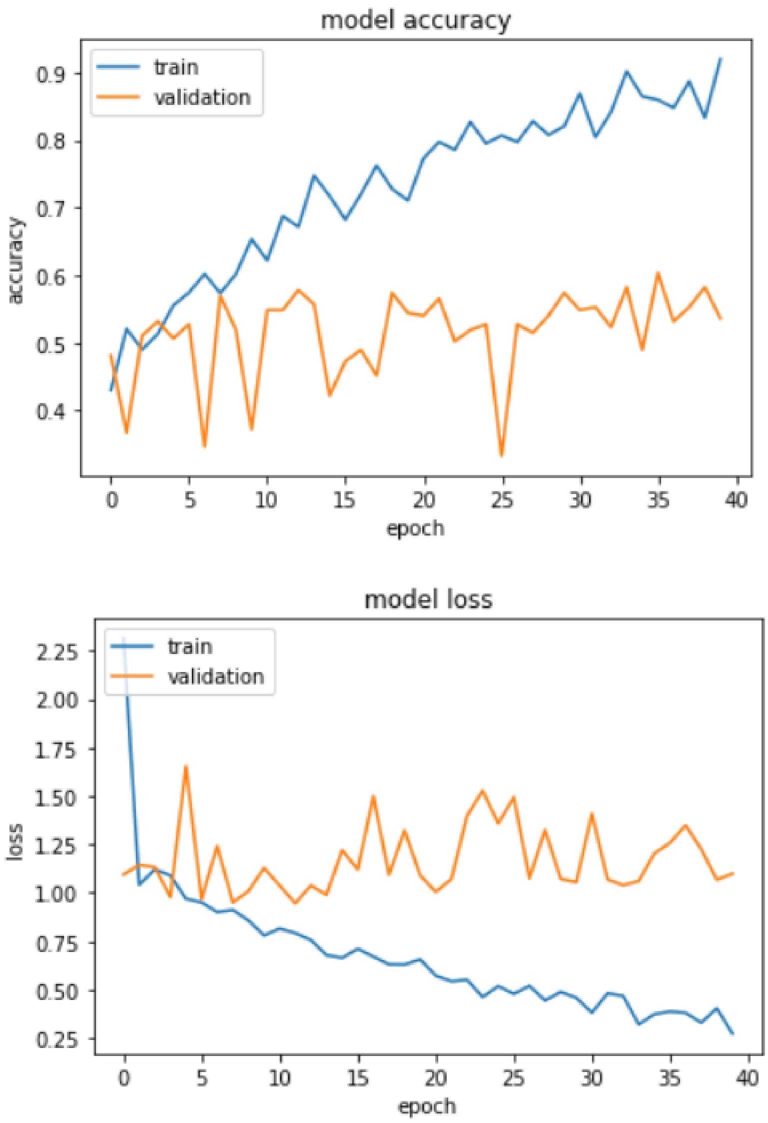
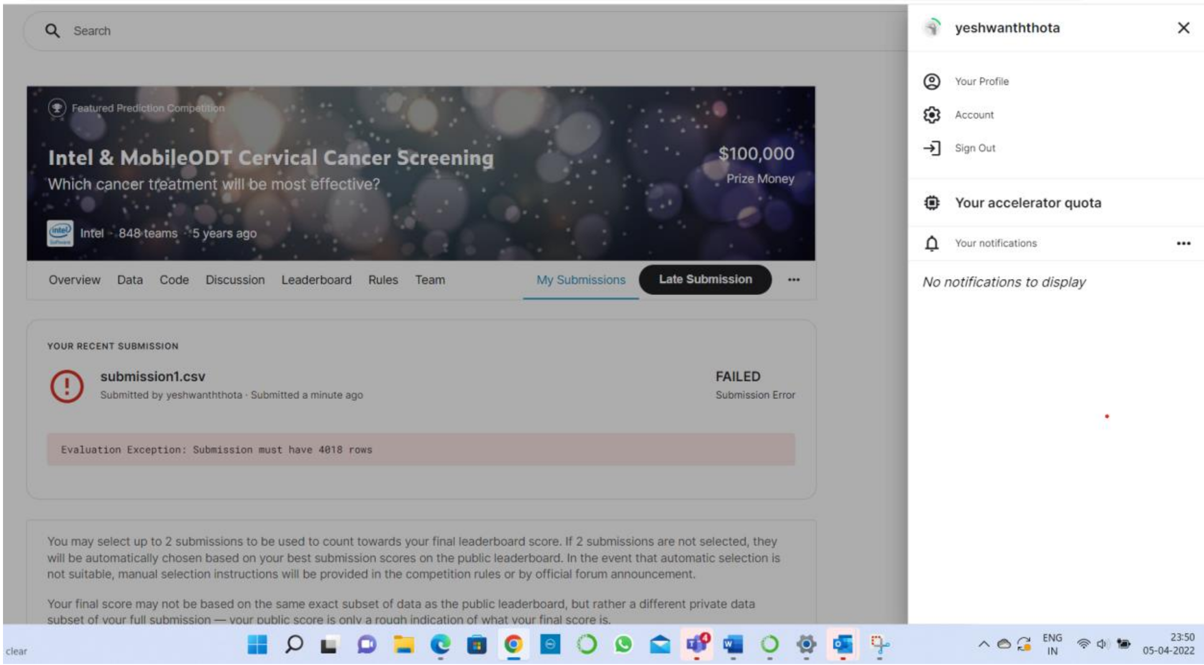


Figure: Accuracy and loss for VGG19 model

The accuracy of both training and validation are in same direction which is a good sign that model is performing better if not best. The loss also decreasing through each epoch even though there are enough divergence during the training.

6 Kaggle Submission Proof

The following attached is the submission proof for Kaggle challenge



7 Future Work

In this paper, deep learning models such as CNN and transfer learning is used to classify the cervix types. Though the models produced good performance the lack of balance between data such as type 2 cervix types of images are more when compared to other two types type1 and type2, this resulted in some bias in data. With balanced data and other features such as using lower dimensional data without information loss can provide best solutions with application of deep learning algorithms.

References

Cömert, K. A. a. S. K. a. O., 2019. Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Systems with Applications,* Volume 115, pp. 557-564.

Ghouti, Z. A. a. L., 2020. A fully-automated deep learning pipeline for cervical cancer classification. *Expert Systems with Applications,* Volume 141, p. 112951.

Hossain, A. G. a. G. M. a. M. S., 2020. Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems,* Volume 102, pp. 643-649.

Mango, L. J., 1994. Computer-assisted cervical cancer screening using neural networks. *Cancer Letters,* Volume 77, pp. 155-162.

Muñoz, N. a. B. F. X. a. d. S. S. a. H. R. a. C. X. a. S. K. V. a. S. P. J. a. M. C. J., 2003. Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer. *New England Journal of Medicine,* Volume 348, pp. 518-527.

Punitha, M. A. D. a. S. R. a. J. V. a. S., 2016. Classification of Cervical Cancer Using Artificial Neural Networks. *Procedia Computer Science,* Volume 89, pp. 465-472.

Rahmadwati and Naghdy, G. a. R. M. a. T. C. a. N. E., 2011. *Cervical Cancer Classification Using Gabor Filters.* San Jose, CA, USA, IEEE.

Won, A. C. a. Y. M. S. A. W. a. D., 2018. Classification of Cervical Cancer Dataset. *CoRR,* Volume abs/1812.10383.

Yessi Jusman, S. C. N. N. A. A. O., 2014. Intelligent Screening Systems for Cervical Cancer. *The Scientific World Journal,* Volume 2014, p. 15.