Communication in Physical Sciences, 2024, 12(2): 175-193

Conceptual Design Of A Hybrid Deep Learning Model For Classification Of Cervical Cancer Acetic Acid Images

Fatima Binta Adamu*, Muhammad Bashir Abdullahi, Sulaimon Adebayo Bashir, and Abiodun Musa Aibinu

Received: 04 November 2024/Accepted: 30 January 2024/Published: 05 February 2025
https://dx.doi.org/10.4314/cps.v12i2.2

Abstract: *Automated image-based cervical cancer detection plays a vital role in diagnosing cervical cancer, particularly through the use of digital cervical images obtained via visual inspection with acetic acid (VIA). Many algorithms have been developed to classify these images by extracting mathematical features. Artificial intelligence (AI) has significantly advanced healthcare by improving disease detection, diagnosis, and prediction of health outcomes. While various cervical cancer screening methods have evolved, VIA remains one of the most feasible options in low-resource settings. However, its effectiveness relies heavily on the examiner's experience, which can be limited due to a shortage of qualified healthcare professionals. This study evaluates the performance of AI image processing techniques for detecting cervical cancer using VIA images. The research compares four traditional machine learning techniques and six deep learning techniques in classifying cervical cancer images, where each model was trained on four randomly selected batches of images (300, 700, 1000, and 1678 images) to assess model performance with an increasing number of training images. The VGG19 model achieved a consistent accuracy of 81% across all training sizes. The Vision Transformer (ViT) model, on the other hand, showed a performance improvement from 57% accuracy with 300 images to 77% accuracy with 1678 images. The hybrid model, combining VGG19 and ViT, demonstrated superior performance with an accuracy of 86.65%, an AUC of 0.85, a sensitivity of 0.832, and a specificity of 0.8485. This study demonstrates that the hybrid model outperforms individual models, offering a promising solution for cervical cancer detection in low-resource environments.*

Keywords *- Artificial Intelligence, Image Processing, Cervical Cancer, Visual Inspection Acetic Acid, Cervical Intraepithelial Neoplasia.*

Fatima Binta Adamu
Department of Computer Science, Federal University of Technology, Minna, Niger State, Nigeria
Email: fatimabinta@futminna.edu.ng
Orcid id: 0000-0001-8803-1618

Muhammad Bashir Abdullahi
Department of Computer Science, Federal University of Technology, Minna, Niger State
Email: el.bashir02@futminna.edu.ng

Sulaimon Adebayo Bashir
Department of Computer Science, Federal University of Technology, Minna, Niger State
Email: bashirsulaimon@futminna.edu.ng

Abiodun Musa Aibinu
Department of Mechatronics Engineering, Federal University of Technology, Minna, Niger State
Email: maibinu@gmail.com

1.0 Introduction

Cervical cancer is the fourth most frequent malignancy in women globally. The World Health Organization (WHO) noted that about 28 million cancer-triggered mortality would have happened by the end of the year 2020. In particular, cancers of the breast and lungs are the leading cause of mortality amongst cancer sufferers globally, followed by cervical cancer (Sharma et al, 2023). Existing cervical cancer screening programs, such as Pap Smear, have limited sensitivity. As a result, many positive cases are missed during the screening procedure. Also, it is not feasible in low-resource settings because of the financial implications of a sophisticated laboratory and medical resources. Cervical cancer visual inspection with acetic acid (VIA) screening with images of the cervix taken during screening and analysed with Computer Aided Diagnostic (CAD) systems, has the potential to considerably improve screening programs, and it can be especially effective in resource-poor areas of the world (Song, et al., 2013). Approximately 90% of cervical cancer deaths occur in underdeveloped nations, owing to the high expense of undertaking regular screening programs, a lack of resources, and a scarcity of experts (Kudva, et al., 2018). Several machine learning and deep learning approaches have been proposed as a means of mitigating this problem (Kudva et al., 2020), but most of them either have poor accuracy or are not feasible for implementation in low-resource settings. Hence, there is a need for more research on artificial intelligence-based approaches that can be deployed for use in low-resource settings. Also, Liang, et al. (2013) highlighted that cervical cancer is a danger to all women. This suggests a need for models that can be integrated into mobile devices (Adamu et al., 2020), and can be used globally as a simple and efficient cervical cancer screening deployment tool. This paper contributes to the literature by providing new information, perspective, and evaluation of machine learning and deep learning techniques when used to classify cervical cancer visual inspection with acetic acid images. It also highlights the performance when two powerful deep learning techniques (VGG19 and Vit) are merged and used for cervical cancer image classification.

2.0 Related Works

Cervical cancer constituted 14.8% of more than 70,327 female cancer-related fatalities in Nigeria in 2018, making it the second most prevalent malignancy following breast cancer (Aina et al., 2018). The application of acetic acid to the cervix causes the whitening of the epithelium known as acetowhitening, which is essential for detecting aberrant regions during cervical cancer screening (Azene, 2021). Notwithstanding its importance in low-resource environments, diagnostic accuracy continues to be problematic due to dependence on examiner proficiency. Several models and approaches have been proposed to improve cervical cancer image classification accuracy during CAD screening.

Azene (2021) and Asiedu et al. (2019) concentrated on the preparation of cervigrams and the extraction of colour and textural information for automated lesion categorization, thereby improving the efficacy of visual inspection with acetic acid (VIA). Their techniques prioritized enhanced diagnostic precision through feature-based methodologies. Balas (2001) also created a multispectral imaging method to measure alterations in the light-scattering characteristics of the cervix, emphasizing neoplasia caused by acetic acid. This method enhanced the identification of cervical intraepithelial neoplasia (CIN).

Das et al. (2014) employed image segmentation techniques to identify malignant cervical lesions. Their research illustrated the promise of effective segmentation for early cancer detection.

Shu (2019) introduced a modified deep convolutional neural network (D-CNN) for classifying cervical pictures from limited datasets, reducing overfitting and showcasing deep learning's versatility in resource-constrained environments. Kaur et al. (2017) highlighted the relevance of colposcopy in cervical cancer diagnosis but stressed the need for colposcopy expertise. They called for computer-assisted diagnostic techniques to improve accuracy and dependability.

Krizhevsky et al. (2012) introduced a breakthrough in deep learning by building a CNN capable of classifying high-resolution photos from ImageNet. Their methods created a framework for employing CNNs in medical imaging, including cervical cancer screening.

Kudva et al. (2018) examined mathematical feature extraction and classification for discriminating between malignant and non-malignant cervical pictures, paving the road for integrating machine learning into early cancer detection. Later, Kudva et al. (2020) demonstrated that hybrid transfer learning employing pre-trained models like VGG-16 achieved great accuracy (up to 91.46%) in cervical cancer detection.

Liang et al. (2013) suggested an automatic method to discover problematic cervical areas using colposcopic picture sequences, combining segmentation with a support vector machine (SVM) classifier for accurate predictions.

Priya (2014) prioritized precise image segmentation to identify cervical cancer lesions, boosting biopsy targeting and overall diagnostic outcomes.

RamaPraba and Ranganathan (2012) employed statistical characteristics and a Bayes classifier to detect lesions in colposcopy pictures. Their preprocessing procedures effectively identified AcetoWhite (AW) regions, which are markers of aberrant cervical cells. Further, RamaPraba and Ranganathan (2013) presented an active contour-based lesion detection method employing wavelet transformations, producing encouraging results in automated lesion diagnosis.

Rouhbakhsh et al. (2012) trained classifiers, including KNN and Neuro-Fuzzy networks, to detect precancerous lesions using texture and colour characteristics, displaying better diagnostic accuracy with optimized feature selection.

Simonyan and Zisserman (2014) examined convolutional network architectures for action recognition in video, revealing ideas that may be applied to evaluating cervical image sequences dynamically.

Also, Sukumar and Gnanamurthy (2016) created a computer-aided method for cervical cancer detection utilizing wavelet transforms and random forest classifiers, exceeding conventional techniques in classification accuracy.

Xu et al. (2017) curated an expert-annotated cervical illness dataset, employing pyramid histogram characteristics such as PLAB, PHOG, and PLBP to better picture categorization.

Xue et al. (2010) focused on detecting mosaic vasculature patterns in cervical images, solving obstacles including blurry boundaries and small artery calibres to aid gynaecologists in identifying anomalies.

Srinivasan (2019) suggested a unified diagnostic strategy for CIN, using Gaussian mixture modelling (GMM) for segmentation and texture-based classification, underlining the relevance of computational models in enhancing diagnostic accuracy. Raifu et al. (2017) studied the sensitivity and specificity of VIA and VILI as diagnostic approaches, demonstrating the potential for computer-assisted diagnostics to boost accuracy across varied situations.

3.0 Method

3.1 Data Collection and Ground Truth

166 cervix images were collected from the National Cancer Institute (NCI) cervical cancer image database which consisted of 92 VIA-negative and 74 VIA-positive images. Some data augmentation techniques were utilized to increase the size of the datasets and to prevent overfitting. The augmentation techniques include vertical and horizontal flips, random brightness, image shifting, random rotation, and image zooming. A total number of 1678 images were obtained after data augmentation consisting of 938 VIA-negative and 740 VIA-positive images. Fig. 1 depicts images of the cervix before pre-processing.

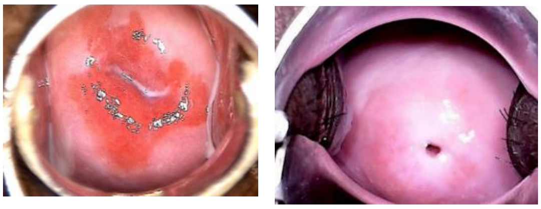
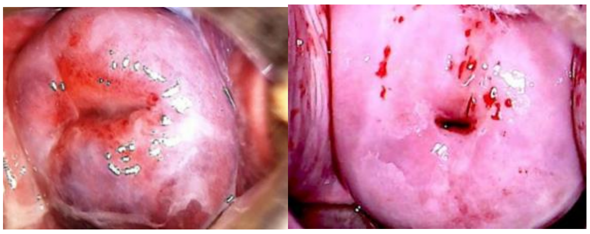


Fig. 1: Cervix image (VIA Negative) & (VIA Positive)

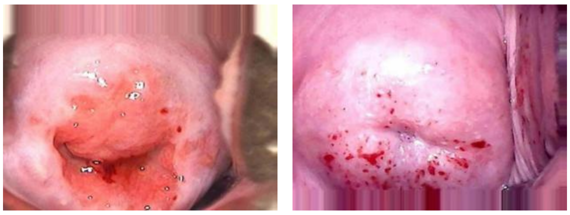
3.2 *Image pre-processing*

To ensure that unnecessary or unwanted features such as the image of the speculum do not affect the accuracy of the cervical image classification, the Region of Interest (RoI) was cropped using a minimal bounding box around the cervical area. Fig. 2 depicts the cropped images, while Fig. 3 presents the augmented images after cropping.



Positive (Malignant) Negative (Benign)

Fig. 2: Cropped images of the dataset



Positive (Malignant) Negative (Benign)

Fig. 3: Augmented images of the dataset

The pre-processed images are then converted to grayscale images for the extraction of features for the machine learning classification.

3.3 Automated Techniques for Cervical Image Classification

The application of acetic acid to the cervix causes textured mosaicism and highlights various vascular patterns which can be manually interpreted for diagnosis (Asiedu et al., 2018), or calculated during automated image analysis using the Haralick textural features (Haralick et al., 1973).

3.3.1Feature Extraction (Haralicks' Features)

Haralick's textural features were used in this study to determine the Gray-Tone Spatial-Dependence Matrix (GSDM) in the grayscale images. The GSDM is a statistical method that calculates the frequency of occurrence of a pair of pixels with specific values and spatial relationships in an image. the GSDM was computed for four different pixel offsets (1, 5, 10, and 15) in four different directions (0, 45, 90, and 135 degrees). A total of 954 textural features were calculated from these GSDMs: contrast, correlation, dissimilarity, energy, and homogeneity (Haralick et al., 1973).

The traditional approaches used in this study seek to identify, sort, and separate the characteristics of each image class (VIA positive and negative), with the predicted procedure in feature identification based on margin, texture, and whiteness information. Feature extraction is a necessary step before employing traditional machine learning algorithms. The technique of decreasing the amount of picture data by extracting required information from the segmented image is known as feature extraction. It is feasible to differentiate between positive and negative cervical VIA using image data-derived characteristics. The classification algorithm's dependability is determined by the retrieved features. The texture features in this study are retrieved using GSDM. Energy, correlation, dissimilarity, homogeneity, and contrast are texture characteristics computed using GSDM, the formula and an explanation of how each characteristic is used are noted below (Haralick et al., 1973):

i. *Energy*: It employs the texture that calculates ordering in an image, yielding the sum of square elements in GDSM. It is not the same as entropy. When the image window, which serves as the sample region for GDSM tabulations and texture calculations is well-organized, the energy value is high. As Energy, the square root of the Angular Second Moment (ASM) texture character is utilized. It has a range of [0 1]. Its value is 1 since it is a constant image. The energy equation is as presented in equation (1).

$$\sum_{i,j=0}^{N-1} P_{i,j}^{2} \qquad (1)$$

ii. *Correlation:* It applies the computation of a pixel's association with its neighbour across the entire image, determining the linear dependency of grey levels on those of neighbouring pixels. The correlation value for a fully positively or negatively linked image is 1 and -1. Its value is NaN in the case of a constant image. The range is [-1,1], and the formula is as depicted in equation (2).

$$\sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right] \qquad (2)$$

iii. *Dissimilarity:* Dissimilarity is a distance measure between two objects (pixels) in the region of interest. The formula for dissimilarity is presented in equation (3).

$$\sum_{i,j=0}^{N-1} P_{i,j}\,|i-j| \qquad (3)$$

iv. *Homogeneity:* It sends the value calculated by the tightness of distribution of the GDSM elements to the GDSM diagonal. The diagonal GDSM has a value of 1 and a range of [0,1]. Homogeneity weight values are the inverse of contrast weight values, with weight decreasing exponentially away from the diagonal. In comparison, the weight used is $(i-j)^2$ and inhomogeneity, it is $1/1+(i-j)^2$. The homogeneity equation is presented in equation (4).

$$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2} \qquad (4)$$

v. *Contrast:* Contrast is also known as the Sum of Square Variance. It postpones the computation of the intensity contrast between a pixel and its neighbour throughout the whole image. When the image is constant, the contrast value is 0. In contrast, when one moves away from the diagonal, the weight grows exponentially (0,1,4,9). The equation for contrast is seen in equation (5).

Range = [0, size (GDSM,1)-1)^2]

$$\sum_{i,j=0}^{N-1} P_{i,j}\,(i-j)^2 \qquad (5)$$

A total number of 954 features were extracted from the converted grayscale images. Fig. 4 presents the grayscale images. The machine learning models were trained on these GDSM features. Fig. 5 is an image of the extracted GDSM features.

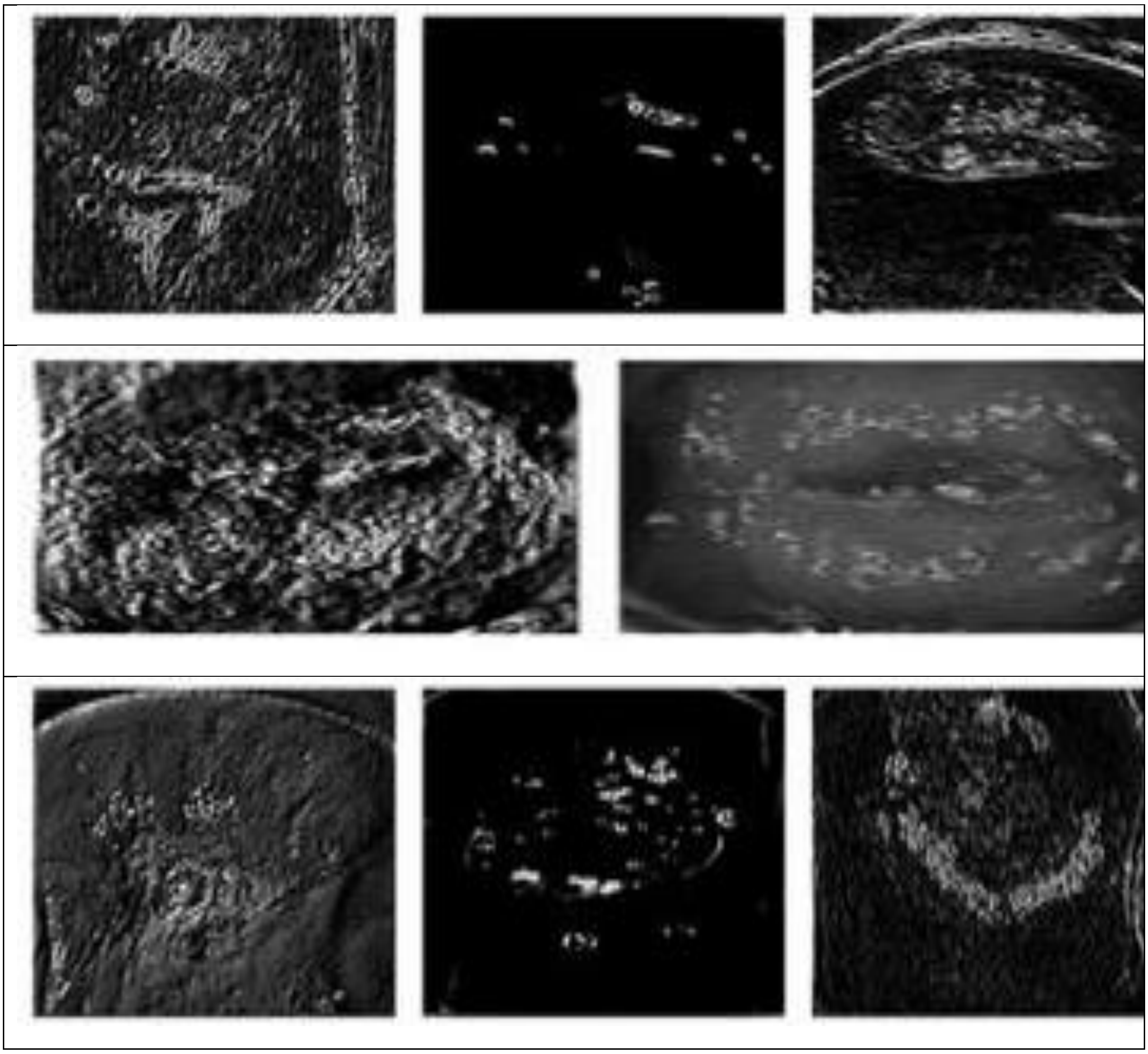


Fig. 4.: Converted cervical cancer grayscale images

3.3.2 Traditional techniques

3.3.2.1 Random Forest Classifier

Random forests, also known as random decision forests, are an ensemble learning method for classification and regression that pre-trains a set number of trees. Because of its inherent variable selection, random forest predictors naturally lead to a dissimilarity measure among the features in data as part of the algorithm-building process; the random forest dissimilarity easily deals with a large number of semi-continuous variables. The images to be used are fed into the numpy reshape function, which transforms them into a specific shape and feature. For repeatable results, a total of 50 decision trees and a random state of 42 are used. We test our model with the test image datasets after training it on our training datasets to see how well it performs on previously unseen image data. The model's accuracy after running on the test data is 48.5%.

3.3.3.2 Support Vector Classifier (SVC)

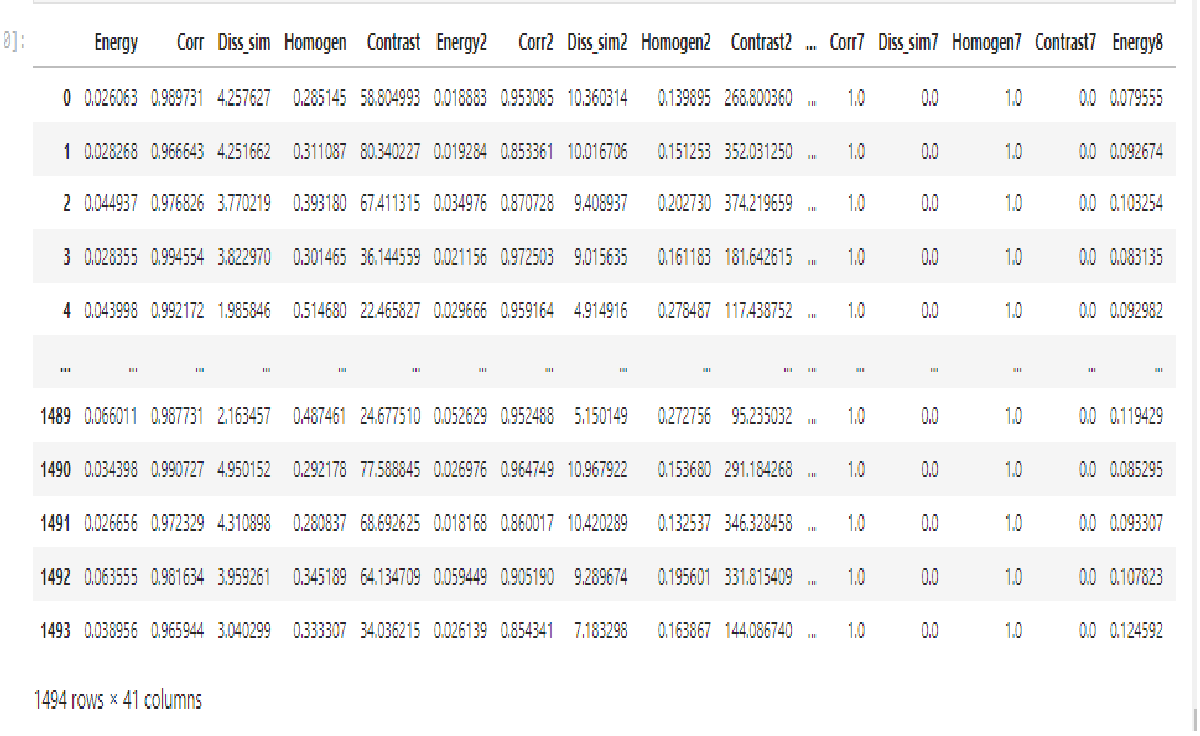
The effectiveness of SVC is based on features like kernels, kernel parameters, and soft margin. It has proven to be fast and effective in a variety of tasks. The model's accuracy is 54.3%, which is slightly better than the Random Forest Classifier.

3.3.2.3Light Gradient Boosting Machine (LightGBM)

LightGBM is associated with many algorithms, one of which is XGBoost, which offers features like sparse optimization, parallel training, multiple loss functions, regularization, bagging, and early stopping. The classification accuracy obtained from the lightGBM model is 60%.

3.3.2.4Artificial Neural Networks

Artificial Neural Networks (ANNs) are networks of artificial neurons that are inspired by the biological neural networks that make up animal brains. An artificial neuron receives the signal, processes it, and communicates with other neurons. Neurons are usually assigned a weight that changes over time as the learner progresses. The accuracy of the ANN model was 57.5%. Table 1 summarizes the classification performance of the traditional machine learning techniques compared in this study.



```
0]:        Energy      Corr  Diss_sim   Homogen   Contrast   Energy2     Corr2  Diss_sim2  Homogen2   Contrast2  ...  Corr7  Diss_sim7  Homogen7  Contrast7   Energy8
      0  0.026063  0.989731  4.257627  0.285145  58.804993  0.018883  0.953085  10.360314  0.139895  268.800360  ...    1.0        0.0       1.0        0.0  0.079555
      1  0.028268  0.966643  4.251662  0.311087  80.340227  0.019284  0.853361  10.016706  0.151253  352.031250  ...    1.0        0.0       1.0        0.0  0.092674
      2  0.044937  0.976826  3.770219  0.393180  67.411315  0.034976  0.870728   9.408937  0.202730  374.219659  ...    1.0        0.0       1.0        0.0  0.103254
      3  0.028355  0.994554  3.822970  0.301465  36.144559  0.021156  0.972503   9.015635  0.161183  181.642615  ...    1.0        0.0       1.0        0.0  0.083135
      4  0.043998  0.992172  1.985846  0.514680  22.465827  0.029666  0.959164   4.914916  0.278487  117.438752  ...    1.0        0.0       1.0        0.0  0.092982
    ...       ...       ...       ...       ...        ...       ...       ...        ...       ...         ...  ...    ...        ...       ...        ...       ...
   1489  0.066011  0.987731  2.163457  0.487461  24.677510  0.052629  0.952488   5.150149  0.272756   95.235032  ...    1.0        0.0       1.0        0.0  0.119429
   1490  0.034398  0.990727  4.950152  0.292178  77.588845  0.026976  0.964749  10.967922  0.153680  291.184268  ...    1.0        0.0       1.0        0.0  0.085295
   1491  0.026656  0.972329  4.310898  0.280837  68.692625  0.018168  0.860017  10.420289  0.132537  346.328458  ...    1.0        0.0       1.0        0.0  0.093307
   1492  0.063555  0.981634  3.959261  0.345189  64.134709  0.059449  0.905190   9.289674  0.195601  331.815409  ...    1.0        0.0       1.0        0.0  0.107823
   1493  0.038956  0.965944  3.040299  0.333307  34.036215  0.026139  0.854341   7.183298  0.163867  144.086740  ...    1.0        0.0       1.0        0.0  0.124592

1494 rows × 41 columns
```

Fig. 5: Extracted GDSM features

3.3.3 Comparison of the traditional Machine Learning (ML) Techniques utilized

Table 1: Performance of the ML techniques utilized

Model	**Accuracy**
Random Forest Classifier (RFC)	48.5%
Support Vector Machine (SVM)	54.3%
Light Gradient Boosting Machine (LGBM)	60%
Artifial Neural Network (ANN)	57.5%

3.3.4 Deep Learning Technique

Deep learning (also known as deep structured learning) is a machine learning method that is based on artificial neural networks and representation learning. Learning can take place in a supervised, semi-supervised, or unsupervised environment(Lecun et al., 2015). Deep learning also has an advantage over traditional transfer learning techniques. Transfer learning is a method used when there is insufficient data or computational power to predict a pre-trained model using a different dataset; the model will be fine-tuned to provide the best performance on the preferred data. The transfer learning models used for this project are VGG16, VGG19, Alexnet, ViT, Efficient Net and Resnet.

3.3.4.1 Resnet34

After preparing all of the necessary libraries, the model's training ratio is set to 90% and passed into the pre-trained model, which is set to true so that the layers used in the original deep learning model are preserved for optimum performance; additionally, the batch size and the number of classes are set to 16 and 2, respectively. After importing the data loader, we iterate once through it, then create a fully connected layer to match our data and classes, after which we set the number of epochs to 50 and create the trainer and evaluator. The fed image data is then fed into the pre-trained model, which is then run.

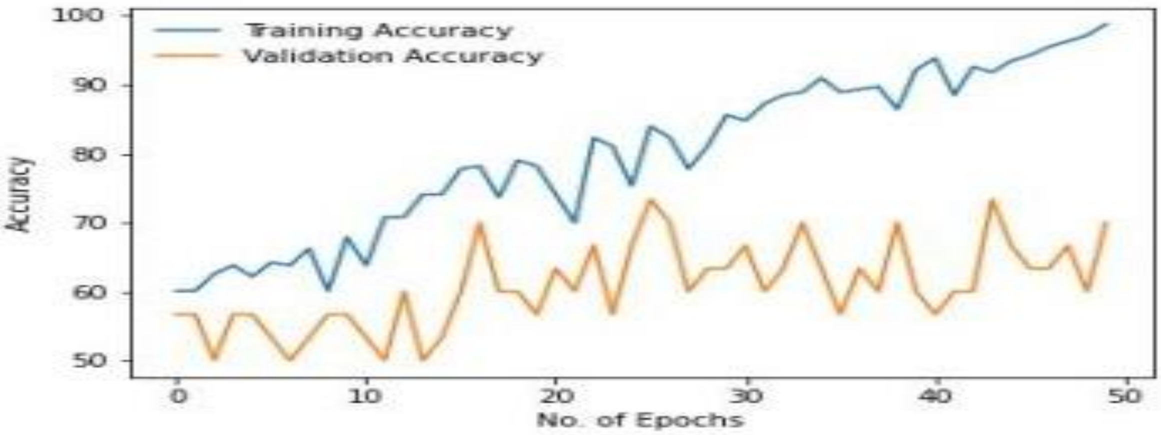


Fig. 6: ResNet training and validation accuracy (300 images)

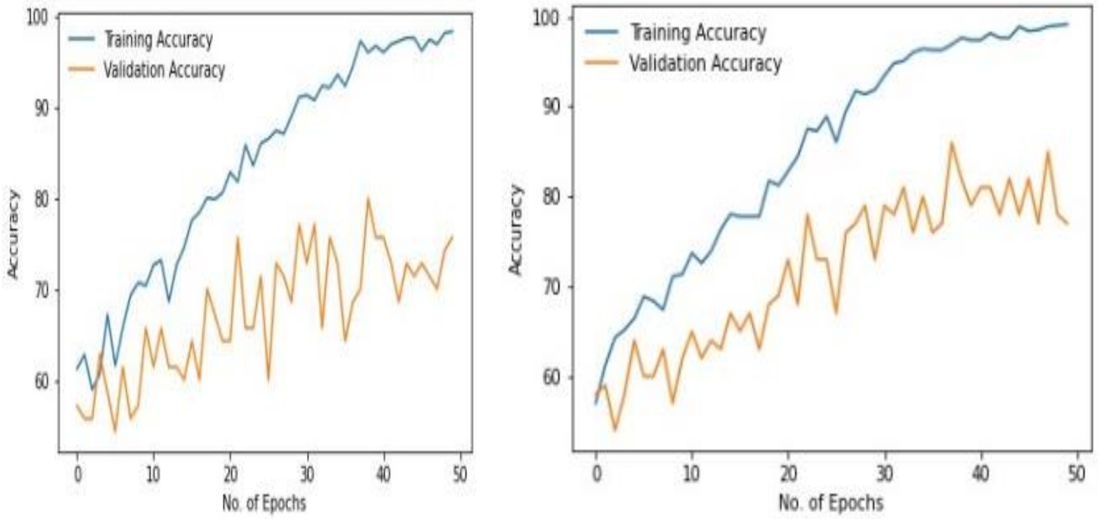


Fig. 7: ResNet training and validation accuracy (700 & 1000 images)

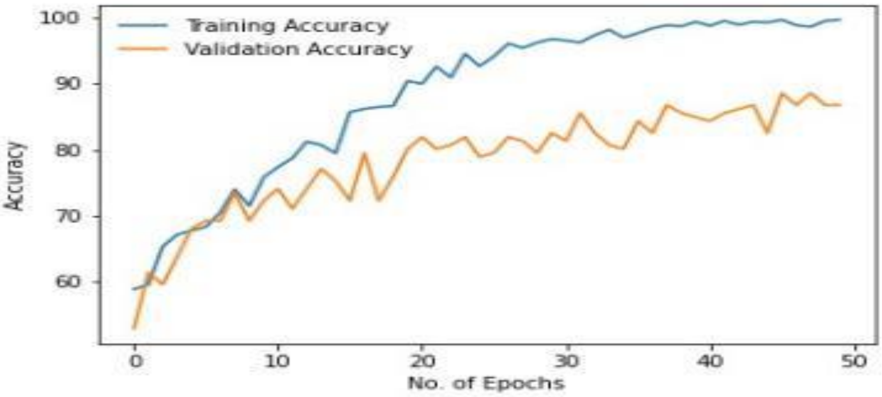


Fig. 8: ResNet training and validation accuracy (1678 images)

After training on our image data, the Resnet34 model achieves an accuracy of 70% as shown in Figs. 6, 7, and 8.

3.3.4.2 Alexnet

AlexNet is the name of a convolutional neural network (CNN) architecture developed by Alex Krizhevsky in collaboration with Ilya Sutskever and Geoffrey Hinton, Krizhevsky's PhD advisor. It has five convolutional neural networks and two max-pooling layers among its eight layers. The training and validation accuracy for this study are as shown in Figures 9 and 10.

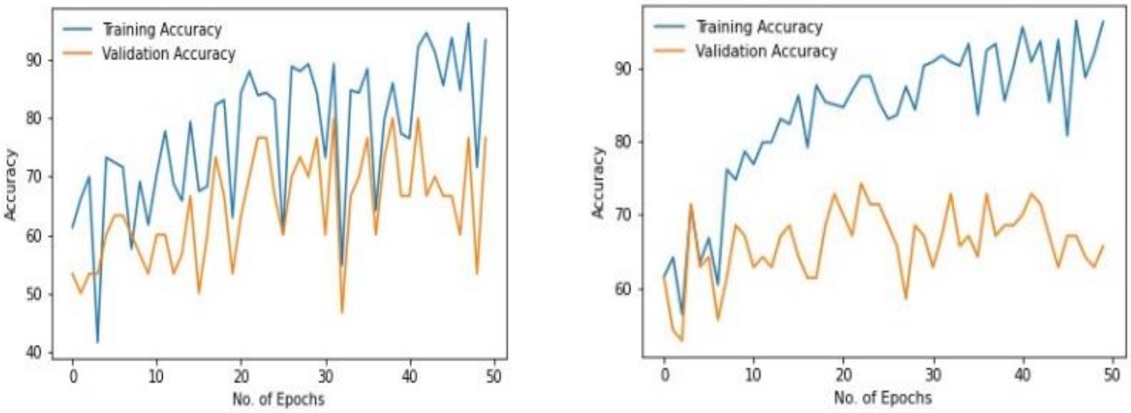


Fig. 9: AlexNet training and validation accuracy (300 & 700 images)

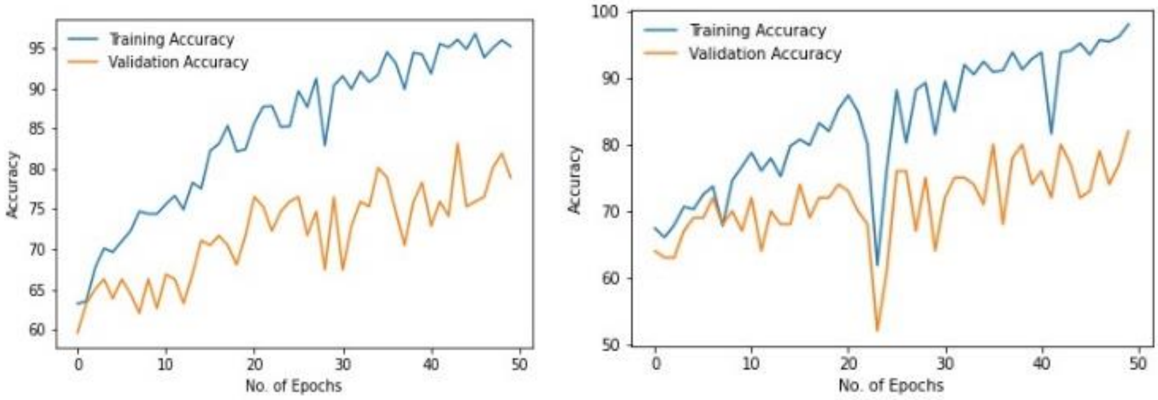


Fig. 10: AlexNet training and validation accuracy (1000 & 1678 images)

3.3.4.3 VGG16

Simonyan & Zisserman (2015), of the University of Oxford created the Visual Geometry Group (VGG-16) convolutional neural network model in 2014. It improved on ImageNet by adding more complex convolutional layers that required more computational power to train. ImageNet contains more than 1.2 million images for training and 50,000 images for testing. The model was built on an NVIDIA Titan Black GPU and ran for weeks before being fully trained. The model's advantages include accurate feature identification in data, high efficiency and convenience in the method of transfer learning, and optimal effectiveness on the data being trained with high accuracy. In this study, Figures 11 and 12 shows the performance of VGG16 as the data increases.

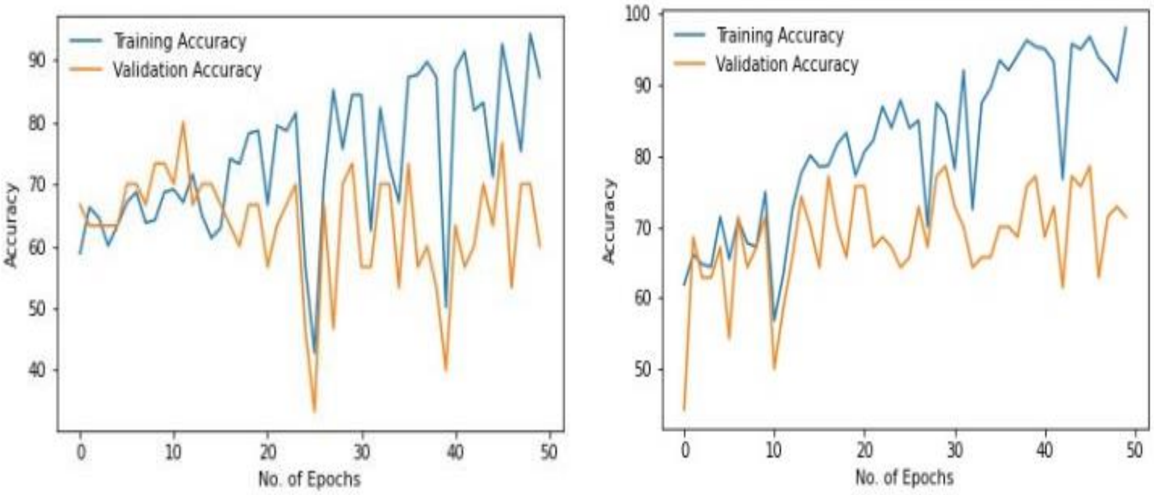


Fig. 11: VGG16 training and validation accuracy (300 & 700 images)

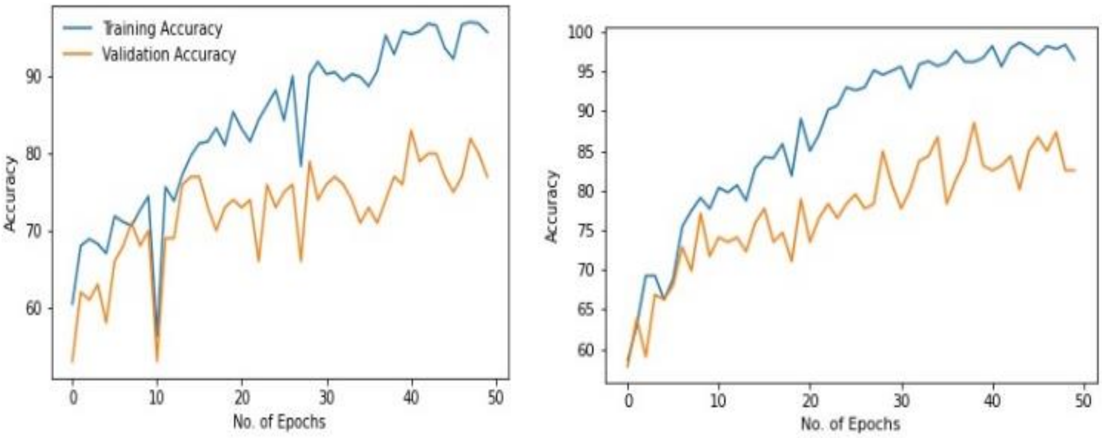


Fig. 12: VGG16 training and validation accuracy (1000 & 1678 images)

3.3.4.4 VGG19

VGG-19 is a 19-layer deep convolutional neural network. The network can classify images into 1000 different object categories, including keyboards, mice, pencils, and a variety of animals. As a result, the network has learned a variety of rich feature representations for a variety of images. The network's picture input size is 224 × 224 Pixels. option to the Image Data Generator constructor, which provides the min and max range as a float indicating a percentage for determining the amount of brightening. Values less than 1.0 darken the image, e.g. [0.5, 1.0], whereas values more than 1.0 brighten it, e.g. [1.0, 1.5], with 1.0 having no impact. Figs. 13 and 14 present the accuracy curve for VGG19 as the data increases in this study.

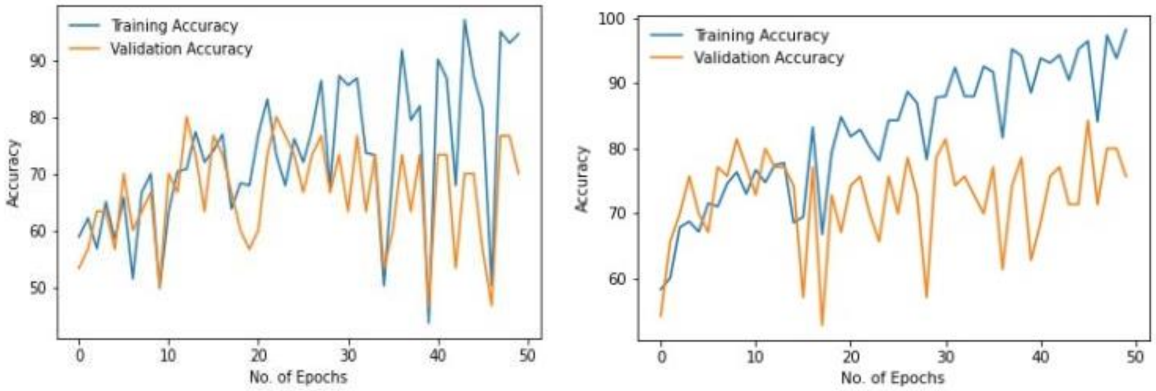


Fig. 13: VGG19 training and validation accuracy (300 & 700 images)

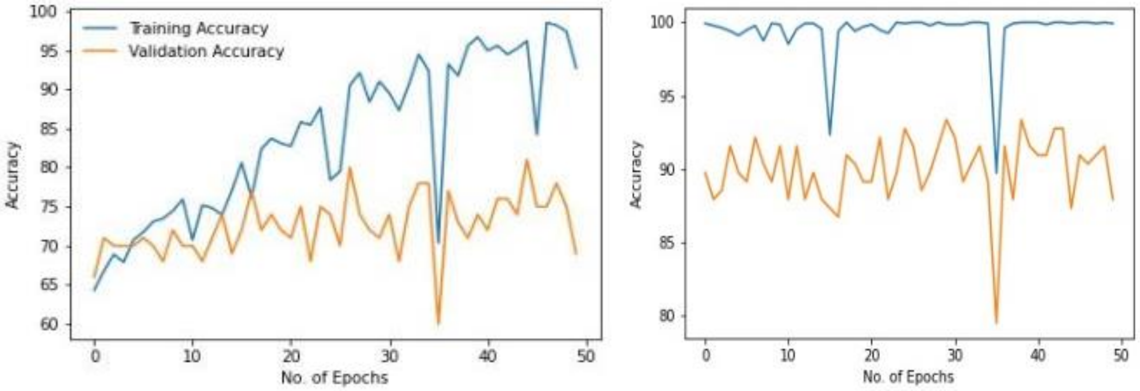


Fig. 14: VGG19 training and validation accuracy (1000 & 1678 images)

3.3.4.5 Vision Transformer (ViT)

The Vision Transformer (Han et al., 2020), is a type of deep neural network that is primarily based on the self-attention mechanism and was first used in the field of natural language processing. It has been used in computer vision tasks because of its strong representation capabilities. Other types of networks, such as convolutional and recurrent networks, performed similarly to or better than it. Dosovitskiy et al. (2020) demonstrated that a pure transformer applied directly to image patch sequences can perform very well on image classification tasks. When compared to state-of-the-art convolutional networks, Vision Transfomer (ViT) achieves excellent results while requiring significantly fewer computational resources to train. Figures 15 and 16 presents the results for ViT in this study as the data increases.

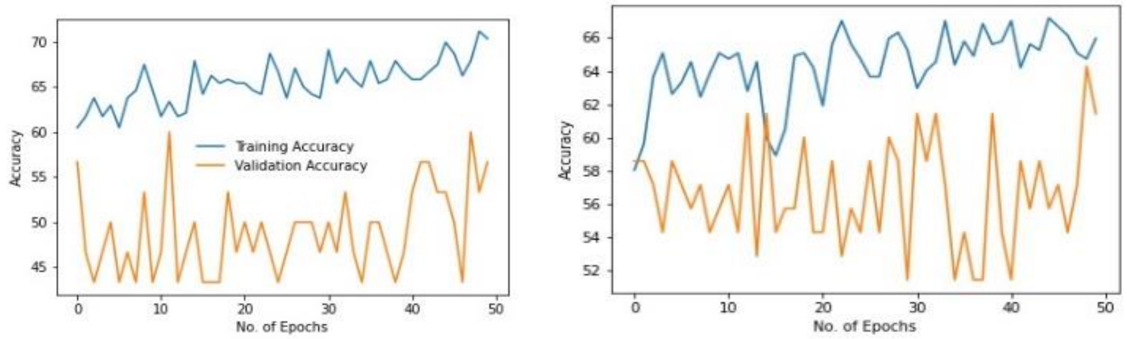


Fig. 15: ViT training and validation accuracy (300 & 700 images)

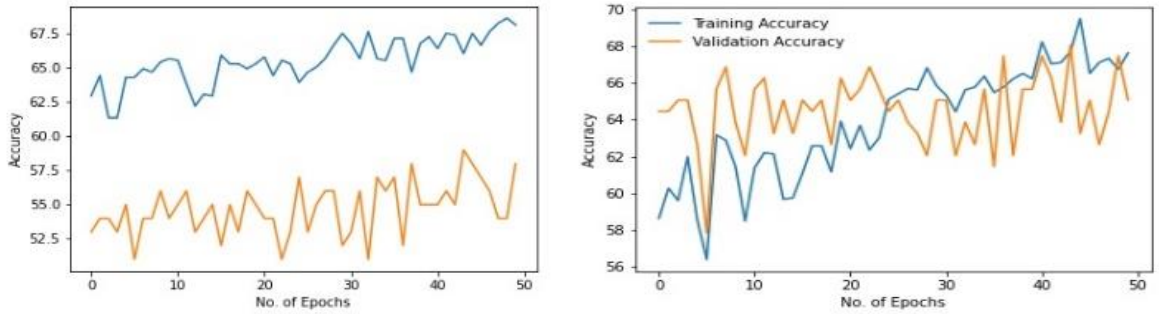


Fig. 16: ViT training and validation accuracy (1000 & 1678 images)

3.3.4.6 EfficientNet

EfficientNet (Tan & Le., 2019), is a convolutional neural network architecture and scaling method that uses a compound coefficient to scale all depth/width/resolution dimensions uniformly. The EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients, unlike conventional practice, which scales these factors arbitrary (Tan & Le, 2019). Figures 17 and 18 shows the accuracy and validation plot for EfficienNet being implemented on 4 varying batch sizes of data.

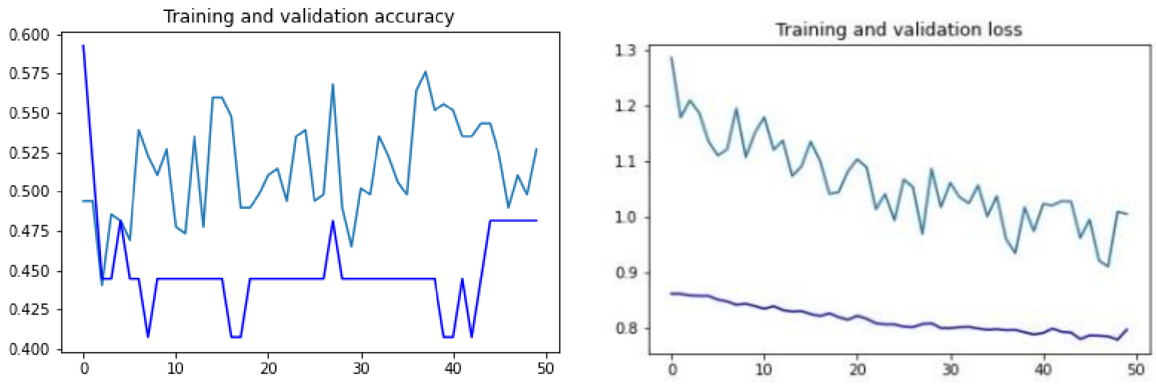


Fig. 17: EfficientNet training and validation accuracy (300 & 700 images)

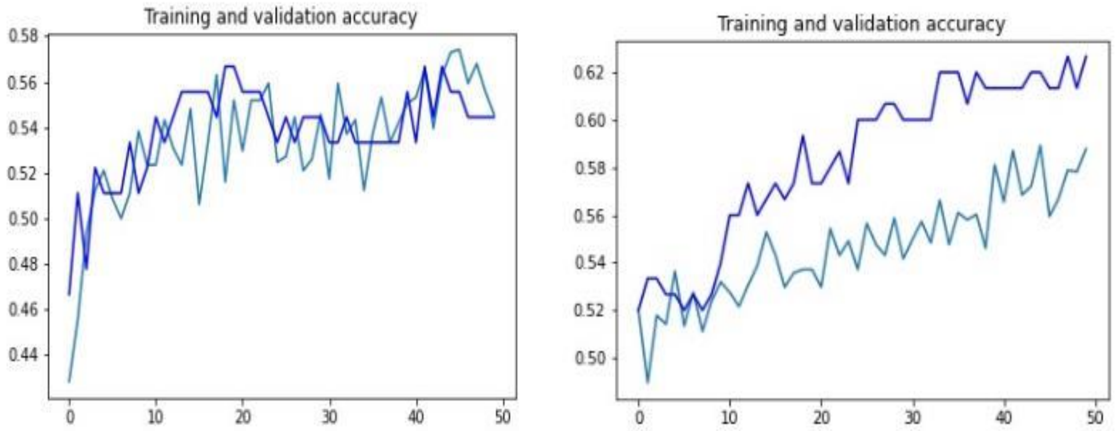


Fig. 18: EfficientNet training and validation accuracy (1000 & 1678 images)

3.4 Machine learning and deep learning techniques

The dataset was divided into four groups of 300, 700, 1000, and 1600 images, with each group being fed into the Deep Learning models used in this project. Fig. 19 – Fig. 22 shows the model's performance as the size of datasets increases.

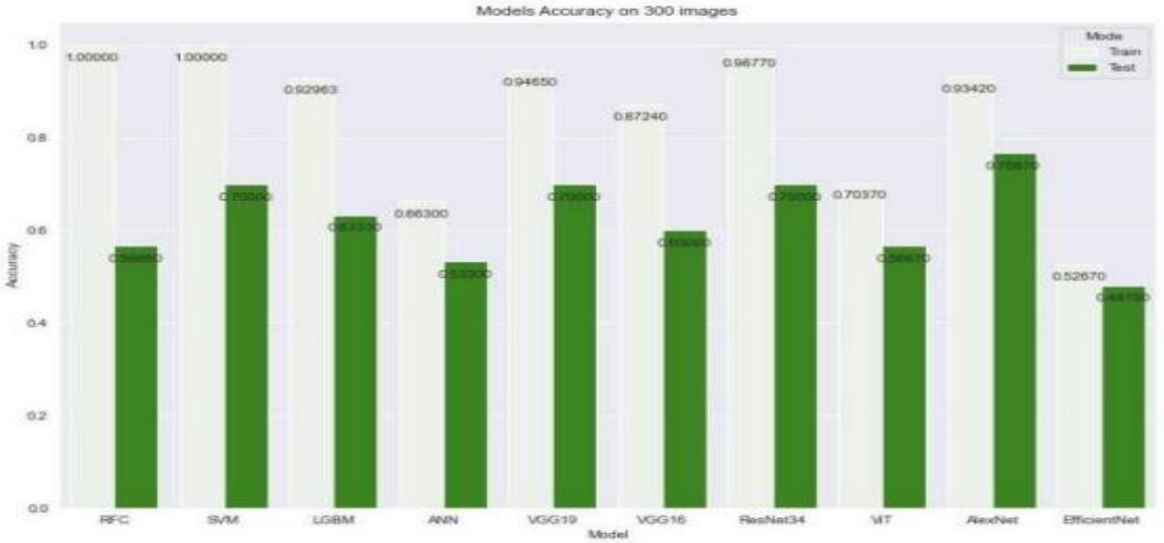


Fig. 19: performance of models on 300 random datasets

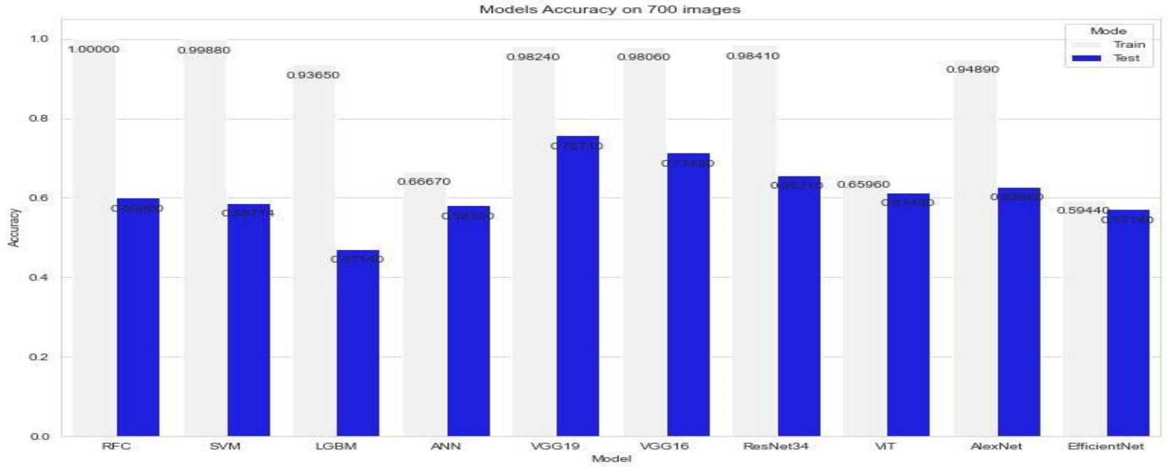


Fig. 20: performance of models on 700 random dataset

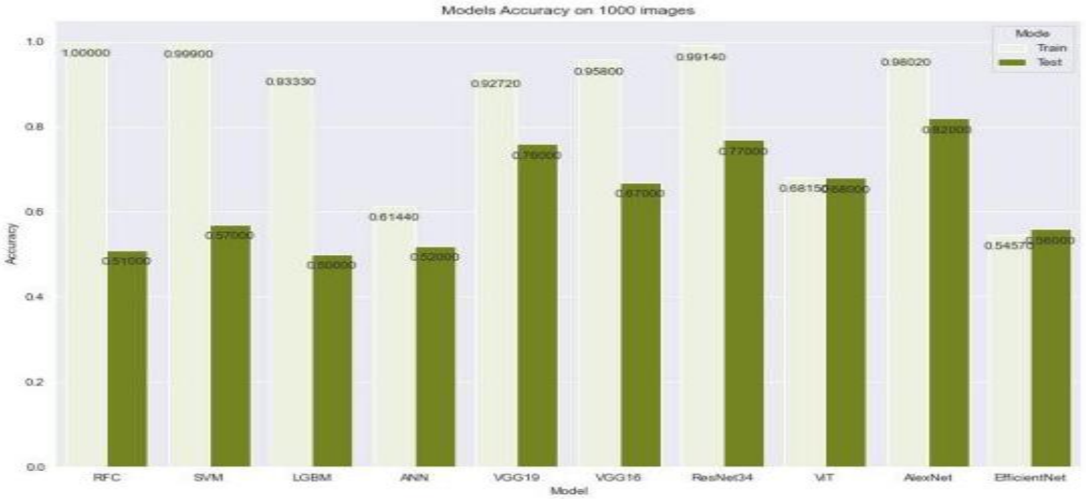


Fig. 21: performance of models on 1000 random dataset

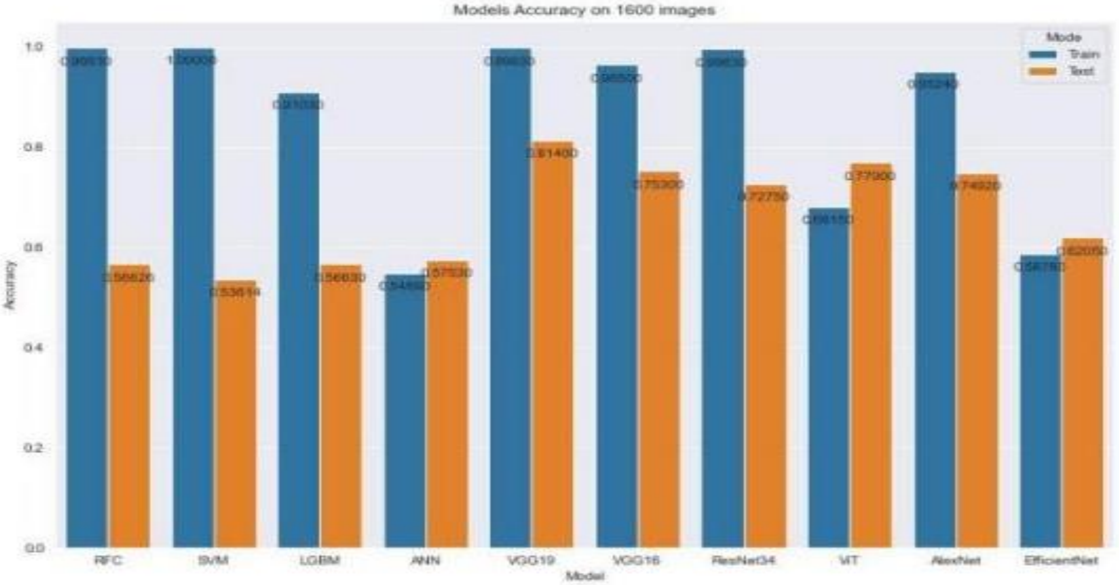


Fig. 22: performance of models on 1678 random dataset

From Fig. 19 -22, six (6) deep learning models were trained on the classification of cervical cancer images. Each of the models were trained on 4 random batches of images (300, 700, 1000, and 1678 images) to observe the model's performance with increase in number of training data. From the result, there is a significant gap between the ML techniques and the DL techniques as expected. As expected, this shows that the DL techniques performs way better than traditional ML approaches in image classification owing to the depth of their network. It can also be seen that VGG19 shows consistently good performance (81% accuracy) and Vision Transformer (ViT) on the other hand shows improving performance (57% to 77% accuracy) as the number of images increases.

As a result of this, a hybrid model is proposed to benefit from the strength of both models.

3.5 Proposed Hybrid Machine Learning Model

A hybridization of the CNN network VGG-19 and the transformer network was proposed. The proposed hybrid model aims to improve the accuracy and efficiency on the predictive analysis of the cervical images.

3.5.1 The VGG19 Model

The VGG-19 model, developed by Simonyan and Zisserman (2014), of the University of Oxford, is a 19-layer (16 conv., 3 fully-connected) CNN that strictly uses 3x3 filters with stride and pad of 1, as well as 2x2 max-pooling layers with stride 2. The VGG-19 is a deeper CNN with more layers than AlexNet. It uses small 3x3 filters in all convolutional layers to reduce the number of parameters in such deep networks (Zheng et al., 2018). The VGG-19 has been trained on over a million images and can classify them into 1000 different object categories, including keyboards, mice, pencils, and a variety of animals. As a result, the model has learned a variety of rich feature representations for a variety of images.

Other deep learning techniques, such as AlexNet, VGG16, ResNet, and EfficientNet, have been proposed to expedite the process and improve accuracy. VGG19, on the other hand, has consistently good performance on the data as the size of dataset increases. Fig. 23 shows the VGG19 model architecture.

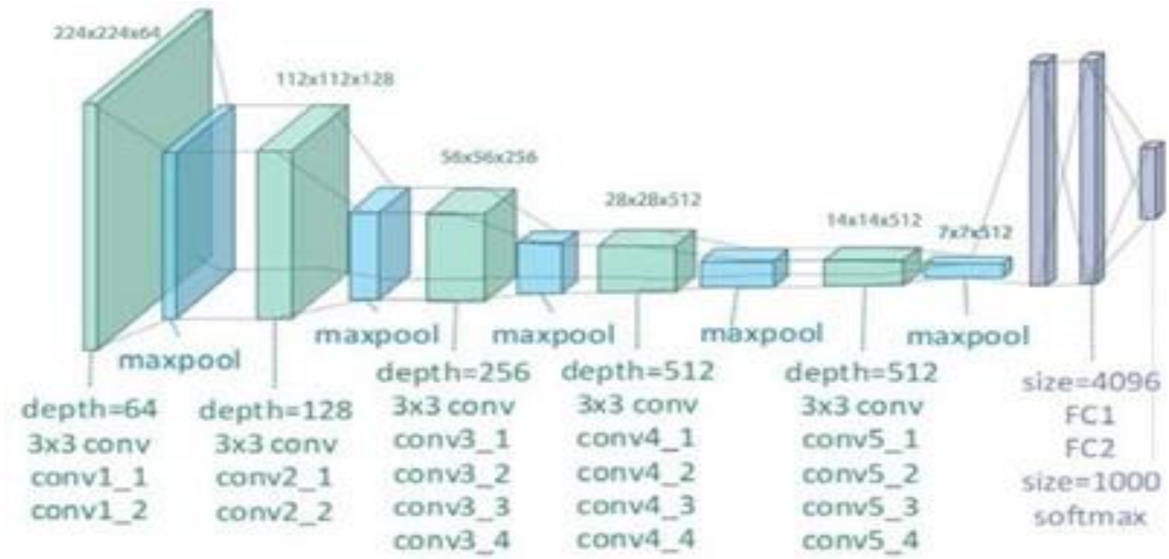


Fig. 23: VGG19 Model Architecture **(Zheng et al., 2018).**

3.5.2 The Vision Transformer Model

A Vision transformer is a deep learning technique that uses the attention mechanism to weight the significance of each element of the input data differently. Its primary applications are in natural language processing (NLP) and computer vision (CV) (Vaswani et al., 2017). Transformers, like recurrent neural networks (RNNs), are built to handle sequential input data like natural language for tasks like translation and text summarization. Transformers, unlike RNNs, do not always process data in the same order. The attention mechanism, on the other hand, provides context for any point in the input sequence.

If the input data is a natural language sentence, for example, the transformer does not need to process the first part of the sentence before the last. Rather, it detects the context that gives each word in the phrase its meaning. Because this feature enables for higher parallelization than RNNs, training times are reduced. Fig. 24 depicts the ViT model architecture. (Vaswani et al., 2017) .

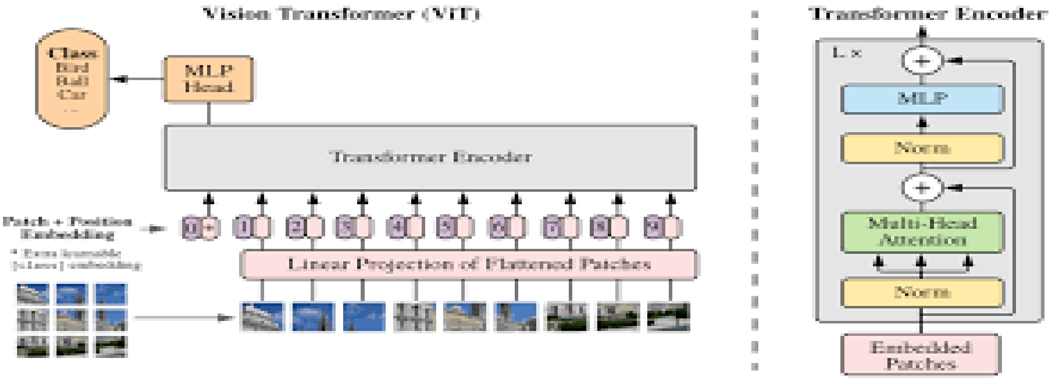


Fig. 24: Vision Transformer Model Architecture

3.5.3 Proposed Hybridized Deep Learning Network

The hybridized network integrates the VGG19 architecture and the Vision Transformer (ViT) model to leverage their complementary strengths in image feature extraction and contextual understanding, as shown in Fig. 25. VGG19 was selected for its consistent high performance across varying dataset sizes, due to its well-established convolutional feature extraction capabilities. Conversely, ViT demonstrates a unique advantage as dataset sizes grow, achieving superior global feature embedding through self-attention mechanisms. This synergy allows the hybrid model to deliver improved classification performance and computational efficiency (*Simonyan and Zisserman, 2014*).

3.5.4 Architecture and Design

The hybrid model architecture comprises two primary components:

3.5.4.1 VGG19 as a Feature Extractor

1. The VGG19 model processes 224 x 224 x 3 input images to extract feature maps. Using convolutional layers with 3 x 3 kernels and stride of 1, spatial features are preserved while non-linearities introduced by ReLU, enhance model robustness against vanishing gradients. Max pooling with 2 x 2 windows down-samples the image, reducing dimensions while retaining critical features.

2. The output of VGG19's final max-pooling layer is a tensor of size 7 x 7 x 512. Fully connected layers are omitted to reduce computational overhead and ensure compatibility with ViT.

$$W \in \mathbb{R}^{768\times 25,088},\ x \in \mathbb{R}^{25,088},\ \text{and}\ y \in \mathbb{R}^{768}.$$

3.5.4.2 Vision Transformer for Global Context

1. The downsampled feature map from VGG19 is flattened into a sequence of 7 x 7 = 49 tokens, each with 512 dimensions. To align with ViT's input requirements, a linear projection as seen in equation 1, maps these tokens into a 768-dimensional space.

$$y = W \cdot x + b \qquad (1)$$

where W= assigned weight, x=input feature map, and b=bias.

2. Positional embeddings are added to the token sequence to encode spatial relationships, ensuring that global attention mechanisms within ViT can utilize spatial context for classification.
3. ViT processes the token sequence through multiple transformer layers, using self-attention to capture inter-token relationships and embedding global information effectively.

3.5.4.3 Theory and Assumptions

The VGG19 model is computationally intensive due to its deep convolutional architecture, which requires significant processing power for training, especially at the fully connected layers. Drawing on research by Karen & Andrew (2015), the hybrid model eliminates these layers and integrates ViT as a lightweight alternative for classification. ViT's design inherently benefits from parallelism and efficient attention mechanisms, addressing the computational bottleneck of fully connected layers.

1. The heavy computational workload of VGG19's fully connected layers is avoided, as their role is replaced by ViT's transformer encoder.
2. Spatial and feature-rich outputs from VGG19 are efficiently processed by ViT, leveraging its ability to handle large datasets and embed global context.

3.5.4.4 Optimizations and Training Efficiency

A series of operations were performed to further enhance computational efficiency and accelerate training. The operations performed are listed as follows:

1. ***Parallelization***

Data parallelism was introduced by processing training batches on multiple GPUs, dividing gradients synchronously across devices, consistent with methodologies described by Simonyan & Zisserman (2014).

This approach reduced training time significantly without compromising accuracy.

2. ***Layer Freezing and Fine-Tuning***

Earlier layers of VGG19 were frozen during initial training to retain pre-trained weights, focusing computational resources on training the projection layer and ViT's attention layers.

3. ***Reduction of Training Complexity***

By removing fully connected layers from VGG19 and feeding its pooling layer output directly into ViT, computational cost was lowered substantially. Techniques such as global average pooling (GAP) and sparse attention mechanisms within ViT further minimized complexity.

4. ***Positional Encoding***

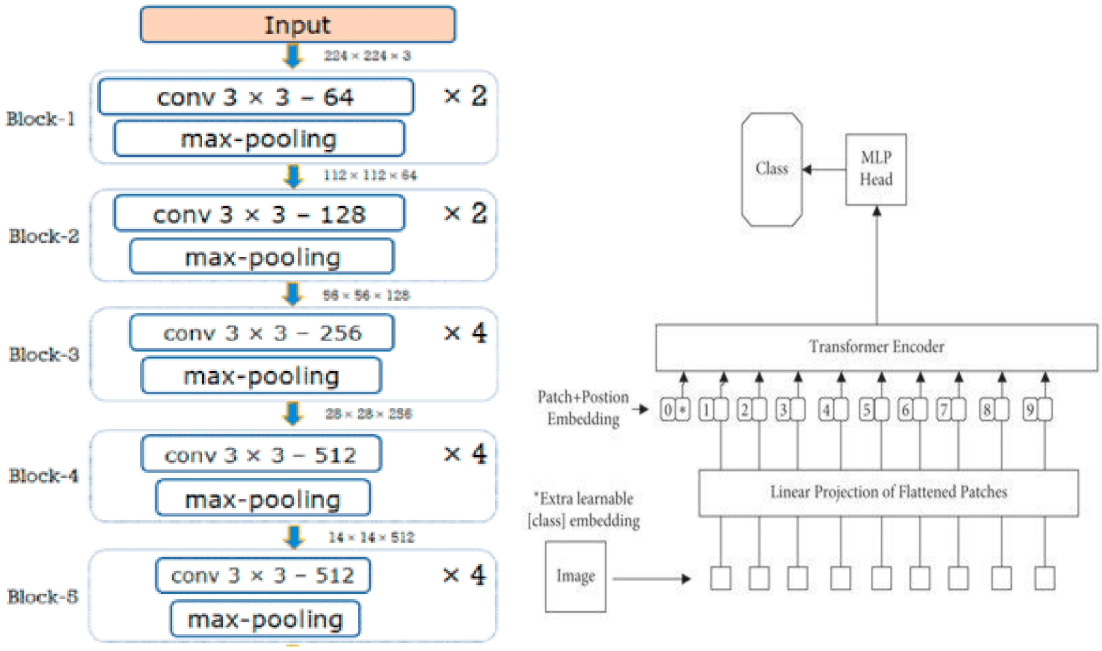
To mitigate the potential loss of spatial information when flattening VGG19 outputs, positional embeddings were incorporated. Using Fourier features as presented in equation 2.

$$P(h,w)=\left[\sin\left(\frac{h}{10000^{2i/D}}\right),\cos\left(\frac{h}{10000^{2i/D}}\right),\sin\left(\frac{w}{10000^{2i/D}}\right),\cos\left(\frac{w}{10000^{2i/D}}\right)\right] \qquad (2)$$

3.5.5 Advantages of the Hybrid Approach

This hybridized system offers significant performance improvements:

1. *Enhanced Accuracy:* The hybrid model achieved an 86.5% classification accuracy, outperforming standalone VGG19 and ViT implementations.
2. *Efficient Training:* Training time was reduced through parallelization and layer-freezing strategies.
3. *Robustness Across Datasets:* While VGG19 excels on smaller datasets, ViT's self-attention mechanisms enhance performance on larger datasets, resulting in a balanced, scalable model.



VGG19 Model *Vision Transformer Model*

Fig. 25: Proposed Hybrid Model Network Architecture

4.0 Results and Discussion

The proposed hybrid model was implemented using the Kaggle notebook environment, leveraging 16 GB of CPU and GPU resources along with over 50 GB of storage capacity. Testing and experimentation were conducted on Google Research AI notebook (Google Colab), ensuring seamless execution of the model. A total of 1678 images were generated using data augmentation techniques to enhance the diversity and representativeness of the training data. The implementation utilized both the VGG19 and Vision Transformer (ViT) models, with a focus on their performance characteristics as the dataset size increased. Table 2 presents the performance metrics of the hybrid model on four batches of randomly selected training images: 300, 700, 1000, and 1678. Initially, with 300 images, the model showed relatively low performance, achieving an accuracy of 56.2%, sensitivity of 0.557, specificity of 0.552, and an AUC of 0.512. These metrics highlight the challenges of training models on limited data, including inadequate feature representation and an increased likelihood of overfitting.

As the training set increased to 700 images, the model exhibited substantial improvement. Accuracy rose to 75.4%, with sensitivity and specificity reaching 0.733 and 0.712, respectively. The AUC also increased to 0.714, indicating better classification performance. Further improvements were observed with 1000 images, where the model achieved an accuracy of 85.4%, sensitivity of 0.8375, specificity of 0.821, and an AUC of 0.8412. These results demonstrate the hybrid model's ability to learn complex patterns and achieve greater generalization as more training data is provided. At 1678 images, the hybrid model achieved its highest performance metrics, with an accuracy of 86.5%, sensitivity of 0.8475, specificity of 0.832, and an AUC of 0.85. Notably, this represents a 6% improvement in accuracy over the VGG19 model, illustrating the hybrid model's superior performance. Additionally, the ViT model exhibited increasing performance with larger datasets, and it is anticipated to surpass the VGG19 model when trained on a significantly larger number of images.

Table 2: Performance of the Proposed Hybrid Model

Number of Images	Accuracy (%)	Sensitivity	Specificity	AUC
300	56.2	0.557	0.552	0.512
700	75.4	0.733	0.712	0.714
1000	85.4	0.8375	0.821	0.8412
1678	86.5	0.8475	0.832	0.85

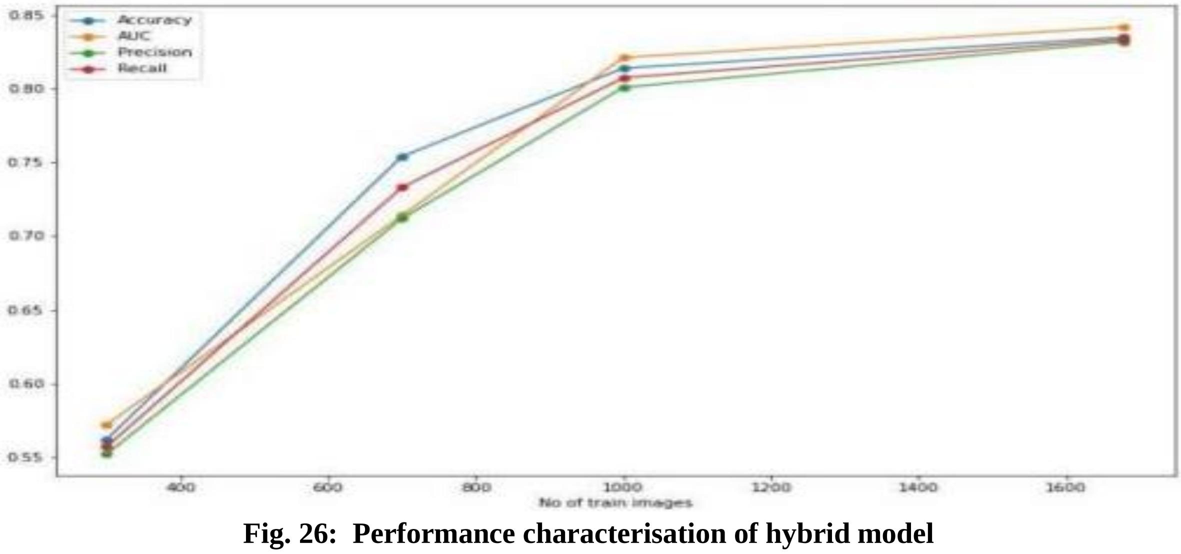


Fig. 26: Performance characterisation of hybrid model

The performance improvements are graphically illustrated in Figure 26, which depicts the accuracy, sensitivity, specificity, and AUC across different training image sizes. The consistent upward trend highlights the critical role of data augmentation and increased dataset size in optimizing the model's predictive power. When compared to previous studies, such as the work by Asiedu et al. (2018), which reported an accuracy of 80% and specificity of 81.3% under low-computational power constraints, the proposed hybrid model demonstrates a significant 8% improvement in accuracy and better overall specificity. This indicates the potential of the hybrid model for point-of-care applications and scenarios requiring low computational resources.

The hybrid model's implementation on computationally limited environments such as Kaggle and Colab further reinforces its practical applicability. By achieving high accuracy with relatively low computational power, the model demonstrates its utility for deployment in resource-constrained settings, such as developing regions or field-based applications.

Finally hybrid model successfully achieves robust performance through effective data augmentation and the complementary strengths of VGG19 and ViT models. Its adaptability to varying dataset sizes and resource environments makes it a promising solution for scalable, high-performance image classification tasks.

5.0 Conclusiom

This study explored the application of both traditional machine learning techniques and deep learning models for the classification of cervical images as VIA-positive or VIA-negative. By leveraging data augmentation and pre-processing techniques, the dataset size was significantly increased, enhancing model robustness and reducing overfitting risks. The feature extraction process, based on Haralick's textural features, provided a solid foundation for traditional machine learning models like Random Forest, SVC, and LightGBM, achieving classification accuracies of 48.5%, 54.3%, and 60%, respectively. Among these, LightGBM demonstrated the highest performance, indicating its suitability for such tasks.

Deep learning approaches, including transfer learning models such as ResNet34, AlexNet, VGG16, and VGG19, were also implemented, achieving higher accuracies compared to traditional methods. ResNet34, for example, achieved an accuracy of 70%, demonstrating the potential of pre-trained models in cervical image classification when combined with sufficient training data. The results underscore the critical importance of effective pre-processing, feature extraction, and advanced machine learning methodologies in medical imaging applications. While deep learning models outperformed traditional methods, further improvements in accuracy could be achieved by combining models, exploring additional feature engineering strategies, or utilizing larger and more diverse datasets. These findings provide valuable insights into leveraging machine learning for enhancing cervical cancer screening, offering a foundation for future studies aimed at improving diagnostic accuracy and accessibility.

To further improve the outcomes of such studies, it is recommended that future research focus on the integration of ensemble learning methods to combine the strengths of different models for improved classification performance. The development of custom deep learning architectures tailored specifically for medical image analysis should also be explored. Additionally, expanding the dataset with more diverse and annotated images can improve model generalization across varied populations. Collaboration with medical experts is critical to ensuring that the models are clinically relevant and interpretable. Finally, the deployment of these models in real-world scenarios, such as point-of-care devices, should be prioritized to assess their practical impact and usability in cervical cancer screening programs.

6.0 References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: transformers for image recognition at scale.* http://arxiv.org/abs/2010.11929.

Habib, G., & Qureshi, S. (2020). Optimization and acceleration of convolutional neural networks: A survey. In Journal of King Saud University - Computer and Information Sciences. King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2020.10.004.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2020). *A survey on vision transformer*. http://arxiv.org/abs/2012.12556.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 6, pp. 84–90. https://doi.org/10.1145/3065386.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 7553, pp. 436–444). https://doi.org/10.1038/nature14539.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. http://arxiv.org/abs/1409.1556.

Tan, M., & Le, Q. V. (2019). *Efficient net: rethinking model scaling for convolutional neural networks*. http://arxiv.org/abs/1905.11946.\

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. http://arxiv.org/abs/1706.03762.

Zheng, Y., Yang, C., & Merkulov, A. (2018). Breast cancer screening using convolutional neural network and follow-up digital mammography. 4. https://doi.org/10.1117/12.2304564

Asiedu, M. N., Simhal, A., Chaudhary, U., Mueller, J. L., Lam, C. T., Schmitt, J. W., Venegas, G., Sapiro, G., & Ramanujam, N. (2019). Development of Algorithms for Automated Detection of Cervical Pre-Cancers with a Low-Cost, Point-of-Care, Pocket Colposcope. *IEEE Transactions on Biomedical Engineering*, *66*, *8*, pp. 2306–2318. https://doi.org/10.1109/TBME.2018.2887208.

Balas, C. (2001). A novel optical imaging method for the early detection, quantitative grading, and mapping of cancerous and precancerous lesions of cervix. *IEEE Transactions on Biomedical Engineering*, 48, 1, pp. 96–104. https://doi.org/10.1109/10.900259.

Das, A., Kar, A., & Bhattacharyya, D. (2014). Detection of abnormal regions of precancerous lesions in digitised uterine Cervix images. *2014 International Electrical Engineering Congress, IEECON 2014*. https://doi.org/10.1109/iEECON.2014.6925937.

Kaur, N., Panigrahi, N., & Mittal, A. (2017). Automated Cervical Cancer Screening Using Transfer Learning. *International Journal of Advance Research in Science and Engineering*, 6, 8, pp. 2110–2119.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2019). ImageNet Classification with Deep Convolutional Neural Networks, pp. 1–9.

Kudva, V., Prasad, K., & Guruvare, S. (2020). Hybrid Transfer Learning for Classification of Uterine Cervix Images for Cervical Cancer Screening. *Journal of Digital Imaging*, *33*(3), 6, pp. 619–631. https://doi.org/10.1007/s10278-019-00269-1.

Kudva, V., Prasad, K., & Guruvare, S. (2018). Automation of detection of cervical cancer using convolutional neural networks. *Critical Reviews in Biomedical Engineering*, *46*(2), 135–145. https://doi.org/10.1615/CritRevBiomedEng.2018026019,

Liang, M., Zheng, G., Huang, X., Milledge, G., & Tokuta, A. (2013). *Identification of abnormal cervical regions from colposcopy image sequences.* 21st International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG 2013 - *Communication Papers Proceedings*, pp. 130–136.

Priya, P. (2014). Detection of Cancerous Lesion By Uterine Cervix Image Segmentation. *ICTACT Journal on Image and Video Processing*, 4, 3, pp. 762–766. https://doi.org/10.21917/ijivp.2014.0110.

Raifu, A. O., El-Zein, M., Sangwa-Lugoma, G., Ramanakumar, A., Walter, S. D., Franco, E. L., Ferenczy, A., Mahmud, S., Nasr, S., Kayembe, P., Rahma, Liaras, J., & Lorincz, A. (2017). Determinants of cervical cancer screening accuracy for visual inspection with acetic acid (VIA) and lugol's iodine (VILI) performed by nurse and physician. *PLoS ONE*, 12, 1, pp. 1–13. https://doi.org/10.1371/journal.pone.0170631.

Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. http://arxiv.org/abs/1409.1556.

Shu, M. (2019). Deep learning for image classification on very small datasets using transfer learning. *Creative Components*, 14–21.

Song, D., Kim, E., Huang, X., Patruno, J., Muñoz-Avila, H., Heflin, J., Long, L. R., & Antani, S. (2015). Multimodal entity coreference for cervical dysplasia diagnosis. IEEE Transactions on Medical Imaging, 34, 1, pp. , 229–245. https://doi.org/10.1109/TMI.2014.2352311.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*(Nips), 5999–6009.

Zheng, Y., Yang, C. K., & Merkulov, A. (2018). Breast cancer screening using convolutional neural network and follow-up digital mammography, *2018-December*(Nips), 1–13.

RamaPraba, P. S., & Ranganathan, H. (2012). Computerized Lesion Detection in Colposcopy Cervix Images Based on Statistical Features Using Bayes Classifier. *Proceedings of the InConINDIA 2012, AISC 132, 597–604.*

RamaPraba, P. S., & Ranganathan, H. (2013). Wavelet Transform Based Automatic Lesion Detection in Cervix Images Using Active Contour. *Journal of Computer Science*, 9, 1, pp. 30-36.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*(Nips), pp. 5999–6009.

Rouhbakhsh, F., Farokhi, F., & Kangarloo, K. (2012). Effective Feature Selection for Pre-Cancerous Cervix Lesions Using Artificial Neural Networks. *International Journal of Smart Electrical Engineering, 1(3), 199-204.*

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *Published as a conference paper at ICLR, 2015*, pp. 1–11.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-scale Image Recognition. *Published as a conference paper at ICLR, 2015, 1–14.*

Sukumar, P., & Gnanamurthy, R. K. (2016). Computer-aided Screening of Cervical Cancer Using Random Forest Classifier. *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, 7, 1, pp. 1521–1529.

Xu, T., Zhang, H., Xin, C., Kim, E., Long, L. R., Xue, Z Antanid, S., & Huanga, X. (2018). Multi-feature based Benchmark for Cervical Dysplasia Classification Evaluation. *Pattern Recognit. 2017 March ;* 63, pp. 468–475. *doi:10.1016/j.patcog.2016.09.027.*

Xu, T., Kim, E., & Huang, X. (2015). Adjustable Adaboost Classifier And Pyramid Features For Image-Based Cervical Cancer Diagnosis. *Department of Computing Sciences, Villanova University, Villanova, PA, USA*, pp. 281–285.

Compliance with Ethical Standards Declaration

Ethical Approval

Not Applicable

Competing interests

The authors declare that they have no known competing financial interests

Funding

The authors discovered no external source of funding

Authors' Contributions

All authors contributed equally to the design, analysis and writing of the paper.