Section A-Research paper
ISSN 2063-5346

Prediction of Ovary Cysts Combining LGBM and Neural network models.

Mrs.Y.Suganya[1,a)] Mrs.Sumathi Ganesan [2] Mrs.P.Valarmathi[3]

[1]*Research scholar, Department of Computer Science , Annamalai university,Tamilnadu, India.*
[a)]suganyasuchithrra@gmail.com

[2] *Assistant professor, Department of Computer Science , Annamalai university,Tamilnadu, India.*
[3] *Assistant professor, Mookambigai college of Engineering– Pudukkottai-622502, Tamilnadu, India.*

Abstract

Cysts of the ovary are prevalent. Typically, people experience little to no difficulty, and the cysts are inconsequential. Ovarian cysts can become distorted and even explode at times. This can result in severe symptoms. The majority of cysts resolve on their own after a couple months, but if left ignored, they can cause serious consequences like cancer. The ovary is hidden deep within the abdominal region. As a result, detecting a bulge or expanded area may become more difficult. Occasionally, doctors are unable to discover an anomaly during a pelvic examination. Imaging techniques are commonly employed to detect malignancies. The healthcare business produces vast quantities of data. The application of machine learning approaches can significantly improve both predictions and therapies. The advancement of algorithms for machine learning enables doctors to diagnose and cure diseases more quickly. Five different classes of ovarian cyst Simple Cyst, Polycystic ovary syndrome (PCOS), Dermoid Cyst, Endometriotic cyst, Hemorrhagic cyst are being classified in this work with three different segmentation techniques each employed in an DNN-LGMB model that is created.

Keywords—Cyst, Ovary, Machine learning, Imaging,Abdominal region, Imaging techniques

I. INTRODUCTION

Ovarian cyst is a condition that affects a woman's uterus, its diagnosis and analysis are carried out by experts by analysing the cyst's volume and properties on an ultrasound gadget. Transvaginal ultrasonography evaluation is acceptable for use in this study since it offers details on the morphology and features of polycystic ovary syndrome (PCOS). In diagnostics, physical analysis by physicians is widely employed. In general, measuring the cyst's size and shape over the course of many days is the most important evaluation technique.

However, this technique strongly relies on acquired knowledge in detecting the morphology and features of the many forms of ovarian anomalies visible in the proper ultrasound scans. Consequently, novice ultrasound operators will constantly struggle to differentiate between different forms of cysts, resulting in a low level of clinical diagnosis.

The majority of ovarian cysts are asymptomatic. Even if the signs are present ,they cannot always be utilised to detect ovarian cysts because they are comparable to those of endometriosis and pregnancy. Symptoms of ovarian cysts may include abdominal bloating or swelling, bowel movement pain, etc. After menopausal, ovarian cysts are much more prone to be malignant than cysts which originate at the pre-menopausal stage.

The foundation of computing technology is machine learning (ML). Modern machine learning enables machines to learn despite leaving their context. The ML evaluates the input, converts it into a format suitable for clinical processes, and identifies the nature of various disorders. This study examines the classification of ovarian cysts by combining a Deep Learning Neural Network (DLNN) model with the Light Gradient-Boosting-Machine (LGBM) method in a hybrid multi-stage concatenation.

II. LITERATURE REVIEW

Researchers in the field of computer science has developed a variety of approaches to identify diseases in humans by utilizing cutting-edge machine learning and image processing methods. These approaches can be found in use today. Subrato Bharati [1] used Univariant feature selection algorithm to find the best feature to predict PCOS, holdout and cross validation are used to detach the data for training from the testing data. This same set of data is placed through an amount of classification models, such as gradient boosting, random forest, logistic regression, and a fusion of random forest and logistic regression (RFLR). RFLR has the best testing accuracy at 91.01%.. Amsy Denny et al developed the i-HOPE system in [2], which is a method that can anticipate and diagnose polycystic ovary syndrome (PCOS) With the help of Machine Learning Methods and SPSS Version 22.0, features are recognised according to the significance they hold. The RFC approach, which has an accuracy of 89.02%, is the one that is best suited and most accurate for predicting PCOS. A strategy which automating the PCOS screening based on medical and biochemical characteristics was outlined by Palak Mehrotra in [3]. The research created a feature representation using the clinical and metabolic variables, and statistically meaningful features for differentiating among healthy and PCOS categories were identified using a two sample t-test. For supervised classification, Bayesian and Logistic Regression (LR) classifiers are used, and their respective levels of accuracy are 93.93% and 91.04%. By using feature extraction technique known also as Gabor Wavelet technique and the Computational Neural Network (CNN), the author of this paper [4] designs a classification system for PCO. The Neural Network was able to achieve the utmost accuracy possible, which was 80.84 percent, and the processing time was 60.64 seconds.

III. METHODOLOGY

This section details about the pre-processing techniques and the machine learning algorithm used to achieve the goal in detail Fig 1.

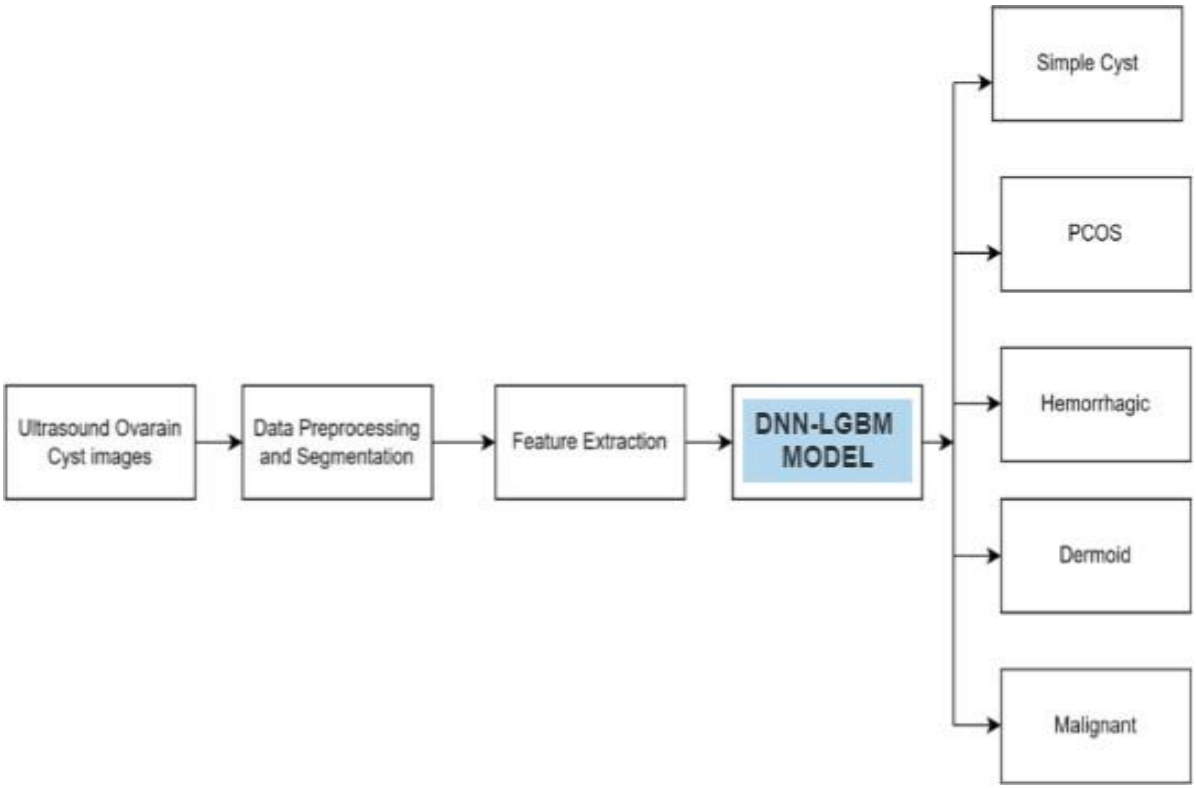


Fig. 1. Flow chart of the proposed methodology

1) *Data collection and description*

The dataset is collected from various private clinical laboratories in and around Nagercoil, Tamil Nadu, India. The dataset consists of 374 images with 5 classes, of which 351 are used for training and 23 are used to test the data.

2) *Data preprocessing*

For any deep learning architecture to operate as intended, a suitable dataset that has been thoroughly cleaned is absolutely necessary. The below preprocessing methods had been applied to the obtained dataset in order to achieve better results and to ensure that the model operates at its most effective level possible.

2.1 *Converting images from BGR format to RGB format*

First, OpenCV is used to convert images that are in the BGR format, which specifies the order of color's as blue, green, and red, to the RGB format, which specifies the order of color's as red, green, and blue. It assists in reducing procedures and also removes the complexity that is linked towards the necessities for calculation.

2.2 *Converting images from RGB format to HSV format*

At this stage, all of the photos that have previously been transformed into the RGB format are now being transformed into the HSV format. The Hue Saturation Value (HSV) level offers a number of interpretations of the images, which match the color labels that are displayed .The finalpreprocessed image can be seen in Fig 2.

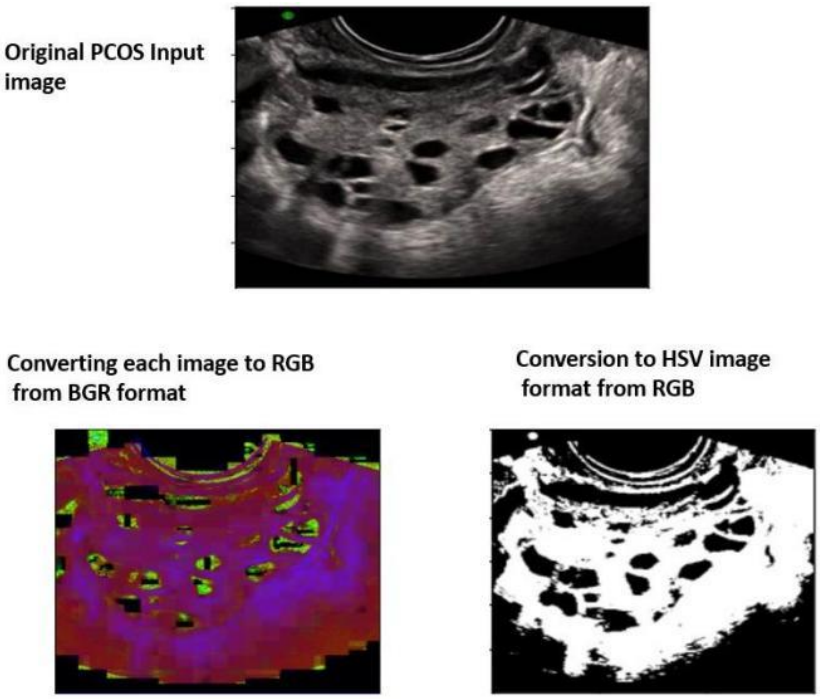


Fig. 2. Preprocessed image

3) *Segmentation*

Using the method of segmentation, the image that has been sent to us in digital format is divided into a number of different visual sections or subsets. When we take this step, the level of difficulty associated with postprocessing or evaluation of the images is significantly reduced. Even though there are various segmentation methods available this work have utilized three segmentation techniques[5].

3.1) *Edge based segmentation*

Using this particular method, the borders between sections are completely unique from one another and from the environment, which enables boundary identification to be dependent on localized interruptions in strength (gray level). To put it another way, it refers to the process of identifying edges in a picture. We already understand how borders comprise of relevant characteristics and carry substantial data, thus this is a very crucial step to comprehending visual features. Edges may be broken down into these two categories. The purpose of edge segmentation is to produce a conclusion of preliminary segmentation, that can then be subjected to area-based identification or any other sort of classification in order to produce the end segmented image.

3.2) Region based segmentation

Each region is defined as a collection of linked dots or pixels that share common characteristics and can be grouped together. The resemblance among pixels could be expressed in a variety of ways, including strength, colour, and so on. In this kind of segmentation, there are some established criteria that are available, and in order for a pixel to be grouped into related pixel sections, that pixel must follow those criteria. When dealing with a noisy picture, it is best to use methods of segmentation that are largely based on regions rather than techniques that are focused on edges. On the basis of the methodologies that they employ, region-based strategies can be further subdivided into two distinct types.

3.3) Cluster based segmentation

Approaches that fall under the category of "clustering based" are those that are used to split an image to groups of pixels that share comparable qualities. This process of grouping data items into groups in a way that the components within a single cluster are less related to one another than they are to components in other groups is known as data clustering. The structured approach and the partition-based technique are the two fundamental categories that make up the clustering method landscape. The idea of trees serves as the conceptual underpinning for the hierarchical organizational approaches. The entire database is represented by the node at the root of the tree, while the individual clusters are shown by the nodes inside the tree. On the other hand, approaches based on partitioning apply optimization methods in an iterative manner in order to reduce the value of an objective function. In the middle of these two approaches lie a number of other algorithms for locating clusters

4) Feature Extraction

Feature extraction is a component of the dimensionality decrease procedure, in which an original collection of unprocessed data is partitioned into more comprehensible categories. Therefore, processing will be simplified. The most crucial aspect of these massive data sets is the high amount of characteristics they contain. Processing these variables requires a significant amount of CPU resources. Feature extraction enables the selection and combination of factors into features, so substantially lowering the volume of data. These features are simple to handle while still accurately and creatively describing the real data set[6].

4.1) Hu moment

By using cv2.HuMoments() method, one could locate the Hu-Moments. It provides seven moments that are translations, spin, and scale independent. Seventh element is invariant to skewed. Before we can estimate the Hu-Moments, we must locate the picture. The image moment of an object are derived using the item's outline. So, we identify the entity's contour before applying the cv2.moments() method to calculate the moments.

4.2) *Haralick texture*

This method has been extensively implemented in image processing purposes, particularly in the clinical sciences. There are two processes involved in feature extraction. The GLCM is generated in the initial stage, followed by the calculation of texture characteristics dependent on the GLCM. In the picture, a matrix is created that counts the co-occurrence of nearby grey levels. Before every characteristic is passed to the classifier, it is normalised.

4.3) *Colour histogram*

Colour Histogram is one of the greatest used method for collecting the colour characteristic of an image. It depicts the picture from a distinct perspective. It is the frequency distribution of colour bins inside a picture. It recognizes and stores comparable pixels. OpenCV () includes the function CV2 for calculating the histogram.

5) *Models used*

5.1 *LGBM Algorithm*

Microsoft created Light-GBM, a networked gradient boosting system. It employs a histogram technique that compromises segmentation precision for learning time and reduces file size. LGBM generates the tree structure that use the leaves method as opposed to the level-wise Fig 3 method employed by many conventional tree-based Machine learning algorithms[7], resulting in greater accuracy rate; nevertheless, it could also cause increased tree structure and generalisation.

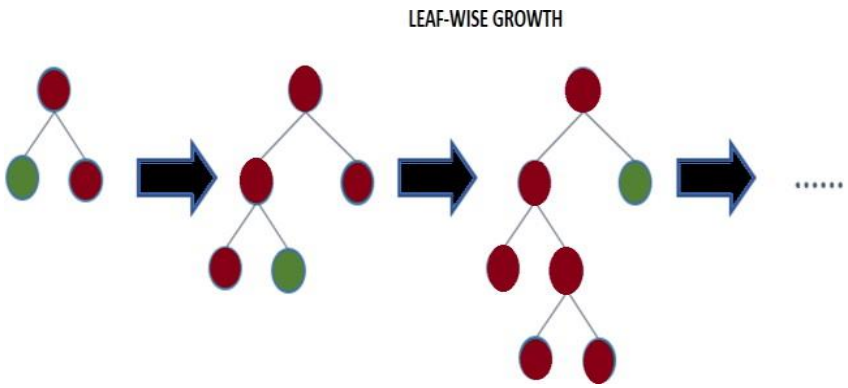


Fig. 3. Representing leaf wise growth of LGBM model

Following is the working methodology of LGBM model:

- It is believed that the learning set contains examples such as A1, A2, and so on down to xn, and that every component of the collection is a vectors of s units in the space A.
- All of the adverse gradients of an error function with regard to the output of the system are labelled as z1, z2, and so forth up to zn in each iteration of a gradient boosting algorithm. This continues until there are no more adverse gradients.
- The decision tree certainly splits evey vertices at its most suggestive point, which is also where the many evidence is gained.

- The variability after segregating can be used toquantify how much the data has improved. Thepreceding equation can be used to describe it: “X=Base_tree(A)-lr*Tree1(A)-lr*Tree2(A)- lr*Tree3(A)”

5.2 *Deep Neural-network*

DNN is an upward or forward-looking neural network algorithm comprised of several hidden layers Fig 4. Various neural modules can be assembled together into a unique discrete module, and the system was trained from beginning to end. A neural network is composed of a sequence of neurons (or vertices), with the input, hidden, and output layers forming its architecture[8]. The neuron accepts and analyses signals or input from neighbouring neurons, as well as transmitting data to output layers.

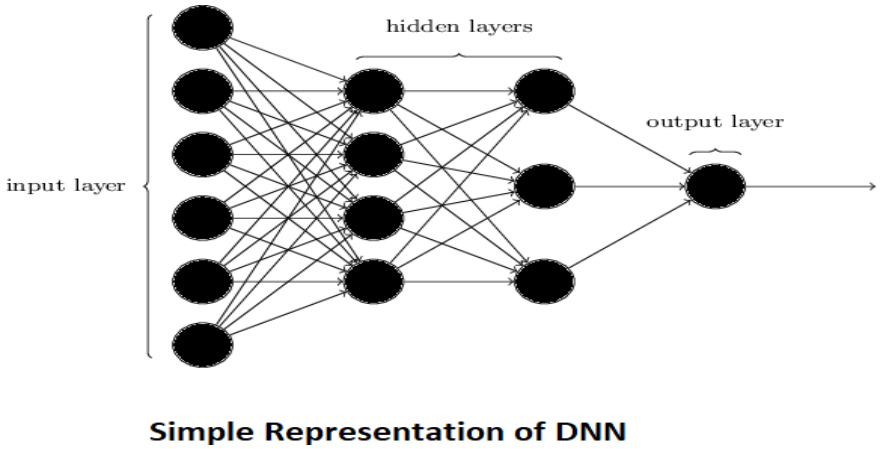


Fig. 4. Multiple layers of DNN algorithm

IV. RESULT AND DISCUSSION

Our primary model is a deep neural network, which consists of three hidden layers and six neuronal connections. After doing some preliminary processing on the data, such as segmentation, feature extraction, and picture enhancement, the input is then passed to the first hidden layer, which employs RELU as its activation function. In the second layer, the Soft-max activation function is used, and the output of the last layer is fed into the LGBM algorithm, a DNN- LGMB model is created by combining both DNN and LGMB Fig 5 algorithms which helps to classify the disease.

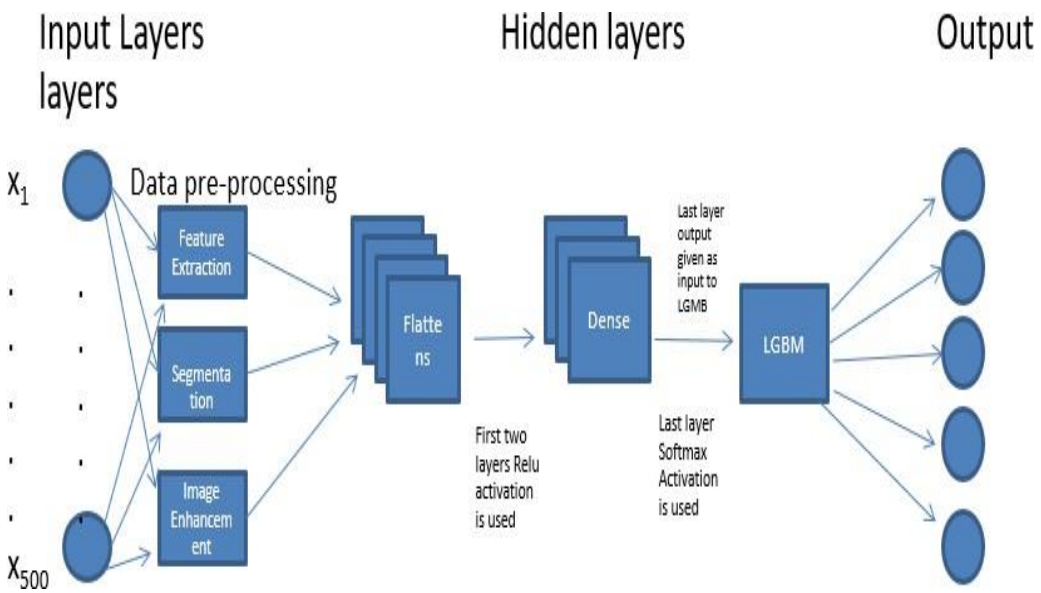


Fig. 5. Representation of proposed I

X & Y representation

0 - Dermoid Cyst
1 - Endrometriotic
2 -Hemorrhagic
3 - Malignant
4 - Pcos

Three segmentation techniques used separately during] e of the results are compared.

a) Edge based segmentation with DNN-LGMB

When edge-based segmentation is used as a pre- an accuracy of around 99.39% is achieved. A detai n be seen in Fig 6.

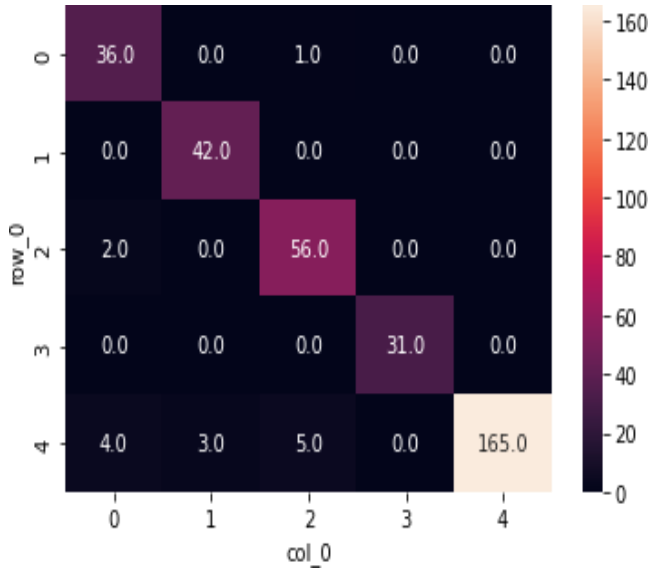
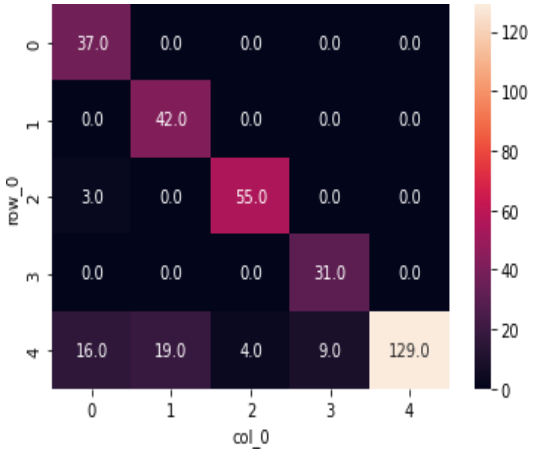


Fig. 6. Correlation matrix representing Edge based segmentation with DNN-LGMB

b) Region based segmentation with DNN-LGMB

97.33% accuracy is achieved when region - based segmentation is used as pre- processing technique. Confusion Matrix for 5 classes in region based segmentation can be observed in Fig 7.



X & Y representation

0 - Dermoid Cyst

1 - Endrometriotic

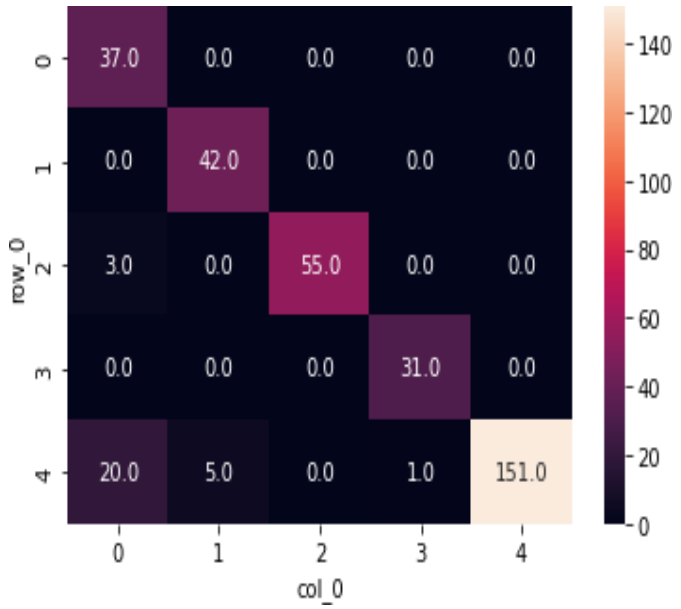
2 -Hemorrhagic

3 - Malignant

Fig. 7. Correlation matrix representing Region based segmentation with DNN-LGMB

c) Cluster based segmentation with DNN-LGMB

Using Cluster based segmentation with the data we arrive at 98.52% accuracy. Fig 8 shows the confusion matrix for Cluster based segmentation.



X & Y representation

0 - Dermoid Cyst
1 - Endrometriotic
2 -Hemorrhagic
3 - Malignant
4 - Pcos

Fig. 8. Correlation matrix representing cluster based segmentation with DNN-LGMB

From Fig 9 it is evident that the model with Edge based segmentation technique outperformed other models with and accuracy of 99.39%.

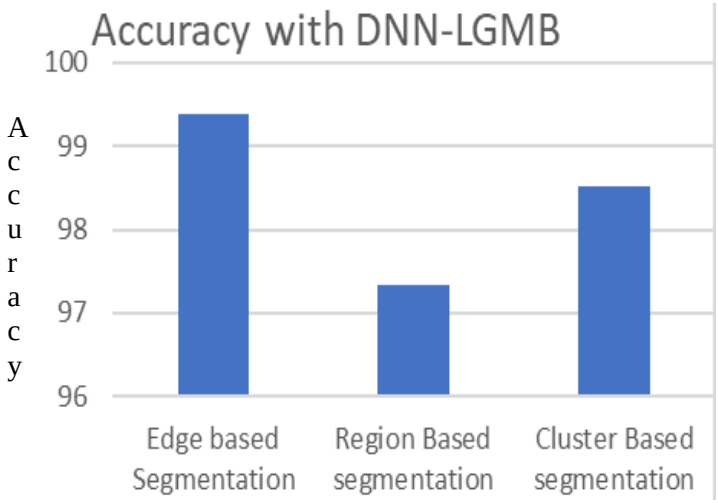


Fig. 9. Comparison of results

REFERENCES

[1]. S. Bharati, P. Podder and M. R. Hossain Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1486-1489, doi: 10.1109/TENSYMP50017.2020.9230932.

[2]. A. Denny, A. Raj, A. Ashok, C. M. Ram and R. George, "i- HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 673-678, doi: 10.1109/TENCON.2019.8929674.

[3]. P. Mehrotra, J. Chatterjee, C. Chakraborty, B. Ghoshdastidar and S. Ghoshdastidar, "Automated screening of Polycystic Ovary Syndrome using machine learning techniques," 2011 Annual IEEE India Conference, Hyderabad, India, 2011, pp. 1 - 5, doi: 10.1109/INDCON.2011.6139331.

[4]. Dewi, R.M. and Wisesty, U.N., 2018, March. Classification of polycystic ovary based on ultrasound images using competitive neural network. In *Journal of Physics: Conference Series* (Vol. 971, No. 1, p. 012005). IOP Publishing.

[5]. Seo, H., Badiei Khuzani, M., Vasudevan, V., Huang, C., Ren, H., Xiao, R., Jia, X. and Xing, L., 2020. Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications. *Medical physics*, *47*(5), pp.e148-e167.

[6]. M. Agarwal, V. K. Bohat, M. D. Ansari, A. Sinha, S. K. Gupta and D. Garg, "A Convolution Neural Network based approach to detect the disease in Corn Crop," 2019 IEEE 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India, 2019, pp. 176-181, doi: 10.1109/IACC48062.2019.8971602.

[7]. Csizmadia, G., Liszkai-Peres, K., Ferdinandy, B., Miklósi, Á. and Konok, V., 2022. Human activity recognition of children with wearable devices using LightGBM machine learning. *Scientific Reports*, *12*(1), pp.1-10.

[8]. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J. and Müller, K.R., 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), pp.247-278.