International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Classification of Mammograms using Various Feature Extraction Methods and Machine Learning

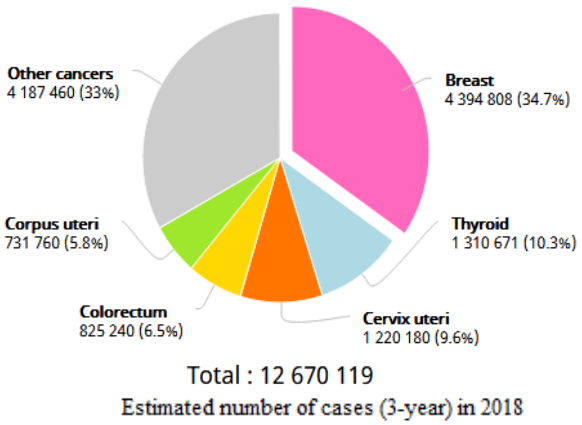
Apanveer kaur, Amit Doegar

Abstract*: Breast cancer is an alarming disease which takes millions of lives every year. In 2018, it was anticipated that 627,000 women died due to breast cancer – which is around 15% of all deaths caused due to different types of cancers among women. Currently, risk factors of breast cancer cannot be avoided, and early detection is the only way of survival. Automated detection of breast cancer with the help of image processing methods and machine learning algorithms helps in giving more accurate results and less human power. In the proposed system, multiple features are extracted using HSV histogram, LBP, GLCM, 2-D DWT. Support vector machine and LIBSVM classifiers are used for the classification of mammogram images if it's benign or malign in nature. For classification, the INbreast dataset have been used which includes 115 cases containing 410 images. The dataset is divided into benign and malign category based upon BI-RAIDS scale. According to this partition we have 243 benign images and 100 malign images present in this dataset and a feature matrix of 595 features in total is generated for balanced and unbalanced datasets respectively and fed into SVM and LIBSVM to distinguish the data. The balanced datasets on LIBSVM gave best results with 92% accuracy, 84% sensitivity, 100% specificity and 91.30% F1 score followed by SVM which gave 75% accuracy, 73.61% sensitivity, 76.66% specificity and 75.8% F1 score.

***Index Terms*: Breast Cancer, LBP, GLCM, DWT, LIBSVM, SVM, Mammogram.**

I. INTRODUCTION

Breast cancer has become one of the main causes of death among women all over the world affecting 2.1 million women every year and causes higher rate of death amongst women. In 2018, it was anticipated that 627,000 women died due to breast cancer – which is around 15% of deaths caused due to different types of cancers among women. Currently, risk factors of breast cancer cannot be avoided, and early detection is the only way of survival. A mammography is one of the breast cancer screening tools that uses low light x-rays to get an enhanced visualization of internal structure of the breast tissues.



Due to huge increment in visual and multimedia data, the retrieval of similar multimedia content is an open problem. In the process of retrieving an image the basic requirement is to arrange the images with similarity on the basis of color, texture and shape. Color, texture and shape are the low level features of image which play vital role in image processing. The representation of an image is in the form of feature vector and different techniques are applied to extract these features that help in classification and recognition of an image. A feature is generally used to capture the visual property of an image such as Color, Shape and texture [1]. An image feature can either be local or global feature. Local feature defines the particular region's object of the image (small set of pixels) while global feature defines the visual content of an entire image.

The texture and shape content in mammograms is rich. Therefore, one of the necessary steps is to extract suitable textural and shape features from mammogram using an appropriate feature extraction method. Some of the most commonly used approaches for feature extraction are **local binary pattern (LBP)** [2] , Gabor filter, **discrete wavelet transform (DWT)** [3], bag of visual words (BoVW), **gray level co-occurrence matrix (GLCM)** [4], fisher vector (FV) etc. After extracting the suitable features as per the requirement, the last step is to analyze the best classifier according to the dataset that can provide most accurate results. The better accuracy of system depends upon extracted features and the type of data used.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Apanveer kaur*, Department of Computer Science & Engineering, NITTR, Sector 26, Chandigarh , India.
Amit Doegar, Department of Computer Science & Engineering, NITTR, Sector 26, Chandigarh – 160019, India.

In the proposed work, an automated system is designed to classify mammogram images on the basis of their texture, color and the nature of the image is determined if its cancerous or normal with the help of Support vector machine (SVM) and library of SVM (LIBSVM)[5]. Our study is focused on mammogram dataset INbreast [6] containing 115 cases in the form of DICOM images.

The other sections of the paper are organized in following manner: Section II presents related work. Preprocessing, Techniques to retrieve features and classifiers for the classification of mammograms are mentioned in Section III. Section IV represents the dataset and performance evaluation metrics. Conclusion section concludes and sums up the experimental results in Section V.

II. RELATED WORK

In recent years, many of the research works has focused on early detection of breast cancer with the help of digital mammograms using CAD (computer aided diagnosis) and various image processing techniques. In this section, we will go through brief description of some recent works that have relevance with the proposed methodology.

In Reference [8] the author developed a CAD system where the suspicious regions were extracted from mammograms and Zernike moments of different order was used to create feature vector and classification of DDSM dataset took place by SVM classifier. Reference [9] presented content based image retrieval (CBIR) system for retrieving images based upon their content. The author applied vector quantization to compress the images by creating codebook and for transformation of image into a fuzzy signature. Then Fuzzy S-tree was used for finding similar fuzzy signatures as compare to given query image and return a list of similar images using Euclidean distance metrics.

A method for the classification of mammograms as normal or abnormal and further into malign or benign was discussed in [10]. MIAS and DDSM database was used to test the system. 2D-DWT was applied to get detailed coefficients from region of interest (ROI) of a mammogram then GLCM was used to generate feature matrix.

The author applied histogram equalization followed by morphological enhancement, then the segmentation of region of interest with the help of Otsu's threshold in [16]. The features of ROI was extracted using GLCM and some statistical measures like entropy, energy, homogeneity etc. were calculated based upon GLCM. At the end different classifiers were applied to classify the data into normal or micro-calcification and further malign or benign micro-calcification. SVM classifier gave best result among all other classifiers.

An optimal DWT algorithm and a revised GLCM technique was used for extracting texture features from segmented mammographic images and indicated good accuracy when using texture features in [17]. Reference [18] introduced a mass detection model on the basis of complicated texture features and achieved great performance. The texture features should therefore be provided more consideration in this context.

Reference [19] represents an analysis on LBP variants proposed in the literature. Taxonomy of the LBP variants was proposed for classifying various methods as per their roles in feature extraction. A total of 40 methods including 32 LBP variants based features were implemented on thirteen datasets and strengths & limitations of LBP variants were evaluated. One of the main advantages was its low computational complexity.

Table 1: Summary of some widely used mammogram datasets

Ref No.	Feature Vector	Technique	Database
[7]	DWT,GLCM	ACFNN classifier	MIAS
[8]	Zernike Moments	SVM classifier	IRMA, DDSM
[9]	Vector Quantization	CBIR using Fuzzy S-tree	DDSM
[10]	2D-DWT,GLCM	BPNN classifier	MIAS, DDSM
[11]	2D-BDWT, GLCM, FOA	SVM classifier	MIAS
[12]	GLCM, LBP, Intensity features, Ranklet features, Fractal features.	Combination of various classifiers	MIAS, INbreast
[13]	2D-PCA	SVM classifier	IRMA
[14]	----------	CNN classifier	INbreast, IMG
[15]	----------	CNN classifier	DDSM, BCDR, INbreast, MIAS

Reference [20] presented a method in which gray level co-occurrence matrix and center symmetric local binary pattern (CSLBP) was used. Local features of an image are extracted using CSLBP and a map is obtained which is observed in different distances and directions using GLCM.

III. PROPOSED METHODOLOGY

The proposed methodology is shown in Fig.1

A. Preprocessing

Preprocessing is done to improvise the features of images and get better data for the image that is to be used in future processing. The Mammogram dataset used for the work had large Dicom images with different sizes that were scaled down to 1000 × 1000 pixels as if the original images were to be used it will take more processing time. Further, the contrast of image pixels was enhanced that were difficult to distinguish visually.

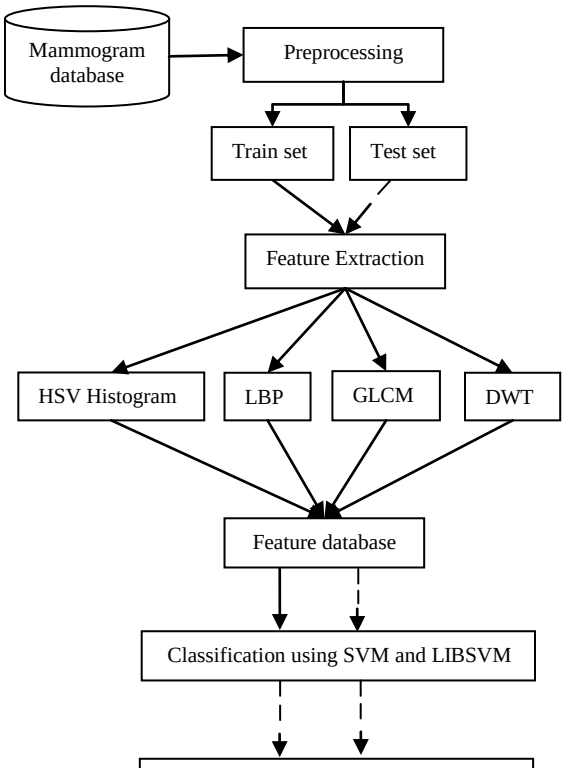


Fig. 1 Flow chart of proposed methodology

B. HSV His

In HSV (Training set / Testing set value) histogram, the given input image is tra... a RGB color space to HSV color space and a histogram is generated for each layer individually. HSV color space is used for extracting detail features that cannot be seen or detected in RGB color space.

C. LBP

LBP is a method for extracting the texture features of an image. It is an uncomplicated and efficient method with computational simplicity. Each pixel of an image is labeled by thresholding the neighbors of all pixels and binary number is considered as a result. Each pixel of image is considered as a center pixel at a time and neighbor pixels are compared with the center pixel. If the value of neighboring pixel is more than or equal to center pixel then it is replaced with '1' else it is replaced with '0'. Fig 2 shows an explanation of LBP method.

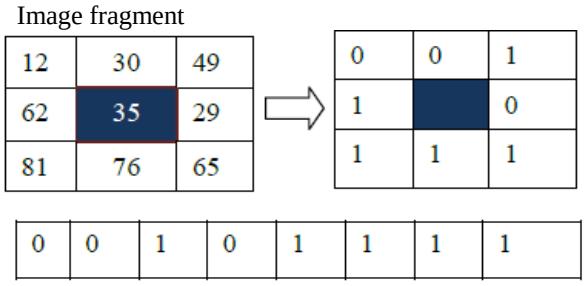


Fig. 2 LBP feature extraction methodology

LBP=0*1+0*2+1*4+0*8+1*16+1*32+1*64+1*128
=244

$$LBP_{P,R} = \sum_{s=0}^{P-1} T_1\,(I_s - I_c) \times 2^s \qquad (1)$$

$$T_1(a) = \begin{cases} 1 & a \geq 0 \\ 0 & else \end{cases}$$

where P and R represents the number of neighbor pixels and radius of neighboring pixels. Center pixel and surrounding pixels are represented as I_c and I_s. The LBP values calculated are the feature vector used to classify the images.

D. GLCM

GLCM is also a method to extract texture feature from an input image. GLCM converts an image into a matrix which represents the relationship between the neighboring pixels. Pixels pairs with mutual occurrences in specific direction 'Θ' and distance 'd' gives GLCM matrix. Value of 'Θ' can be 0˚, 45˚, 90˚, 135˚.

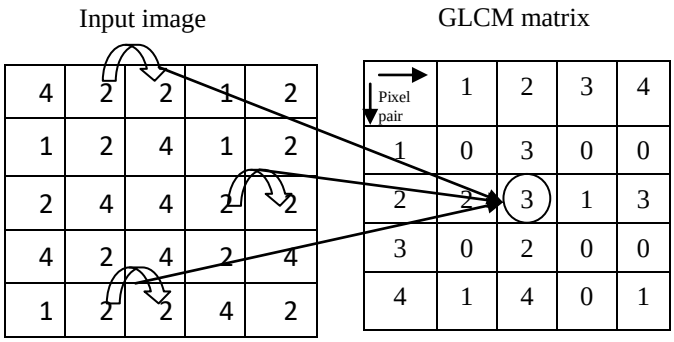


Fig. 3 GLCM example

GLCM for an input image is computed in equation 2 by using distance d_r, d_c for row and column respectively, where value of G (i, j) is the number of occurrences of pair i and j in the image matrix:

$$G(i,j) = (r,c)|\, I(r,c) = i and\, I\left(r + d_r, c + d_c\right) = j \qquad (2)$$

The example in Fig. 3 calculated the GLCM matrix with distance d=1 and Θ=0˚. This GLCM matrix is further used to calculate statistical properties like Contrast, Correlation, Energy, and Homogeneity. These four features are then used as feature vector to classify the data.

$$contrast = \sum_{i,j=0}^{N-1} P_{ij}(i-j)^2 \qquad (3)$$

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

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

$$Homogeneity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2} \qquad (6)$$

where P_{ij} is the value in GLCM matrix and i, j represents the row and column. μ and σ represents mean and variance respectively.

E. Discrete Wavelet Transform

DWT is a powerful tool used for extracting the feature of an image. It captures frequency as well as time information which are useful for classification. It decomposes the input image into sub-bands with pair of wavelet filters (low-pass and high-pass filters). Haar wavelet transform is used for decomposition in our work.

After first level decomposition an input image is divided into four bands i.e. LL representing the approximate image, LH gives horizontal details, HL gives vertical details and HH represent diagonal details of input image. The first level LL band is further divided into second level LL, LH, HL, HH band and LL band of second level is used to calculate the mean and standard deviation which are used as features. Fig.4 shows two level decomposition of original image and the LL band of second level is used for generating the feature matrix for further classification.

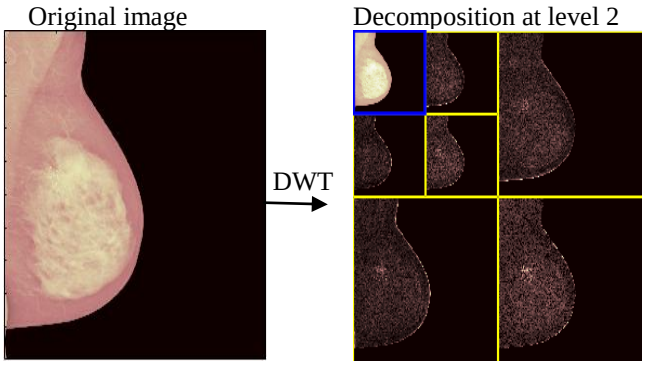


Fig. 4 2-D DWT Decomposition at level 2

F. Classifier

Classification is the process to distinguish the different classes: Benign and Malign classes in our study. Two classifiers have been used in our work: SVM and LIBSVM. SVM, based upon supervised learning technique, construct a hyper-plane between the two classes to classify them appropriately with maximum margin. Another classifier LIBSVM is a library of SVM which includes the source code of the library in C++ and Java. We can specify the LIBSVM options i.e. type of svm & kernel to be used as per our requirement while training it with the dataset. In this study, we have used nu-SVC type of SVM with linear kernel type and it gave best results amongst other with 92% accuracy.

IV. RESULTS AND DISCUSSION

In our research work, we have used INbreast dataset which includes 115 cases containing 410 images. We have divided the dataset into benign and malign category on the basis **of Breast Imaging Reporting and Data System (BI-RAIDS)** scale. There are six categories (1 to 6) of BI-RAIDS on the basis of suspensions where 1 is considered as normal, 2&3 as benign and 4, 5, 6 as malign. According to this partition we have 243 benign images and 100 malign images in this dataset.

The features extracted using GLCM, LBP, HSV Histogram and 2-D DWT were taken into consideration and were used as input to the SVM and LIBSVM classifiers.

A. Performance measures

The performance of the proposed work is evaluated on the basis of accuracy, sensitivity, specificity, F1 score, precision and recall. These are calculates on the basis of confusion matrix shown in Fig. 5.

True Class \ Predicted class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Fig. 5 Confusion Matrix

where TP, FN, FP, TN is true positive, false negative, false positive, true negative respectively.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+} \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Recall(Sensitivity) = \frac{TP}{TP+FN} \quad (9)$$

$$F1\ score = \frac{2 \times (Recall * Precision)}{(Recall + Precision)} \quad (10)$$

$$Specificity = \frac{TN}{(TN+FP)} \quad (11)$$

Initially we have tested our proposed framework on unbalanced dataset including 243 benign and 100 malign images. Table 2 represents the results using SVM and LIBSVM classifiers.

Then we have tested this framework on balanced dataset of 243 benign images and 200 malign images which include 100 malign images generated through augmentation. Table 3 represents the results on basis of SVM and LIBSVM classifiers.

Table 2: Results on unbalanced dataset

	LIBSVM	SVM
Accuracy	52%	69.6%
Precision	66.12%	71.13%
Recall	59.42%	95.83%
F1 score	62%	81.83%
Specificity	35.48%	6.66%

Table 3: Results on balanced dataset

	LIBSVM	SVM
Accuracy	92%	75%
Precision	100%	79.1%
Recall	84%	73.61%
F1 score	91.30%	75.8%
Specificity	100%	76.66%

Fig.5 is the graphical representation of Table 3.

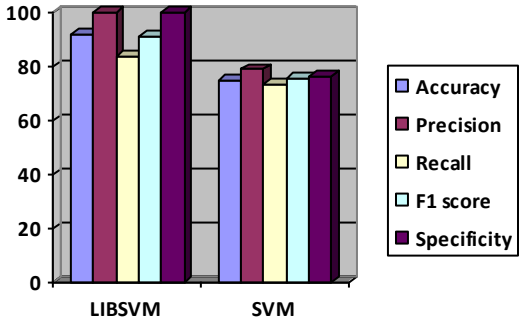


Fig. 5 Graphical representation of Performance of LIBSVM and SVM balanced dataset

V. CONCLUSION AND FUTURE SCOPE

In this paper, the proposed system works on classification of data into benign and malign classes. The study includes various feature extraction methods like LBP, GLCM, and DWT to extract the features of medical images. For classification, the feature vector of 595 features in total is generated and fed into SVM and LIBSVM to distinguish the data. The balanced datasets on LIBSVM gave best results with 92% accuracy, 84% sensitivity, 100% specificity and 91.30% F1 score followed by SVM which gave 75% accuracy, 73.61% sensitivity, 76.66% specificity and 75.8% F1 score. Hence, It is concluded that the classifier give better results if the dataset used is balanced, otherwise it can only classify a particular class which is more in number.

In future, a large and balanced dataset can be used to test the proposed work so that the accuracy of system can be made consistent. As the model is tested using Machine learning algorithm, further we can use deep learning as well to extract more features using CNN model

REFERENCES

1. Mohamed Aly, Peter Welinder, Mario Munich, Pietro Perona, "Automatic discovery of image families: Global vs. local features", *Image Processing (ICIP), 16th IEEE International Conference on*, 2009, pp. 777-780.
2. Mithlesh Arya, Namita Mittal, Girdhari Singh, "Texture-based feature extraction of smear images for the detection of cervical cancer", *IET Computer Vision*, vol. 12, no. 8, pp. 1049-1059,2018.
3. Atish Chaudhary, Vandana Bhattacharjee, "An efficient method for brain tumor detection and categorization using MRI images by K-means clustering & DWT", *International Journal of Information Technology*, pp. 1-8, 2018.
4. N. Varuna Shree, T. N. R. Kumar, "Identification and classification of brain tumor MRI images with feature extraction using DWT and probabilistic neural network", *Brain informatics*, vol. 5, no. 1, pp. 23-30, 2018.
5. Ines C Moreira, Igor Amaral, and Ines Domingues, and Antonio Cardoso, Maria Joao Cardoso and Jaime S Cardoso , "Inbreast: toward a full-field digital mammographic database", *Academic radiology*, vol. 19, no. 2, pp. 23--248, 2012.
6. Chih-Chung Chang and Chih-Jen Lin , "LIBSVM: A library for support vector machines", ACM transactions on intelligent systems and technology, vol. 2, no. 3, pp. 27,2011.
7. Nirase Fathima Abubacker, Azreen Azman, Shyamala Doraisamy, Masrah Azrifah Azmi Murad, "An integrated method of associative classification and neuro-fuzzy approach for effective mammographic classification",, *Neural Computing and Applications*, vol. 28, pp. 3967-3980, Dec 2017.
8. Shubhi Sharma, Pritee Khanna, "Computer-Aided Diagnosis of Malignant Mammograms using Zernike Moments and SVM", *Journal of digital imaging*, vol. 28, no. 1, pp. 77-90, 2015.
9. Jana Nowakova, Michal Prilepok and Vaclav Snasel, "Medical image retrieval using vector quantization and fuzzy S-tree", *Journal of medical systems*, vol. 31, no. 2, pp. 18, 2017.
10. Shradhananda Beura, Banshidhar Majhi, Ratnakar Dash, "Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer", *Neurocomputing* vol. 154, pp. 1-14, April 2015.
11. Figlu Mohanty, Suvendu Rup, Bodhisattva Dash, Banshidhar Majhi, M. N. S. Swamy, "Digital mammogram classification using 2D-BDWT and GLCM features with FOA-based feature selection approach", *Neural Computing and Applications*, pp. 1-15, April 2019.
12. Ehsan Kozegar and Mohsen Soryani,"A cost-sensitive Bayesian combiner for reducing false positives in mammographic mass detection", *Biomedical Engineering*, vol. 64, no. 1, pp. 39-52, 2019.
13. Júlia E.E. de Oliveiraa, Alexei M.C. Machadob, Guillermo C. Chaveza,Ana Paula B. Lopesa, Thomas M. Desernoc, Arnaldo de A. Araújoa, "MammoSys: A content-based image retrieval system using breast density patterns", *Computer methods and programs in biomedicine*, vol. 99, no. 3, pp. 289-297, 2010.
14. J. Sulam, R. Ben-Ari and P. Kisilev,"Maximizing AUC with Deep Learning for Classification of Imbalanced Mammogram Datasets", *VCBM*, pp. 131-135, 2017.
15. Hiba Chougrad, Hamid Zouaki, Omar Alheyane, "Deep Convolutional Neural Networks for Breast Cancer Screening", *Computer methods and programs in biomedicine*, vol. 157, pp. 19-30, 2018.
16. Hayat Mohamed, Mai S. Mabrouk, Amr Sharawy, "Computer Aided Detection System for Micro-calcifications in Digital Mammograms", *Computer methods and programs in biomedicine*, vol. 116, no. 3, pp. 226-235, 2014.
17. Aleel JA, Salim S, Archana S , "Textural features based computer aided diagnostic system for mammogram mass classification", *IEEE International conference on control, instrumentation, communication and computational technologies (ICCICCT)*, 2014, pp 806–811.
18. Tai SC, Chen ZS, Tsai WT, "An automatic mass detection system in mammograms based on complex texture features", *IEEE journal of biomedical and health informatics*, vol. 18, no. 2, pp. 618--627, 2013.
19.] Li Liu, Paul Fieguth, Yulan Guo, Xiaogang Wang, Matti Pietikäinen, "Local binary features for texture classification: Taxonomy and experimental study", *Pattern Recognition*, vol. 62, February 2017, pp. 135-160
20. Manisha Verma, Balasubramanian Raman, "Center symmetric local binary co-occurrence pattern for texture, face and bio-medical image retrieval", *Journal of Visual Communication and Image Representation*, vol. 32, pp. 224-236, 2015.

AUTHORS PROFILE

Apanveer kaur pursuing M.E in Computer Science and Engineering Department from NITTTR, Chandigarh.

Mr.Amit Doegar, M.E from Panjab University Chandigarh, B.E. (Computer Science and Engineering) from Karnatak University, Dharwar. **Research Interests:** Computer Networks, Image Processing, Virtual Learning, Open Source Technology.