Vol. 3, No. 2 | Jul – Dec 2019

Performance Comparison of SVM and its Variants for the Early Prognosis of Breast Cancer

Talha Ahmed Khan[1,2], Muhammad Alam[3,4], Zeeshan Shahid [5] , M.S. Mazliham[6]

Abstract:

Breast cancer has become a leading cause of women death in this era. Breast cancer is very common in various countries including Pakistan. Early identification of the breast cancer or tumor is the only way for the rapid treatment and cure. An imaging approach named as mammography has performed tremendous job in the field of medical to detect the cancer tumors on early basis with less false alarm rate. Breast cancer has two types of tumors a) Benign and b) Malignant. Malignant is acknowledged as cancer tumor as it spread and grow rapidly inside the tissues. Detection of Malignant tumor is very complex in dense breast as it is covered and linked with the milk glands, ducts and other related tissues. Therefore, machine learning and artificial intelligence approaches were needed as mammographic images required edge detection, image enhancement and image processing. Various Artificial Intelligence based algorithms have been applied to the clinical breast cancer data set for the early detection of breast tumor. In this research work the clinical data has been collected from the UCI machine learning repository for the classification of breast cancer tumor a) Benign and b) Malignant. Support Vector machine with its variants Kernel, Gaussian Kernel and Sigmoid Kernel have been applied to the linearly separable breast cancer data set for comparative analysis. Results proved that all the variants of SVM performed better for the breast cancer classification.

Keywords: *Benign, Malignant, Breast cancer, dense breast, fatty breast, Support Vector Machine*

1. Introduction

Digital images based elasto tomography (DIET) was proposed for the robust evaluation of breast tumor. Vibrated breast image was captured and surface motion was recorded to track the tumor. 3D surface motion was accurately evaluated using this approach. DIET based approach could detect upto 100 mm. of tumor [1]. Gene expression pattern data sets were used to find out the similarities and differences between breast tumor and peripheral blood monocular cells [2]. An intelligent system was proposed comprised of

[1] *British Malaysian Institute (BMI), Universiti Kuala Lumpur, Malaysia*

[2] *Usman Institute of Technology, Karachi, Pakistan*

[3] *CCIS, Institute of Business Management, Karachi, Pakistan*

[4]*Malaysian Institute of Information and Technology (MIIT), Universiti Kuala Lumpur Malaysia*

[5] *Electrical Engineering Department, Institute of Business Management, Karachi, Pakistan*

[6] *Malaysian France Institute (MFI), Universiti Kuala Lumpur Malaysia*

neural network classifier and image processing (ISIBC). The features for the classification have been obtained by applying GLCM algorithm. Standard deviation, mean and entropy were calculated. Filtering, edge detection and morphological techniques were applied to the images for the feature's extraction [3]. Mammographic images were collected from the data bases available on the internet [3]. Image processing method were applied to filter the noise and to enhance the mammographic images for the robust investigation of breast cancer. Adaptive histogram equalization, wavelets and data fusion methods were applied to identify micro-calcifications using LabView [4]. Reproduction of cells is the main indication of the breast cancer. Reproduction can usually be examined by the color of the cells for the antigen Ki-67. Artificial Neural network comprising of the development of multi-layer perceptron network with the combination of feed forward neural network was proposed for the accurate and precise classification of three different cancer genes which are linked with Ki-67. Results proved that unique non-recursive method was very successful to provide the complete detailed genes network with the visualizations [5].

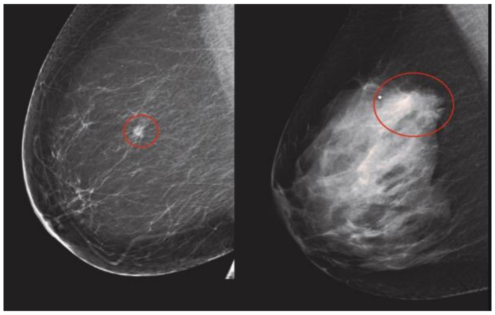


Fig. 1. Mammogram image of breast tumor in fatty and dense breast.

Fig. 1 demonstrated the downloaded image of the mammogram which has been downloaded from the mammographic images databases available on the internet. The figure explains that it is very easy to determine the malignant tumor in a fatty breast compared to the dense breast as it is very complex to identify the malignant tumor in the dense forest. Even a big tumor is very difficult to analyze in a dense breast. Therefore, image processing and filtering techniques are needed to enhance the image quality and features. Mammography is considered as the widely used tool for the detection of the malignant tumor in the breast but in dense breast tissues may overlap each other which causes the confusion for exact investigation of malignant tumor. Tissue overlapping may be minimized by using digital tomosynthesis. Dense breast can be defined as the combination of milk ducts, glands and linked tissues. 4-D virtual prototype breast were designed that were compared to the fourteen imaging approaches [6]. Mammographic filtered images were divided into small grids and features were extracted from these small grids for better classification of Malignant and Benign. Results proved that grid-based pattern evaluation for the doubtful tumor or tissue performed better with the efficiency of 91.67%. The results were also compared with the knowledge-based data base provided by the radiologist [7]. OTSU method was also adopted to reduce the noise and segment the mammogram images to discriminate the changes compared to the normal one [8]. Simplest evolving connectionist system was suggested for the classification of normal, benign and malignant classification. Wisconsin dataset was collected from UCI machine learning repository and sensitivity of 96.02% was achieved [9]. Computer aided diagnostic based approach was developed and 82% sensitivity was achieved [10].

2. Problem Statement

Women normally take the cancer tumor symptoms very lightly due to the the lack of awareness for the malignant tumor which spreads and grow inside the breast tissues. Vigorous classification of malignant and Benign tumor can be acknowledged as very complex due to the similarities between them which creates confusion [11-14]. Mammography is the most widely used technique for the screening of breast cancer but image filtering, processing and machine learning techniques were needed for the accurate results as identification of malignant tumor in dense breast is very difficult. Skin inflammation and breast cancer have some common symptoms. Breast pain, swelling, and change of skin color are very common symptoms of cancer but people ignore it as they take it as normal skin inflammatory problem [15-17].

3. Methodology

3.1. Fundamental Block Diagram

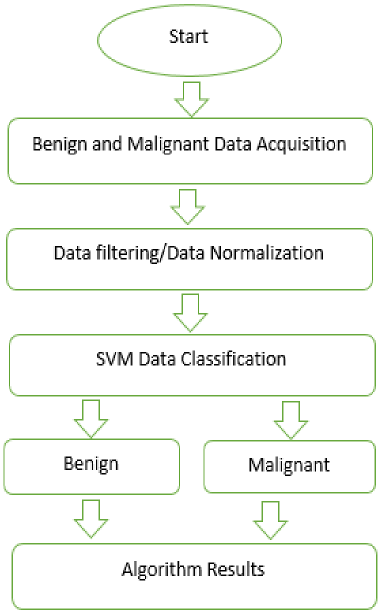
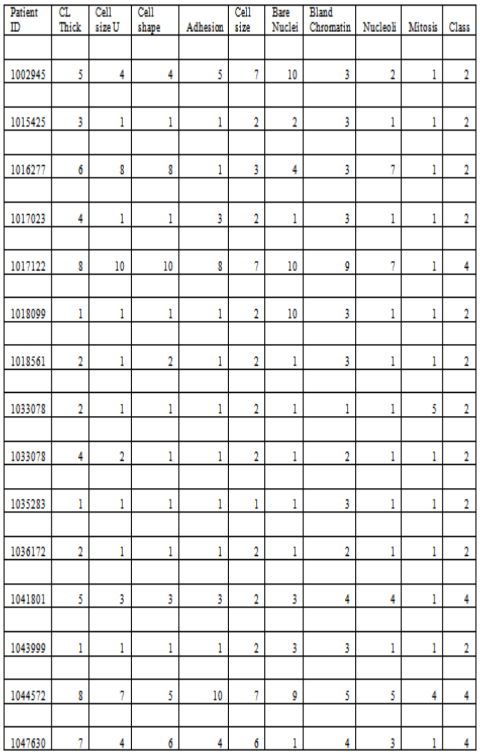

Start
Benign and Malignant Data Acquisition
Data filtering/Data Normalization
SVM Data Classification
Benign
Malignant
Algorithm Results

Fig. 2. Fundamental Block diagram

Fig. 1 demonstrated the fundamental block diagram for the breast cancer classification. Breast Cancer tumor can be classified into Benign and Malignant. Malignant is the breast tumor which can grow with the tissues and spread rapidly. Breast cancer data acquisition was performed and then the data was normalized as the data contained some missed and repetitive values. SVM classifier with the variants of Kernel, Sigmoid and Gaussian Kernel were applied to the breast cancer data set for the classification of Benign and Malignant tumor. The results were compared with each other variants of the SVM. SVM proved to be the competent classifier.

3.2. Breast cancer Dat collection

Table I: Breast cancer data set [18-19]

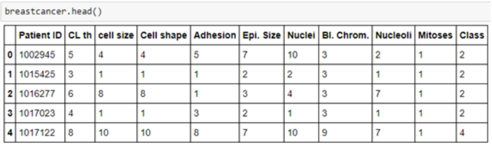


Patient ID	CL Thick	Cell size U	Cell shape	Adhesion	Cell size	Bare Nuclei	Bland Chromatin	Nucleoli	Mitosis	Class
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4

Table no. 1 displayed all the attributes related to the identification of the breast tumor. Attributes can be explained as:

1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10

6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class: (2 for benign, 4 for malignant)



```
breastcancer.head()
```

	Patient ID	CL th	cell size	Cell shape	Adhesion	Epi. Size	Nuclei	Bl. Chrom.	Nucleoli	Mitoses	Class
0	1002945	5	4	4	5	7	10	3	2	1	2
1	1015425	3	1	1	1	2	2	3	1	1	2
2	1016277	6	8	8	1	3	4	3	7	1	2
3	1017023	4	1	1	3	2	1	3	1	1	2
4	1017122	8	10	10	8	7	10	9	7	1	4

Fig. 3. Imported data set in Python

Data set was imported in python and then normalized to filter the data. Data normalization was performed as data contained some missed and repetitive values.

4. Algorithm Implementation

4.1. Support Vector Machine (SVM)

Support Vector Machine classifier algorithm was applied to the breast cancer data set for the accurate classification of brain tumor. Pandas and Numpy libraries were imported in the Python environment for the implementation of SVM. SVM algorithm creates a boundary by using hyperplanes for the classification. Simple SVM is very useful to classify the linear data. Two hyperplanes were created for this data classification.

H1, H2 are the hyper planes:

H1: $w \bullet x_i + b = 2$ (1)

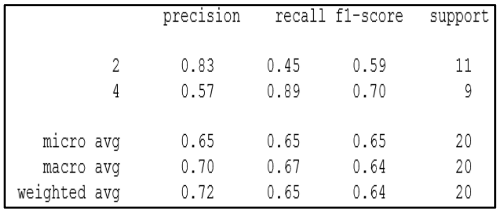
H2: $w \bullet x_i + b = 4$ (2)

SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='auto_deprecated',

kernel='linear', max_iter=-1,

probability=False, random_state=None,

shrinking=True, tol=0.001, verbose=False)



```
              precision    recall  f1-score   support

           2       0.83      0.45      0.59        11
           4       0.57      0.89      0.70         9

   micro avg       0.65      0.65      0.65        20
   macro avg       0.70      0.67      0.64        20
weighted avg       0.72      0.65      0.64        20
```

Fig. 4. SVM Output results

Fig. 4 elaborates that precision of 0.83 and 0.57 have been achieved for the class 2 and class 4 respectively. 0.72 average precision has been achieved by using simple support vector machine.

4.2. Kernel Support Vector Machine (SVM)

Simple Support vector machine is applied to classify the linearly separable data. But for the classification of non-linear data straight line of decision boundary cannot be used. Therefore, a variant of support vector machine named as Kernel was proposed which included the mathematical solution as well for solving the non-linearly separable data. Scikit-learn was needed to implement the Kernel support vector machine with these parameters.

SVC(C=1.0, cache_size=200, class_weight=

None, coef0=0.0,

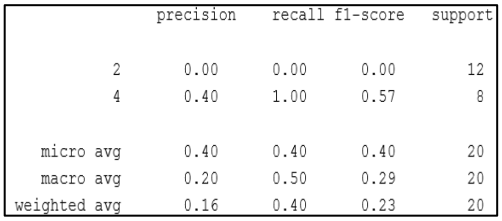
decision_function_shape='ovr', degree=8, ga

mma='auto_deprecated',

kernel='poly', max_iter=-1, probability=False

, random_state=None,

shrinking=True, tol=0.001, verbose=False)



```
              precision    recall  f1-score   support

           2       0.00      0.00      0.00        12
           4       0.40      1.00      0.57         8

   micro avg       0.40      0.40      0.40        20
   macro avg       0.20      0.50      0.29        20
weighted avg       0.16      0.40      0.23        20
```

Fig. 5. Kernel SVM results

Fig. 4 shows that precision, recall. F1-score and support have been calculated to determine the performance of Kernel SVM classifier.

4.3. Gaussian Kernel Support Vector Machine (SVM)

$$k(x, x0) = \exp {}^{3}-\|x - x0\|2/2\sigma2{}' \qquad (3)$$

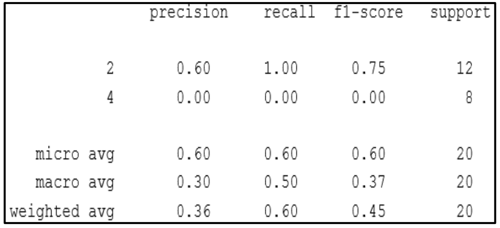
Gaussian Kernel support vector machine was applied to the breast cancer data set and following results were achieved.

SVC(C=1.0, cache_size=200, class_weight= None, coef0=0.0,

decision_function_shape='ovr', degree=3, gamma='auto_deprecated',

kernel='rbf', max_iter=-1, probability=False, random_state=None,

shrinking=True, tol=0.001, verbose=False)



```
              precision    recall  f1-score   support

           2       0.60      1.00      0.75        12
           4       0.00      0.00      0.00         8

   micro avg       0.60      0.60      0.60        20
   macro avg       0.30      0.50      0.37        20
weighted avg       0.36      0.60      0.45        20
```

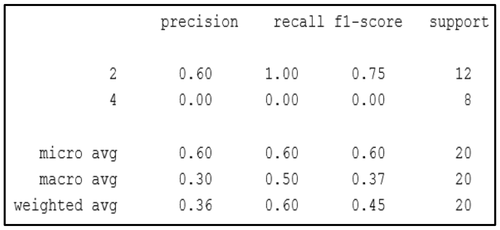
Fig. 6. Gaussian Kernel SVM results

4.4. Sigmoid Kernel Support Vector Machine (SVM)

Sigmoid Kernel can be applied to the data set for the classification by using the following mathematical equation

$$k(x, y) = \tanh(\alpha x^T y + c) \qquad (4)$$

SVC(C=1.0, cache_size=200, class_weight= None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='auto_deprecated', kernel='sigmoid', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)



```
              precision    recall  f1-score   support

           2       0.60      1.00      0.75        12
           4       0.00      0.00      0.00         8

   micro avg       0.60      0.60      0.60        20
   macro avg       0.30      0.50      0.37        20
weighted avg       0.36      0.60      0.45        20
```

Fig. 7. Sigmoid Kernel SVM results

Recall is the measurement of corrected classified values out of the all positive classes. The higher the recall the better the performance. Recall can be measured by using eq. (10).

$$Recall = \frac{Tp}{TP+FN} \qquad (5)$$

F-measure can be acknowledged as the comparative analysis or to know the comparison between recall and precision.

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} (6)$$

The estimation of actual positive out of all positive can be classes can be found be precision.

$$Precision = \frac{TP}{TP+FP} \qquad (7)$$

4.5. Fine Gaussian Support Vector Machine (SVM)

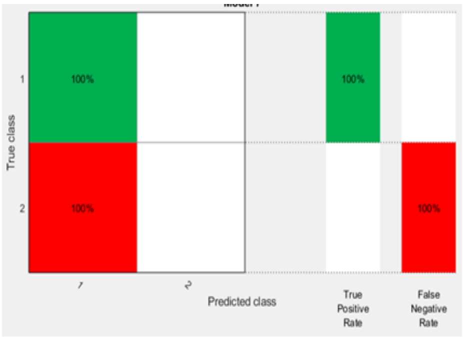


Fig. 8. Confusion Matrix of fine Gaussian SVM

Confusion matrix of fine Gaussian SVM elaborated that the algorithm performed very poor for the prediction of breast cancer as it classified all classes as class 1. Class 2 was not predicted at all therefore false negative rate was found to be 100% and false positive rate for class 2 was found to be 0%.

5. Comparative Analysis

Table II: SVM Parametric Evaluation (CLASS 2)

Classifiers	Precision	Recall	F1-Score	Support
SVM	0.83	0.45	0.59	11
Kernel SVM	0.0	0.0	0.0	12
Sigmoid Kernel SVM	0.60	1	0.75	12
Gaussian Sigmoid Kernel	0.60	1	0.75	12

Table II demonstrated that the SVM and its variants' parametric analysis for the classification of class 2. It can be observed that linear support vector machine achieved better results in terms of precision, recall, F1-score.

Table III: SVM Parametric Evaluation (CLASS 4)

Classifiers	Precision	Recall	F1-Score	Support
SVM	0.57	0.89	0.70	9
Kernel SVM	0.4	1	0.57	8
Sigmoid Kernel SVM	0	0	0	8
Gaussian Sigmoid Kernel	0	0	0	8

Table III demonstrated that the SVM and its variants' parametric analysis for the classification of class 4. It can be observed that linear support vector machine achieved better results in terms of precision, recall, F1-score.

6. Results and discussion

Support vector machine with its variants Kernel, Gaussian Kernel and Sigmoid Kernel have been applied to the data set of breast cancer. All the variants performed better. 0.57 have been achieved for the class 2 and class 4 respectively. 0.72 average precision has been achieved by using simple support vector machine. For the future work SVM can be combined with the Particle swarm optimization in which weights of Support vector machine would be optimized by PSO. The optimization of SVM weights would make the results better and up to the mark for the classification of breast cancer.

REFERENCES

[1] T. Botterill, T. Lotz, A. Kashif and J. G. Chase, "Reconstructing 3-D Skin Surface Motion for the DIET Breast Cancer Screening System," in *IEEE Transactions on Medical Imaging*, vol. 33, no. 5, pp. 1109-1118, May 2014. doi: 10.1109/TMI.2014.2304959

[2] L. He, D. Wang and Z. Guo, "Identification of Potential Non-invasive Biomarkers for Breast Cancer Prognosis and Treatment by Systematic Bioinformatics Analysis," *2015 7th International Conference on Information Technology in Medicine and Education (ITME)*, Huangshan, 2015, pp. 117-120. doi: 10.1109/ITME.2015.28

[3] A. Helwan and R. H. Abiyev, "ISIBC: An intelligent system for identification of breast cancer," *2015 International Conference on Advances in Biomedical Engineering (ICABME)*, Beirut, 2015, pp. 17-20. doi: 10.1109/ICABME.2015.7323240

[4] Meharun Nisa S.P and K. Suresh, "Labview implementation of identification of early signs of breast cancer," *2014 International Conference on Electronics, Communication and Computational Engineering (ICECCE)*, Hosur, 2014, pp. 264-267. doi: 10.1109/ICECCE.2014.7086624

[5] D. Agarwal, M. Kergosien, D. J. Boocock, R. C. Rees and G. R. Ball, "A systems biology approach to identify proliferative biomarkers and pathways in breast cancer," *2014 IEEE International*

Conference on Bioinformatics and Biomedicine (BIBM), Belfast, 2014, pp. 1-7. doi: 10.1109/BIBM.2014.6999240

[6] N. Kiarashi *et al*., "Development and Application of a Suite of 4-D Virtual Breast Phantoms for Optimization and Evaluation of Breast Imaging Systems," in *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1401-1409, July 2014. doi: 10.1109/TMI.2014.2312733

[7] P. Swapnil, E. Pandey, J. R. Yathav, A. Baig and A. Bailur, "Region Marking and Grid Based Textural Analysis for Early Identification of Breast Cancer in Digital Mammography," *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Bhimavaram, 2016, pp. 426-429. doi: 10.1109/IACC.2016.85

[8] J. Kamalakannan, P. V. Krishna, M. R. Babu and K. D. Mukeshbhai, "Identification of abnormility from digital mammogram to detect breast cancer," *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, Nagercoil, 2015, pp. 1-5. doi: 10.1109/ICCPCT.2015.7159454

[9] E. B. Nababan, M. Iqbal and R. F. Rahmat, "Breast cancer identification on digital mammogram using Evolving Connectionist Systems," *2016 International Conference on Informatics and Computing (ICIC)*, Mataram, 2016, pp. 132-136. doi: 10.1109/IAC.2016.7905703

[10] S. Pathan, P. C. Siddalingaswamy, L. Lakshmi and K. G. Prabhu, "Classification of benign and malignant melanocytic lesions: A CAD tool," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, 2017, pp. 1308-1312. doi: 10.1109/ICACCI.2017.8126022

[11] I. E. Magnin, D. Vray and A. Brémond, "Early detection of breast cancer using computer assisted diagnosis," *1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Paris, 1992, pp. 849-850.

[12] R. D. Ghongade and D. G. Wakde, "Detection and classification of breast cancer from digital mammograms using RF and RF-ELM algorithm," *2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech)*, Kolkata, 2017, pp. 1-6. doi: 10.1109/IEMENTECH.2017.8076982

[13] P. Král and L. Lenc, "LBP features for breast cancer detection," *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, pp. 2643-2647. doi: 10.1109/ICIP.2016.7532838

[14] B. Hela, M. Hela, H. Kamel, B. Sana and M. Najla, "Breast cancer detection: A review on mammograms analysis techniques," *10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13)*, Hammamet, 2013, pp. 1-6. doi: 10.1109/SSD.2013.6563999

[15] T. Kao *et al*., "Regional Admittivity Spectra With Tomosynthesis Images for Breast Cancer Detection: Preliminary Patient Study," in *IEEE Transactions on Medical Imaging*, vol. 27, no. 12, pp. 1762-1768, Dec. 2008.

[16] D. A. Woten, J. Lusth and M. El-Shenawee, "Interpreting Artificial Neural Networks for Microwave Detection of Breast Cancer," in *IEEE Microwave and Wireless Components Letters*, vol. 17, no. 12, pp. 825-827, Dec. 2007. doi: 10.1109/LMWC.2007.910466

[17] P. M. Meaney, M. W. Fanning, D. Li, S. P. Poplack, and K. D. Paulsen,"A clinical prototype for active microwave imaging of the breast,"IEEE Trans. Microw. Theory Tech., vol. 48, no. 11, pp. 1841–1853,Nov. 2000doi: 10.1109/MCS.2009.932223

[18] Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193--9196.

[19] Zhang, J. (1992). Selecting typical instances in instance-based learning. In Proceedings of the Ninth International Machine Learning Conference (pp. 470--479). Aberdeen, Scotland: Morgan

[20] T. Khan *et al*., "Foreign objects debris (FOD) identification: A cost effective investigation of FOD with less false alarm rate**,"** 2017 IEEE 4th International Conference on Smart Instrumentation,

Measurement and Application (ICSIMA)*, Putrajaya, 2017, pp. 1-4.*

[21] Talha Ahmed Khan, Kushsairy Kadir, Muhammad Alam, Zeeshan Shahid and M.S. Mazliham, "Geomagnetic Field Measurement at Earth Surface: Flash Flood Forecasting using Tesla Meter", *Proc. of the International Conference on Engineering Technologies and Technopreneurship (IEEE-ICE2T 2017) 18-20 September 2017, Kuala Lumpur, Malaysia*

[22] T. A. Khan, M. Alam, K. Kadir, Z. Shahid and S. M Mazliham, "A Novel Approach for the Investigation of Flash Floods using Soil Flux and CO2: An Implementation of MLP with Less False Alarm Rate," *2018 2nd International Conference on Smart Sensors and Application (ICSSA)*, TBA, Malaysia, 2018, pp. 130-134

[23] Talha Khan, Muhammad Alam, Kushsairy Kadir, Sheroz Khan, M.S Mazliham, Faraz Shaikh, Syed Faiz Ahmed, Zeeshan Shahid" An Implementation of Electroencephalogram signals acquisition to control manipulator through Brain Computer Interface ", *2nd IEEE International Conference on Innovative research and development 2019 (IEEE-ICIRD), 28-30 June 2019, Jakarta, Indonesia*