Scalable Computing: Practice and Experience

Scientific International Journal for Parallel and Distributed Computing

ISSN: 1895-1767

Volume 23(3)

September 2022

Scalable Computing: Practice and Experience

Volume 23, Number 3, September 2022

TABLE OF CONTENTS

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2022 SCPE. Volume 23, Issues 3, pp. 81–90, DOI 10.12694/scpe.v23i3.1993

MACHINE LEARNING BASED TECHNIQUE TO LEARN HIPPOCAMPAL ATROPHY FROM AXIAL MRI FOR ALZHEIMER'S DISEASE DIAGNOSIS

D. K. RAMKUMAR *, N. V. BALAJI † AND T. GENISH ‡

Abstract. Hippocampus (HC) is one of the small brain components and its features majorly take part in diagnosing diseases such as Alzheimer and Dementia. The earlier detection of the size changes of HC leads to take preventive action against Alzheimer disease at initial stage. Thus the HC voxel quantification becomes essential to know the severity of the disease and thus induces computerized segmentation process. Several semi-automatic and automatic HC segmentation techniques proposed earlier. Though, it requires large memory space and high computational cost. This paper reduces the risk of searching a high configuration machine and reduces the cost by utilizing limited number of features. It is to be done by using some strategic features based on mathematical framework of wavelet, statistical features and gray level computations called level set. The features fed as input to the supervised machine learning model called back propagation neural network. A deep study conducted to train the net and analyzed in various views. The results were compared with the similar existing models which were using Random forest, Quicknat and deep learning. The proposed machine learning model produces the higher and similar dice scores of existing model. The validation of the proposed method yields 85% of dice score and 96% of sensitivity and 96% of specificity.

Key words: segmentation, features, back propagation algorithm, hippocampus **.**

AMS subject classifications. 68T05

1. Introduction and examples. Hippocampus (HC) is one of the brain regions and incorporates in memory function and it owns the structure of sea horse [19]. The hippocampus supports to know the age of a person [5]. Hence, the size of HC is notable one because some neurological diseases such as dementia and Alzheimer cause the size loss [7]. Further, the size of HC indicates the diseases like epilepsy and schizophrenia [2]. Alzheimer is one of the brain disorders, which affects memory power and cognitive skills. It easily affects hippocampus (HC) than other brain parts. In the entire world, 90 million peoples have suffered by Alzheimer disease [14]. The volume detection process of HC ensures the progressiveness of Alzheimer and guide to prevent its progression. At present, Magnetic Resonance Imaging (MRI) is non-invasive technology to mimic (capture) the morphology of human organs particularly brain [14]. The manual quantification of HC tissues from MRI images are consuming more times and tired the Radiologists. Hence, the necessity of computer aided methods for automatic segmentation is increasing exponentially. Though, the process of segmenting Hippocampus is very difficult [22].

The difficulties faced by automatic segmentation are noise, resolution constraint and weak boundaries between the brain components. When they are in same intensity, the pixels are not unique in intensity and not defined by a well defined boundary. The hippocampus segmentation methods are broadly categorized into i) Atlas and mullti-Atlas based method, ii) Deformable models such as Active contour models (ACM), Active shape model (ASM) and active appearance model (AAM) which uses local neighbourhood features and iii) Machine learning (ML).

ML is a rising technology which requires a set of features from a dataset to train the network and obtain optimal parameters in training stage, utilizes these parameters to detect objects from unknown set of features in testing stage. The ML is a flashy technology, which is majorly categorized by source of the features. Some networks admit features from external data source to the network. Convolution neural network (CNN) type of networks automatically constructs their own features from the given input images. There are several ML are

*Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India(ramd.kannan@gmail.com).
†Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India (balaji.nv@kahedu.edu.in)
‡KPR College of Arts Science and Research, Coimbatore, India (th.genish@gmail.com)

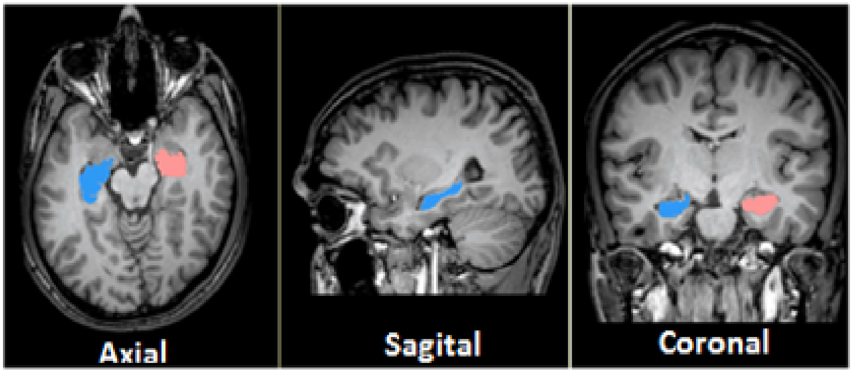


Fig. 1.1: Different MRI image orientation

available such as Support vector machine (SVM), artificial neural network (ANN), CNNetc. and attracted by its accuracy. Several ML methods have been implemented in the task of HC segmentation [3, 4, 12, 24, 25]. Among all, CNN is more compatible for image and employed in HC segmentation from brain MRI [8, 13, 21].

The paper [2] use deep convolution neural network (DCNN) and update the erroneous pixels by error correction steps. Initially, segmentation masks are prepared by using an ensemble models which includes three independent models. The erroneous pixels are corrected by replacing and refining the networks. The method uses several datasets are combined by transfer learning techniques. Thus reduces the time consumption of segmentation process and improves the accuracy of segmentation [2]. In a method, Deep artificial neural network was employed to separate the input image into large patches. The distances between the patches and its regional centroids in compressed images are depicted as feature set [17]. In another CNN model, four gray level patches of a voxel are provided as input to the CNN [10].

The model called Hippodeep proposed by [20] uses CNN and trained in a region of interest (ROI). This model makes use of a single CNN layer, starting with a planar analysis followed by layers of 3D convolutions and shortcut connections. The paper [6] proposed a open source software for the same task. It includes the features obtained from the all orientation MRI images such as coronal, axial and sagital as given in Fig. 1.1. The combined features fed into U-Net FCNN architecture [17] the network updated with RooNet in [9] to combine two-convolution block of same patch image. The results were finalized after the post processing and obtained 90% dice score [6]. A deep learning method with feed forward learning was used in the segmentation of hippocampus subfield. In the method, a canonical geometrical intensity space is used to reduce the time of pre-processing.

The above stated methods yield good results when using good quality of training images and its truth values. The irregular shape and in-significant boundary of HC provide lower accuracy in results [14]. The presented method selects the very relevant limited features to reduce the memory space and employes multilayer perceptron with back propagation learning algorithm. The following sections describe data set, the validation metrics to analyze the significance of segmentation, methodology, results and discussion and conclusion.

2. Data Set and Valuation Metrics. The experiments carried over Harp and clinical datasets which contains randomly selected 50 patients' images. The axial oriented images only consider for the experiment. The efficiency of a segmentation process is confirmed by computing the Dice Similarity Coefficient (DSC) which is used to compare two segmentation methods (Dice 1945). It can reveal the overlap coefficient between the two different segmented volumes S and G, and is defined in [1] as follows:

$$DSC = 2.\frac{|S \cap G|}{|S| + \ |G|} \tag{2.1}$$

S is the region segmented by the automatedmethod and G is the region obtained by manual segmentation. The '$\cap$' operator provides the amount ofpixels common to S and G. The '+' operator, produces the sumof all pixels

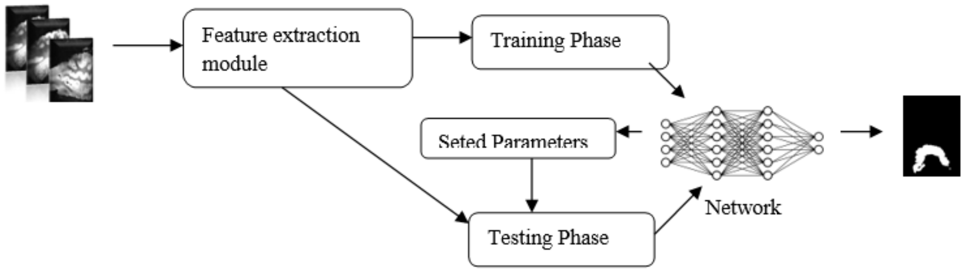


Fig. 3.1: Architectural diagram of presented method

in both regions. The ending valueof the DSC is a normalized computation, acquired by the coefficient between the amount of pixels common to bothregions and the mean of pixels in each region. The '0' DSC represents no one similar pixel exist in S and G. Controversy, '1' represents the both S and G contain similar pixels.

Sensitivity and specificity are calculated using True positive (TP), False positive (TN) and False negative (FN). Let the pixels inside HCare taken in S and G are positive and background are negative, the TP, TN and FN described as TP = Number of positive pixels in S are in G. FP = Number of positive pixels in S are as negative in G. FN = Number of negative pixels in S are positive in G. The expressions to compute sensitivity and specificity are as follows,

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.2}$$

$$Specificity = \frac{TP}{TP + FP} \tag{2.3}$$

3. Proposed Method. Figure 3.1 shows the architectural diagram of the proposed method. In this module, the input and ground truth images are fed into the feature extraction block at training stage. Twenty types of features recognize and pass as a matrix to the next module. There is a multi-layerperceptron, which able to understand the feature and train itself to detect the hippocampus pixel by using the feature set. Further, unknown (not included in training image set) image features given to check its performance. Finally, the net gives output as binary values and are used to construct the output image. The algorithm of the proposed method are given below.

Algorithm
1. start
2. f=featureExtraction (image) // implement the all formulas given in feature extraction section. //The pixel indexes are taken in row and the features are taken in column
3. t=truth(image) // assign binary value from the ground truth image
4. // network definition
5. weights=[random()]
6. // assign net
7. net=(layers =10, nodes per layer= 24, activation function = "sigmoid")
8. x=train(net,f,t)
9. check error values
10. if error>previous output
11. adjust the weight

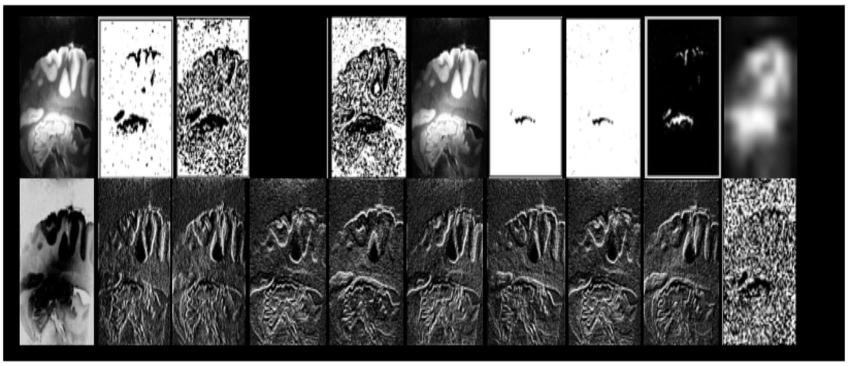


Fig. 3.2: Features of an input image. Row 1: Original image intensity, Skewness using mode, Kurtosis, Mean absolute percentage error, Pixel redundancy, Mean of 5×5 pixel neighbourhood, Variance of 5×5 pixel neighbourhood, Standard deviation of 5×5 pixel neighbourhood, Mean difference and Wavelet feature respectively. Row 2: Energy level, Edge detection filtered images (North, South, East, West, North East, North West, South East, South West) and Standard variation respectively.

12. if (error>previous error)
13. //repeat weight adjustment
14. else
15. // training completed
16. //testing
17. f1=featureExtraction(image)
18. test(net, f1)
19. output the result
20. stop

3.1. Feature Extraction. Features are the numerical or logical representation of image. Each image pixel's association with its neighbourhoods, intensity value and convolution with a filter values are extracted as numerical values and assigned instead of image alone. This method concentrates twenty types of features. The image feature extraction techniques are discussed in the following sections their resulting features are pictorially described in Fig. 3.2.

Skewness: The measure of skewness verifies the symmetry in intensity distribution. The skewness is calculated as follows,

$$f_2 = \frac{\sum_{i,j=1,\ b=1}^{m,n,k} \left(I_{i,j} - \overline{I}_b\right)^3 / N}{S^3} \tag{3.1}$$

Here is skewness obtained from 5×5 neighbourhood of each pixel b. The centre pixel value is replaced with the skewness measure f_1. $I_{i,j}$ referes the image pixel intensity value, $\overline{I}_b$ represents the mean value of 5×5 neighbourhood of each pixel, N is the total number of pixels here it is takes as 25 and S hold the standard deviation of 5×5 neighbourhood of each pixel. Instead of $\overline{I}_b$ mode and median substituted to obtain skewness mode and skewness median features.

Kurtosis: Kurtosis describes the peakness of a frequency distribution and it is defined as follows,

$$f_3 = \frac{\sum_{i,j=1,\ b=1}^{m,n,k} \left(I_{i,j} - \overline{I}_b\right)^4 / N}{\left(\sum_{i,j=1,\ b=1}^{m,n,k} \left(I_{i,j} - \overline{I}_b\right)^2 / N\right)^2} \tag{3.2}$$

The zero kurtosis represents the normal distribution; the high value distribution reflects the abnormal distribution of intensity value.

***Mean Absolute Percentage Error (MAPE)*:** Usually, MAPE utilizes to measure the forecasting error. In image processing, in-homogeneity among a block of pixels derived using MAPE. The error measure by [11]

$$f_4 = \frac{1}{N} \sum_{i=j=1,b=1}^{m,n,k} \frac{\left(\overline{I_b} - \ I_{i,j}\right)}{I_b} \ \times 100 \tag{3.3}$$

The zero result ensures the homogeneity in pixel intensity distribution and high value represent the in-homogeneity proportions with the mean of the block $\overline{I_b}$.

Pixel Redundancy: The redundant pixels represent the similar object in images. Sometimes pixel redundancy is utilized to denoise image. Here, redundant intensity value of a 5×5 neighbourhood replaces the centre pixel of the block.

$$f_5 = |b_i| \tag{3.4}$$

Mean: The mean of 5×5 neighbourhood pixel replaces the centre pixel value. The calculations performed over the each row and column.

Variance: The variance among 5×5 neighbourhood pixel replaces the centre pixel value. The calculations performed over the each row and column.

Standard deviation: The square root of the variance is called the standard deviation.

Wavelet Feature: The feature is extracted by using discrete wavelet transformation. A brain MRI image is decomposed into n levelsuntil get image size as (15 ×10). At this stage, we get a low pass and three high pass filtered images. The low pass filtered image displays distinct intensity at each pixel [11]. At this stage, interpolate the image into its original size of the image.We can see big pixels as given in Fig. 3.1, row 2. The hippocampus and cerebral final fluid (CSF) appearing area looks in high intensity, the others are looks in low intensity as given in [11].

Energy level: The entire work based on the intensity and neighbourhood pixels. The intensity distribution in adjacent regions is categorized as level set. The following mathematical framework helps to extract the energy level of each pixel as defined in [11],

$$f_{12} = \delta\left(\varphi\right) \left[\mu div \left(\frac{\nabla \varphi_{x,y}}{\left|\nabla \varphi_{x,y}\right|}\right) - \left(R\left(x,y\right) - C_1\right)^2 + \left(R\left(x,y\right) - C_2\right)^2\right] \tag{3.5}$$

where, f represents the image, $\delta\left(.\right)$ is representing Heaviside function, C_1 and C_2 represent the average intensity of pixels above 150 (intensity >150) and below 150 (intensity <150) respectively. The energy level is approximated by the by the Euler Lagrange's formulation. It is defined as,

$$div\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right) = \frac{\varphi_{xx}\varphi_y^2 - 2\varphi_{xy}\varphi_x\varphi_y + \varphi_{yy}\varphi_x^2}{\left(\varphi_x + \varphi_y\right)^{3/2}} \tag{3.6}$$

Edge filters: The edge filters showed in Fig. 3.3 convoluted with the image to obtain edges in all directions individually.

3.1.1. Artificial Neural Network. The presented method uses Multilayer Perceptron with back propagation learning algorithm. It has constructed with more number of layers such as input, hidden and output layers and each layer is holding numerous nodes as described in Fig. 3.4. Each node (neuron) in input layer represents a single feature and includes the connection with n-number of nodes in hidden layer. The nodes act as decision makers with the help of activation functions such as sigmoid, tanh, ReLu and etc. Each neuron calculates the weighted sum of its inputs and then applies an activation function to normalize the sum.

Training: Requirement number of layers, neurons and suitable activation function are determining at the stage of training. Input features and the expected outputs are provided to the network. The network

$$\begin{bmatrix}-3 & -3 & 5\\ -3 & 0 & 5\\ -3 & -3 & 5\end{bmatrix}\begin{bmatrix}5 & -3 & -3\\ 5 & & -3\\ 5 & -3 & -3\end{bmatrix}\begin{bmatrix}5 & 5 & 5\\ -3 & 0 & -3\\ -3 & -3 & -3\end{bmatrix}\begin{bmatrix}-3 & -3 & -3\\ -3 & 0 & -3\\ 5 & 5 & 5\end{bmatrix}\begin{bmatrix}-3 & -3 & -3\\ 5 & 0 & -3\\ 5 & 5 & -3\end{bmatrix}\begin{bmatrix}-3 & -3 & -3\\ -3 & 0 & 5\\ -3 & 5 & 5\end{bmatrix}\begin{bmatrix}5 & 5 & -3\\ 5 & 0 & -3\\ -3 & -3 & -3\end{bmatrix}\begin{bmatrix}-3 & 5 & 5\\ -3 & 0 & 5\\ -3 & -3 & -3\end{bmatrix}$$

Fig. 3.3: Edge detection filters, North, South, West, East, South East, North East, South East and North West respectively

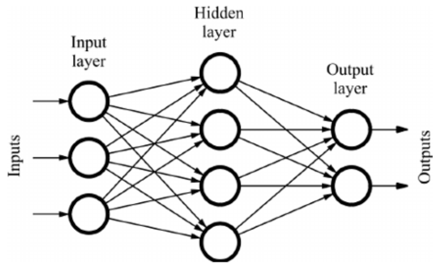


Fig. 3.4: Out sketch of Neural Network

then processes the inputs and compares its outputs against the desired outputs, if fails to give similar output adjust the weights of the network connection until finds the exact results. At the final stage, sets all weights as constant and ready to validate and to run unknown data set.

Testing: At this stage, accepts the unknown image features (not in training set) and set “1” for hippocampus pixels and “0” for non-hippocampus pixels without human intervention.

4. Results and Discussions. The experiment carried over axial brain MRI images data set, which includes three training volumes 1-validation volumes and two testing volumes. The experiment utilizes Harp and clinical dataset and tries to resolve the queries i) finalize the essential features , ii) confirm the fitness of the net iii) Make sure the training package is such that a training image taken from one MRI machine can completely segment the hippocampus from another type of MRI machine.

The experiment commenced with the features skewness, kurtosis, MAPE, pixel redundancy, mean, variance, standard deviation, wavelet features and original intensity. The intensity of hippocampus shows as the Cerebro spinal fluid (CSF) in T2-weighted images. Since the CSF fills in the gap of brain tissues appears as edges, the presented net observes the edges as HC region shown in Fig. 4.1. After the result, we concluded that to include edge features in the training set. For that, the input images convoluted with the eight types of filters individually. When the edge based features included in the training set, the net trained itself and avoid the classification of edge as hippocampus.

Next ensure whether the net is over-fitting or under-fitting, wefitted the net by knowing the accuracy of the training set. Initially, five consecutive images took from each volume and provided at the moment of training which results 0.725 accuracy in 3 epochs.

The architecture of the net determination is the primary problem; minimal number of neuron takes minimal computation time. Initially, one hidden layer consisting of 50 neurons employed for the segmentation which accounts 76% of accuracy. Further, increased the hidden layers as 3 with the same count of neurons as 50. Though, the net is not capacious enough for some complex images. Finally, the neurons are increased for first, second and third layers as 800, 200 and 50 respectively the results are given in Table 4.1. The Table 4.1 shows the highest precision value for the updation of neurons. The Table demonstrated that the net provides lower precision value for 50 neurons and higher precision value for the neuron updated net. Subsequently, the epochs was increased from 5 to 25. At this stage we get lower mean square error value (MSE) 0.0435 and high accuracy

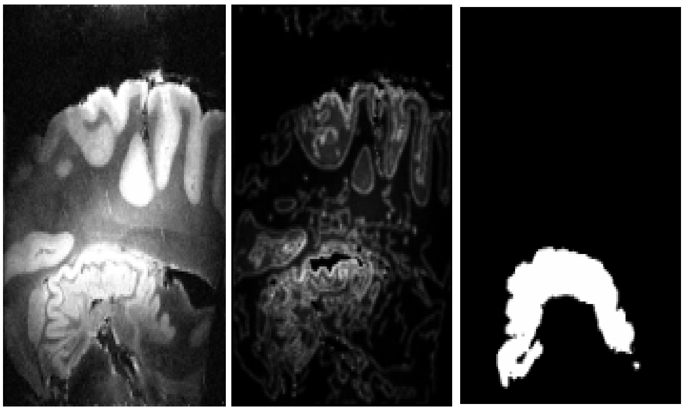


Fig. 4.1: Original image, segmented and ground truth image respectively

Table 4.1: Precision versus Neurons

Training Image	Hidden Layers	Neurons	Highest Precision
10	1	50	76%
15	1	50	70%
20	1	50	62%
10	3	800,200,50	83%
15	3	800,200,50	89%
20	3	800,200,50	94%

as 94%.

The results of modified network by increasing and deleting neurons, layers and epochs have high improvement that is delineated in Fig. 4.2. In Fig. 4.2, the block colour line indicates the performance (DSC score) of the modified net and the gray line indicates (DSC score) of the default net. Y-axis shows the DSC score and X-axis shows the image number.

The sensitivity and specificity of the images illustrated in the Fig. 4.3. In Fig. 4.3, dark line shows the sensitivity and light line shows the specificity. The sensitivity remains the performance of extracting true positive and specificity remains the localization of the work. The obtained average DSC, sensitivity, specificity and standard deviation of the scores are listed in Table 4.2. The sensitivity ranges from 82% to 99% which indicates the proposed features and net can resolve the complex images also. The specificity ranges from 96% to 99% which indicates the net and features incorporate to detect the exact location of hippocampus. The deviations in the results are measured using standard deviation and listed in Table 4.2. In DSC, there are no much deviations among the images of a volume.

The Random forest utilizes the local energy pattern feature set which were proposed in [23] for the classification. The random forest method adopts 20 trees for construction and the depth of each tree is empirically set as 25. The QuickNat network utilizes the convolution neural network for the segmentation process [18] which results lower than the proposed method. The deep learning based method also provided 0.85 DSC score.

Compared to the other method, the proposed feature set and network work similarly and utilizes small data set and less number of feature set. The proposed method takes time for training is 20 minutes and 10 minutes for testing in Dual core 2 machine.

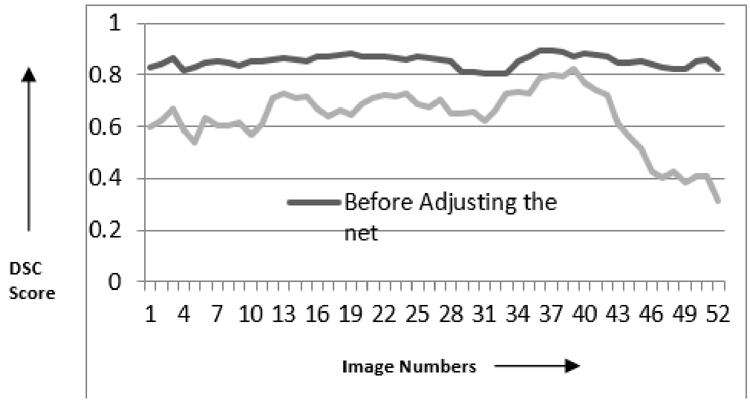


Fig. 4.2: Dice scores of default and modified net

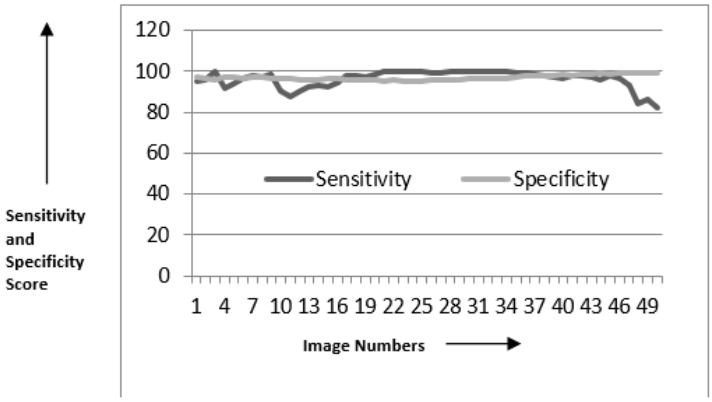


Fig. 4.3: Sensitivity and Specificity of modified net

Table 4.2: Validation measures

	DSC Score	**Sensitivity**	**Specificity**
Average	0.851923	96.18962	96.89855
Standard deviation	0.203303	4.288032	1.236299

Table 4.3: Comparison with similar methods

Method	**DSC Score**
Hippodeep -[20]	0.85
QuickNat -[18]	0.84
Random forest - [3]	0.85
Proposed Method	0.85

The proposed feature set is enough for more images but in two or three images which has connected pixels between CSF and HC lead over-segmentation due to the energy feature.

5. Conclusion. In this paper, we present a multi-layer perceptron framework for hippocampus segmentation from harp and clinical data sets. The network utilizes back propagation learning algorithm. There are twenty features extracted and utilized to set the net. Major advantages of our method lie on that we don't need any time-consuming non-linear registration for pre-processing MR images, and features generated by MDL are consistent with subsequent learning models. The experimental results suggested that the proposed strategically features can boost the performances of hippocampus segmentation and minimize the MSE score. The network model adjusted only based on the MSE error measure and the results are analysed with qualitative and quantitative measures. Sensitivity, specificity, Dice similarity and Accuracy were computed to ensure the outstanding performance of the proposed model. The net model initializes randomly defined weight value for each neuron that increased the training time. In future, initial weight value could be optimized to reduce the overall execution time. As well as more clinical images to be incorporated to achieve the proposed model as reliable one for all configured MRI machine image.

REFERENCES

[1] Karim Aderghal, Karim Afdel, Jenny Benois-Pineau, GwénaëlleCatheline.Improving Alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities,Heliyon, 6 (2020) e05652

[2] Dimitrios Ataloglou, Anastasios Dimou, Dimitrios Zarpalas, Petros Daras, Fast and Precise Hippocampus Segmentation Through Deep Convolutional Neural Network Ensembles and Transfer Learning, Neuroinformatics (2019) 17:563–582.

[3] Cao, L., Li, L., Zheng, J., Fan, X., Yin, F., Shen, H., Zhang, J., 2018. Multi-task Neural Networks for Joint hippocampus Segmentation and Clinical Score Regression. Multimedia Tools & Applications, , 1st77. Springer, pp. 1–18.

[4] Cao, P., Liu, X., Yang, J., Zhao, D., Huang, M., Zhang, J., Zaiane, O., 2017. Nonlinearityaware based dimensionality reduction and over-sampling for AD/MCI classification from MRI measures. Comput. Biol. Med. 91.

[5] Decario L. T. (1997) on the meaning and use of Kurtosis, Psychological methods,2,3, pg. 292-307

[6] Diedre Carmo, Bruna Silva. Alzheimer's disease neuro-imaging initiative,Clariss Yasuda, Leticia Rittner, Roberto Lotufo, Hippocampus segmentation on edpilepsy and alzheimer's disease studies with multiple convolutional neural networks, Heiliyon 7, 2021, e06226.

[7] Du A., Schuff N., Amend D., Laakso M., Hsu Y., Jagust W., Yaffer K., Kramer J., Reed B., Norman D. et al. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and alzheimer's disease. Journal of Neurology, Neurosurgery and Psychiatry,71(4), 2001, 441-47.

[8] Chen, J., Yang, L., Zhang, Y., Alber, M., Chen, D.Z., 2016. Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation. Adv. Neural Inf. Process. Syst. 3036–3044.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp.770–778.

[10] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K.H. Maier-Hein, Brain tumor segmentation and radiomics survival prediction: contribution to the BraTS2017 challenge, in: International MICCAI Brainlesion Workshop, Springer, 2017, pp.287–297.

[11] S. Karthigai Selvi, 2019. DWT based method to locate tumor from T2-w axial head scan, international journal of engineering research and applications, 9,3, pg.63-70.

[12] Liu, M., Zhang, J., Nie, D., Yap, P.T., Shen, D., 2018a. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE J. Biomed. Health Inform. 22 (3.2), 1476–1485.

[13] Liu, M., Cheng, D., Wang, K., Wang, Y., 2018b. Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis, 16. Neuroinformatics, pp. 295–308.

[14] Manhua Liu, Fan Li, Hao Yan, Kundong Wang, Yixin Ma. A multimodel deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease, Neuroimage, 208, 2020, 116459

[15] Andrew O'Shea, Ronald A. Cohen, Eric C. Porges, Nicole R. Nissim, and Adam J. Woods, Cognitive Aging and the Hippocampus in Older Adults. Front Aging Neurosci.(2016).

[16] Todd Mitton, Kaith Vorkink, Equilibrium underdiversification and preference for skewness, The financial studies, 20,4, July 2007,pg.1255-1288.

[17] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp.234–241.

[18] A.G. Roy, S. Conjeti, N. Navab, C. Wachinger, A.D.N. Initiative, et al., Quick-NAT:afully convolutional network for quick and accurate segmentation of neu-roanatomy, NeuroImage 186 (2019) 713–727.

[19] Scoville W.B., Milner B. Loss of recent memory after bilateral hippocampal lesions. The journal of neurophsychiatry and clinical neurosciences, 12(2.1), 2000, 103-a.

[20] B. Thyreau, K. Sato, H. Fukuda, Y. Taki, Segmentation of the hippocampus by trans-ferring algorithmic knowledge for large cohort processing, Med. Image Anal. 43 (2018) 214–228.

[21] Wang, X., Gao, L., Song, J., Shen, H., 2017. Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition. IEEE Signal Process. Lett. 24, 510–514.

[22] Zarpalas D, Gkontra P, Daras P, Maglaveras N (2014) Accurate and fully automatic hippocampus segmentation using subject-specific 3D optimal local maps into a hybrid active contour model. IEEE J Trans Eng Health Med 2:1–16.

[23] Zhang J, Liang J, Zhao H (2013) Local energy pattern for texture classification using self-adaptive quantization thresholds. IEEE Trans Image Process 22(2.1):31–42

[24] Zhang, J., Gao, Y., Gao, Y., Munsell, B., Shen, D., 2016. Detecting anatomical landmarks for fast Alzheimer's disease diagnosis. IEEE Trans. Med. Imaging 35, 2524–2533.

[25] Zhang, J., Liu, M., An, L., Gao, Y., Shen, D., 2017. Alzheimer's disease diagnosis using landmark-based features from longitudinal structural MR images. IEEE J. Biomed. Health Inform. 21 (3.3), 1607–1616.

Edited by: Vinoth Kumar
Received: Sep 4, 2022
Accepted: Aug 21, 2022

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 23, Issues 3, pp. 91–101, DOI 10.12694/scpe.v23i3.2001

OPTIMIZATION OF INTELLIGENT NETWORK INFORMATION MANAGEMENT SYSTEM UNDER BIG DATA AND CLOUD COMPUTING

LI MA *, RAJIV KUMAR GUPTA †, AND EDEH MICHAEL ONYEMA ‡

Abstract. With the economy and society development, big data storage in management has become a problem which can't be ignored. Task allocation management and optimization is a significant factor in the enterprise sustainable development. The enterprise information intelligent management has become management mode and presents information to the business managers in a more effective and lower cost. In order to solve the management problem under big data and cloud computing, a research on the optimization of intelligent network information management system is proposed. Firstly, build the overall model of big data diversion system under cloud computing platform, and design the architecture of Tiny OS operating system for big data diversion; Secondly, S3C2440 is used as the system control core to design the hardware structure of the shunting system; Finally, the software of the system is developed by using the dynamic loading method of functional subroutines in Linux operating system. The power supply voltage is DC 3.3 V and 1.25 v. DC 5 V is used as the overall power supply of the circuit board, and 0.1 is added at both ends of lm1117 chip F and 100 F capacitance for FIR filtering of power supply. Simulation experiments are carried out to verify the performance, which shows the superior performance of the big data shunting system designed in this paper.

Key words: Big data; Cloud computing; Intellectualization; Network; System optimization; Data diversion; Shunting system

AMS subject classifications. 64M14

1. Introduction. In recent years, under the background of the continuous development of computer technology in China, cloud computing and big data technologies have also begun to penetrate into various industries in China. Multimedia teaching has always been one of the hottest topics in recent years. Many relevant personnel are actively exploring how to effectively combine modern information technology with educational technology. Management is a more important part in the development of various fields. Only by strengthening management can we better promote the smooth development of various work. Management informatization is not only an important trend of the development of the times, but also an effective means to optimize management and improve management efficiency. As an important concept of the continuous development of Internet information technology, information management is also an important measure to optimize and improve management in various fields. Under the background of cloud computing and big data era, carrying out management work according to the application characteristics of cloud computing and big data technology can promote the realization of management informatization [1]. Therefore, this paper explores the management informatization strategy in the era of cloud computing and big data. The Big data and network management under the cloud is shown in Figure 1.1.

The acquisition system demand is increasing; it is superior for the data processing performance for the electric information acquisition system improvement leading technology. The data generated by electricity is considerable, therefore the data storage and mining has become the important, including the database selection and the SQL statement optimization. These are precisely research direction. For the Henan electric power company implementation "three years for the quality of rural power enhancement", to support the Henan county company management, to deepen the measurement acquisition application, and measurement acquisition related business city and other work, on the basis of establishment of station expansion principle and electric information acquisition system function up gradation issued by the power grid and realized the management

*School of Information Engineering Yellow River Conservancy Technical Institute, Kaifeng, Henan 475004, China (lima098760@yahoo.com).

† Pandit Deendyal Energy University, Gandhinagar, India (Rajiv.gupta@sot.pdpu.ac.in).

‡ Department of Mathematics and Computer Science, Coal City University, Enugu, Nigeria (michael.edeh@ccu.edu.ng).

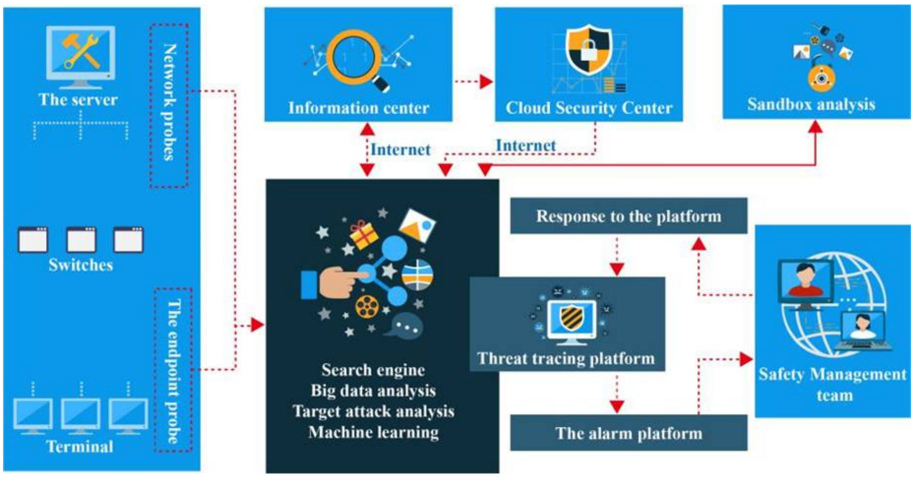


Fig. 1.1: Big data and network management under the cloud

system. The upper controllers improved their work efficiently and in the real time, they can observe the power which is of practical implication. Intelligent information management becomes the model in the age, which changes the way managers works. Firstly, the big data management in enterprises has become a problematic in modern times, while information management find the information more rapidly, and the work efficiency is improved greatly. Secondly, the enterprise data information management is organized and facilitate the company operation. Thirdly, as differentiated and modified management is becoming common, management software is highly effective and easy to operate.

2. Literature Review. Data refers to massive, fast-growing and diversified data, which is characterized by maximizing information resources for all walks of life and people's life. Informational refers to the process of cultivating and developing new productive forces represented by intelligent tools, better creating spiritual and material civilization, serving and benefiting the society. Under the unified planning, organization and leadership of the state, national informatization uses advanced modern technology in industry, agriculture, military, science and technology, integrates various information resources, and accelerates the process of realizing the development of national economy, science and technology. In the national informatization development strategy for 2006-2020, China clearly pointed out that informatization is the development trend of the world, an important force to promote social change and a catalyst for the development of social economy, education and other fields. Especially in the field of education, the application and development of information technology can greatly promote the cultivation of information talents, the construction and management of information resources, promote information exchange and knowledge sharing, continuously improve economic growth and the information development of education, and promote the progress of science and technology. The concept of cloud computing can be traced back to 2006. Kolodziej, J. launched elastic computing cloud EC2 service in March 2006 [2]. In August of the same year, at the search engine Conference (SES San Jose 2006), Google CEO Zhang, B. first proposed the term "cloud computing" [3].

He, X. L. and Song, Y. believe that data application talents will play an important role in the scientific decision-making process of enterprise development in the era of knowledge economy [4]. Jin, S. believes that the most precious information of mathematics in the data age is massive educational data, which is the core cornerstone of the development of intelligent education. Data mining technology and learning analysis technology are important fields under big data and cloud computing [5]. Moser, O. and others studied the management mode of big data discipline construction project with system engineering methodology, and put forward project management suggestions from four dimensions [6]. Tkachenko, O. and others studied the post evaluation method of applied mathematics discipline construction project by using fuzzy analytic hierarchy process, and gave the corresponding three-level evaluation index system [7]. Meena, V. and others believe that the construction project of applied mathematics discipline is discussed from the aspects of organizational structure, existing problems and project management mechanism, and some suggestions are given [8].

Based on the current research, this paper puts forward a research on the optimization of intelligent network information management system. Cloud computing is an important product of the information revolution. Through cloud computing and cloud storage, we can realize the collaborative management and scheduling of massive big data. At present, cloud computing information system is widely used in the scheduling and allocation of big data information resources. The power supply voltage is DC 3.3 V and 1.25 v. DC 5 V is used as the overall power supply of the circuit board, and 0.1 F and 100 F capacitors are added at both ends of LM1117 chip for FIR filtering of the power supply [9].

Author in this paper design a more effective, suitable and the big data management platform. The big data management system is designed firstly and according to the big data processing dataset, interface acquisition modules, platform alerting, marketing analysis and visualization are designed on the basis of communication big data architecture. To realize the data testing environment and experimental environments under the scenario of communication application and presented the big data system mechanism for production for experiments. Then design the corresponding scheduling module architecture process and built the corresponding scheduling rules [13]. Contribution: A research on the optimization of intelligent network information management system is presented in order to solve the management problem under big data and cloud computing. Firstly, build the overall model of big data diversion system under cloud computing platform, and design the architecture of Tiny OS operating system for big data diversion; Secondly, S3C2440 is used as the system control core to design the hardware structure of the shunting system; Finally, the software of the system is developed by using the dynamic loading method of functional subroutines in Linux operating system.

3. Optimization of intelligent network information management system under big data and cloud computing.

3.1. Introduction to concepts related to cloud computing and big data.

3.1.1. Cloud computing. Cloud computing was born in 2007, but after less than six months, it has attracted more attention than grid computing. Cloud computing is a super-computing model. In its data center, there are tens of thousands or even tens of millions of computers or servers. Therefore, cloud computing can even allow users to enjoy the computing power of more than 100000 billion times per second. Such powerful computing power makes - everything impossible [14]. Users can connect to the data center through handheld notebooks, mobile phones and laptops, and then use the cloud computing platform according to their own needs. Cloud computing will completely change the way of working and business model in the future. Of course, due to the application scenarios and social progress, new views on the definition of cloud computing are constantly emerging. In popular terms, cloud computing is a multi data center environment.

3.1.2. Cloud computing architecture. The architecture of cloud computing consists of five parts: resource layer, platform layer, application layer, management layer and user access layer. For users, it mainly obtains services through cloud computing, so the daily question of cloud computing is 3 and the question of four networks is X. The core is service [15]. The cloud computing architecture is shown in Figure 3.1.

3.1.3. What is big data. Big data first appeared in the 1990s. With the continuous development of cloud computing and the Internet of things, the emergence of a large number of data sources has led to the rapid growth of unstructured and semi-structured data, and the data unit has also crossed from TB level to ZB level. These data generated by a large number of informants have far exceeded the scope that human can handle at present. When people think about how to manage and use these data, they gradually explore a new field. The "big" of big data refers not only to the size of capacity, but also to diversity, processing speed and complexity [16]. Whether people have paid attention or not, massive data has affected people's lives, and the era of big data has come.

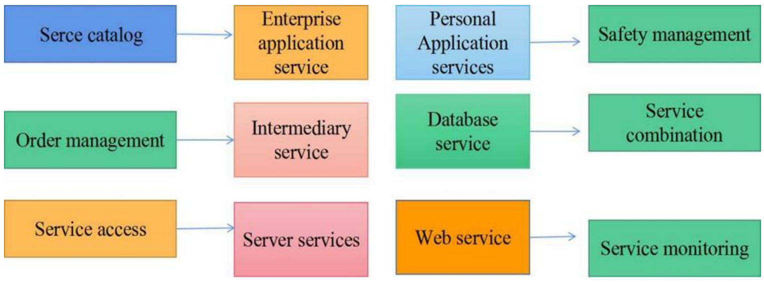


Fig. 3.1: Cloud computing architecture

3.2. Big data features. IBM believes that big data has 3V characteristics, namely diversity, scale and high speed, but these can not reflect the great value of big data. The industry represented by IDC believes that big data needs to have 4V characteristics, that is, add value on the basis of the previous 3V, indicating that although the overall value of big data is high, its value density is very low [17]. At present, it is recognized that big data has the following four characteristics: large data volume, multiple data types, fast processing speed and low data value density [18]. (1) Large data scale With the wide application of Internet technology, it is very easy for users to obtain and share data. In addition, users' clicking, browsing and sharing will quickly generate a lot of data, and the data has jumped from TB level to Pb level. And the amount of data is increasing. (2) There are many kinds of data Big data data types include not only traditional relational data types, but also unprocessed, semi-structured and unstructured information, such as data in the form of documents, audio, web pages, e-mail, video and so on. (3) Processing speed Another important feature of weighing big data is the frequency of data generation and update. One second law, which is the most significant feature that distinguishes big data from traditional data mining. For example, we chat, stock information and other data generated and updated by users all over the world are transmitted at any time, which requires that the data processing speed must be fast [19]. (4) Low data value density The amount of data grows rapidly, generally exponentially, but the growth rate of useful information implied in the data has not been significantly improved. Moreover, greater efforts need to be made to obtain useful data.

3.2.1. Clarify the important role of management informatization. When implementing management informatization in the era of cloud computing and big data, the primary task is to clarify the importance of management informatization construction and actively build a relatively perfect informatization management responsibility system, so as to effectively promote the smooth implementation of management informatization. Management itself is a very complex procedure and task. Assuming that relevant leaders do not pay enough attention to this work and do not vigorously support the orderly development of this work, it is naturally difficult to promote effective communication and cooperation among various departments, further increasing the difficulty of management information construction [20-22]. Therefore, relevant leaders must give full play to their role as leaders, continuously improve the cost investment in management information construction during work practice, and comprehensively build the fundamental project of information construction. During the development period, a representative and authoritative department can be built internally to provide a good guarantee for the construction of management informatization, so as to effectively promote the smooth development and implementation of various work.

3.2.2. Build a reasonable and comprehensive development goal and scheme of management informatization. For the orderly implementation of management informatization in the era of cloud computing and big data, it is necessary to build management informatization development goals and plans that can effectively meet their own development demands according to the documents issued by relevant national departments and the actual situation of the school. Through this way, we can effectively build a highly targeted information development plan, so as to truly lay a good foundation for the development of management information construction. In this process, colleges and universities can actively build a more comprehensive education and teaching information management system, actively cooperate with reasonable education and teaching information management regulations, effectively guide and restrict the work of staff, and further strengthen the use of data information, so as to better promote the implementation of management information [23, 24]. In addition, we can also build a special information management platform in the development process to timely solve the problems related to the construction of management information, so as to provide a good platform for the implementation of management information work, so as to truly and effectively improve the efficiency and quality of management.

3.3. Overall design and functional module analysis of big data diversion system under cloud computing platform.

3.3.1. Overall design of big data diversion system under cloud computing platform. Big data streaming under the cloud computing platform is the basis for improving the parallel computing ability of cloud computing and realizing data clustering and pattern recognition. The big data diversion system based on cloud computing platform realizes the collection, processing and release of big data under cloud computing platform through feature extraction and data partition of multi-source information resources. Big data streaming system is the infrastructure to realize the integration of multi-source information resources under the cloud computing platform. As an open source framework, cloud computing information system can realize the wireless sending and receiving of data and data clustering through the streaming of big data under the cloud computing platform [25]. The big data streaming system under the cloud computing platform designed in this paper has the characteristics of scalability, stability, reliability and openness. Scalability means that the big data streaming under the cloud computing platform is open source, so it faces a wide range of objects. Through the big data streaming, it sends and receives the RF bytes of the upper layer of the notification to realize data serial communication and wireless transceiver; Reliability refers to the distributed processing method adopted in the cloud computing big data information diversion system. When one or several devices fail, it can notify the high-level active message components to carry out alternative work, including the later level data acquisition and processing system, which is scalable and open. The overall model of the big data diversion system under the cloud computing platform designed in this paper is shown in Figure 3.3.

Based on the overall model design of the system, the network design of the system is carried out. The big data shunting system under the cloud computing platform adopts the form of distributed weighted ad hoc network. The nodes in the cluster in the network consume energy when sending the data length to their corresponding cluster hair in each round. ZigBee communication technology is used to design CDMA module [26]. When the data distribution node of the shunting system elects to be the cluster head, set Ru to 0, and the transmission distance is greater than or equal to the threshold d0, the residual energy of the node gradually decreases. The power amplifier adopts the multi-path attenuation channel transmission model, and the energy consumption coefficient is mp. G is the set of nodes that were not selected as candidate cluster heads in the last 1p round. In the distributed weighted ad hoc network, 32-bit ARM processor and embedded gateway are used for data communication. The consumption of sending L B data by nodes exceeds the carrying capacity. 16934400 of them needs to be processed in Mach mini2440 to generate the power consumption of gap aliasing frequency band in dense channel. After compiling, the Linux kernel can be downloaded to the target board for operation to realize data diversion. Among them, the main controller is the core of the whole embedded gateway. The 32-bit RISC ARM920T Network Microcontroller produced by SamSung company in Korea is used as the core, and the 32-bit RISC microprocessor with ARM920T as the core is used to design the main controller of the big data shunting system. Tsxm, tsym, tsxp and tsyp are used as the four wire touch screen interface control signals.

4. Experiment and analysis.

4.1. Cloud computing environment modeling. Generally speaking, cloud computing is a multi data center environment. This paper studies the data migration of two common applications in big data in cloud computing environment. Therefore, this paper focuses on the differences of storage resources in data centers

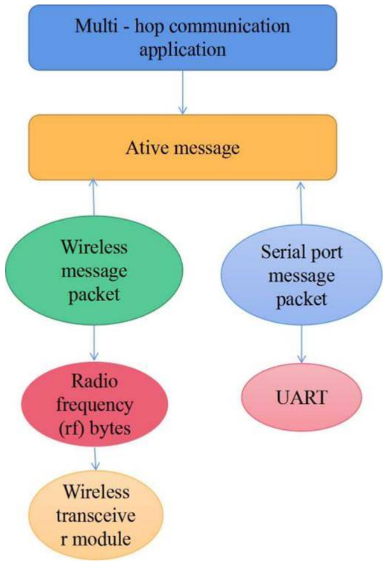


Fig. 3.2: Model of big data diversion system under cloud computing platform

and network bandwidth between data centers. Definition 1 cloud computing environment is represented as a collection of N data centers distributed in different locations. DC ={ dc1,de2... dcn}. These data centers are connected through different networks, and the bandwidth in these networks is different [19]. Where dcn is the data center numbered n. A real physical network can be abstracted as a logical network topology.

4.2. Key technologies of hardware design of big data diversion system under cloud computing platform. On the basis of the above hardware platform design, the system software is developed by using the dynamic loading method of functional subroutines in the Linux operating system. The telosB wireless module of crossbow company is used to build the network coordinator to realize the dynamic networking and data transmission of big data shunting node data. The software development platform adopts the open source Linux operating system [27]. The target board and the host computer are usually connected by 232 serial port, network cable and USB cable. The binary code compiled by GCC compiler is used in Linux. The default target platform of Linu2.6.32.2 is used to become the platform of arm, and the make file in the general directory is modified.

4.3. Functional module design of the system. The big data shunting module under the cloud computing platform designed in this paper mainly includes controller module, power management module, data memory module, transmission and communication module, Ethernet module and display module [28-30]. According to the overall design of the above data diversion system and the design of the data diversion system, TinyOS priority scheduling method is used to collect and schedule the characteristics of the big data information flow, read the characteristic sampling values of the big data, and carry out digital FIR filtering in DSP. As shown in Figure 3.2, the data is shunted to the PC or power amplifier for dynamic data processing, and then the data is sent to the PCI bus for data processing, as shown in Figure 4.1.

While (subci is not empty)
According to CF Δ I calculate the distance between SubCi and the center CenCj
of the data streaming cluster CJ of all cloud computing platforms,
and determine the minimum distance dist (Subck, Cencl), $l \in [1, k]$;
While (the central point of data clustering changes)
Select sound card suprt and add the cluster center subci of cloud computing big data to Cl, subci $\in$ CL;

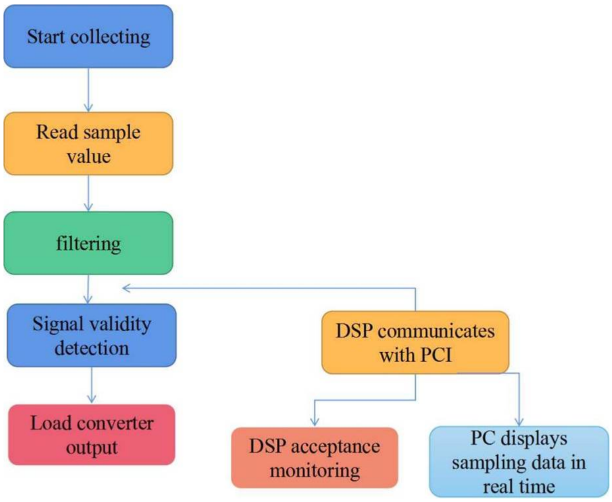


Fig. 4.1: Big data shunting process

$i \leftarrow i+1$
Read in the new data clustering vector subci, subci $\notin$ Cl, $J \in [1,k]$;
function:
mkdir -p /var/lock
mkdir -p /var/run
mkdir -p /var/tmp
Calculate the sub cluster CenCj representing the cluster center in all data streaming clusters Cj,
and generate the root file system to minimize the system transfer function
from each sub cluster in Cj to its corresponding cluster center cencj //Thus, data diversion is realized.

4.4. Big data system control. The power control module is to provide power supply for the big data diversion system under the cloud computing platform. The arm processor is used to design the power control module. In the design of data memory module, one 128 MB flash chip and two SDRAM chips hy57v561620 are extended in parallel to meet the data storage requirements of arm and transferred into SDRAM for use. Since there is no need for external expansion equipment, only USB device control port is used for Linux terminal control through RS 232 to realize streaming cloud storage of big data. The design hardware circuit of data storage function module of big data streaming system is shown in Figure 4.2. Through TinyOS priority scheduling, each port can work in interrupt mode or DMA mode to improve data shunting performance.

4.5. Simulation experiment and result analysis. In order to test the performance of the system designed in this paper, simulation experiments are carried out. The test platform is a general-purpose PC and the CPU is Intel ® Coretm i7 2600 © 3.40 GHz, memory 4 × 4 GB DDR3@1600 99924Firstly, the hardware debugging of the system is carried out. In the debugging process, check is set as a timer to realize the connection between the developed PC and the embedded mobile development equipment; The dm9000 sensor module is used to collect data from the network, and then the dm9000 sensor module is used to transmit the data to the network.

The procedures for data sampling are: ecall ReadStream. postBuffer, pressureSamples, PRESSURE_ SAMPLES

On the basis of the above simulation environment settings, set the reversal parameters of the shunt system, as shown in Table 4.1.

The system adopts PHP and MYSQL to realize data diversion programming simulation. In the device

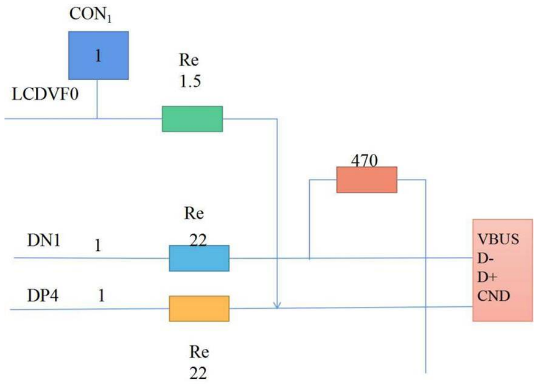


Fig. 4.2: Data storage function module of big data diversion system

Table 4.1: Simulation reference data

Parameter type	Value	Reference type	Value
Routing distribution	201*201	Member node energy	1
Number of topologies	12	Data split duty cycle	15b
Data return location	0.0	Transmission bit rate	3 b/s/Hz

drivers menu, select SD / MMC device to read the kernel image file of data streaming. The data reading interface system adopts PHP and MYSQL to realize the programming and Simulation of data streaming. Finally, the simulation results of the data system designed in this paper are obtained. Using the system designed in this paper, the data diversion is realized by clustering and feature extraction of big data under the cloud computing platform, dynamic loading according to the functional subroutine and TinyOS priority scheduling. The diversion accuracy is high, and the ability of pattern recognition is improved. The experiment of this paper is completed on the simulation platform of cloudsim. Firstly, according to the modeling of cloud computing environment, this chapter simulates the cloud computing environment composed of multiple data centers. This paper creates 600 data centers (the configuration of the data center is as follows: the virtual machine CPU is quad core, the host memory is 16g, the bandwidth is 1000mips, the image size is 100000m, and the processing capacity is 10mips). These data centers are connected through high-speed networks with different bandwidths. Then, according to the modeling of two kinds of special applications of big data, randomly simulate the test examples of two kinds of special applications of big data. These data sets have been stored in these 600 data centers according to Zipf distribution before the task. In this section, two special applications for big data cloud computing proposed in the paper are compared and analyzed. In the experiment, keep the number of overall data sets unchanged, and compare and analyze by increasing the number of tasks, as shown in Figure 4.3.

The large amount of data fusion work is required by the office system and it requires to have a humanization and office effect convenient network, so conducts system performance verification with the cloud computing and ML technique. First, the system's data fusion effect is evaluated and the results are shown in Figure 4.4.

The presented method is advantageous as compared to the existing techniques in data fusion. The data fusion is based on the cloud computing, so the data fusion speed is more beneficial. From the analysis it is seen that the system performs well in terms of data fusion in this paper and can basically meet the daily smart office needs. On this basis, a satisfaction survey is conducted in this paper on the office system and the system operating effect can be reflected through the degree of satisfaction. The statistical results are illustrated Figure 4.5.

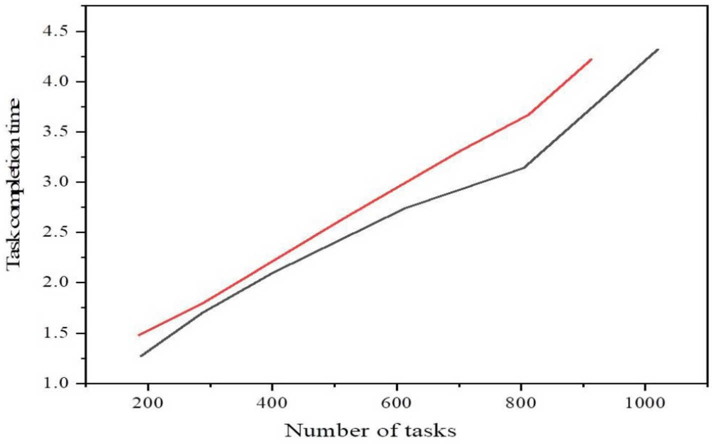


Fig. 4.3: Task completion time

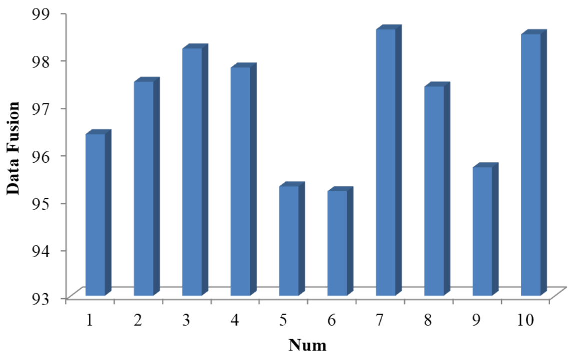


Fig. 4.4: System data fusion effect statistical diagram

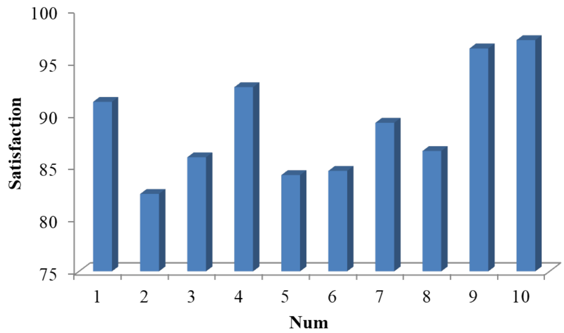


Fig. 4.5: Satisfaction survey of the intelligent office system

5. Conclusion. In the cloud computing environment, the sources of computing resources, storage resources and software resources have multiple attributes, forming the cloud computing and cloud storage of multi-source information resources. It is necessary to shunt the multi-source big data in the cloud computing environment, improve the accuracy of data clustering, and provide the basis for pattern recognition. This paper proposes an optimization design method of big data diversion system under cloud computing platform based on dynamic loading of functional subroutines and Tiny OS priority scheduling. Firstly, build the overall model of big data diversion system under cloud computing platform; S3c2440 is used as the system control core to design the hardware structure of the shunting system; Finally, the software of the system is developed by using the dynamic loading method of functional subroutines in Linux operating system. The experimental results show that the system designed in this paper can effectively realize the shunting processing of big data in cloud computing platform, and the performance is superior. The presented method is advantageous as compared to the existing techniques in data fusion. The data fusion is based on the cloud computing, so the data fusion speed is more beneficial. From the analysis it is seen that the presented system performs well and can basically meet the daily smart office needs. On this basis, a satisfaction survey is conducted on the office system and it can be reflected through the degree of satisfaction.

REFERENCES

[1] Novo, O. Qian, H. , Optimization of intelligent management andmonitoring system of sports training hall based on internet of things, Wireless Communications and Mobile Computing, 2021(2), 1-11., 2021.

[2] Kolodziej, J. , Khan, S. U. , ErolGelenbe, and Talbi, E. G. , *Scalable optimization in grid, cloud, and intelligent network computing – foreword*, Concurrency and Computation Practice and Experience, 25(12), 1719-1721, 2013.

[3] Zhang, B. , *Optimization of fp-growth algorithm based on cloud computing and computer big data*, International Journal of System Assurance Engineering and Management, 12(4), 853-863, 2021.

[4] He, X. L. , Song, Y. , and Binsack, R. V. , *The intelligent task scheduling algorithm in cloud computing with multistage optimization*, International Journal of Grid and Distributed Computing, 9(4), 313-324, 2016.

[5] Jin, S. , Wu, H. , Yue, W. , and Takahashi, Y. , *Performance evaluation and nash equilibrium of a cloud architecture with a sleeping mechanism and an enrollment service*, Journal of Industrial, Management Optimization, 16(5), 2407-2424, 2020.

[6] Moser, O. , Rosenberg, F. , and Dustdar, S. , *Domain-specific service selection for composite services*, IEEE Transactions on Software Engineering, 38(4), 828-843, 2013.

[7] Guo, Y., Guo, Y. , *Intelligent Network Office System Based on Cloud Computing and Machine Learning*. Mobile Information Systems, 2021.

[8] Tai, L., Li, L., Du, J. , *Multimedia based intelligent network big data optimization model.* Multimedia Tools and Applications, 78(4), 4579-4603, 2019.

[9] Qin, Y., Xuan, H., Zhang, B., *Intelligent Management System of Power Network Information Collection Under Big Data Storage*. In MATEC Web of Conferences (Vol. 100, p. 02028). EDP Sciences, 2021.

[10] Hsu, K. , *Big data analysis and optimization and platform components*. Journal of King Saud University-Science, 34(4), 101945, 2022.

[11] Tkachenko, O. , Tkachenko, O. , and Tkachenko, K. , *Actual trends of cloud computing and technologies in optimization of data storage*, Digital Platform Information Technologies in Sociocultural Sphere, 3(2), 192-208, 2020.

[12] Meena, V. , Gireesha, O. , Krithivasan, K. , and Sriram, V. , *Fuzzy simplified swarm optimization for multisite computational offloading in mobile cloud computing*, Journal of Intelligent and Fuzzy Systems, 39 (6), 1-13, 2020.

[13] Zhou, Q. , Zhang, Z. , and Y Wang. , *Research on safety management system optimization of b2c e-commerce intelligent logistics information system based on data cube*, Journal of Intelligent and Fuzzy Systems, 38(4), 1-8, 2019.

[14] Peng-Fei, G. , Wei-Wei, Z. , Wei-Xin, S. , Jun-Quan, Y. , Yi-Qun, Y. , De-Xin, T. , et al. , *Commissioning and operation of bioprocess intelligent optimization system in yangzhou liuwei wwtp*,China Water and Wastewater, 28(9), 5-8, 2012.

[15] Zhang, B. , *Optimization of fp-growth algorithm based on cloud computing and computer big data*, International Journal of System Assurance Engineering and Management, 12(4), 853-863, 2021.

[16] Zhu, D. D. , and Sun, J. Q. , *The path optimization algorithm of car navigation system considering node attributes under time-invariant network*, Mobile Information Systems, 2021(4), 1-20, 2021.

[17] Huang, C. H. ,and Wang, C. H. , *Optimization of preventive maintenance for a multi-state degraded system by monitoring component performance*, Journal of Intelligent Manufacturing, 27(6), 1-20, 2016.

[18] Sarma, P. , *Technology focus: digital data acquisition*, Journal of Petroleum Technology, 72(1), 70-70, 2020.

[19] Kwon, O. , *The potential roles of context-aware computing technology in optimization-based intelligent decision-making*, Expert Systems with Applications, 31(3), 629-642, 2006.

[20] Donciu, C. , Temneanu, M. , and Branzila, M. , *Sustainable irrigation based on intelligent optimization of nutrients applications*, Environmental engineering and management journal, 6(6), 537-540.

[21] Luo, A. , Chao, H. , and Peng, J. , *Adaptive inertia weight based particle swarm optimization for resource scheduling in*

medical cloud system, Journal of Information and Computational Science, 12(2), 589-599, 2015.

[22] Narayanan, M. , Kumar, R. G. , J Jayasundaram, Kumar, R. and Isaac, C. , *Big data analytics and an intelligent aviation information management system*, Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 4328-4340, 2021.

[23] Mishra, B. , Das, H. , Dehuri, S. and Jagadev, A. K. , *studies in big data*, [studies in big data] cloud computing for optimization: foundations, applications, and challenges volume 39 || internet of cloud: security and privacy issues. , 10.1007/978-3-319-73676-1(Chapter 11), 271-301.

[24] Jian, H. , *Food safety and network resource optimization under the environment of cloud computing*, Advance Journal of Food Science and Technology, 11(5), 404-407, 2016.

[25] You, X. M. , Liu, K. , Liu, S. and Wang, Y. M. , *Study on dynamic resource allocation optimization model under cloud computing environment*, Xitong Gongcheng Lilun yu Shijian/System Engineering Theory and Practice, 34(s1), 234-238, 2021.

[26] Peng-Fei, G. , Wei-Wei, Z. , Wei-Xin, S. , Jun-Quan, Y. , Yi-Qun, Y. , De-Xin, T. , et al. , *Commissioning and operation of bioprocess intelligent optimization system in yangzhou liuwei wwtp*, China Water and Wastewater, 28(9), 5-8, 2012.

[27] Reshmy, A. K. and Paulraj, D. , *Data mining of unstructured big data in cloud computing*, International Journal of Business Intelligence and Data Mining, 13(3/4), 147, 2017.

[28] Drissi, S. , *Integration of cloud computing, big data, artificial intelligence, and internet of things: review and open research issues*, International Journal of Web-Based Learning and Teaching Technologies, 6 2021(1548-1093), 10-17, 2021.

[29] Chang, Victor , *Cloud computing for brain segmentation - a perspective from the technology and evaluations*, International Journal of Big Data Intelligence, 1(4), 192, 2014.

[30] Vakilinia, S., Zhang, X., Qiu, D., *Analysis and optimization of big-data stream processing.* In 2016 IEEE global communications conference (GLOBECOM) (pp. 1-6). IEEE, (2016, December).

Edited by: Vinoth Kumar
Received: May 5, 2022
Accepted: Aug 23, 2022

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2022 SCPE. Volume 23, Issues 3, pp. 103–114, DOI 10.12694/scpe.v23i3.2005

MUSIC INFORMATION RETRIEVAL USING SIMILARITY BASED RELEVANCE RANKING TECHNIQUES

KARTHIK VASU* AND SAVITA CHOUDHARY†

Abstract. The purpose of this proposed study activity is to construct a system for the job of automatically assessing the relevance of music datasets, which will be used in future work. Determine item similarity is an important job in a recommender system since it determines if two items are similar. Participants' systems must provide a list of suggested music that may be added to a given playlist based on a set of playlist characteristics, which will work along with the algorithms designed to provide other similar songs. Specifically, in this study, the challenges of detecting music similarity only on the basis of song information and tags given by users have been addressed. The proposed technique has been tested using a variety of machine learning algorithms to see how well it performs. tf-idf and Word2Vec are the methods used to model the dataset and generate feature vectors. It has also been found we that machine learning techniques, including Collaborative Filtering, KNN, Frequent Pattern Growth, and Matrix Factorization, have a greater influence on relevance ranking than traditional methods.

Key words: Music Information Retrieval, Machine Learning, Collaborative Filtering, Spotify

AMS subject classifications. 68P20, 68T05

1. Introduction and examples. It is the search and organization of enormous collection of music, or musical information, as per their relevance to particular queries that is the subject of Music Information Retrieval (MIR). This is especially significant in light of the large amounts of musical information that is now accessible in digital format, as well as the widespread use of music-related digital services. Aside from that, given its apparent commercial appeal, the majority of media content owners as well as distributors (e.g., Philips and Sony), as well as major technology companies (e.g., Apple and IBM), are actively engaged in research in the area, and a growing number of libraries are attempting to integrate some form of support for MIR into their on-line digital services. This results in an analysis of the text against the text data connected with album and& songs, making the system practically identical to any text-based search engine in terms of functionality (e.g. Google, Yahoo). Although this is the case, systems that are capable of receiving ”musical” enquiries, like musical scores, whistling melodies (query by humming), or audio recording segments are necessary due to the nature of the content being retrieved.

The phrase ”song similarity” refers to the measurement of how close two songs are close in terms of how probable it is that user really want listening to them when they are compared side by side. Yes, the process of developing an objective similarity measure is subjective, and researchers have taken two approaches to accomplish this: Both the objective technique, in which likeness is identified related to the raw data, such as spectral or rhythmic analysis of songs, and the subjective approach, wherein user-generated data, such as tags (also known as collaborative filtering), are used, are discussed further. For the sake of this research, we shall use the subjective technique to determine the extent to which two tracks are comparable to one another. We would then create a resemblance level among two songs scale from 0 (totally distinct) to one (identical), and then we will determine it using the co-occurrences of pairs of things in users' histories by using cosine metric to determine how similar two songs are to one another. For more information, please see the following link. This measure will also serve as our model of reality and, as a result, as our source of truth in the future. Such a concept is realistic, as shown by the fact that researchers in the area have utilized it successfully [1].

*Assistant Professor, Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bengaluru, India (karthikv@msrit.edu).

†Associate Professor, Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, India (savitha_cs@sirmvit.edu)

The intended objective of this work is to employ Data Science methodologies in conjunction with Machine Learning algorithms to estimate song popularity on Spotify as well as other music streaming services, which are quite popular and extensively used. Using success measures or criteria, we will evaluate each and every popular Machine Learning algorithm and choose the best algorithm from among those evaluated. It is often the result of some form of calculated miscalculation. The purpose of the best model that has been built is to forecast the popularity of a song based on a variety of current and historical characteristics.

2. Related Work. Music data retrieval, recommendations, and similarity approaches will be discussed in this session. Recommendation system and Relevancy rankings, music information retrieval research, and the tradeoff among short- and long-term involvement in digital sites are all areas of study that we draw upon in our work [2].

2.1. Spotify. The fact that Spotify is in business after three of the world's largest corporations, Google, Apple and Amazon have all entered the music streaming sector is testament to the company's tenacity. In comparison to the other three, they are able to provide superior music recommendations to the user. At first glance, it seems like Spotify has a far superior system than the other three competitors. Spotify recommends music based on three different types of models [3]. The first of them are models of Natural Language Processing. These compare songs based on the words used to describe the music, which may be found in things like articles on the internet. The following is used to propose songs that aren't as popular as the previous. These Content-Based models analyse the real audio and utilise commonalities to suggest comparable songs to you based on your preferences. Collaborative filtering is the name given to the final style. This is accomplished by establishing a user vector for each individual and a song vector for each individual song in the music library [4]. It then compares them in order to propose music that is comparable to each other as well as music that similar people like listening to [4]. As discussed in this article, Collaborative Filtering is the most widely used technique of music recommendation on Spotify and is at the core of Spotify's most popular music suggestion method: Discover Weekly. It will be this algorithm, and more precisely how linear algebra is employed to power it, that will be the focus of our discussion. Notably, other firms, such Last.fm, have taken a similar approach to Spotify's strategy, while another has been promoted by the Netflix Prize, which is modelled after Spotify's concept.

2.2. Collaborative Filtering. Filtering in a Collaborative Way Every week, Spotify creates a Discover Weekly playlist for each of its 140 million customers [5]. Using over 40 million tracks, they discover the songs that are most likely to be a good fit for you and your music collection. If your buddy tells you about a specific music x because they found out that you enjoy song y, Collaborative Filtering aims to mimic that. For this, a latent factor model, a machine learning method, is utilized, which converts unobservable raw data into unobservable latent features. Cost functions are used to calculate the latent components using an alternating least-squares approach. Switching back & forth between user and song factors is the procedure of alternating-least-squares in our example. This approach keeps the cost function converged [6]. A cost function is used to determine how inaccurate an estimate is compared to the true value of the data. In order to solve the problem, you enter a certain set of parameters. Then, using our alternating-least-squares technique, you may fine-tune these characteristics until the cost function converges. When that happens, we'll have a preferences-confidence pair for users to display our choices and the degree to which we're confident in them.

3. Materials and Methods.

3.1. Dataset. The Spotify Million Playlist Dataset (MPD) is a collection of 1,000,000 user-created playlists. From January 2010 until October 2020, these playlists were compiled. Some playlists from original playlists are removed to match the competition's challenge playlist, and all holdout tunes appear in the MPD Test Set 1000 playlists. The initial set of test sets will be the train set that removes rails. When addressing this issue, we refer to "item" as a "song" and "user" as a "playlist". We won't utilise content-based filtering since the dataset doesn't give much information about every song. Our emphasis would be on KNN, Collaborative Filtering and Frequent Pattern Growth as well as matrix factorization [7].

Each of the 1,000 slice files in the Million Playlist Dataset makes up one Million Playlist.

mpd.slice.STARTING PLAYLIST ID - ENDING PLAYLIST ID.json is the name pattern for these files.

Table 3.1: Playlists Field

S.NO	Parameter	Outcome
1.	Playlist	1000000
2.	Tracks	66346428
3.	Unique tracks	2262292
4.	Unique albums	734684
5.	Unique Titles	92944

File mpd.slice.0-999.json contains the MPD's first 1,000 playlists, for example mpd.slice.999000-999999.json contains the latest 1,000 playlists.

Info Field: In the info box, you may get basic information about the slice you're looking at: Slicing is based on the number of slices in a certain file. For example, slices 0-999 indicate the file's version, whereas slices generated on and MPD version 1 indicate whenever the slice was produced. The information and playlists fields in each slice file are both JSON dictionaries.

Playlists Field: Typically, this is a collection of 1,000 playlists. An individual playlist is nothing more than a dictionary, with the following entries:

1. The MPD ID of all this playlist is the pid (integer) ranging from zero to ninety-nine,999.
2. name - string - the playlist's title.
3. The playlist's description, if any (this is an optional string). For the time being, the majority of Spotify's playlists do not include descriptions contributed by its listeners.
4. This playlist's time has just recently been updated in the seconds ever since epoch. Whenever the playlist is reloaded, the timings are calculated based on the current day and time in Greenwich Mean Time (GMT).
5. A playlist's total number of distinct artists may be seen in the num artists field.
6. num albums - the playlist's total number of distinct albums.
7. In the playlist, the number of tracks can be found.
8. num followers - how many followers this playlist had when the MPD was generated. This number does not include the playlist creator's followers.
9. This is the number of distinct editing sessions that have occurred. During a two-hour editing window, any tracks that have been added are considered to be part of the same editing session.
10. A playlist's overall length is represented by duration ms (in milliseconds)
11. A collaborative playlist is one in which all members of the group are actively participating. It is possible for many people to submit tunes to a playlist that is shared.
12. A playlist's "tracks" section contains a slew of data about the songs. This array has a dictionary with following fields for each of its elements:

3.2. Recommender. Products like books, movies, music, and friends would be recommended to customers using a recommender system. Recommenders may fall into one of two categories. Content-based filtering and collaborative filtering are two types of filtering. Content-based systems, on the whole, are easier to use but provide less interesting suggestions. Although collaborative systems may be cumbersome to operate and need a lot of user-generated data, they are the most advanced technology available today [8].

3.2.1. Using content as a filtering mechanism. Products may be manually investigated and labelled by experts or consumers according to several categories or qualities. The more characteristics we have, the easier it is to compute the similarity of one thing to other objects and to find related products.

3.2.2. Collaborative filtering. Filtering in groups: a process in which many people work together to improve the quality of the input. Collaborative filtering, unlike content-based filtering, does not need manual la-

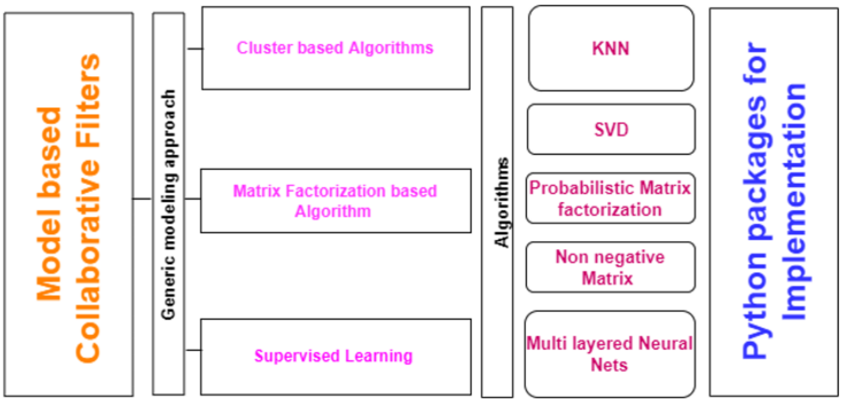


Fig. 3.1: Architecture of Model based Collaborative filtering

belling. To produce product recommendations, the algorithm would seek for other users with similar preferences and habits. There are two forms of collaborative filtering: memory-based and model-based. Clustering and matrix factorization are two machine learning methods that are used in model-based CF in order to determine similarity between user items and their corresponding models.

There are two forms of memory-based filtering: user-item and item-item. Take a certain individual, look for others with comparable ratings to that person, and propose products those people enjoyed. For example, if you're looking for a certain item, you may use item-item filtering to look for additional goods that individuals who like that item also like. Recommendations are generated from things that are entered into the system. "Customers who purchased this item also purchased. It's simple "Customers who are identical to you also enjoyed." and "Customers who like this item also liked..." are the same thing when it comes to Customer Feedback (CF).

The implementation of memory-based algorithms is simple, and the quality of the predictions they make is acceptable. In real-world applications, memory-based CF has limitations since it cannot handle the well-known cold start issue, which arises when new users or items join the system. In contrast to memory-based models, model-based CFapproaches are scalable and therefore can handle greater sparsity levels, but suffers when new members or things that do not have ratings join the system.

Matric factorization is the foundation of all model-based filtering and model-based CF. There are several applications where matrix factorization (CF) is preferred over Memory-based CF, such as recommendation systems.

Users' preferences and the properties of objects may be learned from past ratings using Matrix factorization. Then, using the dot product of these latent features, Matrix factorization can forecast ratings for new items. If you have a user-item matrix, factorization may reorganise it into a low-rank structure that can be expressed as the product of the two low-rank matrices, each of which contains the latent vector. In order to get as close as possible to your original matrix, multiply the low-rank matrix together, which completes the original matrix's missing members.

3.3. Playlist Recommender. The terms "user" as "playlist" and "item" as "song" will be referred for the sake of this discussion. Since the data will not really contain enough detail about every song, we can't utilise content-based filtering. As a consequence, we'd only be able to focus on one thing. KNN, Collaborative Filtering, Word2Vec., Frequent Pattern Growth, ID3, Customized kernel SVM [16].

3.3.1. Collaborative Filtering.

Playlist-based CF: The proposed model will infer the current "rating" from the similarities between each playlist or how other playlists "rate" (include or exclude) a tune. Song-based CF: The model will also estimate the current "rating" from the similarity of each song and how the current playlist "rates" other

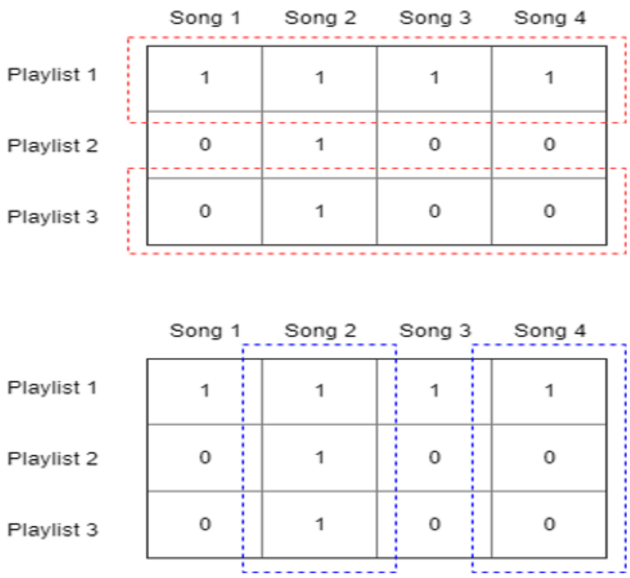


Fig. 3.2: Similarity between song-song or playlist-playlist

songs. This approach has been followed in Collaborative Filtering [17].

Song-based CF: I can estimate the current ”rating” from the similarity of each song and how the current playlist ”rates” other songs. We followed the approach in Collaborative Filtering.

Step1. Create a playlist-song matrix in which ”1” indicates that the song is in the playlist & ”0” indicates that it is not. For instance, playlist 1 includes songs 2 and 3, and song 2 is also included in playlist 2.

Step2. To begin, the model divides the data into testing and training sets.

Step3. Determine the degree of similarity among songs or playlists. In playlist-playlist similarities, every row is treated as a vector, but in song-song similarity, each column is treated as a vector.

Cosine Similarity formula is given below in equation 3.1:

$$similarity = \cos(\theta) = \frac{A.B}{\|A\|\,\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.1}$$

A_i and B_i are components of vector A and B. We make predictions on the testing set based on similarity matrix and. According to our algorithm, we can predict that song S will be included in any given playlist p by calculating a total of the weighted sums of all other playlists that include the same song. After that, the data is normalised.

$$\hat{r}_{ps} = \frac{\sum_{p'} sim(p, p') r_{p's}}{\sum_{p'} |sim(p, p')|} \tag{3.2}$$

Song-song tries to replace the similarity matrix of playlist with that of songs.

$$\hat{r}_{ps} = \frac{\sum_{s'} sim(s, s') r_{ps'}}{\sum_{s'} |sim(s, s')|} \tag{3.3}$$

Table 3.2: Singular Value Decomposition Equation: **X, U, S, V**^{T}

$\mathbf{X}_{11}$	$\mathbf{X}_{12}$	**...**	$\mathbf{X}_{1M}$		$\mathbf{U}_{11}$	$\mathbf{U}_{12}$	**...**	$\mathbf{U}_{1R}$		$\mathbf{S}_{11}$	**0**	**...**	**0**		$\mathbf{V}_{11}$	$\mathbf{V}_{12}$	**...**	$\mathbf{V}_{1N}$
$\mathbf{X}_{21}$					X_{21}					0	S_{22}				V_{21}			
. . . . .				=	. . . . .					. . . . .					. . . . .			
											.	.						
$\mathbf{X}_{M1}$			X_{MN}		U_{M1}			U_{Mr}		0			S_r		V_{M1}			V_{MN}

Table 3.3: Feature Vector

S.NO	Parameter	Outcome
1.	Playlist	1000000
2.	Tracks	66346428
3.	Unique tracks	2262292
4.	Unique albums	734684
5.	Unique Titles	92944

3.3.2. K Nearest Neighbor. Playlist-based (like user-based) :

1. Create a similarity matrix for playlists (cosine, euclidean, Pearson correlation)
2. Px for each playlist
3. 1 = n
4. While total track is less than 500
5. Find the n-th most relevant Px playlist, termed Pr
6. Move K (or all) songs from Pr to Px.
7. Increase n by one.

Song-centered (like item-based) :

1. (The user space is the space of similarity)
2. Create a song similarity matrix (cosine, euclidean, Pearson Correlation)
3. For each playlist, Px:
4. Determine the ”cluster centre” by averaging all matches.
5. Obtain K = 500 nearest neighbours and add them to existing songs.

3.3.3. Model Based. Probabilistic factorization (PMF), Orthogonal factorization (SVD) or non-negative factorization are all methods for matrix factorization (NMF). Collaborative Filtering via Singular Value Decomposition (SVD) could be defined as the process of estimating a matrix X through svd. The team that wins in the Netflix Award contest generated product suggestions using SVD matrix factorization models. The generic equation is as follows: $\mathbf{X} = \mathbf{U} \times \mathbf{S} \times \mathbf{V}^T$. The transpose is shown in Table 3.2.

Now that we hide feature vectors for playlists and song, we could inject them into any Machine Learning algorithms for result prediction (Table 3.3).

Figure 3.3 shows the distribution of Playlist length, Number of Albums / Playlist, Number of Artist / Playlist, Number of edits / Playlist, Number of Followers / Playlist, Number of Tracks / Playlist. Song based relevance ranking is described in the figure 3.5. Top 20 songs are listed as per the similarity check using machine learning technique. This describes the song based and artists based relevance ranking. could improve the success of automatic ranking. Based on this intuition, a new method has been introduced for the selection

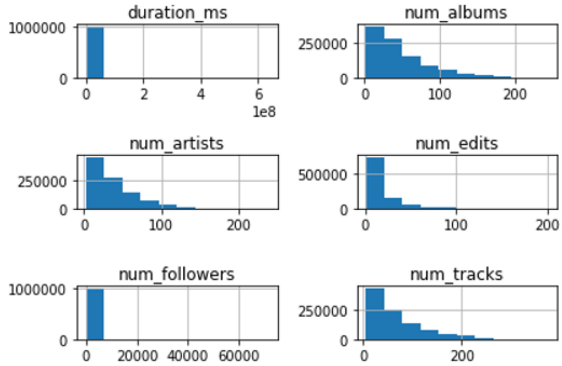


Fig. 3.3: Distribution of Playlists

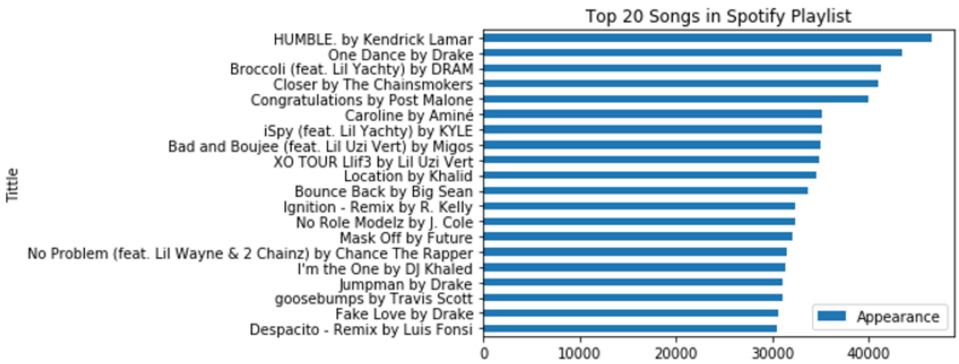


Fig. 3.4: Song Based Relevance ranking

of systems to be used for data fusion. For this purpose, bias concept is utilized that measures the deviation of a system from the norm or majority and employ the systems with higher bias in the data fusion process. This approach provides even higher correlations with the human based results. This demonstrates that the proposed solution outperforms the proposed automatic ranking methods shown in figures 3.4 and 3.5 [20,21,23].

The figures 3.4 and 3.5 show the relevance ranking of songs and artists respectively. The list is of the year 2020. For the search of Top 20 Songs, the list obtained has used the similarity based relevance ranking technique, which is discussed in Section 3. The same applies to the search of Top 20 Artists in the playlist available.

4. Results and Discussions.

4.1. Metrics. The following metrics will be used to evaluate submissions. In other words, all metrics would be analysed at both the level of tracks (the identical track should be replicated) and also at the artist levels (any track by that artist is a match). G's ”ground truth” set of music and R's ”ordered lists of recommended tunes” has been employed in the following discussion. From: to-subscripts are used in order to index a list, which is signified by ∥. On individual metrics, previous entries are scored higher when there are multiple submissions

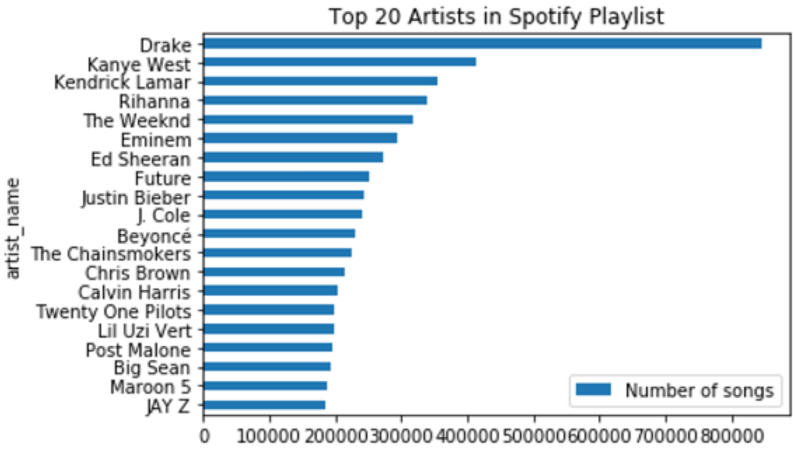


Fig. 3.5: Artists Based Relevance ranking

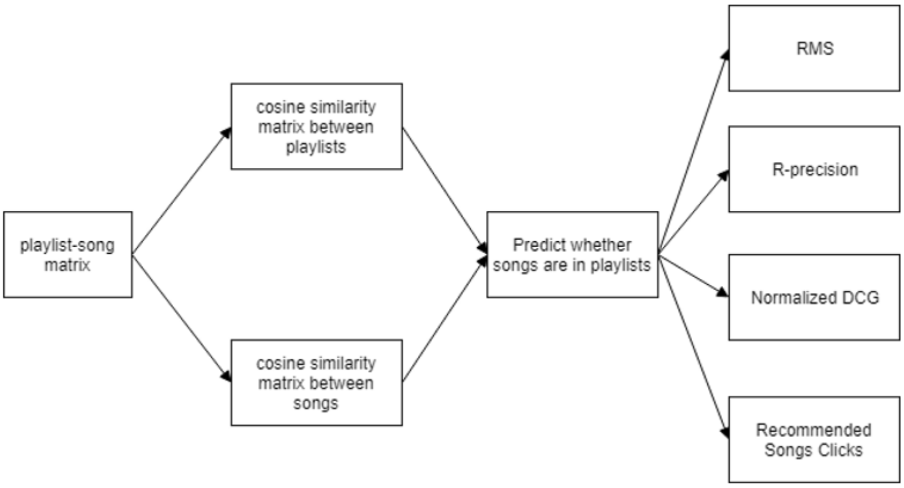


Fig. 4.1: Sample: Similarity between song-song or playlist-playlist

that are tied. Figure 4.1 shows the Similarity between song-song or playlist-playlist [20,21].

4.2. R-precision. It is defined as the number. of essential tracks that were found divided by number of tracks that were already known to be relevant (i.e., the number of tracks that were not found).

$$R - precision = \frac{|G \cap R_{1:|G|}|}{|G|} \tag{4.1}$$

To get this metric, you need to look at all playlists inside the challenge set. This metric is about how many tracks were found that were relevant (regardless of order).

5. Normalized discounted cumulative gain. Recommendation tracks with a higher DCG ranking are more likely to be of interest to users. When calculating the DCG, divide by the optimal DCG wherein the

Table 7.1: Comparison of Performance with Machine Learning algorithms vs Customized Model

Algorithms	Variable	Accuracy/ R-precision	NDCG	Duration
KNN	Song-features	87.04	81.44	36.25
FP growth	Song-features	81.25	76.42	41.28
Support Vector Machine	Song-features	78.23	74..21	57.11
ID3- Iterative Dichotomiser 3	Song-features	76.17	81.77	49.23
Navie Bayes	Song-features	74.20	79.45	44.18
Customized kernel based Support Vector Machine	Song-features	89.21	84.25	31.56

suggested tracks are ideally ranked:

$$DCG = rel_1 + \sum_{i=2}^{|R|} \frac{rel_i}{\log_2(i+1)} \tag{5.1}$$

In this situation the ideal IDCG & DCG are :

$$IDOG = 1 + \sum_{i=2}^{|G \cap R|} \frac{1}{\log_2(i+1)} \tag{5.2}$$

The DCG is zero if the intersection of R & G is empty. The following formula is used to compute NDCG:

$$NDCG = \frac{DCG}{IDCG} \tag{5.3}$$

6. Recommended Songs clicks. Using Spotify's Suggested Songs function, you can get a list of 10 songs to add to an existing playlist based on the music in it. Ten more tracks could be generated by refreshing the list. The quantity of refresh required before the suitable track is found is called "Recommended Songs clicks." It's done like this:

$$clicks = \left\lfloor \frac{\arg min_i\{R_i : R_i \in G|\} - 1}{10} \right\rfloor \tag{6.1}$$

A value of 51 (i.e. 1 + the greatest clickthrough conceivable) is chosen if the metric (i.e. the necessary track in R) does not exist.

7. Aggregation of Rank. The Borda Count voting procedure will be used to compute the final rankings. The highest ranked system receives p points for each of the p participant rankings based on NDCG, R-precision &, the next system gets p-1 points, Recommended Songs clicks etc. The player with more overall points wins. Top-down comparative assessment is used to determine which systems have the most first-place position in the case of a tie. The result of item and user based search is shown in Table 7.1 and the outcome of an item-item/ customer- items based search is shown in Table 7.2.

It can be concluded from figures 7.1, 7.2 and 7.3 that the custom model has outperformed the above mentioned algorithms, KNN, FP growth, support vector machine, ID3 and Navie bayes. The Accuracy in Figure 7.1 has been proven to be better than the comparison algorithms by 2.17%, while in Figure 7.2, the time taken for a similar match has reduced by about 13%. The NDCG has also seen a growth about 2.81% which can be discerned from Figure 7.3.

8. Conclusion. We are able to produce suggestions of music that are identical to a music in our playlist by using idea of cosine similarity. As a result, you can use the procedures outlined above to create your own Spotify playlist recommender. Possible consequences of machine learning also include individualized, immersive

Table 7.2: The outcome of an item-item / customer- items based search

Procedure	**Variable**	**Accuracy/ R-precision**	**NDCG**	**Clicked songs**	**Duration**
Playlist-based baseline	Playlist	0.7766	1.601	0	41.42
Song-based baseline	Song	0.7847	0.7975	0	4183
Word2Vec + Song-based	100-200-300 dimension	0.003	0.004	10.35	69.13
Word2Vec + Playlist-based	min_fre = 3, dimension 50	0.0171	0.015	8.086	83.25
Word2Vec + Playlist-based	min_fre = 3, dimension 100	0.019	0.0172	7.805	88.35
Playlist-based CF (get top K rating songs)	Song	0.8045	0.8011	0	approx 12000

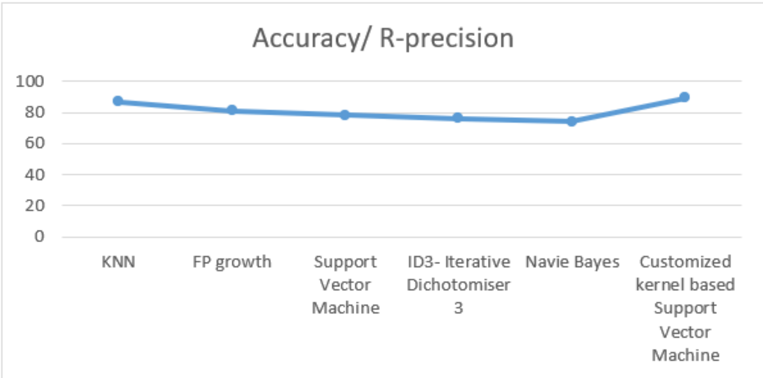


Fig. 7.1: Comparison of Accuracy with Machine Learning algorithms vs Customized Model

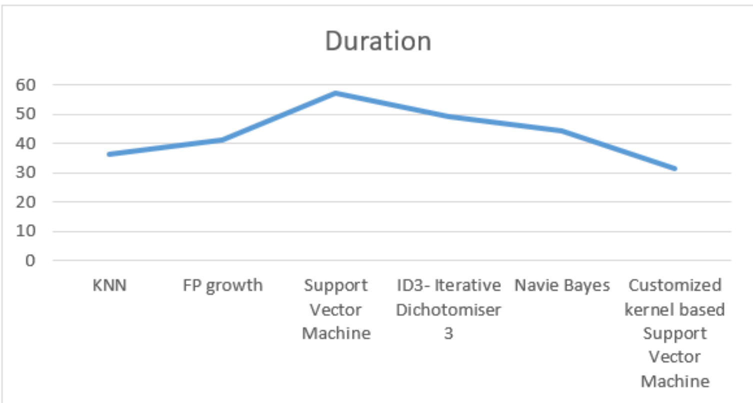


Fig. 7.2: Comparison of Time taken with Machine Learning algorithms vs Customized Model

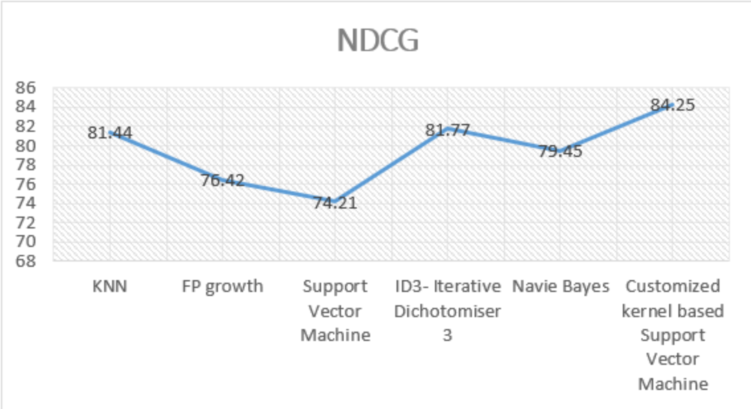


Fig. 7.3: Comparison of NDCG with Machine Learning algorithms vs Customized Model

Table 7.3: RMS and Accuracy Result

Procedures	RMS	Accuracy/R-precision
CF that is playlist-based (used to generate songs having k rating)	0.00716	0.621
Song that is based on CF	0.00014	0.76
Matrix Factorization	0.00045	0.69

guest experiences having increasingly complex features, in parallel with developments in technology and media affordances. Choosing recommendation methods produces better results for the full dataset, however for large datasets, processing techniques are required using big data technologies to fetch the data within the less time which will be handled in future improvements and analyses.

9. Future Scope. With the advent of various similar algorithms, the proposed model has shown promise by being relatively better at obtaining similar songs at the quickest. For huge datasets, the time taken for the searching process can be further reduced using Spark and Hadoop frameworks. There is hope to be a milestone and a reference point for future researches done in the same field as well as for others.

REFERENCES

[1] Cunha, R. L., Caldeira, E., Fujii, L., *Determining Song Similarity via Machine Learning Techniques and Tagging Information.*, arXiv preprint arXiv:1704.03844

[2] Anderson, Ashton, et al., *Algorithmic effects on the diversity of consumption on spotify*, Proceedings of The Web Conference, 2020.

[3] Bateira, José Lage, *Spotify-ed-Music recommendation and discovery in Spotify.*, 2014.

[4] Pichl, Martin, Eva Zangerle, and Günther Specht., *# nowplaying on# spotify: leveraging Spotify information on Twitter for artist recommendations.*, International Conference on Web Engineering, Springer, Cham, 2015.

[5] Ouiem Bchir, Mohamed M. Ben Ismail and Sara Algarni, *Support Kernel Classification: A New Kernel-Based Approach.*, International Journal of Advanced Computer Science and Applications(IJACSA), 11(**??**), 2020.

[6] Belattar Sara, Abdoun Otman and El khatir Haimoudi, *New Learning Approach for Unsupervised Neural Networks Model with Application to Agriculture Field.*, International Journal of Advanced Computer Science and International Journal of Advanced Computer Science and Applications(IJACSA), 11(5), 2020.

[7] T V Divya and Barnali Gupta Banik, *An Empirical Study on Fake News Detection System using Deep and Machine Learning Ensemble Techniques.*, International Journal of Advanced ComputerScienceandApplications(IJACSA),12(12),2021.

[8] Lan Phuong Phan, Hung Huu Huynh and Hiep Xuan Huynh, *Recommendation using Rule based Implicative Rating Measure.*, International Journal of Advanced Computer Science and Applications(IJACSA),2018.

[9] Kim, J., Urbano, J., Liem, C.C.S. et al, *One deep music representation to rule them all?*, A comparative analysis of different representation learning strategies., Neural Comput & Applic 32, 1067–1093 (2020).

[10] Pampalk, Elias.,, *Computational models of music similarity and their application in music information retrieval.*, Diss. 2006.

[11] Mitra, Bhaskar, and Nick Craswell, *Neural models for information retrieval.*, arXiv preprint arXiv:1705.01509 (2017).

[12] Schedl, Markus, Emilia Gómez Gutiérrez, and Julián Urbano, *Music information retrieval: Recent developments and applications.*,Foundations and Trends in Information Retrieval. 2014 Sept 12; 8 (2-3): 127-261. (2014).

[13] Amudha, S., and I. Elizabeth Shanthi, *Phrase based information retrieval analysis in various search engines using machine learning algorithms.*,Data Management, Analytics and Innovation. Springer, Singapore, 2020. 281-293.

[14] Lin, Jimmy, et al., *Supporting interoperability between open-source search engines with the common index file format.*,Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.

[15] Mitra, Bhaskar, and Nick Craswell, *An introduction to neural information retrieval.*,PBoston, MA: Now Foundations and Trends, 2018.

[16] Schindler, Alexander, *Multi-Modal Music Information Retrieval: Augmenting Audio-Analysis with Visual Computing for Improved Music Video Analysis.*,arXiv preprint arXiv:2002.00251 (2020).

[17] Myna, A. N., K. Deepthi, and Samvrudhi V. Shankar, *Hybrid Recommender System for Music Information Retrieval.*,Journal of Computational and Theoretical Nanoscience 17.9-10 (2020): 4145-4149.

[18] Knees, Peter, Markus Schedl, and Masataka Goto, *Intelligent user interfaces for music discovery.*,Transactions of the International Society for Music Information Retrieval 3.1 (2020).

[19] Castellon, Rodrigo, Chris Donahue, and Percy Liang, *Codified audio language modeling learns useful representations for music information retrieval.*,arXiv preprint arXiv:2107.05677 (2021).

[20] Karthik, V., and Savita Choudhary, *TaCbF-"Trending Architecture for Content based Filtering using Data Mining.*,Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Springer, Singapore, 2021.

[21] Gupta, Rahul, Jayesh Yadav, and Cheshtha Kapoor, *Music information retrieval and intelligent genre classification.*,2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), IEEE, 2017.

Edited by: Vinoth Kumar
Received: May 5, 2022
Accepted: Aug 23, 2022

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2022 SCPE. Volume 23, Issues 3, pp. 115–127, DOI 10.12694/scpe.v23i3.1992

TRANSFER LEARNING ASSISTED CLASSIFICATION OF ARTEFACTS REMOVED AND CONTRAST IMPROVED DIGITAL MAMMOGRAMS

PARITA OZA*, PAAWAN SHARMA, AND SAMIR PATEL †

Abstract. Mammograms are essential radiological images used to diagnose breast cancer well in advance. However, an accurate diagnosis also depends on the quality of mammogram images. Therefore, removal of artefacts and mammogram enhancement are necessary pre-processing steps. Artefact removal helps exclude unsolicited regions in the mammograms and limits the search for suspicious regions without excessive impact from the background. Mammogram enhancements improve apparent visual details and improve some features of an image. In this paper, we propose a method for mammogram pre-processing. These pre-processed mammograms are then fed into Deep Convolutional Neural Network for the classification process. Two approaches are used and compared to classify mammograms; Training model from scratch and Transfer Learning. Transfer Learning is an excellent approach to dealing with the small-sized training set, allowing us to consume the extendibility of deep learning entirely. By employing VGG16 as a pre-trained network on the pre-processed MIAS dataset, we improved training accuracy (96.14%) compared to the model developed from scratch and other strategies described in the literature.

Key words: Mammogram, Artefacts,Thresholding, CLAHE, Classification, Deep CNN

AMS subject classifications. 68T05

1. Introduction. Among the female population worldwide, breast cancer is the most commonly diagnosed cancer [1, 2]. Breast cancer has become a widespread and significant health issue across Indian cities [3]. In the South-East Asia regions, Breast cancer ranks second as per the number of deaths by cancer site, and region [4]. A study [4] presented to estimate the burden of breast cancer in India using parameters like YLL (years of life lost), YLD (Years of healthy life lost due to disability), and DALY (Disability-adjusted life years) for the years 2016, 2021 and 2026. According to this study, breast cancer would increase total DALYs in India over the years. Hence, there is a need to initiate actions to control this disease in the country. Early detection of breast cancer can improve the chances of survival. A radiologist uses various imaging modalities to detect and diagnose this disease, including mammography, Breast MRI, Ultrasound, etc [5]. Out of all, mammography is the most traditional and successful imaging modality as it uses very low dose X-rays to capture the image and is easily affordable by people [6].

Mammogram pre-processing is an essential initial step for developing computer-aided breast cancer diagnosis. Lots of work have been done to process and analyze mammograms. A significant number of image processing techniques are used for processing mammograms, such as pixel-based transformation, edge-based transformations, and region-based transformation, and this field is still evolving [7]. Though mammography symbolizes a significant technological advancement in breast imaging, many artefacts and other noises are commonly perceived with this imaging modality [8]. Artefacts can be defined as any mammographic density variability. These artefacts may affect the accuracy rates and the inference knowledge capabilities of underlying techniques of breast cancer diagnosis. Hence artefact removal is an essential mammogram pre-processing task.

Sometimes mammograms have poor contrast. A large amount of stress exists on the eyes to achieve an appropriate focus due to poorly contrasted mammograms [9]. In addition, poor contrast may result in misdiagnosis if the breast region appears as almost monotonous grey [10]. This can slow down the learning process of the model.

The fields of Artificial Intelligence (AI), such as Machine Learning (ML) [11], [12] as well Deep Learning (DL), have also achieved tremendous success in various mammogram analysis tasks along with the detection and

*PhD Scholar at Pandit Deendayal Energy University, Gandhinagar, India, Nirma University, Ahmedabad, India
†Pandit Deendayal Energy University, Gandhinagar, India

classification [13] of breast abnormalities. Nowadays, the current research focus is on deep learning approaches due to the availability of a wide range of deep learning networks such as VGGnet [14], AlexNet [15], ResNet [16] etc. The research community of the domain widely adopts all these networks. VGGNet has many learnable parameters, but it is still most successful due to its uniform and simple architecture.

1.1. Research Contribution. The main scope of this paper is to develop a Deep convolutional neural network (Deep CNN) based model that can classify breast abnormalities. This model can serve as a second view tool to assist radiologists. The article also presents mammogram pre-processing techniques to eliminate unwanted mammogram regions and improve the mammogram's visual information contents. The main contribution of the paper is as follows:

- To provide a methodology for artefacts removal from the mammograms
- To perceive the effect of mammogram enhancement techniques
- To examine the power of the transfer learning approach for medical imaging data over CNN architecture implemented from scratch

1.2. Paper Organization. The paper is organized as follows: Section 2 presents existing literature in the domain. The proposed methodology is presented in section 3. Section 4 gives a detailed result analysis followed by conclusion in section 5. Figure 1.1 shows the taxonomy of the paper.

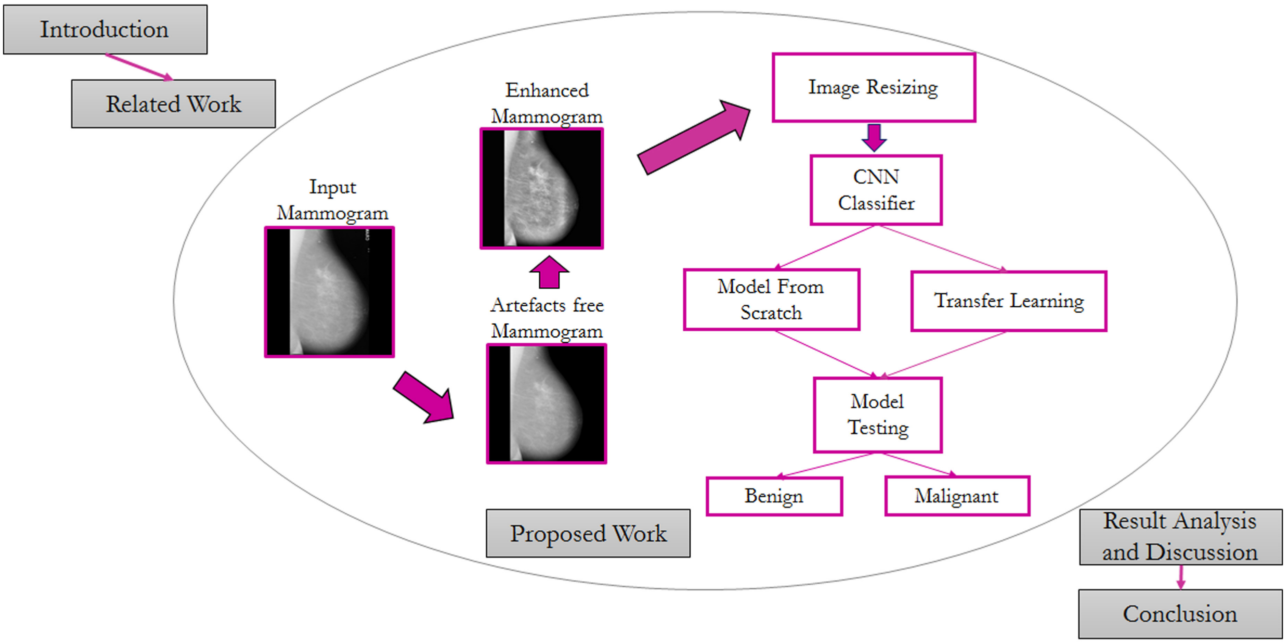


Fig. 1.1: Taxonomy of Paper

2. Previous Work. Deep learning techniques have been widely used in various application fields, including healthcare, personalization, fraud detection, image classification [30], pixel restoration, and many more. Moreover, this technique is also quite successful in various healthcare applications such as classification-Segmentation Pipeline for MRI [31], Retinal Blood Vessel Segmentation [33], abnormality detection, classification from medical images, and mammogram image analysis [34]. Dhivya et al. [17] proposed a model for enhanced tumor classification using pretrained neural network VGG16. The model is evaluated on various benchmark datasets. Authors could achieve an accuracy of 69.85% which was subsequently increased to 88% and 94% by conventional data augmentation and generative adversarial networks, respectively. Filtering techniques and morphological operations were applied on a mammogram as a part of pre-processing. Vaira et al. [6] developed a CNN model with eight convolutional, four max pool layers, and two fully connected layers. The model is compared with pre-trained nets such as AlexNet and VGG16. The proposed model achieved better results. Authors also have applied adaptive histogram equalization to improve the quality of mammograms. Sushreeta and Tripti [18] presented contrast enhancement techniques to detect the breast tumor boundaries from mammograms of the MIAS dataset. Authors have used thresholding and contrast limited adaptive histogram equalization (CLAHE) methods. Results were compared using the contrast improvement index (CII). It is concluded that the thresholding technique encouraged breast enhancement and segmentation results. Arefan et al. [19] presented a method to remove artefacts from a mammogram. Authors applied thresholding to get a binary image followed by connected component labeling to identify connected areas in an image. The largest connected areas in the image are identified, and the other regions are removed. Authors also have removed pectoral muscles from mammograms. Finally, breast tissues are extracted to classify mammograms such as Fatty, Glandular and Dense. Classification accuracy recorded was 97.66% with eight hidden layers neural network. Li et al. [20] presented a benign and malignant classification of a mammogram. Mammograms are pre-processed using zero-mean normalization, and data enhancement is used to prevent overfitting. The authors have modified DenseNet to replace the first convolutional layer with the Inception structure. Then, the pre-processed mammograms are trained and tested on various pre-trained CNN models and altered versions of DenseNet. Modified DenseNet could perform well as compared to other pre-trained models. The proposed work achieves 94% of overall classification accuracy. Another work is presented by Enas, and Nisreen [21] to classify malignant and nonmalignant mammograms. Pre-processing on mammograms was done using image filtration and CLAHE. The authors used two approaches to address the classification problem; using patches of ROI and whole images. The model is tested on different mammogram datasets. A model achieved higher performance on all the datasets.

3. Proposed Work. The focus of the proposed work is to classify breast abnormalities. Deep CNN model such as VGG16 is used as a classifier. The model is tested on the MIAS benchmark dataset. The initial step of the proposed work is to pre-process mammograms, which is a primary task for any medical imaging application. Pre-processing task includes artefact removal and mammogram enhancement. Thresholding transformations are used to remove artefacts such as high and low-intensity labels, markers, scratches, etc. Artefact removal is an essential pre-processing task that can limit the search for abnormal regions in the image without any impact from the background of the mammograms. The quality of mammograms is improved by applying the popular Contrast Limited AHE (CLAHE) method. Mammogram enhancement helps improve the visual quality and some features in the image. Figure 3.1 presents processing flow of proposed model. We present methodology of proposed model using activity diagram in figure 3.2.

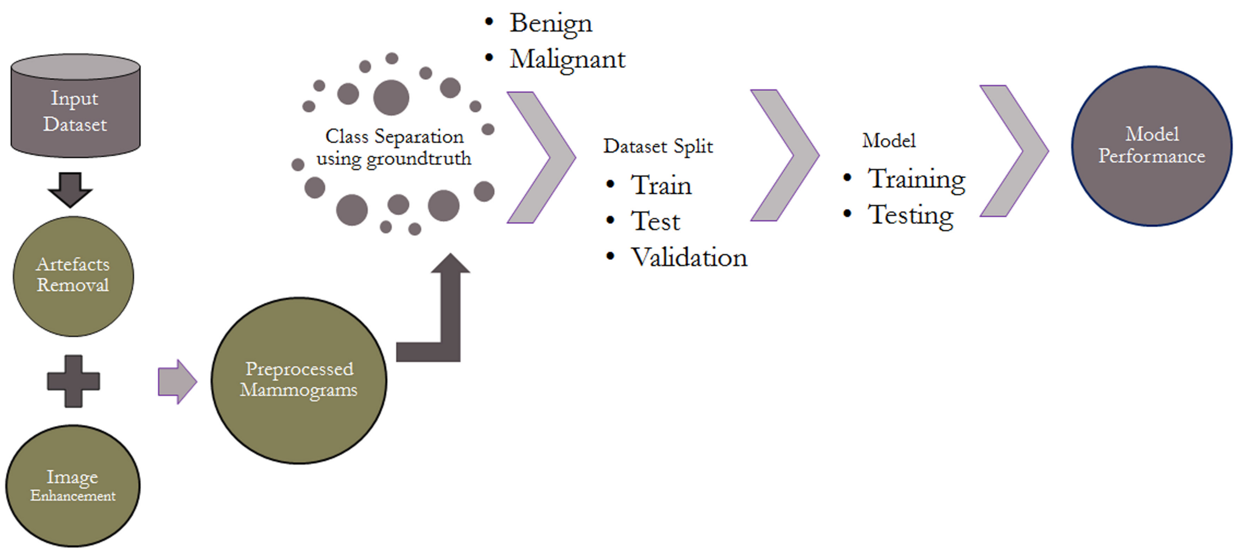


Fig. 3.1: Processing Pipeline of Proposed Model

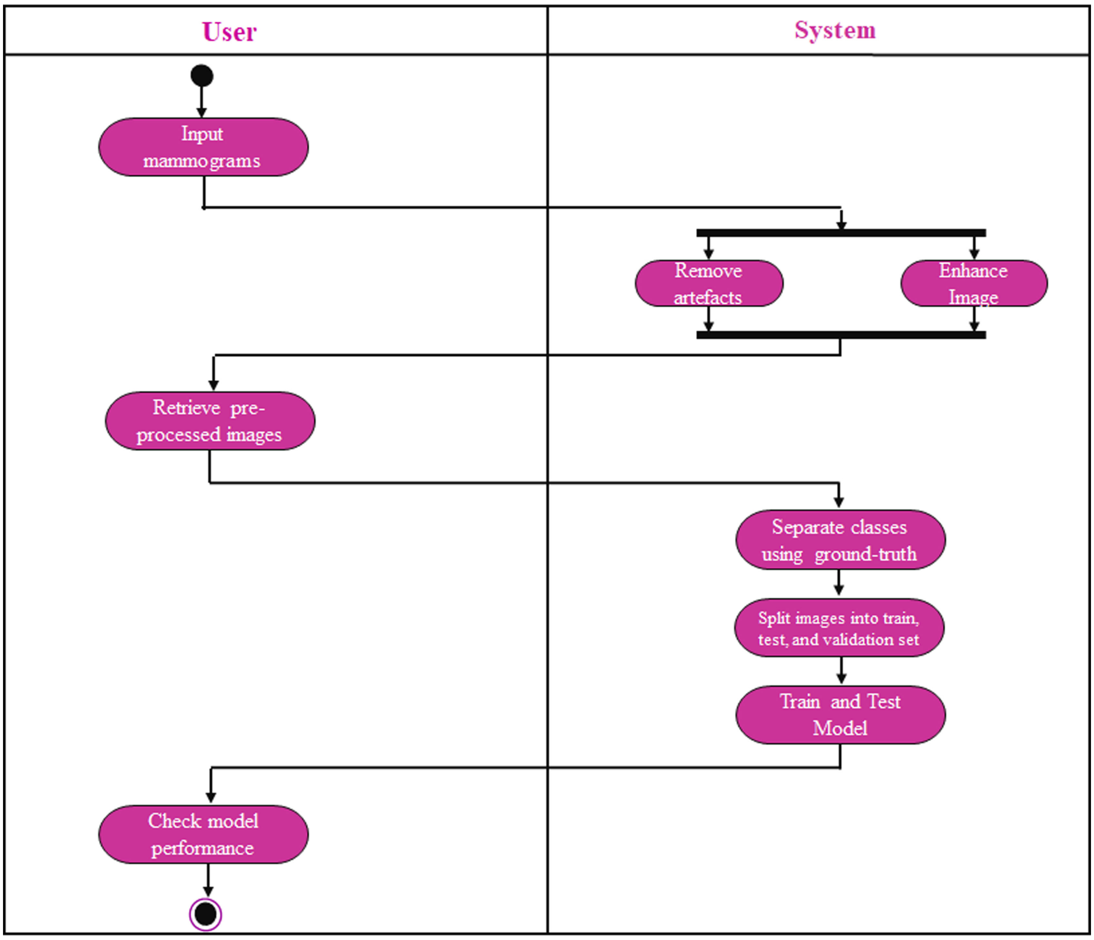


Fig. 3.2: Activity diagram of proposed model

3.1. Material and Methods.

3.1.1. Dataset. An organization of the UK research community, The Mammographic Image Analysis Society (MIAS [22]), has developed a mammogram repository. MIAS is one of the old benchmark mammogram datasets. Due to the dataset's widespread use among academics and researchers studying breast cancer, we employed MIAS in our research. Moreover, in comparison to other datasets like DDSM [27], CBIS-DDSM [28], and INbreast [29], which have very high-resolution images and are also accessible with variable sizes, all the images in the MIAS are the same size, 1024×1024, which will minimize the processing time to a higher extent. The dataset comprises 322 images categorized into three classes; Normal, Benign, and Malignant. Dataset also presents abnormality types such as microcalcifications, architectural distortion, circumscribed masses, speculated masses, ill-defined masses, and bilateral asymmetries. Details of all the images are given in a separate file. Ground truth is available as x, y image coordinates of the center of abnormality and approximate radius (in pixels) of a circle enclosing the abnormality. Table 3.1 shows classification of abnormalities of MIAS dataset.

3.1.2. Mammogram Pre-processing.
Why preprocessing?: Pre-processing is the first and essential step for any detection or classification process. It improves the quality of input data and makes it easier for the classifier to learn patterns and features from the processed information.

Table 3.1: MIAS Dataset

Types of Abnormaly	**Total Images**
Spiculated Mass	19
Circumscribed Mass	24
Microcalcification	25
Breast Asymmetry	15
Architectural distortion	19
Ill-defined mass	15
Class of Abnormality	**Total Images**
Benign	64
Malignant	51
Normal	207

Common problems with raw mammography images:

- Mammograms come with floating artefacts in the background. The other noise observed in mammograms is low and high-intensity labels, markers, and other tape artefacts.
- Some mammograms come with extra boundaries, which may create ambiguous features in the mammogram that the classifier might learn.
- Mammograms may have poor contrast. Small-sized lesions or calcifications may be obscured by the poor contrast and complex structure of density variations in the breast tissues.
- Low contrast abnormal regions may cause marginal visual threshold between suspicious and normal breast tissues.

Figure 3.3 shows a labeled mammogram with artefacts and other breast tissue.

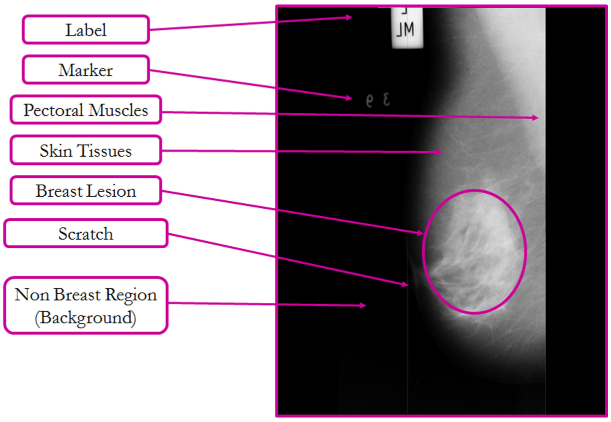


Fig. 3.3: Mammogram with Artefacts and Other Breast Tissues

Artefacts removal from mammograms: Lots of techniques are used to remove artefacts from mammograms. More broadly, these techniques are categorized as Region Segmentation and Edge-based Segmentation. In Edge-based Segmentation, edge pixels are detected and linked to form a contour. Edges do not completely enclose the entire object; it also needs post-processing to link all the detected edges that belong to the same boundary. As compared to edges, regions cover more image pixels; hence, we can have more detailed information to characterize those regions. Further, this region segmentation can be categorized as region-based and thresholding (also known as pixel-based segmentation). Region-based segmentation is an iterative and computationally expensive task. Hence, we have used thresholding transformations to remove artefacts, which is computationally very fast and simple [23],[24]. These methods achieve better results for images with uniform intensity and contrasting backgrounds. We applied two methods of thresholding; Global and OTSU [25]. Global thresholding is manual. This method selects an initial threshold value T to create clusters of classes for extracting the object from the image background. For a given image $f(a,b)$, the thresholded image can be defined by the following equation,

$$G(a,b) = \left\{ \begin{array}{ll} 1, & if(a,b) > T \\ 0, & if(a,b) \leq T \end{array} \right\} \tag{3.1}$$

The output of thresholding transformation is a binary image with a pixel value of 1 (object) and 0 (background).

OTSU is auto thresholding. It separates two clusters by the threshold calculated as an outcome of minimization of the weighted variance of those classes, which can be defined as $\sigma_w^2(t)$. The following equation can describe this computation.

$$\sigma_b^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t) \tag{3.2}$$

Here $w_1(t)$ and $w_2(t)$ are the class probabilities divided by the threshold T. The initial threshold can be defined in varieties of ways. As per [25] there are two ways; minimize the intra-class variance, $\sigma_w^2(t)$, and maximize the inter-class variance,

$$\sigma_b^2(t) = w_1(t)w_2(t)[\mu_1(t) - \mu_1(t)]^2 \tag{3.3}$$

where μ_i is a mean of class. The methodology adopted to get artefacts free mammograms is shown in algorithm 1.

Algorithm 1: Algorithm 1

Input: Mammograms
Output: Artefacts free mammograms

```
for all mammograms do
    Convert image to grayscale and median blur
    Function Thresholding Transformation(Grayscaled and Blurred Image):
        Apply thresholding transformation and get the threshold
        Get a binary image with the threshold
        return Binary Image

    Function Morphological Operations(Binary Image):
        Morph close to get a smooth image
        Dilate to enhance contour
        return Smooth Image

    Function Generating Mask(Smooth Image):
        Extract convex regions
        Find the largest convex region and eliminate other components; this largest object is breast
          contour
        Generate mask using this largest convex region
        return Binary Mask

    Perform bitwise AND operation between input mammogram and generated mask
```

Why should pectoral muscle not be removed while pre-processing mammograms?: In a mammogram, the pectoral muscle looks like a triangular opacity across the upper posterior border of the breast region. Detection and removal of pectoral muscle is per se a research theme with many publications. However, removing the pectoral muscle is medically incorrect as this region might have an abnormality. This argument is supported by figure 3.4 labeled with pectoral muscle and associated abnormality.

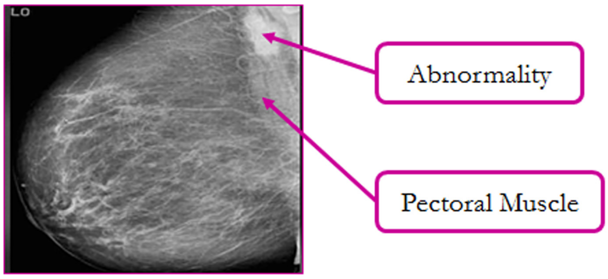


Fig. 3.4: Mammogram with Abnormality in Pectoral Muscle

Mammogram Enhancement: Image enhancement is an essential pre-processing step that allows changing the visual features of an image. A lack of good contrast in an image puts lots of strain on the eyes to get proper focus [9]. Various operations like Histogram Equalization, Image smoothing, Image sharpening, and Greyscale Multiplication are used to improve the contrast of an image. In our work, we used histogram-based methods to get enhanced mammograms. We applied both global histogram and CLAHE on mammograms and recorded the results.

3.1.3. Breast Abnormality Classification. In the last several decades, deep convolution neural networks (DCNNs)-based computer-aided diagnostic (CAD) systems for breast cancer have operated as a decision-making system [32]. Many pre-trained deep CNNs have been utilized in the literature to diagnose breast cancer. All these models, along with their parameter-based comparison, strength, and limitations, are covered in one of our prior works [26]. We used VGG16 as a base architecture to categorize breast anomalies as benign or malignant based on comparisons from our earlier research. The model was first proposed by Simonyan, and Zisserman [14] in 2014. The authors have made this model uniform by using the same kernels throughout the network, consisting of 13 convolutional layers, each with 3×3 filters. The model has 2×2 max pooling layers following every convolutional layer. It is practically demonstrated in a model that multiple small-sized filters can perform better than a single large kernel. This increased network depth can learn more intricate features nicely. To further explore the efficacy of transfer representation learning, we utilized VGG16 in two ways; implementation from scratch and transfer learning. To implement from scratch, we used five blocks (convolutional followed by max pool layer) of VGG16, two fully connected layers of 4096 units followed by one dense softmax layer with two units (to classify mammogram as benign or malignant). RELU activation is used for the dense layer with 4096 units to add non-linearity to the network. The model was trained with 100 epochs and optimized with an Adam optimizer using a learning rate of 0.001. The ”ImageDataGenerator” class of Keras pre-processing is used for loading data. We also used Transfer Learning on VGG16 to deal with a small-sized training set and used pre-trained weights of VGG16 (trained on ImageNet). We created a base model and populated it with pre-trained weights. All the layers in the base model are then frozen by setting trainable = False. A new model is then created on top of the output of the base model. Figure 3.5 presents a process of TL used in our work. We show hyperparameters used to train the models in table 3.2.

The dataset we used (MIAS) has all the images of size 1024×1024. As per the input layer of VGG16, these images are resized to 224×224 before training. We also used the early stopping method from Keras. Early stopping is essential to stop the model's training if there is no improvement in the learning or no increment in the parameters. Therefore, we monitored validation accuracy to do early stopping.

4. Results and Discussion. We used two methods to remove artefacts from mammograms and compared their results; Global thresholding and OTSU Binarization. OTSU binarization is auto thresholding and is

Table 3.2: Hyper Parameters for Training

Hyper Parameters for Model Training	
Epoch	100
Activation Function	RELU
Learning Rate	0.001
Batch Size	64
Optimizer	Adam
Loss Function	Binary Cross Entropy

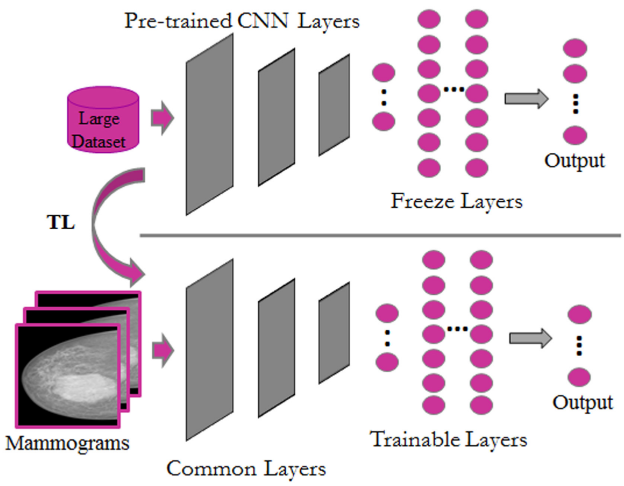


Fig. 3.5: Transfer Learning based training process

usually operated on grayscale images. The technique works by minimizing within-class variance and maximizing between-class variance. OTSU finds a threshold that keeps class clusters as tight as possible. In addition, it tries to reduce the overlap between two class clusters. Figure 4.1 presents the results obtained by OTSU thresholding as per the steps mentioned in the algorithm 1.

Global thresholding is manual, where the initial threshold (T) is selected; it then segments the image using the specified threshold and creates two-pixel groups. Finally, the average of pixels in both groups is computed to find a new threshold. This process repeats till there is no change in the threshold value. Figure 4.2 presents results obtained by Global thresholding as per the steps mentioned in the algorithm 1.

We have compared (see fig. 4.1 and 4.2) these two methods for two images of the MIAS dataset (mdb163 and mdb174) to showcase both the worst case and best case. As can be seen from image 4.1, there is a lot of pixel loss in the final processed image (mdb163) when applied OTSU binarization. On the other hand, from figure 4.2, We can see that global thresholding performs better as far as it concerns pixel loss. The detailed analysis of the thresholding technique is presented in table 4.1.

We applied histogram-based techniques to improve the quality of the mammogram. Figure 4.3 shows the performance comparison of the global histogram and CLAHE. Global histogram can enhance image contrast, but sometimes we may lose important details and features due to over brightness. The reason is that histograms do not restrict to a particular region. CLAHE can do better here; images are divided into small chunks known as "tiles ."Each of these tiles is then histogram equalized. With the help of contrast limiting, for any histogram bin above a specified contrast limit, those pixels will be clipped and scattered uniformly to other bins. Then, histogram equalization is applied. Finally, bilinear interpolation is used to remove artefacts in the borders of tiles. We noticed that over brightness due to global histogram for some images has resulted in the loss

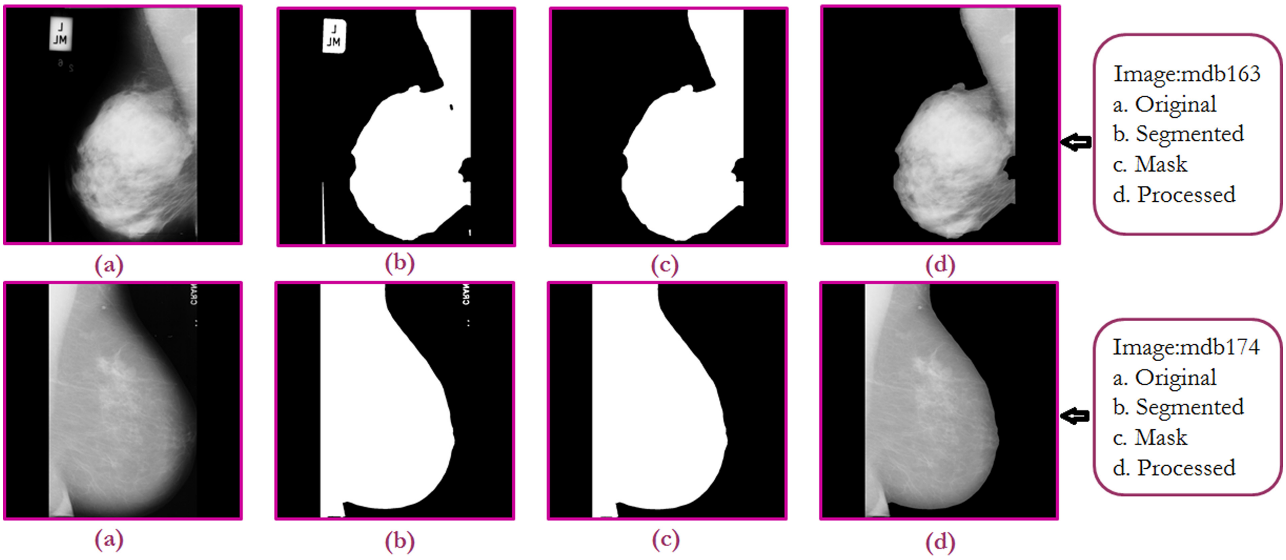


Fig. 4.1: Artefacts removal: OTSU Binarization

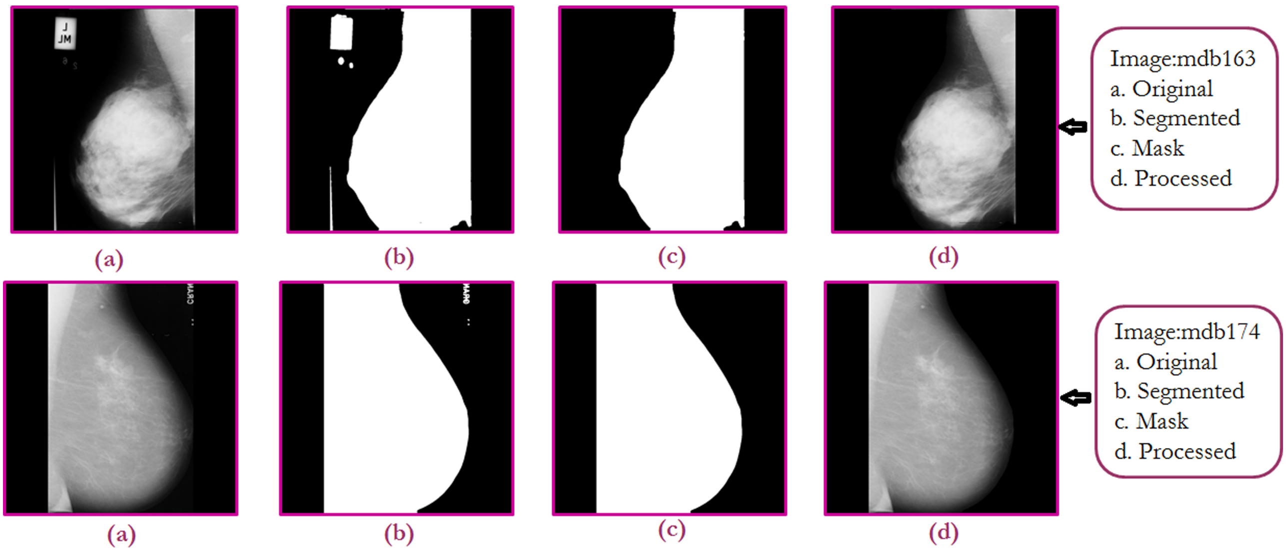


Fig. 4.2: Artefacts removal: Global Thresholding

of scattered benign calcifications. However, these calcifications were very well seen after applying CLAHE to enhance the mammogram.

Finally, these pre-processed mammograms are used to train the models. We have trained two models on the pre-processed MIAS dataset and recorded their results; the first model was built from scratch (based on VGG16 block structure), and the second model was built using Transfer Learning (Using pre-trained VGG-16). We evaluated the two models on the test dataset using precision, recall, sensitivity, specificity, and F1-score as performance indicators. We also provide validation and training accuracy and loss outcomes to clearly understand the model performance. The F1 score reflects the classifier's performance better than the traditional

Table 4.1: Result Analysis: Thresholding Techniques

Method	**Threshold Value**	**Acceptable**	**Unacceptable**
	10	306	16
	15	314	8
	20	317	5
	26	318	4
	28	319	3
	30	321	1
Global	Exceptional Case : mdb274 (T value 69)		
Otsu	-	322	0

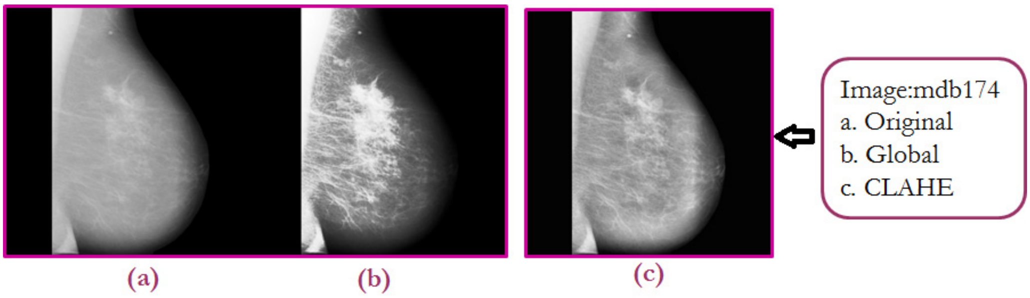


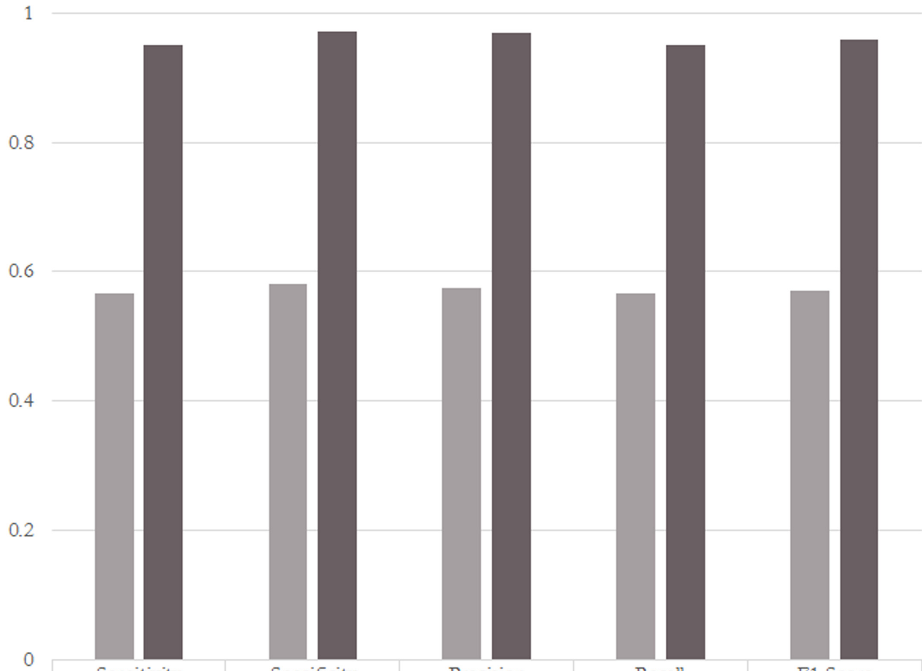
Fig. 4.3: Results: Mammogram Enhancements. (a) Artefact free mammogram (b) shows over brightness as compared to (c).

accuracy metric. We tried to examine the power of the transfer learning approach for medical imaging data over CNN architecture implemented from scratch. The results are presented in figure 4.4 and 4.5 respectively.

It can be seen that the use of transfer learning for the small-sized dataset can indeed help to improve the model performance and also help to mitigate the effect of overfitting. Table 4.2 shows a comparative analysis of methods available in the literature for pre-processing mammograms and abnormality classification with the proposed work. The result indicates that Transfer Learning improves accuracy rates because the model has already learned a lot and trained to find some patterns and features.

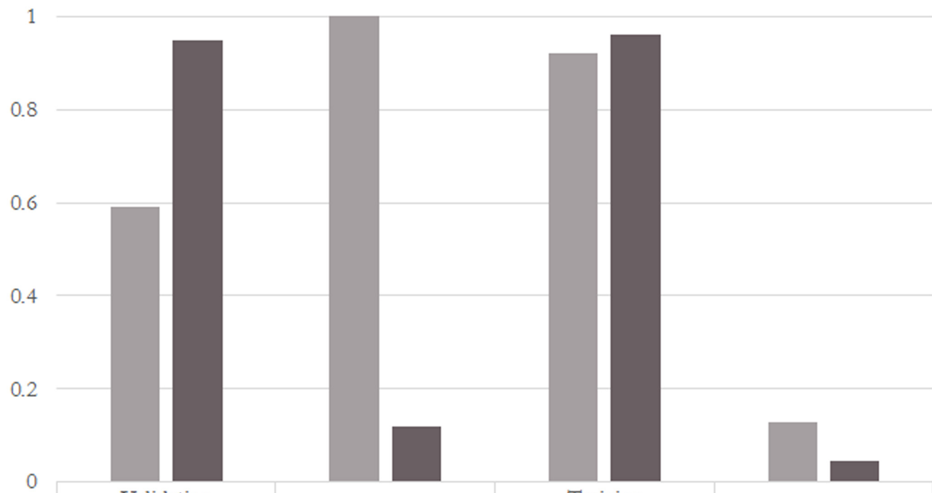
5. Conclusion. Artefacts may affect deep learning techniques' accuracy rates and inference knowledge capabilities. So, having artefact-free mammograms can improve the model's extendibility onto a wide range of datasets. We presented a methodology to pre-process images of MIAS to remove artefacts. First, two thresholding techniques are implemented for image binarization, and their results are compared. Global thresholding could do better if it concerns pixel loss in the image. Next, images of the datasets are enhanced using histogram-based methods. We implemented two techniques; global histogram and CLAHE. CLAHE could properly enhance the mammogram without obscuring small-sized lesions and benign calcifications. This work also presents a CNN-based abnormality classification from mammograms, where the abnormalities are classified as benign and malignant. The model was trained and tested on the pre-processed MIAS images. We used two versions of VGG16 to classify the mammograms, VGG16 trained from scratch and Transfer Learning. Due to a small-sized training set, the first model started to overfit after certain epochs. We achieved better results using Transfer Learning for the performance measures like precision (96.83%), sensitivity (95.11%), specificity (97.09%), recall (95.11%), F1-score (95.96%), and accuracy (96.14%).

Acknowledgements. The authors like to thank Dr. Rajiv Oza (Consultant Radiologist) for providing them with medical advice on radiological imaging techniques.



	Sensitivity	Specificity	Precision	Recall	F1-Score
VGG 16 (From Scratch)	0.5667	0.5812	0.5737	0.5667	0.5701
VGG 16 (TF)	0.9511	0.9709	0.9683	0.9511	0.9596

Fig. 4.4: Model comparison based on various Performance Measures



	Validation Accuracy	Validation Loss	Training Accuracy	Training Loss
VGG-16 (From Scratch)	0.59	1.69	0.92	0.13
VGG-16 (TF)	0.9469	0.1189	0.9614	0.047

Fig. 4.5: Accuracy and Loss measures of the models

Table 4.2: Comparison with other Work

Work	Pre-processing Technique	Model	Task Performed	Dataset	Training Accuracy
[6]	Adaptive histogram equalization	VGG 16	Mass Classification	MIAS	48.67%
[6]	Adaptive histogram equalization	Proposed CNN	Mass Classification	MIAS	92.54%
[17]	Filter Techniques and Morphological Operations	VGG 16	Mass Classification	MIAS	69.85%
[17]	Filter Techniques and Morphological Operations	VGG 16	Mass Classification	MIAS (Augmented)	94%
Our Work (1)	Thresholding, CLAHE and Morphological Operation	VGG 16 (From Scratch)	Mass Classification	MIAS	92%
Our Work (2)	Thresholding, CLAHE and Morphological Operation	VGG 16 (Transfer Learning)	Mass Classification	MIAS	96.14%

REFERENCES

[1] Bray, Freddie, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 68.6 (2018): 394-424.

[2] Viale, Pamela Hallquist. The American Cancer Society's facts & figures: 2020 edition. Journal of the Advanced Practitioner in Oncology. 11.2 (2020): 135.

[3] Mathur, Prashant, et al. Cancer statistics, 2020: report from national cancer registry programme, India. JCO Global Oncology. 6 (2020): 1063-1075.

[4] Kunnavil, Radhika, et al. Estimation of burden of female breast cancer in India for the year 2016, 2021 and 2026 using disability adjusted life years. Int J Community Med Public Health. 3 (2016): 1135-40.

[5] Oza, Parita, et al. A bottom-up review of image analysis methods for suspicious region detection in mammograms. Journal of Imaging. 7.9 (2021): 190.

[6] Gnanasekaran, Vaira Suganthi, et al. Deep learning algorithm for breast masses classification in mammograms. IET Image Processing. 14.12 (2020): 2860-2868.

[7] Bhateja, Vikrant, Mukul Misra, and Shabana Urooj. Non-Linear Enhancement Techniques for Mammograms.Non-Linear Filters for Mammogram Enhancement. Springer, Singapore, 2020. 55-62.

[8] Chaloeykitti, L., M. Muttarak, and K. H. Ng. Artifacts in mammography: ways to identify and overcome them. Singapore medical journal. 47.7 (2006): 634.

[9] Gu, Ke, et al. The analysis of image contrast: From quality assessment to automatic enhancement. IEEE transactions on cybernetics. 46.1 (2015): 284-297.

[10] Moradmand, Hajar, et al. Comparing the performance of image enhancement methods to detect microcalcification clusters in digital mammography. Iranian journal of cancer prevention. 5.2 (2012): 61.

[11] Pillai, Rohan, Parita Oza, and Priyanka Sharma. Review of machine learning techniques in health care. Proceedings of ICRIC 2019. Springer, Cham, 2020. 103-111.

[12] Oza, Parita, Paawan Sharma, and Samir Patel. Machine learning applications for computer-aided medical diagnostics. Proceedings of Second International Conference on Computing, Communications, and Cyber-Security. Springer, Singapore, 2021.

[13] Oza, Parita, Yash Shah, and Marsha Vegda. A Comprehensive Study of Mammogram Classification Techniques. Tracking and Preventing Diseases with Artificial Intelligence. Springer, Cham, 2022. 217-238.

[14] Simonyan, Karen, and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556 (2014).

[15] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 25 (2012).

[16] He, Kaiming, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[17] Dhivya, S., et al. GAN based Data Augmentation for Enhanced Tumor Classification. 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP). IEEE, 2020.

[18] Tripathy, Sushreeta, and Tripti Swarnkar. Unified preprocessing and enhancement technique for mammogram images. Procedia Computer Science. 167 (2020): 285-292.

[19] Arefan, D., et al. Automatic breast density classification using neural network. Journal of Instrumentation. 10.12 (2015): T12002.

[20] Li, Hua, et al. Benign and malignant classification of mammogram images based on deep learning. Biomedical Signal

Processing and Control. 51 (2019): 347-354.

[21] El Houby, Enas MF, and Nisreen IR Yassin. Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks. Biomedical Signal Processing and Control. 70 (2021): 102954.

[22] Suckling, J., et al. Mammographic Image Analysis Society (MIAS) database.v1.21 [Dataset]. (2015) https://www.repository.cam.ac.uk/handle/1810/250394

[23] Bhateja, Vikrant, Mukul Misra, and Shabana Urooj. Non-Linear Enhancement Techniques for Mammograms. Non-Linear Filters for Mammogram Enhancement. Springer, Singapore, 2020. 55-62.

[24] Al-Bayati, Moumena, and Ali El-Zaart. Mammogram images thresholding for breast cancer detection using different thresholding methods. Scientific Research Publishing. (2013).

[25] Otsu, Nobuyuki. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 9.1 (1979): 62-66.

[26] Oza, Parita, et al. Deep convolutional neural networks for computer-aided breast cancer diagnostic: a survey. Neural Computing and Applications. (2022): 1-22.

[27] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, “Current status of the digital database for screening mammography,” in Digital mammography, pp. 457–460 , Springer, 1998.

[28] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, “A curated mammography data set for use in computer-aided detection and diagnosis research,” Scientific data, vol. 4, no. 1, pp. 1–9, 2017.

[29] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, “Inbreast: toward a full-field digital mammographic database,” Academic radiology, vol. 19, no. 2, pp. 236–248, 2012.

[30] Leppioja, Markus, Pasi Luukka, and Christoph Lohrmann. ”Image based classification of shipments using transfer learning.” In KNOWCON, pp. 37-44. 2021.

[31] Duong-Trung, Nghia, Dung Ngoc Le Ha, and Hiep Xuan Huynh. ”Classification-Segmentation Pipeline for MRI via Transfer Learning and Residual Networks.” In RICE, pp. 39-43. 2021.

[32] Oza, P., Sharma, P., Patel, S. (2022). A Drive Through Computer-Aided Diagnosis of Breast Cancer: A Comprehensive Study of Clinical and Technical Aspects. In:Recent Innovations in Computing. Lecture Notes in Electrical Engineering, vol 832. Springer, Singapore. https://doi.org/10.1007/978-981-16-8248-3_19

[33] Li, Ming, Qingbo Yin, and Mingyu Lu. ”Retinal blood vessel segmentation based on multi-scale deep learning.” In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1-7. IEEE, 2018.

[34] Oza, P.; Sharma, P.; Patel, S.; Adedoyin, F.; Bruno, A. Image Augmentation Techniques for Mammogram Analysis, J. Imaging 2022, 8, 141. https://doi.org/10.3390/jimaging8050141

Edited by: Luciano Caroprese
Received: Apr 5, 2022
Accepted: Aug 31, 2022

AIMS AND SCOPE

The area of scalable computing has matured and reached a point where new issues and trends require a professional forum. SCPE will provide this avenue by publishing original refereed papers that address the present as well as the future of parallel and distributed computing. The journal will focus on algorithm development, implementation and execution on real-world parallel architectures, and application of parallel and distributed computing to the solution of real-life problems. Of particular interest are:

Expressiveness:

- high level languages,
- object oriented techniques,
- compiler technology for parallel computing,
- implementation techniques and their efficiency.

System engineering:

- programming environments,
- debugging tools,
- software libraries.

Performance:

- performance measurement: metrics, evaluation, visualization,
- performance improvement: resource allocation and scheduling, I/O, network throughput.

Applications:

- database,
- control systems,
- embedded systems,
- fault tolerance,
- industrial and business,
- real-time,
- scientific computing,
- visualization.

Future:

- limitations of current approaches,
- engineering trends and their consequences,
- novel parallel architectures.

Taking into account the extremely rapid pace of changes in the field SCPE is committed to fast turnaround of papers and a short publication time of accepted papers.

INSTRUCTIONS FOR CONTRIBUTORS

Proposals of Special Issues should be submitted to the editor-in-chief.

The language of the journal is English. SCPE publishes three categories of papers: overview papers, research papers and short communications. Electronic submissions are preferred. Overview papers and short communications should be submitted to the editor-in-chief. Research papers should be submitted to the editor whose research interests match the subject of the paper most closely. The list of editors' research interests can be found at the journal WWW site (`http://www.scpe.org`). Each paper appropriate to the journal will be refereed by a minimum of two referees.

There is no a priori limit on the length of overview papers. Research papers should be limited to approximately 20 pages, while short communications should not exceed 5 pages. A 50–100 word abstract should be included.

Upon acceptance the authors will be asked to transfer copyright of the article to the publisher. The authors will be required to prepare the text in LaTeX 2_ε using the journal document class file (based on the SIAM's `siamltex.clo` document class, available at the journal WWW site). Figures must be prepared in encapsulated PostScript and appropriately incorporated into the text. The bibliography should be formatted using the SIAM convention. Detailed instructions for the Authors are available on the SCPE WWW site at `http://www.scpe.org`.

Contributions are accepted for review on the understanding that the same work has not been published and that it is not being considered for publication elsewhere. Technical reports can be submitted. Substantially revised versions of papers published in not easily accessible conference proceedings can also be submitted. The editor-in-chief should be notified at the time of submission and the author is responsible for obtaining the necessary copyright releases for all copyrighted material.