Prediction Of Breast Cancer in Mammography Images Using Quantum-Based Swarm Intelligence Clustering Techniques and Deep Learning Approaches

Akshaya Kumar Dash[1, 2], Nibedita Adhikari[1], Chittaranjan Mohapatra[1]

[1]Utkal University, Vanivihar, Bhubaneswar, Odisha, India, 751004
[2]Silicon Institute of Technology, Patia, Bhubaneswar, Odisha, India, 751024.

Abstract. Mammography is the most efficient approach for breast cancer (BC) screening. However, mammogram-based breast biopsy's limited positive predictive value results in around 70% of needless biopsies with benign outcomes. Several computer-aided diagnostic (CAD) solutions have been proposed in recent years to help reduce the high rate of unneeded breast biopsies. The purpose of the research is to define a fuzzy c-means (FCM) approach based on quantum-inspired particle swarm optimisation (QPSO) for clustering multidimensional data. The specified QPSO technique is used to establish cluster centres for a dataset to address this disadvantage. Pre-processing, clustering, feature extraction, and classification are the four stages of this process. Initially, z-score normalisation is used to perform pre-processing. Then, using the QPSO-based FCM approach, the cluster centres for a dataset are generated. Additionally, the principal component analysis (PCA) technique is suggested for extracting features from the given data. Finally, a deep learning based classifier is presented to accurately predict BC images. The proposed scheme is compared to several well-established methods. Experiments demonstrate that the proposed quantum technique has an 87.72% F-measure and 89.59% accuracy compared to other approaches currently in use.
Keywords: Breast Cancer, mammography, z-score normalization, QPSO, Fuzzy C-Means, Principal Component Analysis, Deep LearningI First Section

1. Introduction

BC has evolved to become one of the leading reasons for the demise of women worldwide during the last several decades [1]. Despite the increased incidence of BC in women, the survival rate of BC patients remains high five years following diagnosis due to effective treatment and diagnostics [2]. Thus, the data demonstrate that early detection and treatment of this disease can significantly reduce global mortality rates. As a result, it is essential to recognise this life-threatening condition early [3]. Predicting breast cancer is a significant clinical issue in this disease. Thus, the primary focus should be on understanding the kind of BC, as this is when most breast cancer states occur [4]. Therefore, this technique may aid in predicting the type of BC that patients may have, thereby saving substantial time and perhaps assisting in significantly lowering the mortality rate.

Breast density cannot be used as a substitute for the brief data encoded [5] in the mammogram [6]. Whereas previous research examined automated methods for assessing

breast density, these efforts reduced the mammographic input to a few statistics primarily related to glandular tissue volume, which are insufficient to distinguish patients at risk of developing breast cancer from those who will not [7].

The paper is organised as follows. The Section 2 discusses recent methodologies on breast cancer classification. The Section 3 describes the developed methods and algorithms. The Section 4 covers the simulation approach and findings. Finally, Section 5 summarises the conclusions of the research..

2. Literature Review

The advantages and disadvantages of some recent techniques for predicting breast cancer are discussed. Amrane et al. [8] compareed the Naive Bayes (NB) classifier and the k nearest neighbour (KNN) classifier for breast cancer classification. The results indicated that KNN classifiers achieve the maximum accuracy (97.51%) with the lowest error rate, followed by NB classifiers (96.19%). The Support Vector Machine (SVM) and KNNs method were used in Islam et al. [9]. These approaches obtained an accuracy of 98.57% and 97.14% respectively.

Golagani et al. [10] designed a breast cancer data set encompassing the concentrations of biomolecules in breast cancer cells as determined by mass spectrometry. This explored a variety of characteristic collection and filtering techniques. This incorporated optimal features subcategories into multiple learning categorization methods, such as KNN, PNN (probabilistic neural network), and SVM classifiers. The SVM sorter achieved the maximum classification efficiency for diagnosing chest cancer with 99.17% connecting diameter and durability attributes.

Rathi et al. [11] developed a machine learning technique for predicting the target class using a set of attributes. This technique determined the type of cancer and forecasted the likelihood of cancer developing in near future. Kaur et al. [12] enhanced the weight-based algorithm for classifying usual and diseased cells. The weight-based classification process is time consuming. This is enhanced by using a Bayesian classifier using mass, density, and margin features.

Ahmad et al. [13] devised a genetic algorithm-based multi-objective optimisation of an artificial neural network classifier, called GA-MOO-NN, for the automated detection of BC. The algorithm had an average classification accuracy of 98.85% and 98.10%, respectively. AttyaLafta et al. [14] defined an approach for recognising cancer that relies on a hybrid neurogenetic system to categorise a well-known WBCD dataset. The system's accuracy has been verified to be 100% in the best trial and greater than 97% in another research.

3. Proposed Methodology

The proposed architecture includes Pre-processing, Clustering, Feature Extraction, and Classification units as shown in Figure 1. This section detailed the units from the input to the predicted result

3.1. Pre-processing

Data should be unambiguous, accurate, and complete, as classification performance is highly dependent on data quality. Pre-processing minimises inconsistencies in the data set and fills in missing values. However, the data set compiled from many sources,

contains redundant and unnecessary information. Hence the Z-score approach is used to normalize the given dataset.

Z-Score Normalization: Each experiment's raw intensity data were normalized by first computing the average intensity for each dataset and then computing the average of the averages. The average of all normalised data equalled the grand average [16]. On a standard distribution curve, a z-score can be plotted. Z-scores vary from -3 up to +3 standard deviations (SDs).

Concretely, let xi (i = 1, 2, · · ·, D) indicate the i-th component of each feature vector x ∈ R D. We first compute the mean and the SD of these D components:

$$\mu_x = \frac{1}{D}\sum_{i=1}^{D} x_i, \sigma_x = \sqrt{\frac{1}{D}\sum_{i=1}^{D}(x_i - \mu_x)^2} \quad (1)$$

Z-score normalization is then applied as

$$x^{(zn)} = ZN(x) = \frac{x-\mu_{x1}}{\sigma_x} \quad (2)$$

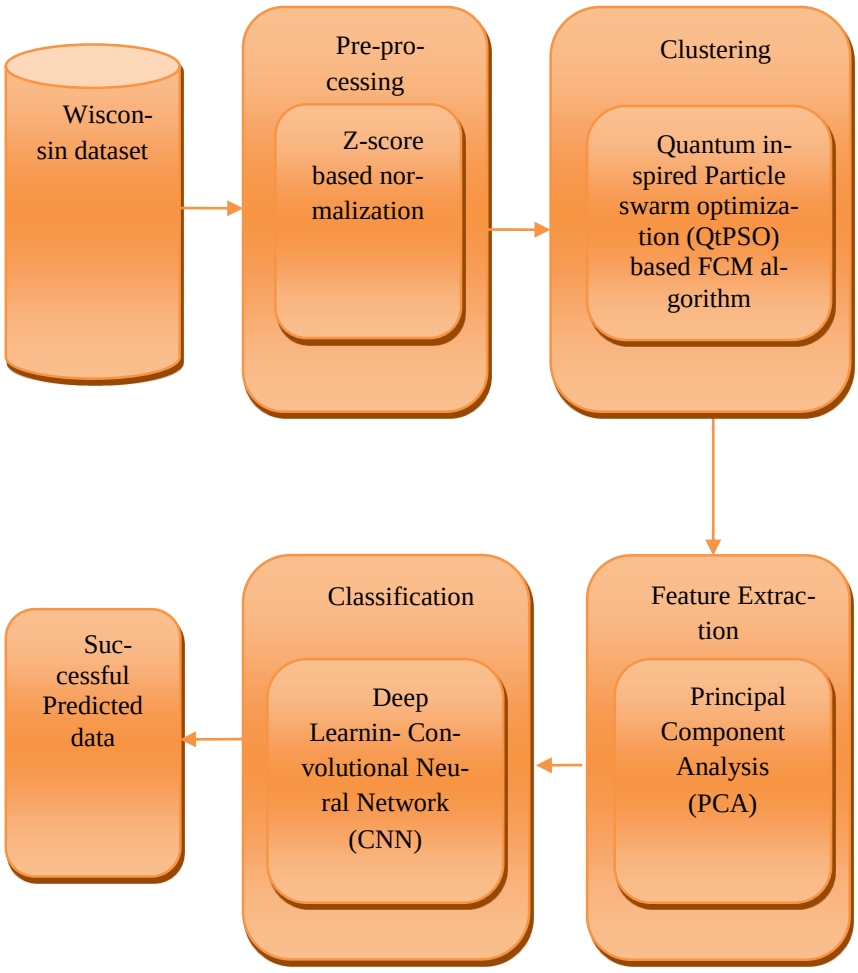


Fig.1. The overall process of the proposed methodology

From these calculations, z-score normalisation first projects the original feature vectors along the 1 vector to a hyperplane that is perpendicular to √1. These vectors are then scaled to the same length as D, i.e., the final normalised vectors lie on a hypersphere with a radius of √D. After pre-processing the given data, the feature selection process is carried out, which is described in the below section.

3.2. Clustering

A modification of PSO-based c-means clustering approach [18] relying on qubits is presented in this section. This technique creates optimal cluster centres from a given dataset. The created optimum cluster centres are utilised as the cluster centres for the FCM approach, which clusters the dataset effectively. Fuzzy C-Means (FCM) is a fuzzy clustering algorithm that allows data points to have partial membership in multiple clusters, rather than being exclusively assigned to one cluster. FCM assigns fuzzy membership values to data points, which represent the degree of belongingness of each data point to each cluster.

It iteratively updates the cluster centroids and membership values based on minimizing a fuzzy objective function, considering both the distances between data points and cluster centroids as well as the membership values.

FCM provides more flexibility in clustering, particularly when data points have ambiguous membership or when they lie near the boundaries of multiple clusters.

Modified QPSO (MQPSO) and the FCM Approach: This original version of PSO created a new inertia weight factor (ω) in this study [20]. The velocity of particle i denoted as $V_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$ where D is the dimension of the vector. The location of particle i denoted as $(x, x_{i2}, \ldots, x_{iD})$, the optimal location of particle i denoted as $p_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$, it is also called p_{best}. The optimum global location of all particles $p_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$, it is also referred to as g_{best}. The upgrade equation of velocity is shown in Equation (3) and (4):

$$v_{id} = w \times v_{id} + c_1 \times rand() \times (p_{id} - x_{id}) + c_2 \times Rand() \times (p_{gd} - x_{id}) \quad (3)$$

$$(X_{id} = x_{id} + v_{id}) \quad (4)$$

The steps of the MQPSO is defined as follows.

A Initialization of population

I. Let α is a matrix of qubit-based particles of dimension is q × n × m, here m denotes population's particle count. After that, β is generated by Equation 3.

II. A new matrix, Y, of dimension q×n×m is generated With the random integers between [0,1]

III. δ is a binary coded matrix

$$\delta_{ij} = \{1, if Y_{ij} < \alpha_{ij}^2\ 0\ otherwise \quad (5)$$

where (i, j) specifies the population's position.

IV. By Equation (3),

$$\rho_{ij} = \{\alpha_{ij}, if\ \delta_{ij} = 1\ \beta_{ij} if\ \delta_{ij} = 0 \quad (6)$$

V. Assume the best particle for each individual particle and preserve it in the appropriate Tik.

VI. ρ is applied to identify the fitness value of every particle.

B For every Generation

I. Upgrade the swarm and save in α

II. β is upgraded, by Equation (6).

III. Repeat step 1.4 to 1.6 For every particle

Detect its fitness level by the Beni and I indices.

If the present value is superior than Tik. A) Preserve the present location in Tik.

Otherwise, no modification in Tik is allowed.

Preserve the location of the particle holding the optimal fitness value in Tgk.

ωPDD, ce1 and ce2 are updated.

The velocity and location of every particle is upgraded by Formulae (5) and (6).

C The particle having optimal fitness value is monitored.

D If the ending condition is fulfilled a) Terminate process.

E Else, return to Step 2.

F Post finishing of the suggested approach, the global optimized cluster centres are acquired.

G The acquired centres are employed as the input.

3.3. Feature Extraction

The third unit extracts valuable features from the given data set and reduces the data set's dimensionality. Determining a subset of the components or attributes that are input to the system is called Feature Selection. PCA (Principal Component Analysis) are composed of the original attributes in a linear combination and they must be orthogonal to one another [21]. They must preserve the most considerable variation in the data. PCA transforms the input into a lower-dimensional characteristics region by the correlation matrix's biggest eigenvectors

3.4. Deep learning-based classification - Convolutional Neural Network (CNN)

Deep learning-based classification using Convolutional Neural Networks (CNNs) has become one of the most popular and powerful techniques for image and video classification tasks. CNNs are specifically designed to handle visual data and have been highly successful in various computer vision tasks, such as image classification, object detection, segmentation. CNNs are a class of deep learning models inspired by the visual processing in the human brain. They consist of multiple layers, including convolutional layers, activation functions, pooling layers, and fully connected layers.

Convolutional layers are the core components of CNNs. They use learnable filters (also called kernels) to scan over the input mammography images and extract local features. CNN is a feedforward neural network that uses convolution, and pooling layers [22]. The essential characteristics are the following:

1. Kernel size (N): Each kernel has a specified window size. The convolution is performed on a region corresponding to its size in the input and display the outcomes in its activation map.
2. Stride (S): It specifies how many pixels the kernel will relocate for the following location. When s8et to 1, each kernel performs convolution process throughout the input volume and then shifts one pixel at a time until it hits the input's selected boundary.
3. Zero-Padding (P): It specifies the number of zeros to pad the input's boundary. Weights shared and local connectivity dramatically reduces the network's overall parameter count.

4. Results and Discussion

The outcomes achieved using the proposed method are compared to those acquired using previously proposed algorithms for breast cancer detection. This new architecture demonstrates promising findings for the medical field and generates confidence for earlier and more efficient diagnoses of such patients. Apart from classification accuracy, some statistical measurements specified in equations (7)– (10) are used to evaluate the classifier, as are the average results for the classifiers.

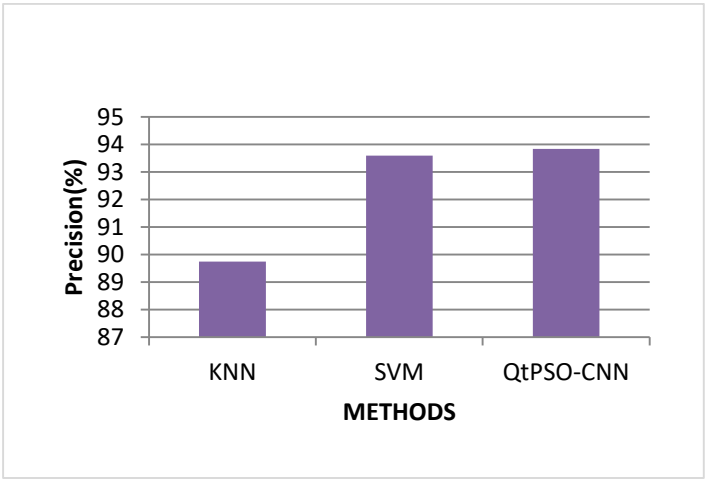
$$\text{Precision} = TP / TP + FP \quad (7)$$

$$\text{Recall} = TP/TP+FN \quad (8)$$

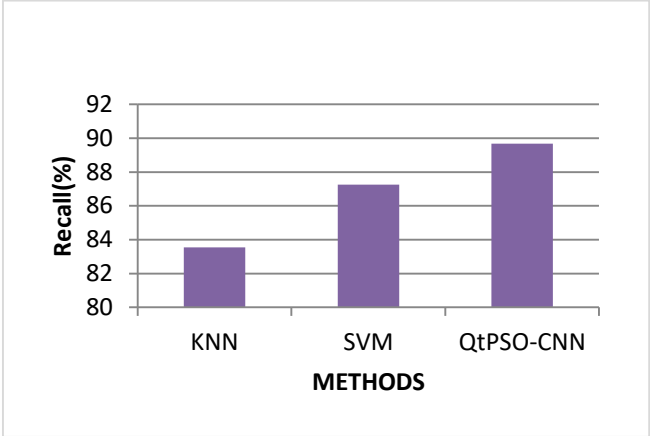
$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (9)$$

$$\text{Accuracy} = (TP+FP) / (TP+TN+FP+FN) \quad (10)$$

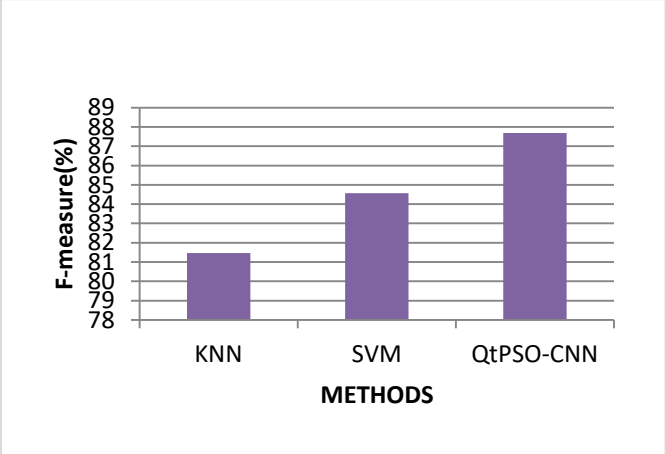
The ratio of accurately identified positive observations to all expected positive observations is precision as in Equation 7. The ratio of accurately identified positive observations to the number of observations is called sensitivity or recall as in Equation 8. The weighted rate of Precision and Recall is defined as the F – measure in Equation 9. Accuracy is measured using Positives and negatives in the following manner as per Equation 10.



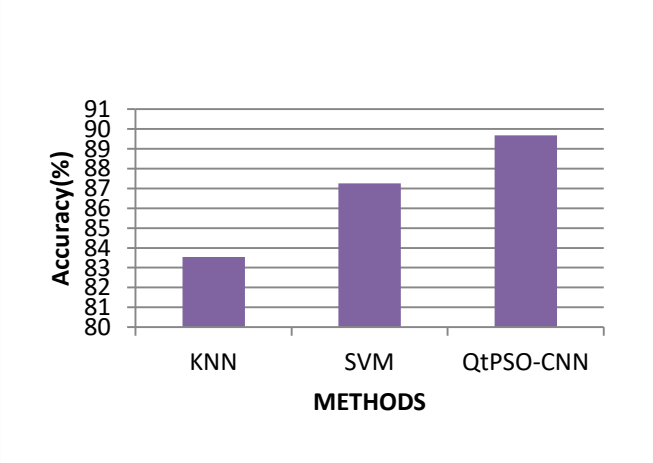
(a)


92
90
88
86
84
82
80
Recall(%)
KNN
SVM
QtPSO-CNN
METHODS

(b)


89
88
87
86
85
84
83
82
81
80
79
78
F-measure(%)
KNN
SVM
QtPSO-CNN
METHODS

(c)


91
90
89
88
87
86
85
84
83
82
81
80
Accuracy(%)
KNN
SVM
QtPSO-CNN
METHODS

(d)

Fig.2. (a) A precision comparison between the suggested QtPSO-CNN methodology and existing BC prediction methods. (b) Recall outcomes of the suggested QtPSO-CNN technique and traditional methods. (c) F-measure outcomes of the suggested QtPSO-CNN technique and traditional methods. (d) Accuracy outcomes of the suggested QtPSO-CNN technique and traditional methods

Figure 2 (a) illustrates the suggested QtPSO-performance CNNs in terms of precision comparisons. The proposed approach is a highly successful strategy for tackling the categorization issue. The performance results of the suggested QtPSO-CNN approach are shown in Figure 2 (b). Thus, the new method achieves a recall rate of 89.68%, while the existing techniques achieve 87.25% and 83.54%, respectively. As illustrated in Figure 2 (c), the suggested QtPSO integrated with the CNN classifier provides remarkable performance of 87.72% in terms of disease prediction rate, considerably above that of the SVM and KNN. The quantitative analysis's F-measure results are consistent with the qualitative analysis's machine learning methodologies. The proposed QtPSO-CNN is compared to traditional categorization approaches for BC prediction accuracy. As illustrated in Figure 2 (d), the proposed QtPSO-CNN-based classifiers outperform the existing classifier in terms of accuracy of 89.59%. The findings demonstrate that the proposed technique is exceptionally effective. As a result, the classifiers' accuracy will be greater than that of other classifiers constructed on previously developed models and new data equivalent to the entire dataset.

5. Conclusion

In conclusion, a deep learning approach that utilises full-field mammography in addition to conventional risk variables outperforms the competition. The purpose of this project is to develop a QPSO technique. FCM became jammed at local minima because of incorrect initial choices of cluster centres. The planned QtPSO technique is used to produce cluster centres for a dataset to address this disadvantage. The well-known PSO algorithm is used to assess the effectiveness of quantum computing. The results established that full-field pictures and traditional risk variables include complementing information, as indicated by the suggested deep learning model's performance increase with F-measure of 87.72% and accuracy of 89.59%. These techniques may eventually supplant conventional risk identification techniques. In future work, we will investigate which risk variables are absorbed by the image, with the expectation that this approach may be scaled to include large gene panels.

References

1. Losurdo, L., Fanizzi, A., Basile, T. M. A., Bellotti, R., Bottigli, U., Dentamaro, R., ... & La Forgia, D. (2019). Radiomics analysis on contrast-enhanced spectral mammography images for breast cancer diagnosis: A pilot study. Entropy, 21(11), 1110.
2. Hernández, A., Miranda, D., & Pertuz, S. (2021). Algorithms and Methods for Computerized Analysis of Mammography Images in Breast Cancer Risk Assessment. Computer Methods and Programs in Biomedicine, 106443.
3. Li, J., Szekely, L., Eriksson, L., Heddson, B., Sundbom, A., Czene, K., ... & Humphreys, K. (2012). High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. Breast Cancer Research, 14(4), 1-12.
4. Son, J., Lee, S. E., Kim, E. K., & Kim, S. (2020). Prediction of breast cancer molecular subtypes using radiomics signatures of synthetic mammography from digital breast tomosynthesis. Scientific reports, 10(1), 1-11.
5. Massafra, R., Bove, S., Lorusso, V., Biafora, A., Comes, M. C., Didonna, V., ... & La Forgia, D. (2021). Radiomic Feature Reduction Approach to Predict Breast Cancer by Contrast-Enhanced Spectral Mammography Images. Diagnostics, 11(4), 684.

6. Fanizzi, A., Losurdo, L., Basile, T. M. A., Bellotti, R., Bottigli, U., Delogu, P., ... & La Forgia, D. (2019). Fully automated support system for diagnosis of breast cancer in contrast-enhanced spectral mammography images. Journal of clinical medicine, 8(6), 891.
7. Yang, J., Wang, T., Yang, L., Wang, Y., Li, H., Zhou, X., ... & Huang, L. (2019). Preoperative prediction of axillary lymph node metastasis in breast cancer using mammography-based radiomics method. Scientific reports, 9(1), 1-11.
8. Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE.
9. Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2017, December). Prediction of breast cancer using support vector machine and K-Nearest neighbors. In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 226-229). IEEE.
10. Golagani, P. P., Mahalakshmi, T. S., & Beebi, S. K. (2021). Supervised Learning Breast Cancer Data Set Analysis in MATLAB Using Novel SVM Classifier. In Machine Intelligence and Soft Computing (pp. 255-263). Springer, Singapore.
11. Rathi, M., & Pareek, V. (2016). Hybrid approach to predict breast cancer using machine learning techniques. International Journal of Computer Science Engineering, 5(3), 125-136.
12. Kaur, N., & Sharma, A. S. (2018, September). Detection of breast cancer on mammograms using neural network approach. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 544-547). IEEE.
13. Ahmad, F., Isa, N. A. M., Hussain, Z., & Sulaiman, S. N. (2013). A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis. Neural Computing and Applications, 23(5), 1427-1435.
14. AttyaLafta, H., KdhimAyoob, N., & Hussein, A. A. (2017, March). Breast cancer diagnosis using genetic algorithm for training feed forward back propagation. In 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT) (pp. 144-149). IEEE.
15. Patro, P., Kumar, K., & Kumar, G. S. (2019). Various Classifiers Performance Based Machine Learning Methods. 3, 305–310. https://doi.org/10.35940/ijrte.C1069.1183S319
16. Fei, N., Gao, Y., Lu, Z., & Xiang, T. (2021). Z-Score Normalization, Hubness, and Few-Shot Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 142-151).
17. Patro, P., Kumar, K., & Suresh Kumar, G. (2020a). Neuro fuzzy system with hybrid ant colony particle swarm optimization (Haso) and robust activation. Journal of Advanced Research in Dynamical and Control Systems, 12(3 Special Issue), 741–750. https://doi.org/10.5373/JARDCS/V12SP3/20201312
18. Li, S. S., Long, G. L., Bai, F. S., Feng, S. L., & Zheng, H. Z. (2001). Quantum computing. Proceedings of the National Academy of Sciences, 98(21), 11847-11848.
19. Patro, P., Kumar, K., & Suresh Kumar, G. (2020b). Optimized Hybridization of Ant Colony Optimization and Genetic Algorithm (HACOGA) Based Interpretable Intuitive and Correlated-Contours Fuzzy Neural Network Classifier for Abalone. Journal of Computational and Theoretical Nanoscience, 17(6), 2755–2762. https://doi.org/10.1166/jctn.2020.9117
20. Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization. Swarm intelligence, 1(1), 33-57.
21. Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.
22. Pramoda Patro, Krishna Kumar, G. Suresh Kumar, Gandharba Swain, Similarity and wavelet transform based data partitioning and parameter learning for fuzzy neural network, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part B, 2022, Pages 3424-3432, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.06.003.
23. Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5), 1285-1298.