European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 11, 2020

A MODE FUZZY WEIGHT BASED CANONICAL POLYADIC (MFWCP) AND ADAPTIVE NEURO FUZZY INTERFACE SYSTEM (ANFIS) FOR MISSING VALUE IMPUTATION IN BREAST CANCER PREDICTION

[1]S.R.Lavanya, [2]Dr.R.Mallika

[1]Research Scholar, Research & Development Center, Bharathiar University, Coimbatore

[2]Assistant Professor, Department of Computer Science, CBM College, Coimbatore

Abstract: Most of the women are affected by BC (Breast Cancer) which is one of the dreadful diseases in the entire world and considered as subsequent threatening reason of cancer death in women. The likelihood of death can be significantly reduced by means of early detection and prevention. Hybrid Bayesian frameworks were utilized previously for breast cancer prediction and handling missing values in patient's characterization. WCP (Weight Canonical Polyadic) algorithms manage continuous missing values in the data by using least squares recursively. A main bottleneck in using WCP is the unfolding of multiple relationships of discovered modes (N). The complexity increases when the value of N is large. This paper uses imputations in attribute dependencies for enhancing BC detections. This work divides the dataset into discrete and continuous subsets where discrete fields are assigned values using BN (Bayesian Networks) followed by Tensor factorization on an integrated dataset using MFWCP (Mode Fuzzy Weight based Canonical Polyadic). The new dataset is created from full/missing value subsets for assigning values to fields with missing values. MFWCP operations result in operations where N value of WCP is greater than three. This third order is reduced to first order WCP by applying Khatri-Rao product. DT (Decision Trees), KNN (K-Nearest Neighbors) and ANFIS (Adaptive Neuro Fuzzy Inference System) classifiers are combined to classify BC and the proposed hybrid method is evaluated using defined performance measures enhanced imputation accuracy.

KEYWORDS: Breast cancer, Missing value imputation, Classification, Tensor factorization, Bayesian network, Optimal Fuzzy weight based Canonical Polyadic.

1. INTRODUCTION

Breast Cancer (BC) is one of the mainly general cancers between women worldwide, indicating the common of new cancer cases and cancer-related deaths related to global statistics, making it an important public health issue in today's society [1]. The cancer cells elimination is done by initial treatment, but some of the cells may escape the treatment and survive. Apart from this, now-a-days MLTs (Machine Learning Techniques) are also utilized for BC prediction. Though, a huge amount of patient data is gathered in medical datasets. To benefit from the collected data of patients and increase the accuracy of prediction, a number of researchers have utilized data mining and Machine ML methods for predicting breast cancer [2].Existing MLTs use open source or local databases in their applications [3]. Creating a dataset for BC is again a challenging task and there are no standard predictors for BC. Sensitivity is a compromising factor though high accuracy outcomes are often achieved. Missing data and class imbalances are rarely handled making BC predictions complex. Inappropriate performance metrics contribute to this complexity and thus BC prediction is considered to be an open problem. The primary nature of this work is to nullify missing values by imputations and combine multiple techniques for improving BC predictions and is detailed as points below:

- BN imputes missing discrete values.
- Tensor factorization using MFWCP for imputations.
- Three classifiers namely DT, KNN and ANFIS are combined to predict BCs.

This paper is organized as detailed below. An overview of related works on BC predictions is presented in section 2 while section 3 details on the proposed methodology. Section 4 is results from experimentations while section 5 concludes this research work.

2. LITERATURE SURVEY

DM (Data Mining) techniques are utilized previously for BC predictions through developing models. In [4], BC prognosis is achieved through feature selection. The classification algorithm improvement can be attained through appropriate attribute selection technique. Poor prediction may ensue due to the attributes with less data

in datasets and thus reducing classification accuracies. In [5], BC prediction is achieved through ClassRBM (Classification Restricted Boltzmann Machine). It also predicts BC from symptoms. Dropping, a general probabilistic framework for learning Boltzmann machines with masks is explained in the study. The mask generation may create dissimilar learning methods, i.e., Dropout, Drop Connect. Drop Part, a generalization of Drop Connect is utilized in this research.

The study in [6], exploits EHR's (Electronic Health Records) narratives for developing a model in order to recognize BC . MetaMap is exploited for feature extraction from clinical narratives. The structured clinical data from EHRs are also retrieved. SVM (Support Vector Machine) then trains on these features for recognizing BC patients. In [7], SEER (Surveillance, Epidemiology and End Results) Public-Use Data 2005 is exploited for predicting BC using DM techniques. A novel data pre-classification technique is presented and a possible solution is obtained for obtaining BC information from SEER data. Numerous algorithms are investigated after dataset pre-processing. It is revealed that c5 algorithm possess the best performance of accuracy. The study in [8], utilized three classification techniques for BC prediction and their outcomes were analysed. SVM was regarded as best predictor with the test dataset supported through ANN (Artificial Neural Networks) and DTs. More variables can be utilized for performance improvement and for selecting a longer follow-up duration

The study in [9] deployed MLP (Multi-Layer Perceptron) classifier generating different outputs for BC prediction along with DNN (Deep Neural Networks) for feature extractions. At last, SVM is greatly utilized in this research for classification. The outcomes are compared for the above method. RNNs (Rough Neural Networks) with two outputs produced high accuracy with lowered variances. In [10], Eight popular DM methods including Clustering and classification methods were used to find an efficient predictor algorithm for assessing BC recurrences. The identified parameters were verified on medical datasets to predict disease occurrence in future.

Missing values were imputed in [11] using their proposed MIAEC (Missing Value Imputation algorithm based on the Evidence Chain). The method mined missing values in each data row and built an evidence chain by combining discovered columns for further estimations. Map-Reduce programming extended the proposed MIAEC for processing voluminous data and parallel distributions. Tolerance sets were used in [12] which proposed MBOI, an incomplete data clustering algorithm. A set of constraint tolerant sets were defined for incomplete categorical variables of datasets. These sets estimated overall differences in incomplete data objects and missing data values were imputed by MBOI which was found to be accurate in its imputations in a reduced time frame.

Imputation methods for longitudinal data were compared in [13] where 12 different methods were studied. Their analysis on simulated data using MAR (Missing At Random) mechanisms showed MI methods were less biased in their estimates while using linear regression. Some methods estimated better with regression mixed with linear random intercepts. The study observed an inverse association between a child's BMI and health based on imputed data. The study in [14] examined the effect of noise on imputation methods while proposing GMDH (Group Method of Data Handling) to handle noisy incomplete data. The proposed GMDH was benchmarked with RIGB (Robust imputations based on GMDH). The experimental results showed noises affect efficiency of imputation techniques where RIBG was found to be more robust to noise than other compared methods.

BC prediction is regarded as an open challenge for almost a decade. A combination of ML techniques has become a prominent solution for this challenge when compared to other approaches. The significant aspect of 52:35 is not yet completely addressed for BC prediction which is considered as critical problem for strategic imputations. Datasets with more patient records help in balancing values and thus pave the way for establishing new ML algorithms or their explorations.

3. PROPOSED METHODOLOGY

This research work proposes a new framework for detecting diseases and has four significant parts. Initially vertically divide, the entire dataset into two subsets: a subset with discrete missing attributes and a subset with continuous missing attributes. Then Bayesian network is proposed which computes the probability of possible values for discrete missing value. After estimating non-numerical missing values, integrated dataset from discrete imputed subset and the subset of numerical missing data are reconstructed using tensor with Mode Fuzzy Weight based Canonical Polyadic (MFWCP); and continuous missing values are also imputed using this reconstruction. That is, the difference between successive estimated values is minimized via the Mean Squared Deviation (MSD) and no change can be occurred in the next iterations. Finally classification models (Decision Tree (DT), K Nearest Neighbor (KNN) and Adaptive Neuro Fuzzy Inference System (ANFIS)) are applied to complete imputed dataset. The various performance metrics of the classifier such as accuracy, precision, recall, and F-measure are computed for estimating the suggested algorithm. Figure 1 shows the proposed framework for missing values imputation and especially for disease diagnosis.

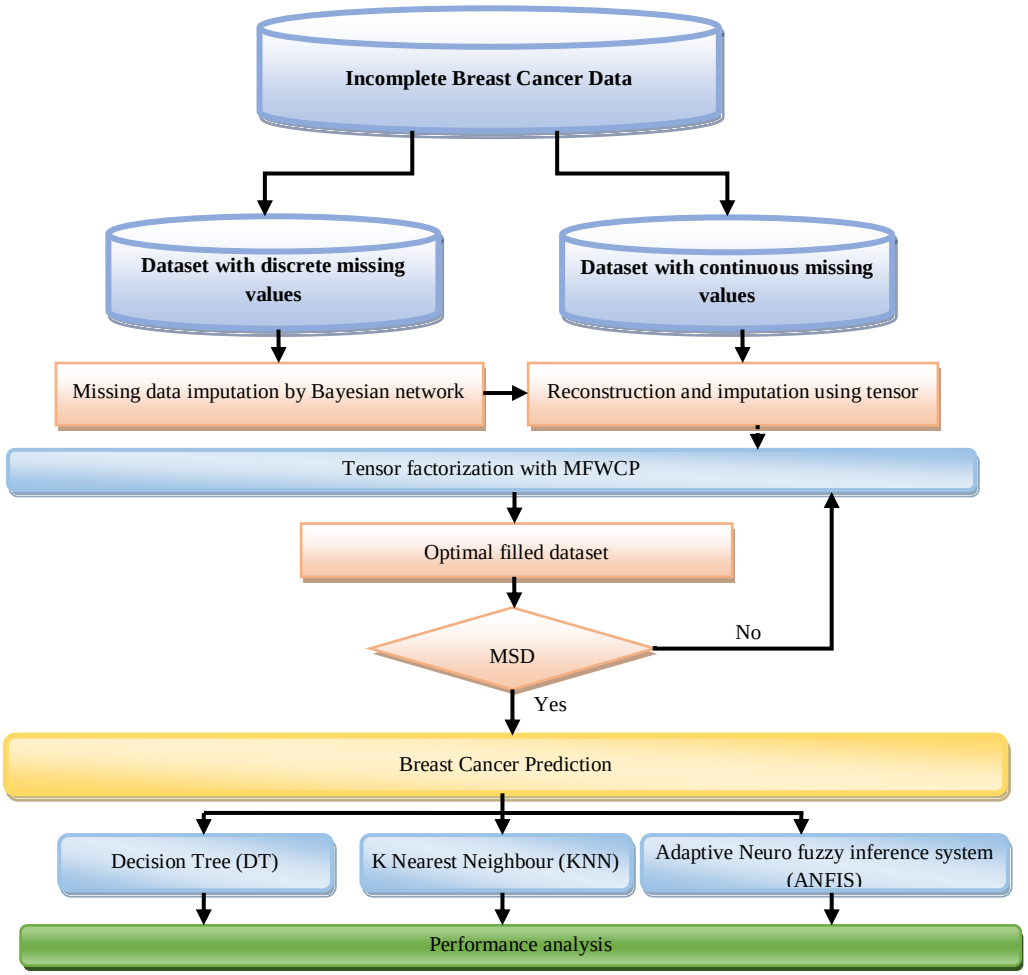

FIG. 1 – PROPOSED BC DETECTION ARCHITECTURAL FRAMEWORK

3.1. INCOMPLETE BC DATA

This research work exploits the UCI Machine Learning Repository by using its commonly available BC datasets namely WDBC (Wisconsin Diagnosis BC), and WOBC (Wisconsin Original BC).

3.2. IMPUTATION METHODS

Classifier accuracy is altered when data values are missing making imputations of these data values a necessity. BN imputes missing values in this work due to its capacity to model uncertainties from causal relationships between variables. One major aim of this work is to propose an imputation method for better predictions of BC from datasets. MAR techniques attach missing values to an instance of the observed data. Hence, this work uses MAR mechanism through BN.

BNs are probabilistic graphic models [15]. DAG (Directed Acyclic Graph) links nodes to attributes for representing dependency of attributes and a joint probability distribution (Pr_M) in discrete dataset variables. BNs can be represented as a triple $M = (\mathcal{g}, X, P)$, where $\mathcal{g} = (V_{\mathcal{g}}, E_{\mathcal{g}})$ is the dependency graph of X variables with m nodes and a set of edges E as variable's dependencies;P is a set of conditional probabilities [23]:

$$Pr_M(X_i|PA_i) \text{ (1)}$$

Where PA_i- nodes which $X_i$depends called parents of X_i and can be a empty or a subset of variables $V_{\mathcal{g}}$. Markov structure is the key capability of a BN where every attribute $X_i$is not conditionally dependent on non-descendants while having its parent(pa_i). Thus, BNs give joint probability distribution given by equation (2):

$$Pr_M(X_1, \ldots, X_m) = \prod_i Pr_M(X_i|PA_i) \text{ (2)}$$

Complicated relationships between random variables is learnt by BN and utilized for ensuing approximations or classifications. BN learning is based on the network structure and its parameters. DAG is used in this work to

learn parameters of conditional probability distributions from the dataset. EM (Expectation Maximization) algorithm is an efficient BN method [15] which recursively estimates highest probability in two steps namely an Expectation step which calculates log probability of data and based on this calculated log, current structure and networks parameters are characterized. In the next step, maximization step, it starts finding parameters that maximize previous step probabilities. The procedure is iterated till the network attains equilibrium or there are no parameters. This results in successful training of missing data.

3.3. RECONSTRUCTION AND IMPUTATION USING TENSOR VIA MFWCP

Multidimensional arrangements are termed as tensors and the degree of a tensor refers to the number of its dimensions. Higher-degree tensors confer to multidimensional arrangements with three or more dimensions [16,17]. Tensor factorizations yield higher precisions but consume more computational time [18, 19]. Tucker and CP (Canonical Polyadic) are known models in tensor factorizations [18]. CP bifurcates tasks into one-rank tensors and then multiplies using Khatri-Rao product for reconstructing the primary tensor. This work uses MFWCP where tensor dimensions are greater than three and reduced to single mode by CPD and followed by Khatri-Rao product as defined in equation (3)

$$A * B = \left(A_{ij} \otimes B_{ij}\right)_{ij} \quad (3)$$

in which the ij[th] block is the $m_i p_i \times n_j q_j$ sized Kronecker product of **A** and **B blocks**, when row and column partitions of the matrices are equal. The size of the product is $(\Sigma_i m_i p_i) \times (\Sigma_j n_j q_j)$. Missing values are reconstructed linearly by combining other feature values. If x is a three-rank tensor of size $I \times J \times K$ and R is the number of broken matrices or rank of tensor, CP factorization is created by factor matrices A, Band C of size $I \times R, J \times R$ and $K \times R$ respectively such that the following equation holds for all values of $i = 1 \ldots I, j = 1 \ldots J$ and $k = 1 \ldots K$:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \quad (4)$$

The main aim of MFWCP factorization is to minimize errors in reconstruction of primary tensor and the aggregate of one rank tensors has minimum t difference with the original tensor where f has the lowest value [20] as described in equations (5-6),

$$f(A,B,C) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(x_{ijk} - \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}\right)^2 \quad (5)$$

$$f(A,B,C) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left\{fw_{ijk} * mode\left(x_{ijk} - \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}\right)\right\}^2 \quad (6)$$

where fw_{ijk} - non-negative fuzzy weight tensor initialized as for all $i = 1 \ldots I, j = 1 \ldots J$ and $k = 1 \ldots K$ and computed using a TFMF (Trapezoidal Fuzzy Membership Function) with a lower limit a, upper limit d, lower support limit b, and upper support limit c and $a < b < c < d$. Non-Negative fuzzy weight tensor using fuzzy membership values are computed using TFMF given inn equation (7),

$$f\left(x_{ijk}, a, b, c, d\right) = \max\left(\min\left(\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}, 0\right)\right)\right) \quad (7)$$

$$fw_{ijk} = \begin{cases} 1 \ if x_{ijk} \ is \ know \\ \frac{x-a}{b-a} if x_{ijk} \ is \ not \ know \ and \ other \ 70\% \ of \ attributes \ know \\ \frac{d-x}{d-c} if x_{ijk} \ is \ not \ know \ and \ other \ 60\% \ of \ attributes \ know \\ 0 \ if x_{ijk} \ is \ not \ know \end{cases} \quad (8)$$

a=lower value of attribute, b= minimum value of attribute, c= medium value of attribute, and d= maximum value of attribute [21]. The missing data imputation with high percentage of missingness might not be achieved by WCP factorization. Hence MFWCP model is suggested for weight tensor with same size like original tensor which is deliberated for missing values imputation. In the equation (6), mode is computed via the frequency of the attribute. The most frequent value of the corresponding attribute is computed via the equation (9),

$$mode = \begin{cases} 1 \ most \text{ frequent value of the attribute} \\ 0.5 \ middle \text{ frequent valueof the attribute} \\ 0 \ no \ frequent \ value \text{of the attribute} \end{cases} \quad (9)$$

These mode and weight values are updated to equation (6).

3.4. Mean Squared Deviation (MSD)

Reconstructions are done till a convergence condition is met or MSD is attained and computed using equation (9) [22]:

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(Output_i(t) - Output_i(t-1))^2 \quad (10)$$

Where $Output(t)_i$ - estimated value in t[th] iteration and n – count of missing values.

3.5. Predictive Models

ML techniques analyze data based on predictions or decisions. BC incidences are predicted from the datasets using DT, KNN and ANFIS.

3.5.1. Decision Tree (DT)

DT has two main parts namely nodes and rules. It draws a flowchart from the root node to the topmost node and non-leaf depicts a test for a single/multiple attribute of a leaf node (final result). DT algorithm is significant in DM techniques, since they have been widely regarded as an efficient classification tool [21]. The following factors define the reasons for utilizing decision trees in data mining and classification:

- Generation of comprehensible rules: These rules are regarded as user-friendly algorithm for the end user in data mining; besides commence associations within dataset attributes in an easily understandable way.
- Provision of explicit indication to significant attributes: This is a vital part of establishing rules between attributes, through which the level of importance of each one has indicated.
- Requiring less computation: Decision tress need comparatively less computation than other classification algorithms (e.g. mathematical formulae)

C4.5 is a DT used on discrete and continuous data [24]. It constructs DTs from a set of training data using entropy. If $S = (s_1, .. s_i)$ is a training set of classified samples, then each sample s_i consists of a p-dimensional vector$(x_{1,i}, ... x_{p,i})$, where x_j are attribute values of the sample and its class s_i. Data attributes are split into subsets and belong to one class or the other. This division gains information or entropy and highest entropy leafs are chosen for the decision. The rules of C4.5 are listed below [24]:

- When all cases belong to one class, the tree is a leaf and leaf is labelled with the class and returned;
- Calculate potential information from a test on each attribute while calculating information gains.
- Based on a selection, find the attribute to branch.

Information Gain (IG)

Entropy is actually a measure of data disorders and calculated using equation (11),

$$Entropy(\vec{y}) = -\sum_{j=1}^{n}\frac{|y_j|}{|\vec{y}|}\log\frac{|y_j|}{|\vec{y}|} \quad (11)$$

Repeating for all values of $\vec{y}$. Conditional Entropy is depicted in equation (12)

$$Entropy(j|\vec{y}) = \frac{|y_j|}{|\vec{y}|}\log\frac{|y_j|}{|\vec{y}|} \quad (12)$$

and Information Gain is computed using equation (13)

$$Gain(\vec{y}, j) = Entropy(\vec{y}) - Entropy(j|\vec{y}) \quad (13)$$

The aim is to maximize Gain and divide it by overall entropy due to split argument $\vec{y}$ by value j.

Pruning

Pruning has significance in removing outliers in data as they are improperly defined data instances. They change the accuracy of predictions due to their values and need to be eliminated. On creation of a DT the tree is pruned to improve classifier accuracy in this study.

3.5.2. K Nearest Neighbors (KNN)

KNN, a supervised ML method is applied for regression and classification as it does not assume data distributions and thus is efficient while predicting or analyzing patterns. KNN collects the nearby data points for each new data point. The distance amidst data points get impacted when they vary heavily in terms of their values [25,26]. Subsequently, the algorithm categorizes closest data points nearer to the arrival data point. Although there are many ways to estimate this distance, Euclidian distance is preferred [25], where the Euclidean Distance amid X_1 and X_2 points, which are also two related BC dataset.

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad (14)$$

This work considers specific number of data points with lesser distances for categorizations. Thus, KNN selects an odd number for K and stops when the number of classes is 2 where it includes maximum number of data point in the category of a new data point. DM algorithms like KNN have greatly utilized for cancer detections. *k* is a user-defined constant and test points are classified by assigning most frequent labels in *k* training samples nearest to that data point.

3.5.3. Adaptive Network-Based Fuzzy Inference System (ANFIS)

ANFIS is a set of IF–THEN fuzzy rules which can learn by approximation of nonlinear functions. ANFIS network comprises five layers that include individual functional nodes, where each node executes specific function. In addition, the hybrid methodology learns from ANFIS by combining BGPD (Backward Propagation Gradient Descent) and LSE (Least-Squares Estimator) approaches. This gives the technique neural network capabilities with fuzzy logic explorations. ANFIS uses a feed-forward network to search for fuzzy decision rules for its tasks [27]. In a dataset, ANFIS generates FIS membership function parameters which are tuned using the back propagation algorithm or by combining back propagation with LSE allowing a fuzzy learning for modelling.

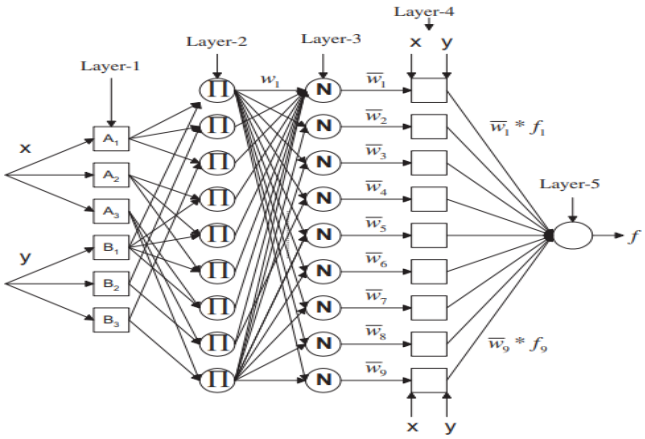

FIGURE 2. ANFIS ARCHITECTURE OF TWO INPUTS AND NINE RULES

The fuzzy inference system is simplified by using two inputs (x and y) and one output (z). The first order Sugeno fuzzy model's rule set or fuzzy if–then rules can be expressed as equation (15),

$$\text{if x is } A_1 \text{ and y is } B_1 \text{ then } f_1 = p_1 x + q_1 y + r_1 \quad (15)$$

where p, r, and q are linear output parameters. ANFIS inputs and output are shown in Figure 2. The proposed architecture is formed with 5 layers and 9 if–then rules:

Layer-1: Every node i in this layer is a square node with a node function. It is defined in equation (16-17),

$$O_{1,i} = \mu_{A_i}(x) \; for \; i = 1,2,3 \quad (16)$$

$$O_{1,i} = \mu_{B_{i-3}}(y) \; for \; i = 4,5,6 \quad (17)$$

where x and y are inputs to node i, and A_i and B_i are linguistic labels for inputs. In other words, $O_{1,i}$ is the membership function of A_i and B_i. Usually $\mu_{A_i}(x)$ and $\mu_{B_i}(x)$ are chosen to be bell-shaped with maximum equal to 1 and minimum equal to 0, such as in equation (18)

$$\mu_{A_i}(x), \mu_{B_{i-3}}(y) = \exp\left(\left(-\frac{(x_i - c_i)}{a_i}\right)^2\right) (18)$$

where a_i, c_i is the parameter set. These parameters in this layer are referred to as premise parameters.

Layer-2: Every node in this layer is a circle node labeled P which multiplies the incoming signals and sends the product out. For instance,

$$O_{2,i} = w_i = \mu_{A_i}(x).\mu_{B_i-3}(y) \;\; i = 1,2,,\ldots,9 \quad (19)$$

Layer-3: Every node in this layer is a circle node labeled N. The i[th] node calculates the ratio of the i[th] rules firing strength to the sum of all rule's firing strengths:

$$O_{3,i} = \overline{w}_i = \frac{w_i}{w_1+\cdots w_9}, \;\; i = 1,2,,\ldots,9 \quad (20)$$

Layer-4: Every node i in this layer is a square node with a node function

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i), \;\; i = 1,2,,\ldots,9 \quad (21)$$

where w_i is the output of layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set. Parameters in this layer will be referred to as consequent parameters.

Layer-5: The single node in this layer is a circle node labeled P that computes the overall output as the summation of all incoming signals:

$$O_{5,i} = Overall\ Output = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (22)$$

The parameter input and rules are possessed by this domain attained from the experimental data. In addition, experimental data input fuzzified with an appropriate membership function are used to acquire antecedent parameters (ANFIS input). The premise parameters adjustments as well as subsequent parameters are captured as an epoch. The following steps abstract the proposed framework.

ALGORITHM 1. BC PREDICTION FRAMEWORK

INPUT: BC dataset with n × d size and consisting of missing values

OUTPUT:

1. Missing data imputed dataset with size n × d
2. Prediction results

STEP 1: Split the dataset into continuous and discrete value sets.

STEP 2: Use BN to impute subset with discrete missing values.

STEP 3: Combine imputed discrete and continuous missing values subsets

STEP 4: Reconstruction and imputation using tensor

- Convert the dataset to tensor and rank them
- Compute non-negative weight by trapezoidal membership function
- Compute mode to tensor
- Impute continuous missing values by MFWCP
- Repeat till convergence condition is achieved

STEP 5: Return missing data imputed dataset

STEP 6: Implement classifiers such as DT, KNN and ANFIS for the imputed data set.

4. EXPERIMENTS AND RESULT DISCUSSION

During the experiments, the scrutinies of the impact of various imputation models based on the accuracy of subsequent prediction methods have been focused. Based on the comprehensive description of real-dataset, experiment has initiated to evaluate the performance and analysed.

4.1. DATASETS

The datasets used in the study are WDBC, and WOBC from (UCI) ML Repository.

TABLE 1. DATASETS USED FOR EXPERIMENTATION

DATASETS	RECORDS	ATTRIBUTES	MISSING VALUES	CLASS DISTRIBUTION
WDBC	569	32	No	357 benign, 212 malignant
WOBC	699	10	Yes	458 benign, 241 malignant

WDBC: Dataset, the features that were computed from a digitized image have included which belongs to Fine Needle Aspirate (FNA) of a breast mass. The characteristics of the cell nuclei existed in the image have defined through the aforementioned features. The dataset being made up of 569 data points, among which 212 belongs to Malignant; 357 belongs to Benign. Based on the following features, the dataset has classified into ten categories, i.e. i) radius, ii) texture, iii) perimeter, iv) area, v) smoothness, vi) compactness, vii) concavity, viii) concave points, ix) symmetry, and x) fractal dimension. In each feature, three significant details, namely Mean, Standard Error, and "worst"/largest (mean of the three largest values) have estimated. Consequently, the number of dataset features will be 30.

WOBC: Dataset includes 699 samples, which have been gathered from UCI repository [28]. In that, the number of benign samples is 458, and malignant samples are 241. Besides, the dataset includes 10 features as well as 1 class. The class level is binary class as benign and malignant. Also, there is missing value in the dataset. The features such as 1. Sample code number: id number, 2. Clump Thickness: 1 – 10, 3. Uniformity of Cell Size: 1 – 10, 4. Uniformity of Cell Shape: 1 – 10, 5. Marginal Adhesion: 1 – 10, 6. Single Epithelial Cell Size: 1 – 10, 7. Bare Nuclei: 1 – 10, 8. Bland Chromatin: 1 – 10, 9. Normal Nucleoli: 1 – 10, 10. Mitoses: 1 – 10 and Class: (2 for benign, 4 for malignant)

4.2. EVALUATION METRICS

All missing data was obtained, the following indicators are used for determining precision, recall, f-measure, sensitivity, specificity and accuracy.

To estimate the missing value, the proposed approach has evaluated, and the accuracy of the imputation techniques have compared with the help of Normalized Root Mean Square Error (NRMSE). This estimation has expressed by the following equation [18],

$$NRMSE = \frac{1}{max-min}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2} \quad (23)$$

In which, the real value has denoted by x_i, and imputed value has signified by x_i'.

Precision refers to the ratio of positive samples, which have properly classified. Estimation of this metric can be formulated as,

$$Precision = \frac{TP}{FP+TP} \quad (24)$$

Recall defines the positive samples that have been designated to the total number of positive samples, which can estimated as follows,

$$Recall = \frac{TP}{TP+FN} \quad (25)$$

F-measure is also called as F 1-score. It refers to the harmonic mean of precision and recall as expressed below,

$$F - measure = \frac{2*(Recall * Precision)}{(Recall + Precision)} \quad (26)$$

Specificity is significant to recognize the appropriate and inappropriate decisions made by the respective classifier, which can be expressed as follows,

$$Specificity = \frac{TN}{FP+TN} \quad (27)$$

Accuracy is a measure that has considered being one of the widely-regarded metric to analyse the classification performance, which has executed in this work for cancer prediction.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (28)$$

Here, True Positive, True Negative, False Positive, and False Negative have signified by TP, TN, FP and FN. For instance, assume that there are two classes in a dataset, the number of appropriate classifications that from the first class can be denoted by true positive, besides the number of appropriate classifications that possessed by the second class can be signified by true negative. Conversely, the number of incorrectly predicted instances in the first class, which are corresponding to the other class, can be defined as false positive. At the same time, the number of incorrectly predicted instances in the second class, but they are actually from the first class, can be referred to as false negative.

4.3. RESULTS COMPARISON

During the experiments, the proposed missing data imputation MFWCP method has compared with three imputation approaches, such as KNN [10], DT [8], and ANFIS. In the proposed method for missing value imputation, to evaluate the performance, 10% of random missing values is inserted to the datasets even WDBC has no missing values and WOBC have a minimal percentage of missing values. Results demonstrate the optimal performance of the proposed ANFIS, which is superior to the results obtained by KNN, DT. Because, due to the inconsistency of DT, the minor change in the data could extensively change the framework of the optimal decision tree. Similarly, in KNN, the increasing number of samples or/and independent variables/predictors hampers the speed of the algorithm.

TABLE 1. RESULTS COMPARISON OF CLASSIFIERS VS. DATASETS

Imputation Methods	Classifiers	WDBC Results (%)					NRMSE
		Precision	Recall	F-measure	Specificity	Accuracy	
WCP	KNN	75.67	81.74	82.30	81.88	74.85	0.3153
	DT	86.27	93.58	89.08	93.57	85.38	0.3824
	ANFIS	94.14	95.16	95.32	95.13	94.15	0.2418
MFWCP	KNN	87.16	87.51	87.33	87.71	88.89	0.3333
	DT	89.41	91.14	90.17	91.14	91.23	0.3244
	ANFIS	95.08	97.41	96.09	97.41	96.49	0.1873
Imputation Methods	Classifiers	WOBC Results (%)					NRMSE
		Precision	Recall	F-measure	Specificity	Accuracy	
WCP	KNN	65.35	65.92	65.57	65.92	68.57	0.6249
	DT	95.12	92.21	91.41	92.21	89.52	0.3842
	ANFIS	96.12	93.51	94.66	93.51	93.33	0.2582
MFWCP	KNN	69.85	69.15	69.57	69.12	70.57	0.5606
	DT	98.21	97.53	97.86	97.53	98.10	0.1380
	ANFIS	98.57	98.23	98.40	98.23	98.57	0.0598

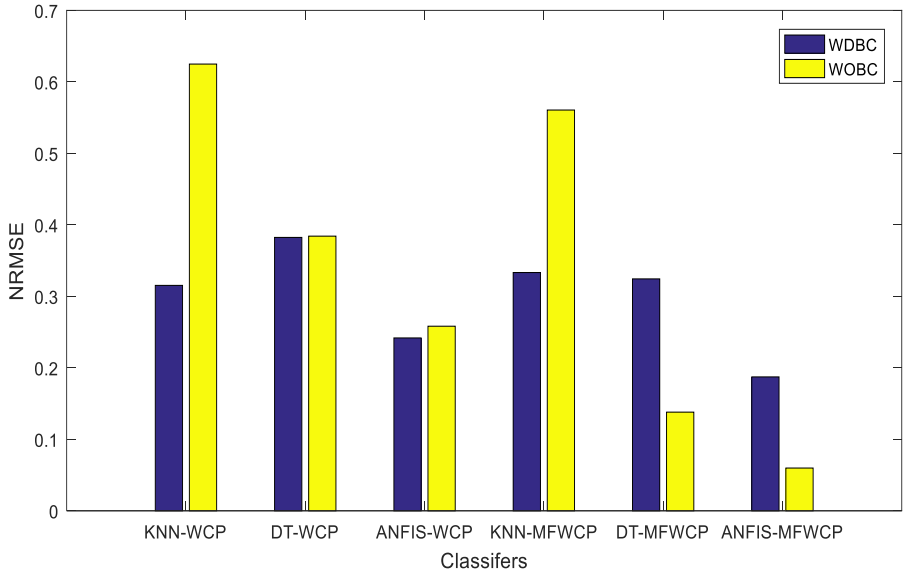

FIGURE 3. NRMSE RESULTS COMPARISON VS. CLASSIFIERS

Figure 3 shows the evaluations of NRMSE comparison with classifiers (WCP-KNN, WCP-DT, WCP-ANFIS, MFWCP-KNN, MFWCP-DT and MFWCP-ANFIS) with respect to WDBC and WOBC datasets. For both WDBC and WOBC datasets, the proposed MFWCP-ANFIS method produces the NRMSE values of 0.1873 and 0.0598, respectively. Proposed model includes Trapezoidal fuzzy membership function and mean value that implies a simplified mathematical model, through which data-in sufficiency of the tensor can get tackled, thereby NRMSE rate get diminished. In Table 1, the numerical outcomes of NRMSE have provided. The other methods such as WCP-KNN, MFWCP-KNN, WCP-DT, MFWCP-DT, and WCP-ANFIS give the NRMSE of 0.3153, 0.3333, 0.3824, 0.3244 and 0.2418 respectively at the WDBC dataset. From the results it concludes that the classifier with MFWCP gives lesser NRMSE than the traditional classifiers.

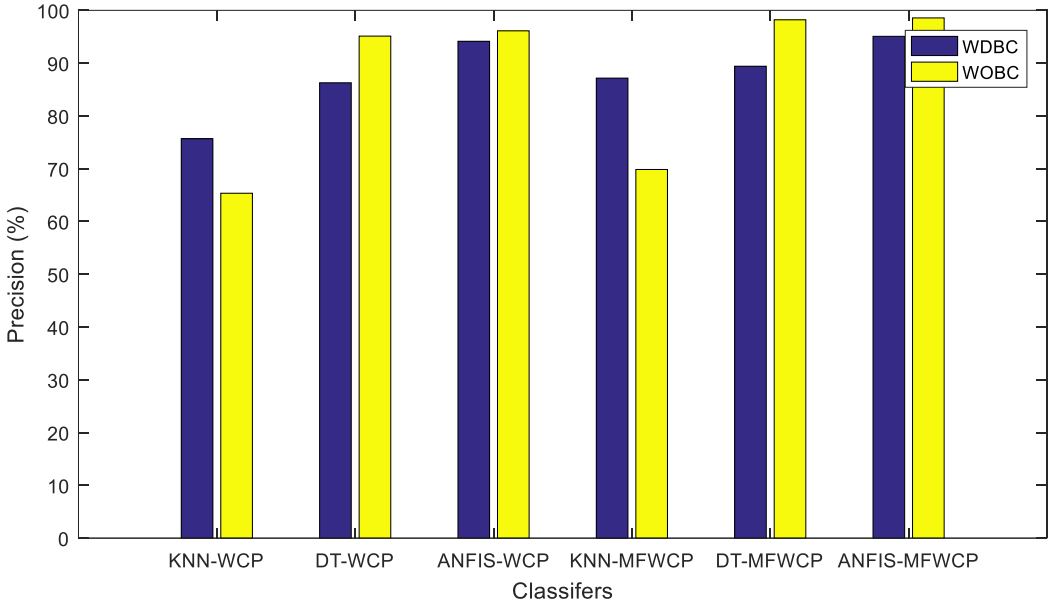

FIGURE 4. PRECISION RESULTS COMPARISON VS. CLASSIFIERS

Figure 4 shows the precision results comparison with classifiers (WCP-KNN, WCP-DT, WCP-ANFIS, MFWCP-KNN, MFWCP-DT and MFWCP-ANFIS) with respect to WDBC and WOBC datasets. For both WDBC and WOBC datasets, the proposed MFWCP-ANFIS method produces the precision values of 95.08% and 98.57%, respectively. The other methods such as WCP-KNN, MFWCP-KNN, WCP-DT, MFWCP-DT, and WCP-ANFIS gives the precision of 75.67%, 87.16%, 86.27%, 89.41% and 94.14% respectively at the WDBC dataset (See Table 1). From the results it concludes that the classifier with MFWCP gives higher precision

value when compared to traditional classifiers. Consequently, it can be concluded that the proposed method can be an appropriate preference to obtain the accurate cancer prediction accompanying higher precision rate.

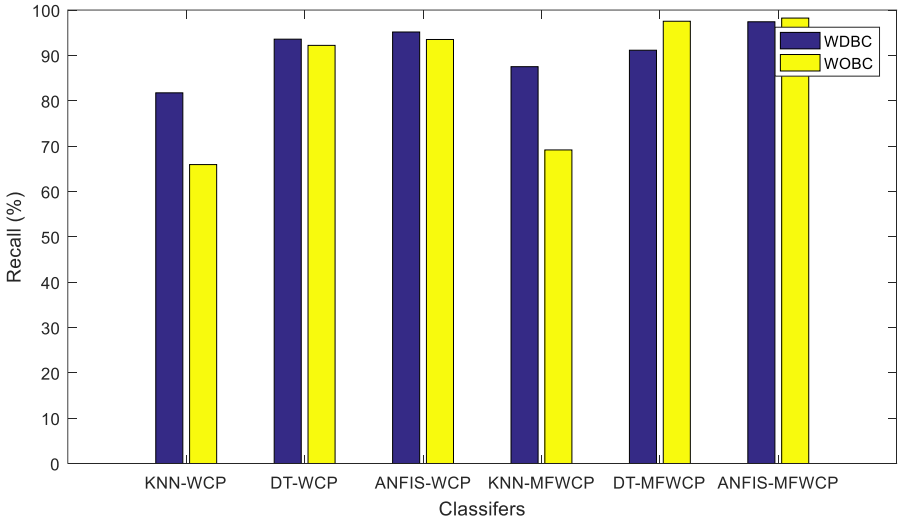

FIGURE 5. RECALL RESULTS COMPARISON VS. CLASSIFIERS

Figure 5 shows the recall results comparison with classifiers (WCP-KNN, WCP-DT, WCP-ANFIS, MFWCP-KNN, MFWCP-DT and MFWCP-ANFIS) with respect to WDBC and WOBC datasets. For both WDBC and WOBC datasets, the proposed MFWCP-ANFIS method produces the recall values of 97.41% and 98.23%, respectively. The other methods such as WCP-KNN, MFWCP-KNN, WCP-DT, MFWCP-DT, and WCP-ANFIS gives the recall of 81.74%, 87.51%, 93.58%, 91.14% and 95.16% respectively at the WDBC dataset (See Table 1). From the results it concludes that the classifier with MFWCP gives higher precision value when compared to traditional classifiers. It concludes that the proposed approach can be an efficient approach for obtaining the accurate cancer prediction, alongside the maximum recall rate. Empirical findings depict that the proposed methodology proves to be efficient than other prevailing techniques. Compared to BN or Tensor method, the proposed ANFIS- MFWCP method provides optimal performance, since it resolves the flaws of both methods during the missing data estimation and generates the optimal filled dataset.

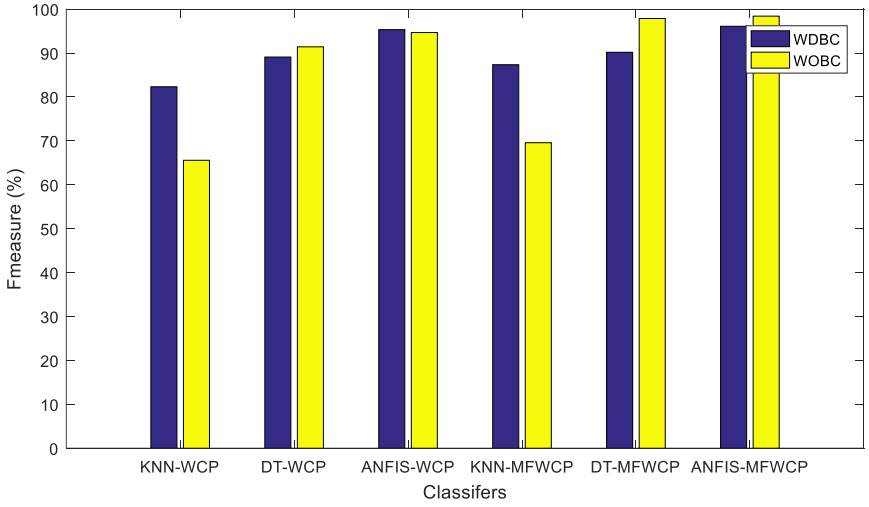

FIGURE 6. F-MEASURE RESULTS COMPARISON VS. CLASSIFIERS

Figure 6 compares the F-measure values of classifiers (WCP-KNN, WCP-DT, WCP-ANFIS, MFWCP-KNN, MFWCP-DT and MFWCP-ANFIS). The graphs demonstrate the efficiency of the proposed approach gives 96.09% of F-measure for WOBC and 98.40% for WDBC, which is comparatively higher than the other methods. The other methods such as WCP-KNN, MFWCP-KNN, WCP-DT, MFWCP-DT, and WCP-ANFIS gives the F-measure of 82.30%, 87.33%, 89.08%, 90.17% and 95.32% respectively at the WDBC dataset (See Table 1). Accordingly, it can be declared that the proposed approach can be an appropriate preference for attaining the accurate cancer prediction, as it delivers optimal precision rate.

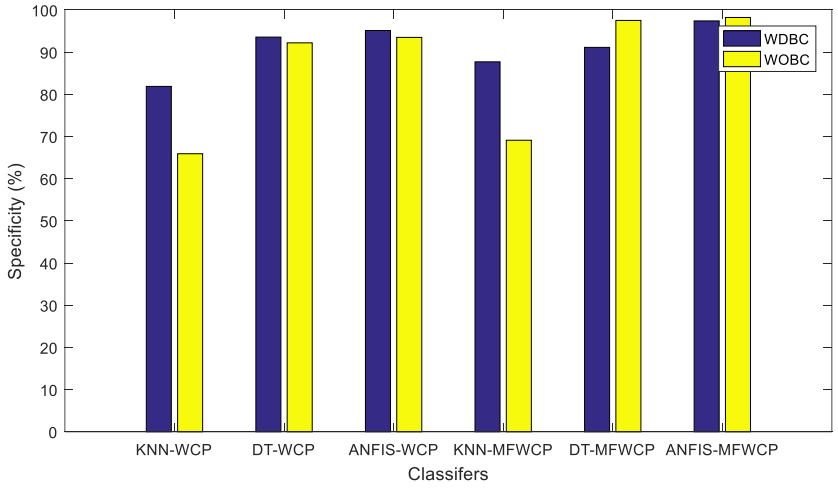

FIGURE 7. SPECIFICITY RESULTS COMPARISON VS. CLASSIFIERS

Figure 7, the outcomes of classifiers (WCP-KNN, WCP-DT, WCP-ANFIS, MFWCP-KNN, MFWCP-DT and MFWCP-ANFIS) in terms of specificity. The graphs represent the capability of the proposed approach gives 97.41% of specificity for WOBC and 98.32% of specificity for WDBC, through which it can be concluded that the proposed approach can be an appropriate preference for obtaining the accurate cancer prediction. Empirical findings depict that the proposed methodology proves to be optimal in WOBC when compared to other techniques. Whereas, the reason for the minimal specificity value of the proposed model in WDBC is that the Bayesian network is only as useful as this prior knowledge is trustworthy. Thus, the extremely positive/negative anticipation regarding the quality of these prior convictions misleads the overall network, besides nullifying the outputs. The other methods such as WCP-KNN, MFWCP-KNN, WCP-DT, MFWCP-DT, and WCP-ANFIS gives the specificity of 81.88%, 87.71%, 93.57%, 91.14% and 95.13% respectively at the WDBC dataset (See Table 1).

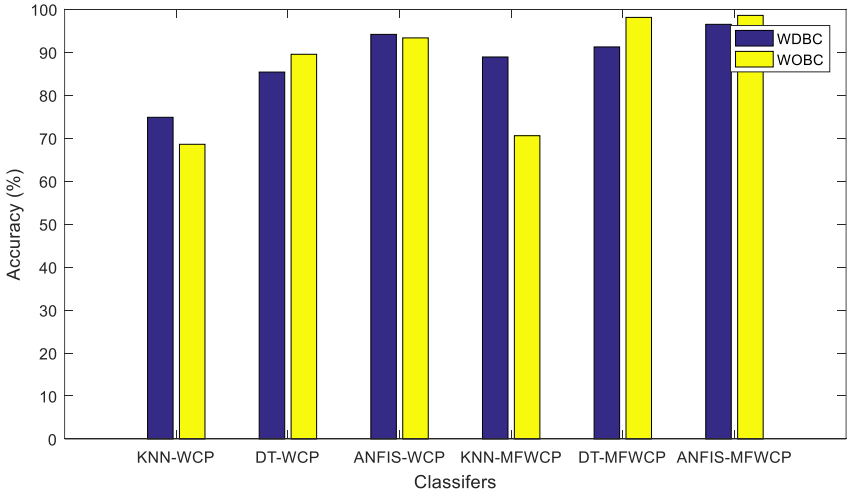

FIGURE 8. ACCURACY RESULTS COMPARISON VS. CLASSIFIERS

Figure 8 compares the accuracy rates obtained by classifiers (WCP-KNN, WCP-DT, WCP-ANFIS, MFWCP-KNN, MFWCP-DT and MFWCP-ANFIS), in which the graphs exemplify the proficiency of the proposed approach gives 96.49% of accuracy for WOBC, and 98.57% of accuracy for WDBC (See Table 1). Empirical findings demonstrate that the proposed approach proves to be efficient than other existing techniques. Besides, the proposed MFWCP-ANFIS classifier achieves great predictive accuracy for high-dimensional problem, as it applies the ANFIS to generate highly correlated features. The other methods such as WCP-KNN, MFWCP-KNN, WCP-DT, MFWCP-DT, and WCP-ANFIS gives the accuracy of 68.57%, 70.57%, 89.52%, 98.10% and 93.33% respectively at the WDBC dataset (See Table 1).

5. CONCLUSION AND FUTURE WORK

In recent days, missing values estimations have received attention of researchers in healthcare. This study resolves class imputation issue by proposing an enhanced missing data imputation method using MFWCP.

Missing data imputation is done by using a Tensor method where non-negative fuzzy weight tensor is used in incomplete data. This Non-Negative fuzzy weight tensor is computed using Trapezoidal fuzzy membership function. MFWCP model is suggested for weight tensor with same size like original tensor which is deliberated for missing values imputation. In addition, mode is also used for the finding frequency of the attribute in the tensor model. Finally classification models DT, KNN and ANFIS use the imputed dataset. These classifiers are applied over BC datasets and evaluated using performance metrics of Precision, Recall, F-measure, Specificity, Accuracy and NRMSE. The performance of proposed approach demonstrates its appropriateness, which is superior to other state-of-the-art methods. During the process, this model diminishes the cost of utilizing additional tools. In future, this study can be further extended through focusing on the enhancement of the proposed method using Swarm intelligence, in order to evade its inappropriate function caused by the data noise during the process of missing values estimation.

REFERENCES

1. Sun, Y.S., Zhao, Z., Yang, Z.N., Xu, F., Lu, H.J., Zhu, Z.Y., Shi, W., Jiang, J., Yao, P.P. and Zhu, H.P., 2017. Risk factors and preventions of breast cancer. International journal of biological sciences, 13(11), p.1387.
2. Choi, S., Jiang, Z., 2010. Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique. Comput. Biol. Med. 40 (1), 8–20.
3. Abreu, P.H., Santos, M.S., Abreu, M.H., Andrade, B. and Silva, D.C., 2016. Predicting BC recurrence using machine learning techniques: a systematic review. ACM Computing Surveys (CSUR), 49(3), pp.1-40.
4. Pritom, A.I., Munshi, M.A.R., Sabab, S.A. and Shihab, S., 2016, Predicting BC recurrence using effective classification and feature selection technique. In 2016 19th International Conference on Computer and Information Technology (ICCIT) ,pp. 310-314
5. Tomczak, J.M., 2013. Prediction of BC recurrence using Classification Restricted Boltzmann Machine with Dropping. arXiv preprint arXiv:1308.6324..
6. Zexian, Z., Ankita, R., Xiaoyu, L., Sasa, E., Susan, C., Seema, K. and Yuan, L., 2018, Using Clinical Narratives and Structured Data to Identify Distant Recurrences in BC . In 2018 IEEE International Conference on Healthcare Informatics (ICHI) ,pp. 44-52.
7. Fan, Q., Zhu, C.J. and Yin, L., 2010, Predicting BC recurrence using data mining techniques. In 2010 International Conference on Bioinformatics and Biomedical Technology (pp. 310-311.
8. Ahmad, L.G., Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M. and Razavi, A.R., 2013. Using three machine learning techniques for predicting BC recurrence. J Health Med Inform, 4(124), pp.1-3.
9. Jafarpisheh, N., Nafisi, N. and Teshnehlab, M., 2018. BC relapse prognosis by classic and modern structures of machine learning algorithms. In 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) ,pp. 120-122.
10. Ojha, U. and Goel, S., 2017, A study on prediction of BC recurrence using data mining techniques. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence ,pp. 527-530.
11. Xu, X., Chong, W., Li, S., Arabo, A. and Xiao, J., 2018. MIAEC: Missing data imputation based on the evidence chain. IEEE Access, 6, pp.12983-12992.
12. Wu S., X.-D. Feng, Z.-G. Shan, "Missing data imputation approach based on incomplete data clustering", Chin. J. Comput., vol. 35, no. 28, pp. 1726-1738, 2012.
13. Huque, M.H., Carlin, J.B., Simpson, J.A. and Lee, K.J., 2018. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC medical research methodology, 18(1), pp.1-16.
14. Zhu, B., He, C., &Liatsis, P. (2012). A robust missing value imputation method for noisy data. Applied Intelligence, 36(1), 61-74.
15. Franzin, A., Sambo, F., Di Camillo, B., 2017. bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. Bioinformatics 33 (8), 1250–1252.
16. Tan, H., Yang, Z., Feng, G., Wang, W. and Ran, B., 2013. Correlation analysis for tensor-based traffic data imputation method. Procedia-Social and Behavioral Sciences, 96, pp.2611-2620.

17. Mørup, M., 2011. Applications of tensor (multiway array) factorizations and decompositions in data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), pp.24-40.

18. Dauwels, J., Garg, L., Earnest, A. and Pang, L.K., 2012, Tensor factorization for missing data imputation in medical questionnaires. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ,pp. 2109-2112.

19. Yang, F., Shang, F., Huang, Y., Cheng, J., Li, J., Zhao, Y. and Zhao, R., 2017. Lftf: A framework for efficient tensor analytics at scale. Proceedings of the VLDB Endowment, 10(7), pp.745-756.

20. Wang, H., Zhang, Q., Yuan, J., 2017. Semantically enhanced medical information retrieval system: a tensor factorization based approach. IEEE Access 5, 7584– 7593.

21. Ali, O.A.M., Ali, A.Y. and Sumait, B.S., 2015. Comparison between the effects of different types of membership functions on fuzzy logic controller performance. International Journal, 76, pp.76-83.

22. Vazifehdan, M., Moattar, M.H. and Jalali, M., 2019. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve BC recurrence prediction. Journal of King Saud University-Computer and Information Sciences, 31(2), pp.175-184.

23. Chen, X., He, Z., Chen, Y., Lu, Y., & Wang, J. (2019). Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model. Transportation Research Part C: Emerging Technologies, 104, 66-77.

24. Agrawal, G.L. and Gupta, H., 2013. Optimization of C4. 5 decision tree algorithm for data mining application. International Journal of Emerging Technology and Advanced Engineering, 3(3), pp.341-345.

25. Kramer, O., 2013. K-nearest neighbors. In Dimensionality reduction with unsupervised nearest neighbors (pp. 13-23). Springer, Berlin, Heidelberg.

26. Lin, Y., Nelson, B.L. and Pei, L., 2019. Virtual statistics in simulation via k nearest neighbors. INFORMS Journal on Computing, 31(3), pp.576-592.

27. Azadeh, A., Saberi, M. and Asadzadeh, S.M., 2011. An adaptive network based fuzzy inference system–auto regression–analysis of variance algorithm for improvement of oil consumption estimation and policy making: The cases of Canada, United Kingdom, and South Korea. Applied Mathematical Modelling, 35(2), pp.581-593.

28. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html].