Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3211 – 3225
Received 20 January 2021; Accepted 08 February 2021.

An Enhanced Filter Based Heterogeneous Data Classification Framework on Mixed Breast Cancer Databases

Anusha Derangula[1]*, Prof. SrinivasaReddy Edara[2]

[1,2]Department of Computer Science and Engineering,
AcharyaNagarjuna University, Andhra Pradesh, India
* Corresponding author's Email: d.anusha21@gmail.com

Abstract

Medical disease classification is one of the major challenges to scientific and medical researchers. Due to its high dimensional feature space and data imbalance, most of the medical databases contain many homogeneous or heterogeneous features. It is difficult to predict the disease label due to class imbalance. Feature transformation, feature ranking and data classification are the essential approaches used to classify the high dimensional data with a high true positive rate. Feature transformation helps to improve the feature ranking process in high dimensional feature space. Most traditional feature transformation approaches such as min-max variance, probabilistic normalization and min-max normalization, etc., are independent of data distribution and multi-class labels. In this work, a novel feature selection based random forest classifier is proposed to improve the efficiency of the medical datasets. Practical results proved that the proposed heterogeneous classification framework has better accuracy, nearly 98.6 % accuracy compared to the traditional nominal data classification models.

Keywords: Breast cancer, heterogeneous databases, classification model, feature ranking.

1. Introduction

Machine learning is the process of identifying and analyzing the unknown hidden patterns and their relationships on large uncertain databases. Machine learning is the art of computer science without being programmed explicitly. Machine learning has allowed us to recognize practical speech, self-driven cars, efficient web research, and humans' perception enormously improved in the early decade. This is so persistent today that it may be applied without knowledge several times a day. Most researchers are involved and believe it is an excellent way to develop artificial intelligence towards the human level. Cancer research is one of the main fields of research in the field of medicine[18]. The predictability of different cancer types is important for better treatment and minimization of severity for patients[1]. Therefore, microarrays can be used for the classification of different cancers and for the prediction. In the classification of lymphoma, leukemia, breast cancer, and liver cancer, for example, gene expression data have been employed to obtain good results. Microarray technology provides a new tool for automating the diagnostic work and improving the precise traditional diagnosis techniques[2]. The expression of thousands of genes can be examined at once with microarrays. Higher expression testing of certain genes can help cancer predict. The problem in the analysis of microarrays, however, is that gene expression data are ultra-highly dimensional (microarray image)[16].

Microarrays' high dimension makes it extremely difficult to process them and their complexity in time and space. Therefore, it is important to reduce the data dimensionality before further processing in order to make processing microarray feasible.

As mentioned, a number of methods for classifying cancer types with gene expression are developed and researched. However, most of the studies were confined to the problem of binary gene selection and very few considered feature selection and classification in multiple classes. This is because the selection and classification of multi-class genes are significantly more difficult than binary problems. Notably, the most effective and useful classifiers for the accurate diagnosis of cancer by microarray gene expression data are multi-class cancer classifiers such as random forest and support vector machine. SVMs can only be used for binary classification tasks in the first generation. Most real-life diagnostic tasks, however, particularly cancer and not binary classes. In recent years, several algorithms have been implemented to classify the multiple classes along with statistical approaches, the Evolutionary Algorithm, near-K(KNN), naive bays (NB), neural networks (NN)[19], and decision tree (DT), SVM (Support Vector Machine), ELM(Extreme Learning Machine) and so on. In this area, many algorithms for gene selection were proposed [3]. The gene selection process is challenged by a large number of genes and the small number of samples. The T-score between classes and gene expressions is common in the analysis of gene expression in microarrays[4]. The t-test of two samples is a test parameter that tests whether two datasets from the same distribution have been sampled (or have the same mean). In the context of the analysis of difference of expression, the values of expression across two classes for a specific gene are assumed to be of unequal sample size with an unequal difference[5]. Therefore, an unpaired t-test on expression array data is usually carried out. The t statistics are directed to the medium differences between the interclass and inverse to the inter-classes' standard deviations. Small standard deviations in the intrinsic classes and a large interclass mean difference show a good class gene (small p-value)[6]. Based on the overlap of distributions, a p-value is determined. The microarray gene classification of cancer is a major problem.

As the size of the medical databases increases, traditional machine learning models such as decision tree, SVM, neural networks, naive bayes, fuzzy ensemble learning [7], etc. become difficult to process the patterns due to noise, high dimensionality and non-relational instances in the medical databases. Also, the major challenge of the existing models includes disease pattern discovery and quality services [17]. Feature selection and classification are the essential requirements for most of the medical disease pattern discovery models. Generally, the SVM classification scheme is based upon the characteristics of the statistical learning mechanism. The classification of the SVM classifier supports structural risk minimization to carry out the whole process of classification smoothly and effectively.

Another classification method is called bagging. In this, the classifier is able to give an output of a category guess. Each prediction done is considered as a single vote. If a given class gains the majority of the votes, it is then considered the classification's output. The bagging aggregates the classes based on the number of votes. There are some other classifiers that have been derived by improving this bagging method. This is best performed using trees. This is because the structure of trees is easy to interpret[8].

High dimensionality and imbalance are the key problems of medical datasets. Traditional classifiers of machine learning consider a subset of classification and disease prediction characteristics with a high true negative rate and error rates. The medical field is considered one of the most information-intensive domains in which clinical-related data and knowledge are regularly developed. An example of such a complex system is the development of an integrated healthcare system model[9]. Healthcare information systems collect and segregate patients ' clinical history, including attributes, patient demographic data, critical functionalities, test inferences, and unstructured data such as audio and video records. For the medical field and for patients, proper analysis of such information is vital. Intelligent analysis of such aggregated data such as rapid diagnosis of disease, optimum treatment selection for patients, duration of patient treatment and its outcomes, complex risk determination and other optimization of the use of medical resources can perform various tasks. Complete computerization of disease diagnosis and treatment in recent decades allows rapid and effective aggregation. A healthcare disease dataset may contain numerous attributes, and many of these attributes may not contribute to an algorithm's classification accuracy during diagnosis. In addition, due to the presence of such foreign attributes that affect the accuracy of the disease prediction, there is a considerable calculation time-consuming. Therefore, attribute selection is an optimizing agent where a subset of attributes is selected to filter and remove less relevant and noisy attributes for more precise and effective data representation. There are numerous possible solutions in the search space during problem-solving. The aim is to choose a solution that optimizes the processing and produces the best possible output. Problems with optimization are those types of problems that are used to determine the best solution among all possible solutions. Optimization issues can be categorized into two types, depending on the attribute types. These issues may be referred to as combinatorial issues when considering discrete-valued attributes. If continuous attributes are taken into account, they are referred to as constrained or multimodal problems. By using this reduced attribute subset on a classifier to detect the presence of disease, the classification performance of disease diagnosis is improved.

2. Proposed Model

In the proposed framework, an advanced filter-based machine learning model is designed and implemented on the medical databases, as shown in figure 1. In this framework, different types of medical datasets are taken to find the outliers and the data transformation process. After performing the data transformation, different traditional classification models such as Naive Bayes, Logistic Regression, Multilayer Neural Network, K-Nearest Neighbour, Adaboost, C4.5, Random Forest, and also the proposed Random Forest models were implemented. Finally, statistical measures are used to find the performance of the proposed model to the conventional models.

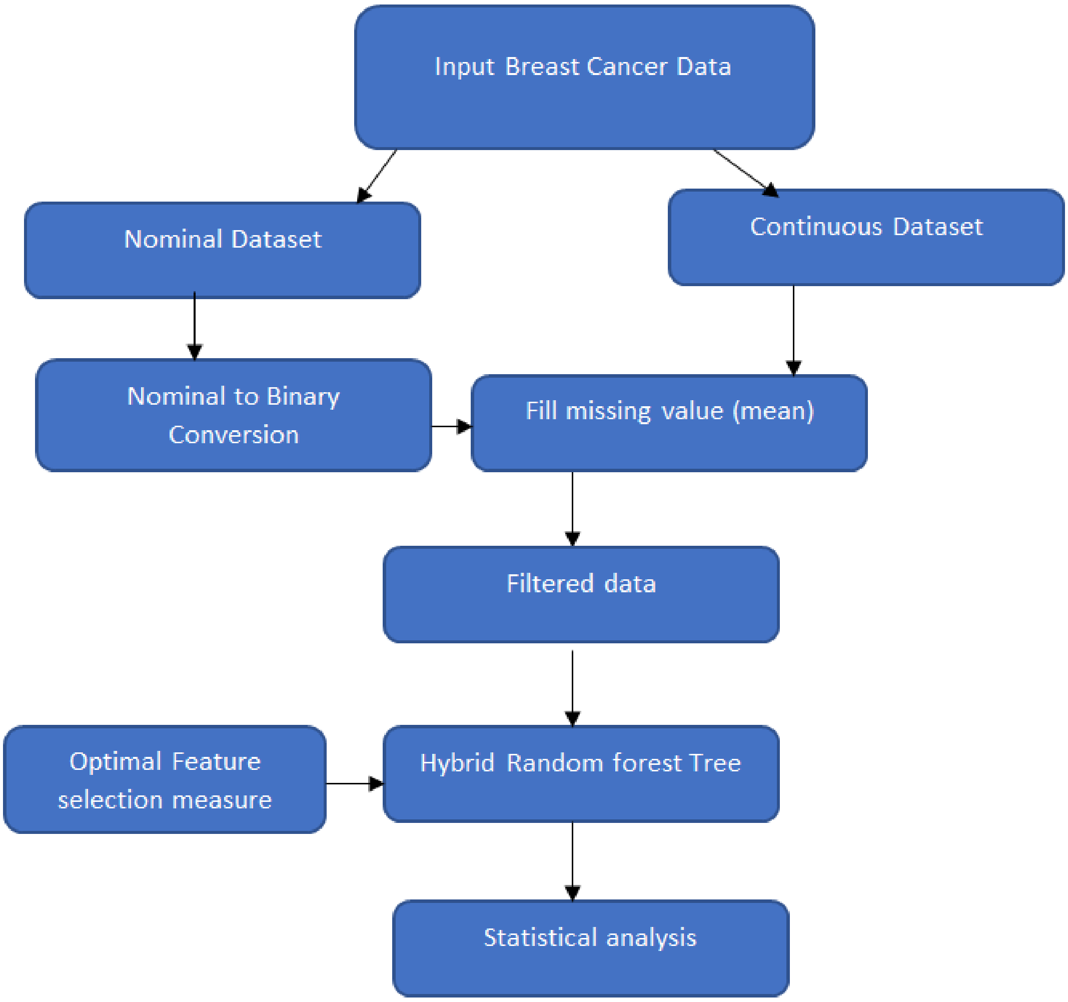

Input Breast Cancer Data
Nominal Dataset
Continuous Dataset
Nominal to Binary Conversion
Fill missing value (mean)
Filtered data
Optimal Feature selection measure
Hybrid Random forest Tree
Statistical analysis

Figure 1: Proposed heterogeneous medical data classification framework

In the proposed framework, heterogeneous datasets are used to find the class prediction using the proposed disease prediction model. Initially, input data is prepared based on different class attributes. Nominal attributes are converted to binary attributes and then missing values are filled with mean values. These filtered data are given to the classification problem for a better disease prediction rate. Here, different nominal attributes are used as class labels for the decision-making process.

Algorithm 1: A Proposed Random forest:

Step 1: Input file (Filtered anomaly data)

Step 2: Preprocess anomaly data for missing values.

Step 3: Data transformation for unequal distribution as

```
    For each attribute $A_i$ in DB
    Do
    If($A_i$.type==numerical)
    Then
        $A_i.\mathrm{value} = \frac{A_i.value + G.\mathrm{M}(\mathrm{A}_i)}{(\mathrm{A}_i .value.Max - \mathrm{A}_i .value.Min)} * (ScaleMax - ScaleMin)$
    End if
    Else
    Continue;
Step 4: For each randomized sample $S_i$
    Do
```

Kernel Probability: The kernel probability is used to estimate the conditional variance of input data features by using the gaussian estimator.

$$\mathrm{B_f} = \mathrm{uniqueCV(D);//\,Unique\ column\ values}$$

$$\mathrm{HB_f} = \mathrm{Histobins[]} = \mathrm{histogrambin(D)}$$

$$\mathrm{GaussianKernel:GK}(\phi,\theta) = \mathrm{e}^{-\theta^2} / (2*\phi^2)$$

$$\psi = \mathrm{gkv} = \mathrm{GK}(\sum \mathrm{HB_f}, \sum \mathrm{B_f});$$

$$\mathrm{Kernel\,Pr\,obability} = \mathrm{KP(D)} =| \mathrm{HB_f} \ / (\sum \psi * \mathrm{HB_f}) \,|$$

$$\mathrm{GaussianEntropy:GE(d_i)} = -\mathrm{GK}(\sum_{\mathrm{i}} \mathrm{d_i} .\log(\mathrm{d_i}), \mu_{\mathrm{d}})$$

In the above equations, the Gaussian entropy is used to check the feature entropy value based on the Gaussian estimator.
Proposed entropy formula:

$$\mathrm{PE} = \mathrm{e}^{-\mathrm{D}^2} / (2*\mathrm{D_1}^2)*|\,\mathrm{HB_{D_1}} \ / (\sum \psi * \mathrm{HB_{D_1}})\,| + \mathrm{e}^{-\mathrm{D}^2} / (2*\mathrm{D_2}^2)*|\,\mathrm{HB_{D_2}} \ / (\sum \psi * \mathrm{HB_{D_2}})\,|$$

```
For each sample in test data
    If(PE>0)
    Then
    $S' = Classify((\mathrm{D}_i, \mathrm{D}_j));$
    Else
```

Received 20 January 2021; Accepted 08 February 2021.

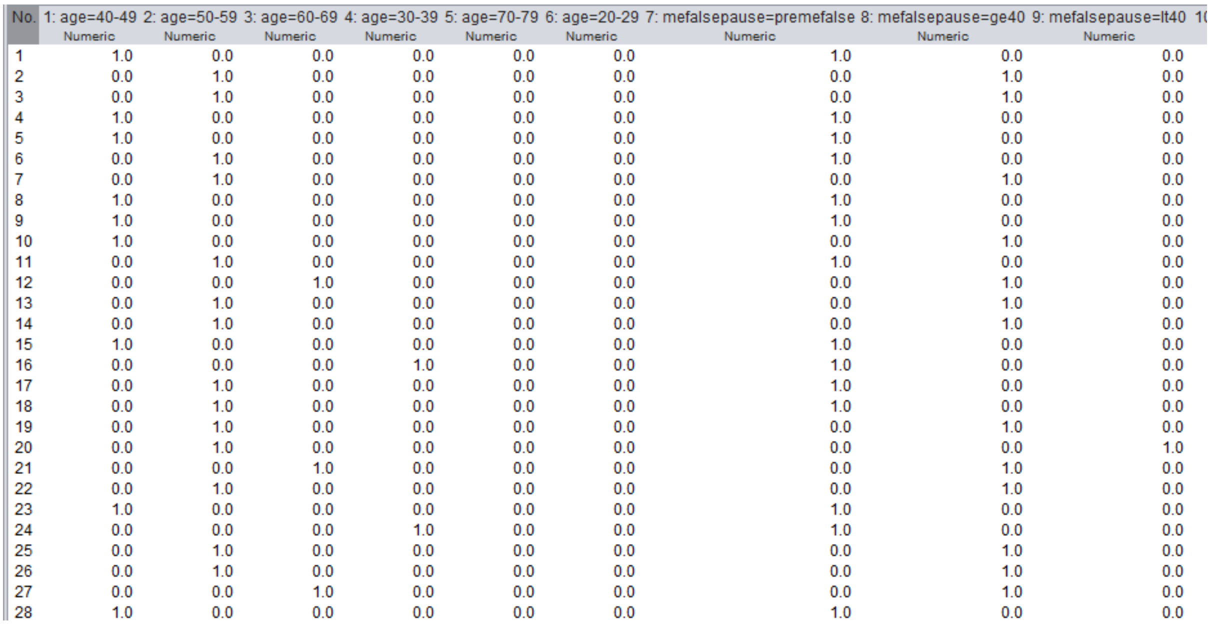
```
        Continue;
        End for
```

In the above ensemble based anomaly detection model, each attribute is checked against the data distribution. If the attribute is not uniform distributed then it was transformed to uniform format. For each attribute in the uniform distributed dataset, instancesare partitioned into set of sub-partitions based on classes. After that, similarity computation was applied on the sub-partitions to find the relevant relational anomaly features.

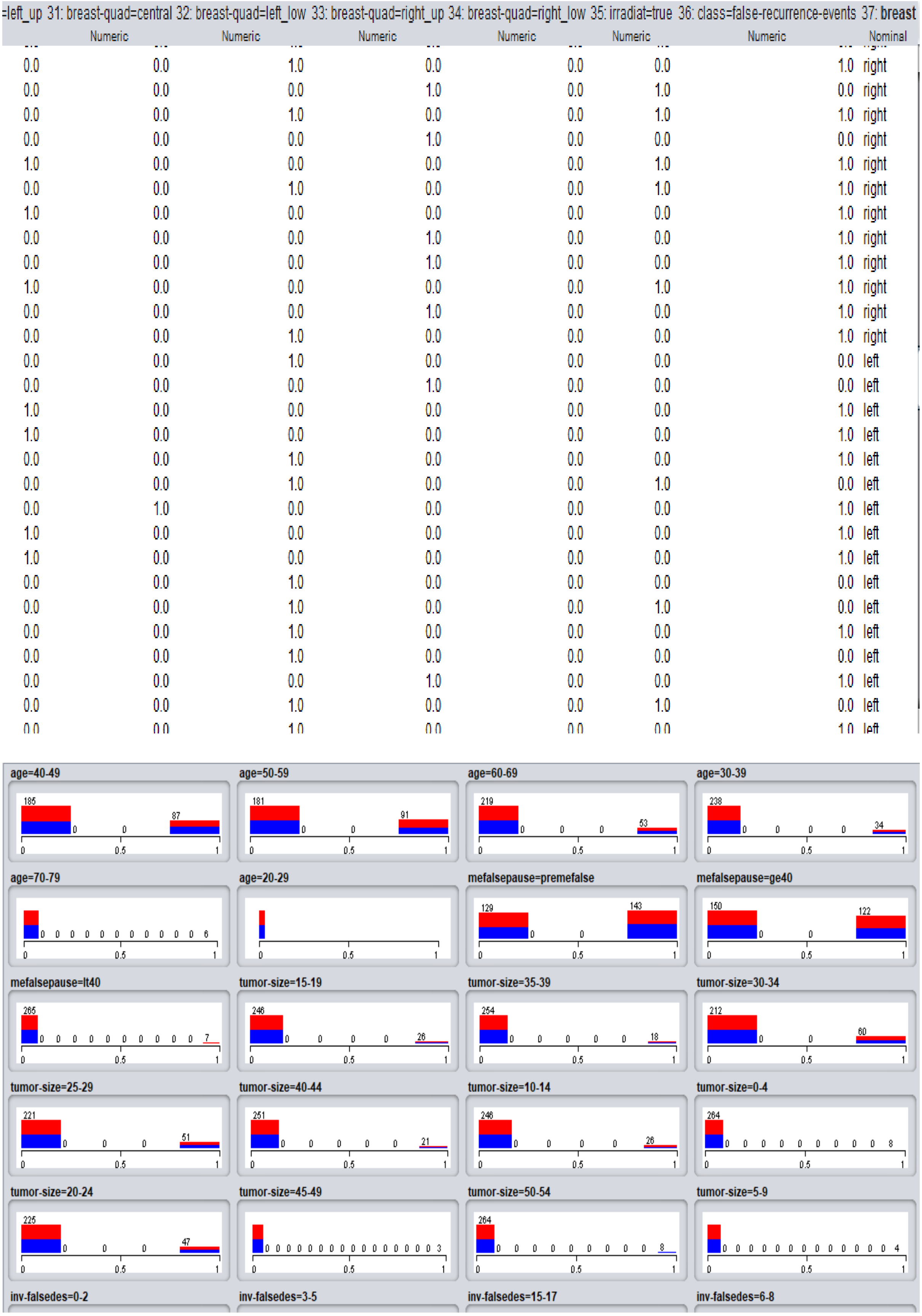
3.Experimental results

Experimental results are simulated in a python environment with third party libraries. In the proposed work, standard numerical breast cancer and nominal breast cancer datasets with a large number of feature spaces are taken for experimental study. Initially, nominal datasets are filtered using the nominal to binary conversion. In this process, each attribute is verified against the missing value. If the attribute contains a missing value, then it is replaced with the mean of the attribute. Later, these filtered data are given to the proposed classification algorithm for disease prediction and decision-making process. Table 1 and table2 represent the heterogeneous nominal to converted binary datasets for the data classification problem. Table 3 represents the standard breast cancer dataset.

Table 1: Heterogeneous breast cancer dataset 1 with class label breast type.



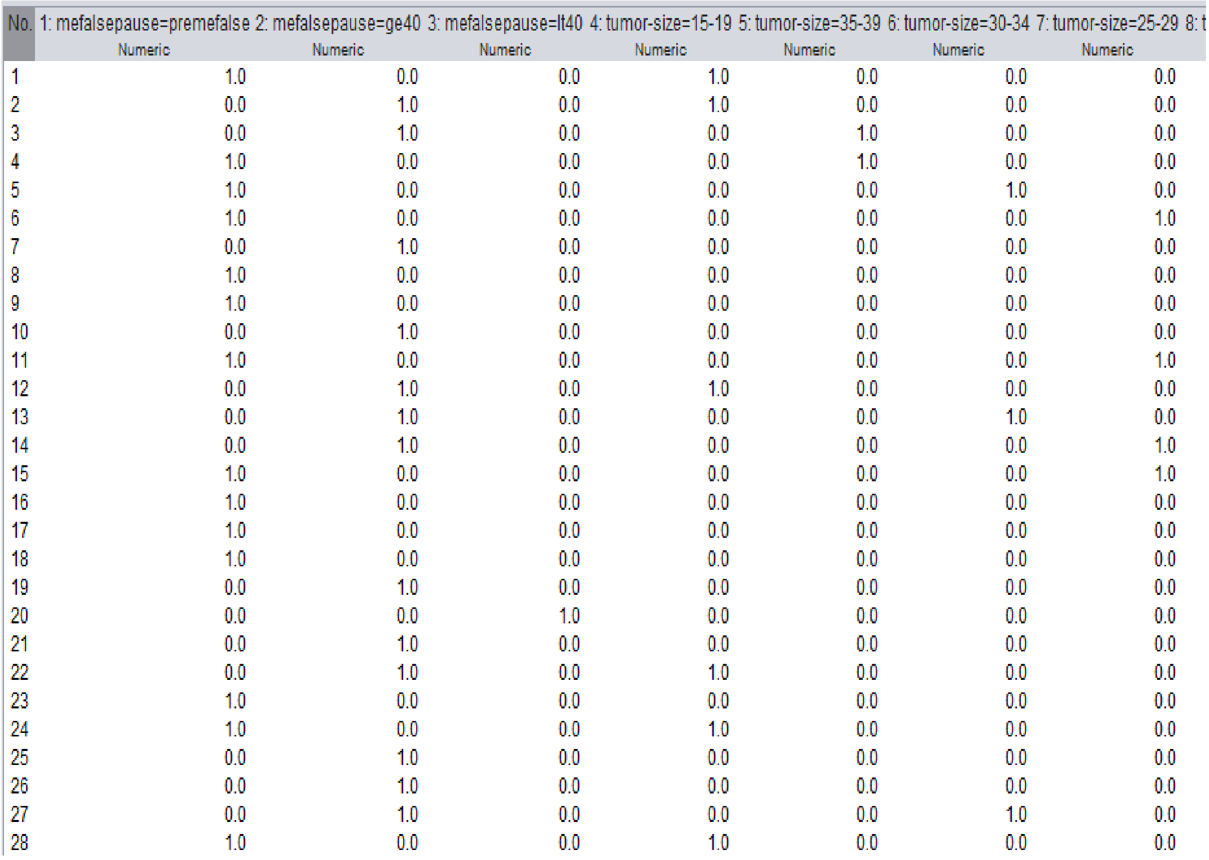
No.	1: age=40-49	2: age=50-59	3: age=60-69	4: age=30-39	5: age=70-79	6: age=20-29	7: mefalsepause=premefalse	8: mefalsepause=ge40	9: mefalsepause=lt40	1(
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	
1	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
3	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
4	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
5	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
6	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
7	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
8	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
9	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
10	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
11	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
12	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	
13	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
14	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
15	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
16	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	
17	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
18	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
19	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
20	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	
21	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	
22	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
23	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
24	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	
25	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
26	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
27	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	
28	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	



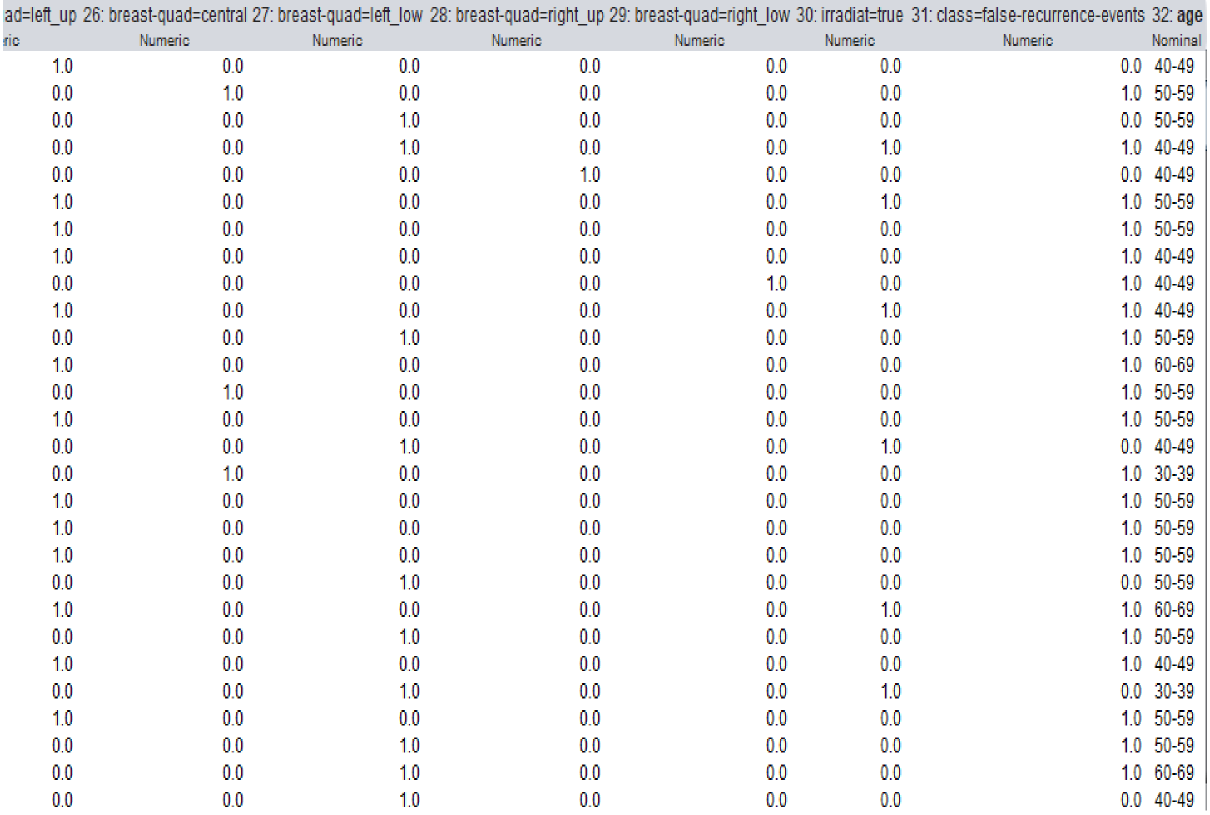
=left_up	31: breast-quad=central Numeric	32: breast-quad=left_low Numeric	33: breast-quad=right_up Numeric	34: breast-quad=right_low Numeric	35: irradiat=true Numeric	36: class=false-recurrence-events Numeric	37: breast Nominal
0.0	0.0	1.0	0.0	0.0	0.0	1.0	right
0.0	0.0	0.0	1.0	0.0	1.0	0.0	right
0.0	0.0	1.0	0.0	0.0	1.0	1.0	right
0.0	0.0	0.0	1.0	0.0	0.0	0.0	right
1.0	0.0	0.0	0.0	0.0	1.0	1.0	right
0.0	0.0	1.0	0.0	0.0	1.0	1.0	right
1.0	0.0	0.0	0.0	0.0	0.0	1.0	right
0.0	0.0	0.0	1.0	0.0	0.0	1.0	right
0.0	0.0	0.0	1.0	0.0	0.0	1.0	right
1.0	0.0	0.0	0.0	0.0	1.0	1.0	right
0.0	0.0	0.0	1.0	0.0	0.0	1.0	right
0.0	0.0	1.0	0.0	0.0	0.0	1.0	right
0.0	0.0	1.0	0.0	0.0	0.0	0.0	left
0.0	0.0	0.0	1.0	0.0	0.0	0.0	left
1.0	0.0	0.0	0.0	0.0	0.0	1.0	left
1.0	0.0	0.0	0.0	0.0	0.0	1.0	left
0.0	0.0	1.0	0.0	0.0	0.0	1.0	left
0.0	0.0	1.0	0.0	0.0	1.0	0.0	left
0.0	1.0	0.0	0.0	0.0	0.0	1.0	left
1.0	0.0	0.0	0.0	0.0	0.0	1.0	left
1.0	0.0	0.0	0.0	0.0	0.0	1.0	left
0.0	0.0	1.0	0.0	0.0	0.0	0.0	left
0.0	0.0	1.0	0.0	0.0	1.0	0.0	left
0.0	0.0	1.0	0.0	0.0	0.0	1.0	left
0.0	0.0	1.0	0.0	0.0	0.0	0.0	left
0.0	0.0	0.0	1.0	0.0	0.0	1.0	left
0.0	0.0	1.0	0.0	0.0	1.0	0.0	left
0.0	0.0	1.0	0.0	0.0	0.0	1.0	left

Figure 2: Visualization of Heterogeneous breast cancer dataset 1

Table 2: **Heterogeneous breast cancer dataset 2 with class label age**



No.	1: mefalsepause=premefalse Numeric	2: mefalsepause=ge40 Numeric	3: mefalsepause=lt40 Numeric	4: tumor-size=15-19 Numeric	5: tumor-size=35-39 Numeric	6: tumor-size=30-34 Numeric	7: tumor-size=25-29 Numeric	8: t
1	1.0	0.0	0.0	1.0	0.0	0.0	0.0	
2	0.0	1.0	0.0	1.0	0.0	0.0	0.0	
3	0.0	1.0	0.0	0.0	1.0	0.0	0.0	
4	1.0	0.0	0.0	0.0	1.0	0.0	0.0	
5	1.0	0.0	0.0	0.0	0.0	1.0	0.0	
6	1.0	0.0	0.0	0.0	0.0	0.0	1.0	
7	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
8	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
10	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
11	1.0	0.0	0.0	0.0	0.0	0.0	1.0	
12	0.0	1.0	0.0	1.0	0.0	0.0	0.0	
13	0.0	1.0	0.0	0.0	0.0	1.0	0.0	
14	0.0	1.0	0.0	0.0	0.0	0.0	1.0	
15	1.0	0.0	0.0	0.0	0.0	0.0	1.0	
16	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
17	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
18	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
19	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
20	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
21	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
22	0.0	1.0	0.0	1.0	0.0	0.0	0.0	
23	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
24	1.0	0.0	0.0	1.0	0.0	0.0	0.0	
25	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
26	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
27	0.0	1.0	0.0	0.0	0.0	1.0	0.0	
28	1.0	0.0	0.0	1.0	0.0	0.0	0.0	



ad=left_up ric	26: breast-quad=central Numeric	27: breast-quad=left_low Numeric	28: breast-quad=right_up Numeric	29: breast-quad=right_low Numeric	30: irradiat=true Numeric	31: class=false-recurrence-events Numeric	32: age Nominal
1.0	0.0	0.0	0.0	0.0	0.0	0.0	40-49
0.0	1.0	0.0	0.0	0.0	0.0	1.0	50-59
0.0	0.0	1.0	0.0	0.0	0.0	0.0	50-59
0.0	0.0	1.0	0.0	0.0	1.0	1.0	40-49
0.0	0.0	0.0	1.0	0.0	0.0	0.0	40-49
1.0	0.0	0.0	0.0	0.0	1.0	1.0	50-59
1.0	0.0	0.0	0.0	0.0	0.0	1.0	50-59
1.0	0.0	0.0	0.0	0.0	0.0	1.0	40-49
0.0	0.0	0.0	0.0	1.0	0.0	1.0	40-49
1.0	0.0	0.0	0.0	0.0	1.0	1.0	40-49
0.0	0.0	1.0	0.0	0.0	0.0	1.0	50-59
1.0	0.0	0.0	0.0	0.0	0.0	1.0	60-69
0.0	1.0	0.0	0.0	0.0	0.0	1.0	50-59
1.0	0.0	0.0	0.0	0.0	0.0	1.0	50-59
0.0	0.0	1.0	0.0	0.0	1.0	0.0	40-49
0.0	1.0	0.0	0.0	0.0	0.0	1.0	30-39
1.0	0.0	0.0	0.0	0.0	0.0	1.0	50-59
1.0	0.0	0.0	0.0	0.0	0.0	1.0	50-59
1.0	0.0	0.0	0.0	0.0	0.0	1.0	50-59
0.0	0.0	1.0	0.0	0.0	0.0	0.0	50-59
1.0	0.0	0.0	0.0	0.0	1.0	1.0	60-69
0.0	0.0	1.0	0.0	0.0	0.0	1.0	50-59
1.0	0.0	0.0	0.0	0.0	0.0	1.0	40-49
0.0	0.0	1.0	0.0	0.0	1.0	0.0	30-39
1.0	0.0	0.0	0.0	0.0	0.0	1.0	50-59
0.0	0.0	1.0	0.0	0.0	0.0	1.0	50-59
0.0	0.0	1.0	0.0	0.0	0.0	1.0	60-69
0.0	0.0	1.0	0.0	0.0	0.0	0.0	40-49

Received 20 January 2021; Accepted 08 February 2021.

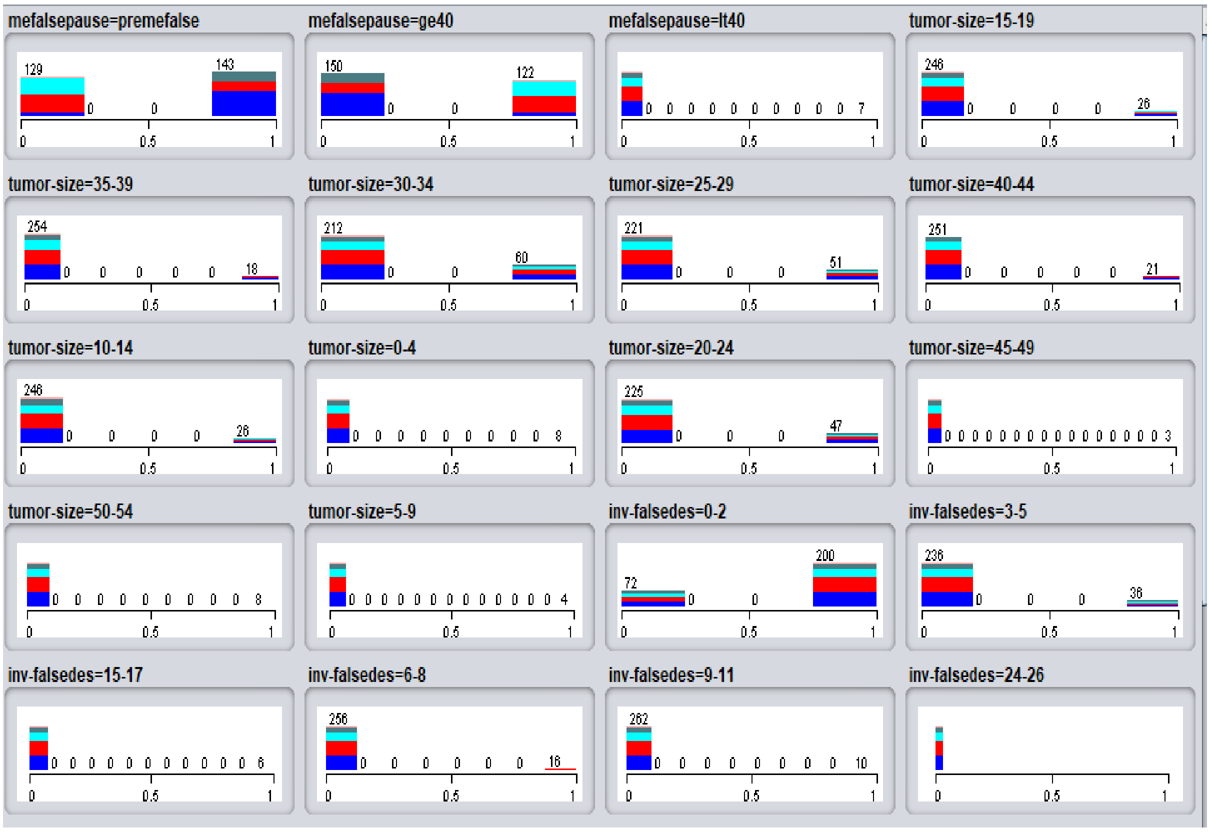
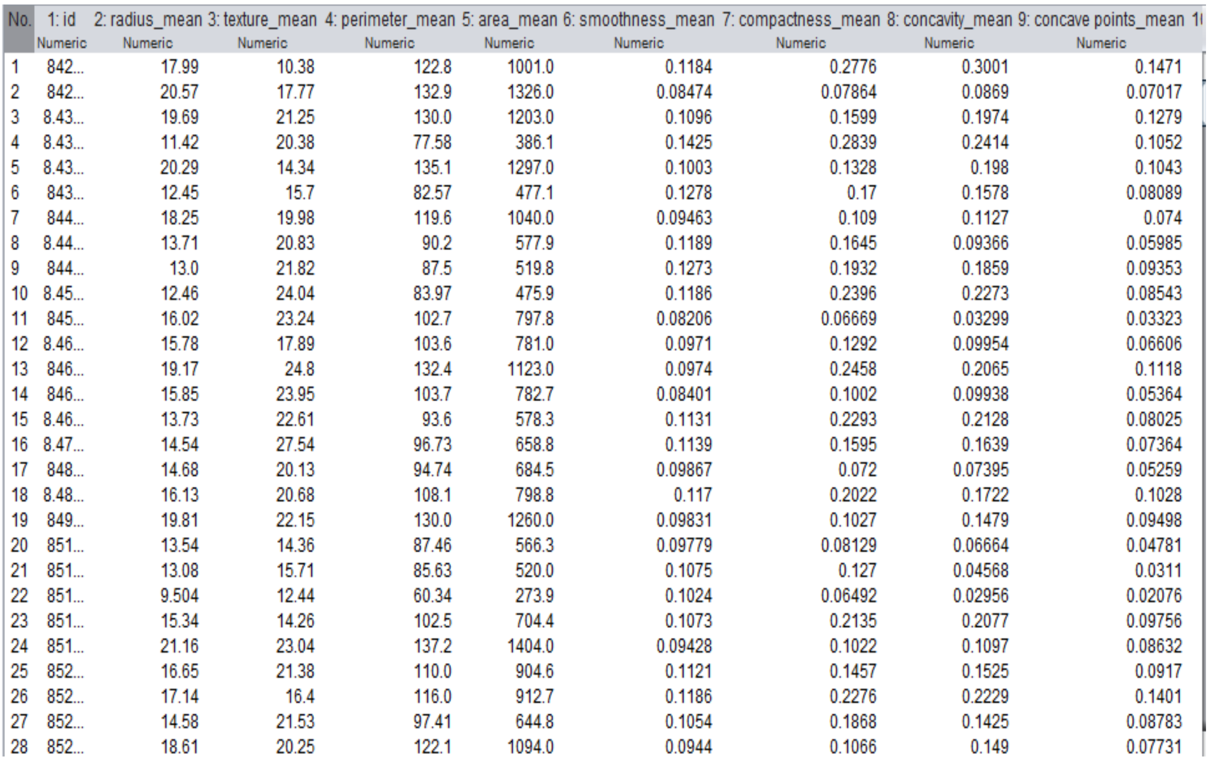


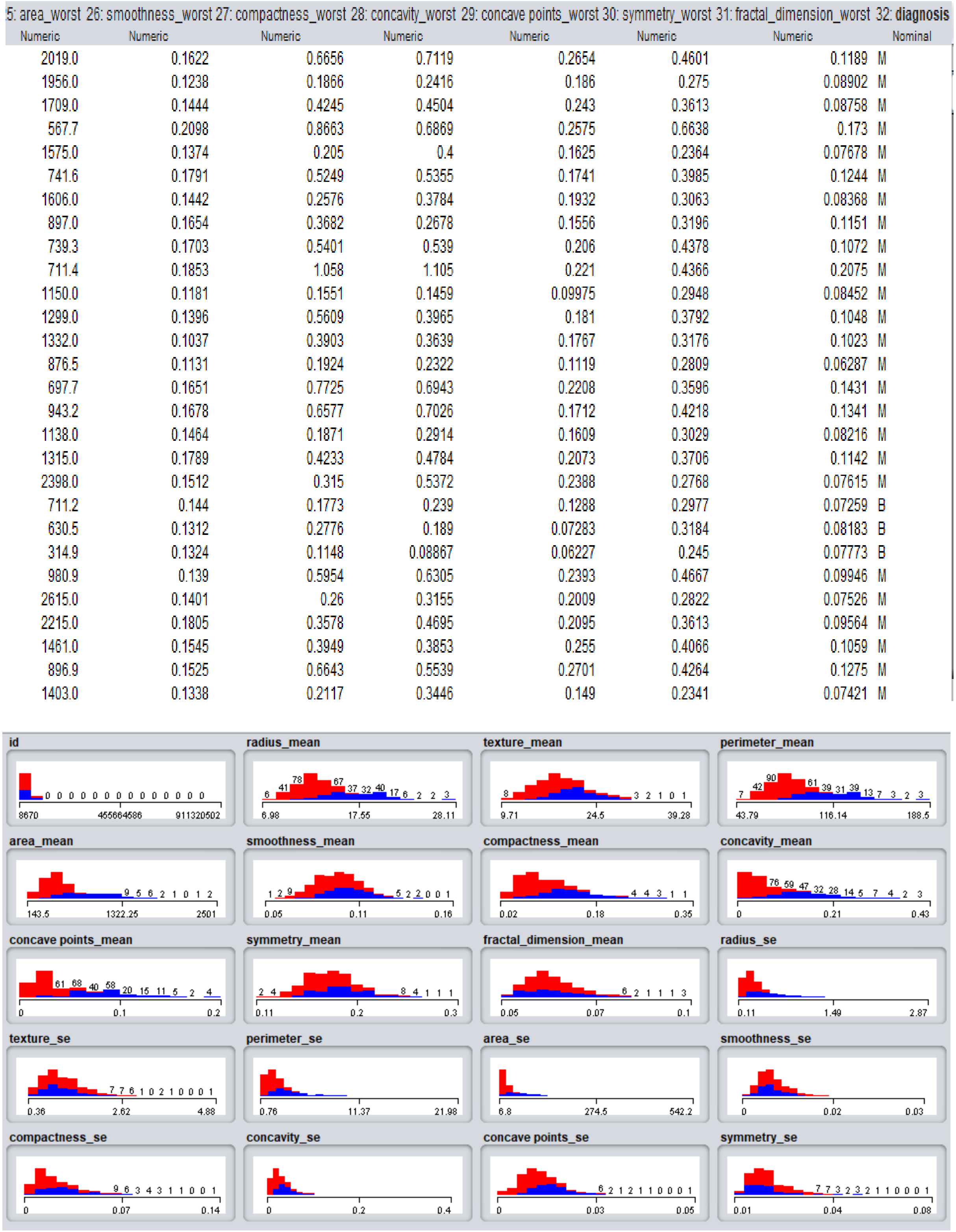
Figure 3: Visualization of Heterogeneous breast cancer dataset 2

Table 3: Standard breast cancer dataset with numerical attributes



No.	1: id Numeric	2: radius_mean Numeric	3: texture_mean Numeric	4: perimeter_mean Numeric	5: area_mean Numeric	6: smoothness_mean Numeric	7: compactness_mean Numeric	8: concavity_mean Numeric	9: concave points_mean Numeric	1(
1	842...	17.99	10.38	122.8	1001.0	0.1184	0.2776	0.3001	0.1471	
2	842...	20.57	17.77	132.9	1326.0	0.08474	0.07864	0.0869	0.07017	
3	8.43...	19.69	21.25	130.0	1203.0	0.1096	0.1599	0.1974	0.1279	
4	8.43...	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	
5	8.43...	20.29	14.34	135.1	1297.0	0.1003	0.1328	0.198	0.1043	
6	843...	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	
7	844...	18.25	19.98	119.6	1040.0	0.09463	0.109	0.1127	0.074	
8	8.44...	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	
9	844...	13.0	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	
10	8.45...	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	
11	845...	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	
12	8.46...	15.78	17.89	103.6	781.0	0.0971	0.1292	0.09954	0.06606	
13	846...	19.17	24.8	132.4	1123.0	0.0974	0.2458	0.2065	0.1118	
14	846...	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	
15	8.46...	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	
16	8.47...	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	
17	848...	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	
18	8.48...	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	
19	849...	19.81	22.15	130.0	1260.0	0.09831	0.1027	0.1479	0.09498	
20	851...	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	
21	851...	13.08	15.71	85.63	520.0	0.1075	0.127	0.04568	0.0311	
22	851...	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	
23	851...	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	
24	851...	21.16	23.04	137.2	1404.0	0.09428	0.1022	0.1097	0.08632	
25	852...	16.65	21.38	110.0	904.6	0.1121	0.1457	0.1525	0.0917	
26	852...	17.14	16.4	116.0	912.7	0.1186	0.2276	0.2229	0.1401	
27	852...	14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	
28	852...	18.61	20.25	122.1	1094.0	0.0944	0.1066	0.149	0.07731	

Received 20 January 2021; Accepted 08 February 2021.



5: area_worst	26: smoothness_worst	27: compactness_worst	28: concavity_worst	29: concave points_worst	30: symmetry_worst	31: fractal_dimension_worst	32: diagnosis
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.1189	M
1956.0	0.1238	0.1866	0.2416	0.186	0.275	0.08902	M
1709.0	0.1444	0.4245	0.4504	0.243	0.3613	0.08758	M
567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.173	M
1575.0	0.1374	0.205	0.4	0.1625	0.2364	0.07678	M
741.6	0.1791	0.5249	0.5355	0.1741	0.3985	0.1244	M
1606.0	0.1442	0.2576	0.3784	0.1932	0.3063	0.08368	M
897.0	0.1654	0.3682	0.2678	0.1556	0.3196	0.1151	M
739.3	0.1703	0.5401	0.539	0.206	0.4378	0.1072	M
711.4	0.1853	1.058	1.105	0.221	0.4366	0.2075	M
1150.0	0.1181	0.1551	0.1459	0.09975	0.2948	0.08452	M
1299.0	0.1396	0.5609	0.3965	0.181	0.3792	0.1048	M
1332.0	0.1037	0.3903	0.3639	0.1767	0.3176	0.1023	M
876.5	0.1131	0.1924	0.2322	0.1119	0.2809	0.06287	M
697.7	0.1651	0.7725	0.6943	0.2208	0.3596	0.1431	M
943.2	0.1678	0.6577	0.7026	0.1712	0.4218	0.1341	M
1138.0	0.1464	0.1871	0.2914	0.1609	0.3029	0.08216	M
1315.0	0.1789	0.4233	0.4784	0.2073	0.3706	0.1142	M
2398.0	0.1512	0.315	0.5372	0.2388	0.2768	0.07615	M
711.2	0.144	0.1773	0.239	0.1288	0.2977	0.07259	B
630.5	0.1312	0.2776	0.189	0.07283	0.3184	0.08183	B
314.9	0.1324	0.1148	0.08867	0.06227	0.245	0.07773	B
980.9	0.139	0.5954	0.6305	0.2393	0.4667	0.09946	M
2615.0	0.1401	0.26	0.3155	0.2009	0.2822	0.07526	M
2215.0	0.1805	0.3578	0.4695	0.2095	0.3613	0.09564	M
1461.0	0.1545	0.3949	0.3853	0.255	0.4066	0.1059	M
896.9	0.1525	0.6643	0.5539	0.2701	0.4264	0.1275	M
1403.0	0.1338	0.2117	0.3446	0.149	0.2341	0.07421	M

Figure 4: Visualization of standard breast cancer dataset

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3211 – 3225
Received 20 January 2021; Accepted 08 February 2021.

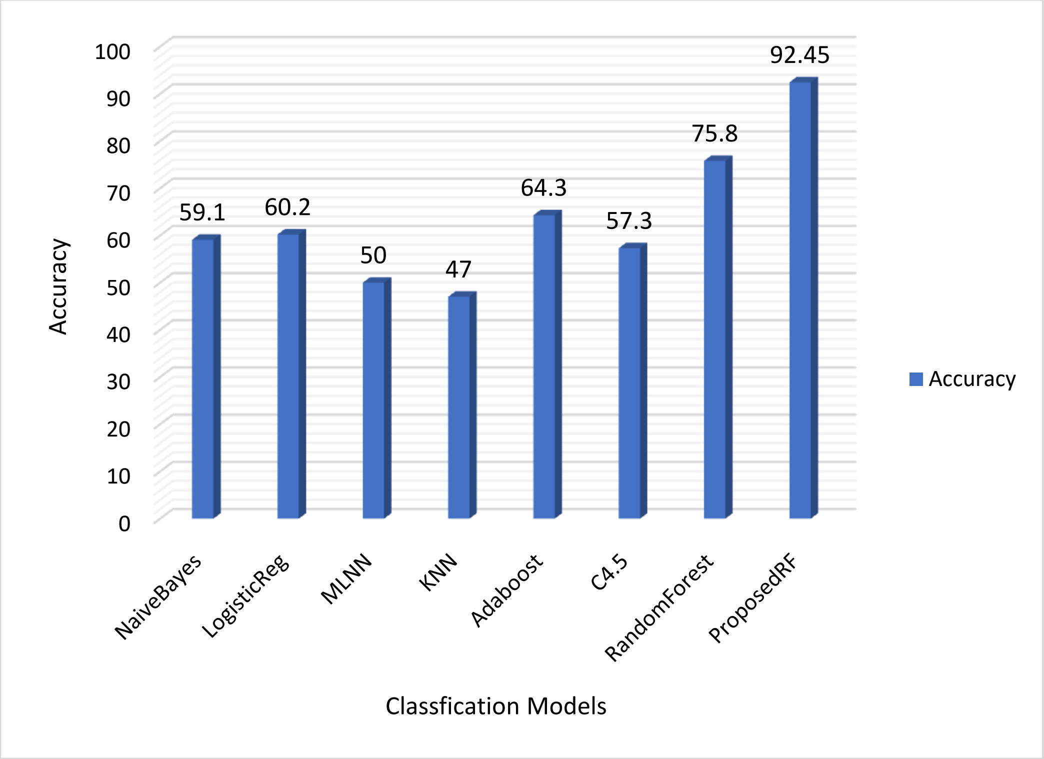


Figure 5: Performance analysis of proposed heterogeneous classification model to the conventional classification models on breast cancer dataset 1.

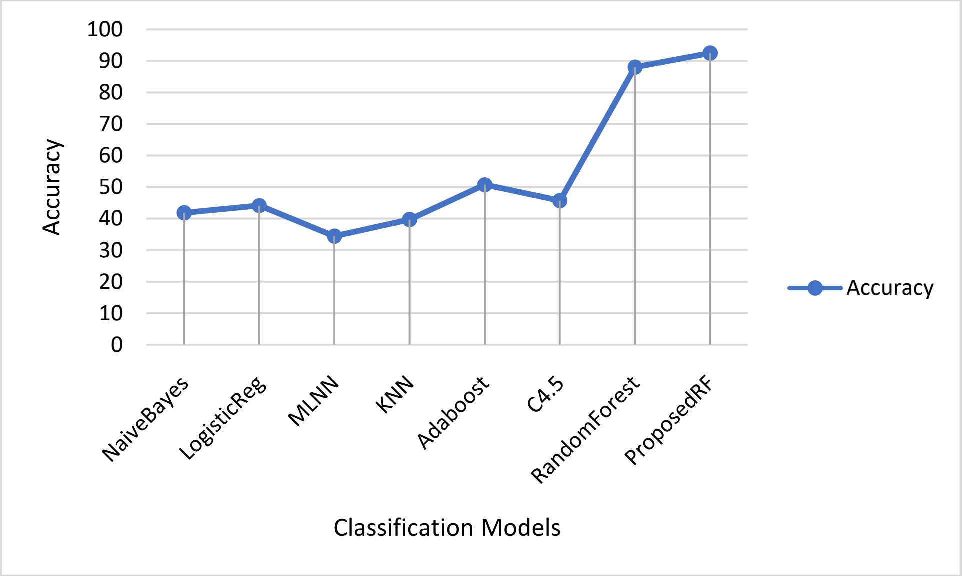


Figure 6: Performance analysis of proposed heterogeneous classification model to the conventional classification models on breast cancer dataset 2.

Received 20 January 2021; Accepted 08 February 2021.

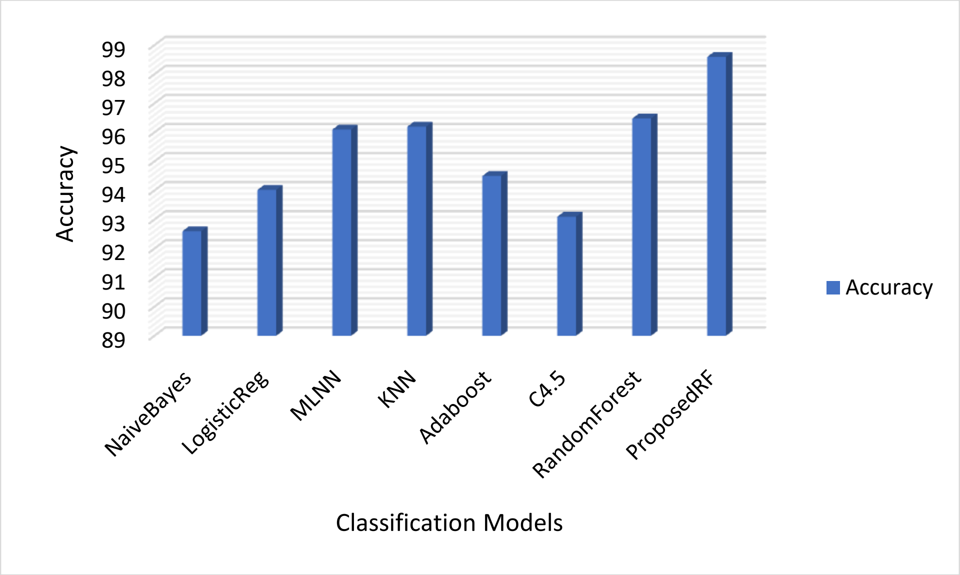


Figure 7: Performance analysis of proposed heterogeneous classification model to the conventional classification models on standard breast cancer dataset.

4.Conclusion

In this paper, advanced machine learning approaches are implemented on the medical databases for better decision making. Since most of the conventional approaches are independent of outliers and data size, the proposed model has better efficiency in outliers, filtering and data classification problems. In this paper, a novel feature selection based random forest classifier is proposed to improve the efficiency of the medical datasets. Experimental results show that the proposed heterogeneous classification framework has better accuracy, nearly 98.6% accuracy than the traditional nominal data classification models.

References

[1] J. Kennedy and R. Eberhart, "Particle swarm optimization," Proceedings of ICNN'95 - International Conference on Neural Networks, Perth, WA, Australia, 1995, pp. 1942-1948 vol.4, doi: 10.1109/ICNN.1995.488968.

[2] Owolabi, Ibrahim. (2018). Diagnosis of Breast Cancer Using Power Swarm Optimization with SVM. 10.13140/RG.2.2.12398.31045.

[3] S. B. Sakri, N. B. Abdul Rashid and Z. Muhammad Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," in *IEEE Access*, vol. 6, pp. 29637-29647, 2018, doi: 10.1109/ACCESS.2018.2843443.

[4] M. Kumar, S. K. Khatri and M. Mohammadian, "Review on Breast Cancer Disease Predictive Modelling using Swarm Intelligence," 2020 International Conference onComputational Performance Evaluation (ComPE), Shillong, India, 2020, pp. 523-530, doi: 10.1109/ComPE49325.2020.9200117.

[5] Zamani, Hoda&Nadimi-Shahraki, Mohammad H.. (2016). Swarm Intelligence Approach for Breast Cancer Diagnosis. International Journal of Computer Applications. 151. 40-44. 10.5120/ijca2016911667.

[6] RozillaJamiliOskouei, NasroallahMoradiKorandSaeidAbbasiMakeki, ”Data mining and medical world: breast cancers diagnosis, treatment, prognosis and challenges”, *American Journal of Cancer Research*, Vol.7, No.3, pp 610-627, 2017.

[7] V. Bolon-canedo, N. Sanchez- Marono, A.Alonso-Betanzos, J.M. Benitez, F. Herrera, “A review of microarray datasets and applied feature selection methods”, Elsevier,Infornation Sciences 282, pp. 111-135, 2014.

[8] Y. Xiao, J. Wu, Z. Lin and X. Zhao, "Breast Cancer Diagnosis Using an Unsupervised Feature Extraction Algorithm Based on Deep Learning," 2018 37th Chinese Control Conference (CCC)s, Wuhan, 2018, pp. 9428-9433, doi: 10.23919/ChiCC.2018.8483140.

[9] H.-J. Xing and W.-T. Liu, “Robust AdaBoost based ensemble of one-class support vector machines,” Information Fusion, vol. 55, pp. 45–58, Mar. 2020, doi: 10.1016/j.inffus.2019.08.002.

[10] V. Christou, M. G. Tsipouras, N. Giannakeas, A. T. Tzallas, and G. Brown, “Hybrid extreme learning machine approach for heterogeneous neural networks,” Neurocomputing, vol. 361, pp. 137–150, Oct. 2019, doi: 10.1016/j.neucom.2019.04.092.

[11] C.-L. Huang, H.-C. Liao, and M.-C. Chen, “Prediction model building and feature selection with support vector machines in breast cancer diagnosis,” Expert Systems with Applications, vol. 34, no. 1, pp. 578–587, Jan. 2008, doi: 0.1016/j.eswa.2006.09.041.

[12] P. Gupta and S. Garg, “Breast Cancer Prediction using varying Parameters of Machine Learning Models,” Procedia Computer Science, vol. 171, pp. 593–601, Jan. 2020, doi: 10.1016/j.procs.2020.04.064.

[13] C. Liangjun, P. Honeine, Q. Hua, Z. Jihong, and S. Xia, “Correntropy-based robust multilayer extreme learning machines,” Pattern Recognition, vol. 84, pp. 357–370, Dec. 2018, doi: 10.1016/j.patcog.2018.07.011.

[14] Xing, B. and Gao, W.-J. 2014. Imperialist Competitive Algorithm. in Innovative Computational Intelligence: A Rough Guide to 134 Clever Algorithms, ed Cham: Springer International Publishing. pp. 203-209.

[15] N. P. Pérez, M. A. Guevara Lopez, A. Silva, and I. Ramos, “Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography,” Artificial Intelligence in Medicine, vol. 63, no. 1, pp. 19–31, Jan. 2015, doi: 10.1016/j.artmed.2014.12.004.

[16] M S Kumar et.al,“Deep learning based image processing approaches for image deblurring”, Materials Today: Proceedings (Science Direct), https://doi.org/10.1016/j.matpr.2020.11.076 , 26 December 2020.

[17] V. A. Natarajan et.al,"Detection of disease in tomato plant using Deep Learning Techniques”, International Journal of Modern Agriculture, Volume 9, No.4, 2020 ISSN: 2305 -7246 pp: 525-540.

[18] Sreedhar B et.al, “A Comparative Study of Melanoma Skin Cancer Detection in Traditional and Current Image Processing Technique”, IEEE Xplore: DOI: 10.1109/I-SMAC49090.2020.9243501 10 November 2020. pp: 654 -658.

[19] V. A. Natarajan et.al, "Segmentation of Nuclei in Histopathology images using Fully Convolutional Deep Neural Architecture",IEEE Xplore September 2020, pp. 1-7, doi: 10.1109/ICCIT-144147971.2020.9213817.

[20] Manikandan, R and Dr.R.Latha (2017). “A literature survey of existing map matching algorithm for navigation technology. International journal of engineering sciences & research technology”, 6(9), 326-331.Retrieved September 15, 2017.

[21] A.M. Barani, R.Latha, R.Manikandan, "Implementation of Artificial Fish Swarm Optimization for Cardiovascular Heart Disease" International Journal of Recent Technology and Engineering (IJRTE), Vol. 08, No. 4S5, 134-136, 2019.

[22] Manikandan, R., Latha, R., & Ambethraj, C. (1). An Analysis of Map Matching Algorithm for Recent Intelligent Transport System. Asian Journal of Applied Sciences, 5(1). Retrieved from https://www.ajouronline.com/index.php/AJAS/article/view/4642

[23] R. Sathish, R. Manikandan, S. Silvia Priscila, B. V. Sara and R. Mahaveerakannan, "A Report on the Impact of Information Technology and Social Media on Covid–19," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 224-230, doi: 10.1109/ICISS49785.2020.9316046.